Unified and Multimodal Learning for Gaze Prediction in Naturalistic Settings

Thèse n. 1234 2025
présentée le 27 juillet 2025
à la Faculté des sciences et techniques de l'ingénieur
laboratoire de l'IDIAP
programme doctoral en génie électrique
École polytechnique fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Anshul Gupta

acceptée sur proposition du jury :

Prof. Pascal Frossard, président du jury
Dr. Jean-Marc Odobez, directeur de thèse
Dr. Xi Wang, rapporteur
Prof. Andreas Bulling, rapporteur
Dr. Sean Andrist, rapporteur

Lausanne, EPFL, 2025

To my family, friends and everyone who was a part of this journey

Acknowledgements

I would first like to thank my advisor, Jean-Marc. He has always been available for discussions and has provided invaluable feedback on writing, presenting, and approaching research in general. I am deeply grateful to him for giving me the opportunity to pursue a PhD.

I would also like to thank my jury members—Prof. Pascal Frossard, Dr. Xi Wang, Prof. Andreas Bulling, and Dr. Sean Andrist—for taking time out of their busy schedules to review my thesis. I really enjoyed our discussion during the oral exam and it helped spark many ideas about potential future directions.

Next, I want to thank my lab members, both past and present, who have been not only collaborators but also friends. Thank you, Angel, Rémy, Samy, Pierre, Arya, Michael, Mattia, Marco, and Mitch.

I am especially grateful to my friends at Idiap, who made this journey so enjoyable. Moving from India to Switzerland was a big change, but thanks to them, I was able to make it my new home. I have many fond memories of our lunches, hikes, trips, and other activities together. Special thanks to members of the Old-G Visp group, Tilak, Fabio, Andrei, Neha, Laurent, Eklavya, Chloé, Florian (German), Florian (Belgian), François, Julian, Suraj, Parvaneh, Teja, Amir, Zohreh, Pavan, Enno, Roberto, Pablo and Suhan. Thanks also to my newer friends at Idiap, Darya, Alina, Olena, Haruki, Shashi, Karl, Maxime, Sergio, Alejandra, Ajinkya, Laura, Luis, Mingchi, Ravinithesh and Alice. A big thank you as well to my EPFL friends, Parth, Dhruti and Megh. I apologize if I have missed any names, but I appreciate all of you.

I would also like to thank my manager and colleagues at Meta for giving me the opportunity to experience AI in industry, and generally life in the US. Thank you Calvin, Vamsi, Ruohan, Yijun, and Ishwarya. It was a great experience, and I learned a lot. Shoutout to Weipeng, Pranay, Apoorv, Sargam, Dhananjay, Lesly, Nikos, Amanda, Connie and Valts who made my time in the US so much more delightful and reminded me of just how special the Idiap bond is.

Last but not least, I want to thank my parents and sister for always being there as emotional support. They have seen me through ups and downs and have always done their best to support me despite the distance.

Lausanne, July 27, 2025 Anshul

Abstract

Gaze is a powerful cue for understanding attention, intention, and social interaction. This thesis presents a comprehensive study of gaze prediction in naturalistic settings, with a focus on developing models, datasets and evaluation protocols that go beyond spatial localization to capture the semantic and social dimensions of gaze behavior. We address key limitations in prior work and advance gaze prediction along several axes.

First, we introduce new datasets and annotations to support multimodal and multi-task learning. These include ChildPlay-audio, which augments child–adult interactions with speaking status; VSGaze, a unified benchmark with annotations for gaze following and social gaze tasks; and new semantic gaze annotations for the RLR-CHAT corpus to enable ego–exo gaze modeling. We also propose new evaluation protocols that extend beyond location-based metrics to assess semantic and socially grounded performance.

Second, we develop new architectures for gaze prediction. These include multimodal gaze following models that incorporate depth and pose; unified frameworks that jointly model gaze following and social gaze behaviors; and approaches to egocentric gaze estimation that leverage exocentric context. We further explore the use of foundation models and vision–language models to extract robust features for these tasks.

Finally, we demonstrate the feasibility of applying these models to child–adult interaction videos in the context of early language learning, where gaze plays a crucial role. Taken together, these contributions lay the groundwork for gaze models that are not only accurate but also semantically meaningful, capable of leveraging complementary contextual and task information, and applicable to real-world settings.

Keywords: gaze following, social gaze prediction, egocentric gaze prediction, multimodal behaviour understanding

Résumé

Le regard est un indice puissant pour comprendre l'attention, l'intention et l'interaction sociale. Cette thèse propose une étude approfondie de la prédiction du regard dans des environnements naturalistes, en mettant l'accent sur le développement de modèles, les jeux de données et de protocoles d'évaluation allant au-delà de la simple localisation spatiale afin de capturer les dimensions sémantiques et sociales du comportement visuel. Nous abordons les principales limitations des travaux antérieurs et faisons progresser la prédiction du regard selon plusieurs axes.

Premièrement, nous introduisons de nouveaux jeux de données et annotations pour soutenir l'apprentissage multimodal et multi-tâche. Il s'agit notamment de ChildPlay-audio, qui enrichit les interactions enfant–adulte avec des annotations de statut de parole ; de VSGaze, un benchmark unifié comportant des annotations pour la détection du regard et les tâches de regard social ; ainsi que de nouvelles annotations sémantiques sur le corpus RLR-CHAT pour permettre la modélisation du regard égocentrique et exocentrique. Nous proposons également de nouveaux protocoles d'évaluation qui vont au-delà des métriques de localisation pour évaluer les performances selon des critères sémantiques et socialement ancrés.

Deuxièmement, nous développons de nouvelles architectures pour la prédiction du regard. Cela inclut des modèles multimodaux intégrant la profondeur et la pose corporelle ; des cadres unifiés modélisant conjointement la poursuite du regard et les comportements sociaux du regard ; ainsi que des approches pour l'estimation du regard en vision égocentrique tirant parti du contexte exocentrique. Nous explorons également l'utilisation de modèles fondamentaux et de modèles vision–langage pour extraire des représentations robustes adaptées à ces tâches.

Enfin, nous démontrons la faisabilité de l'application de ces modèles à des vidéos d'interactions enfant–adulte dans le contexte de l'apprentissage du langage, où le regard joue un rôle essentiel. Pris ensemble, ces travaux développent les bases de modèles de regard à la fois précis, sémantiquement pertinents, capables d'exploiter des signaux contextuels et informationnels complémentaires, et utilisables dans des contextes réels.

Mots-clés : détection du regard, prédiction du regard social, prédiction du regard égocentrique, compréhension multimodale du comportement

Contents

1 Introduction

Human gaze is a subtle yet profound channel of nonverbal communication. From early social interactions in childhood to complex group dynamics in adult settings, where one looks conveys intentions, attention, and engagement. For example, the ability to interpret where others are looking—referred to as *gaze following*—is known to support early language acquisition (Brooks and Meltzoff, 2005, 2008). Gaze has also been shown to play a key role in regulating turn-taking behavior during conversations (Kendon, 1967). Conversely, abnormal gaze patterns are strongly linked to neuro-developmental disorders like Autism Spectrum Disorder (ASD) (Senju and Johnson, 2009; Mundy et al., 1990).

For this reason, the computer vision task of gaze estimation has found wide-ranging applications. These include human–robot interaction (Sheikhi and Odobez, 2015), where interpreting user gaze helps guide the robot's social behavior; consumer behavior analysis (Tomas et al., 2021), where gaze reveals attention in marketing contexts; and clinical diagnosis (Chong et al., 2017), where gaze behavior may serve as a marker for developmental conditions (see Figure 1.1). While gaze estimation encompasses a range of tasks, this thesis focuses on predicting a person's gaze target (*e.g.* 2D pixel location, semantic label) within the scene from both third-person and first-person views, as well as modeling social gaze behaviour.

In third-person settings, earlier approaches to predicting a person's gaze target often relied on specialized hardware, such as multiple calibrated cameras, or required prior knowledge of the scene layout (Otsuka et al., 2018; Bai et al., 2019). While these setups enabled accurate gaze estimation, they limited scalability and everyday applicability. More recent deep learning-based methods, notably introduced by Recasens et al. (2015), have shifted the paradigm toward inferring gaze targets directly from monocular RGB images, optionally augmented by other inputs such as depth and body pose. This approach adopts the term gaze following from developmental psychology and enables robust gaze target prediction in unconstrained scenes (Chong et al., 2020b; Fang et al., 2021; Lian et al., 2018), even under challenging conditions such as low visual clarity or visual clutter.

Beyond single-person gaze target prediction, researchers have also explored social gaze in

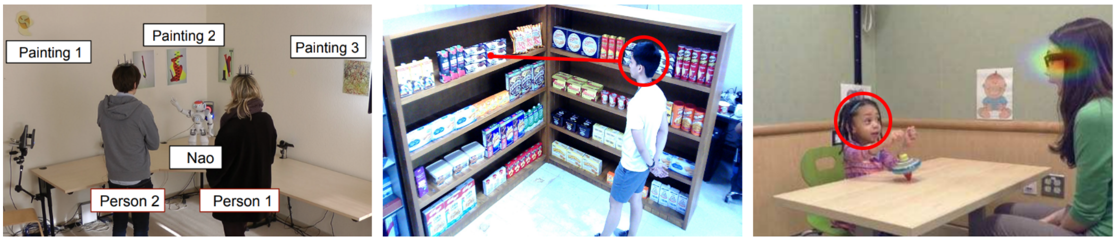


Figure 1.1: Gaze estimation has important applications ranging from human-robot interaction where the robot can serve as a guide in a museum (left, image from Sheikhi and Odobez (2015)), consumer behavior analysis to study shopping patterns (middle, image from Tomas et al. (2021)), and clinical diagnosis of autism (right, image from Chong et al. (2020b)).

multiparty settings. These models detect high-level interactive behaviors such as eye-contact (Marin-Jimenez et al., 2019) or shared attention (Fan et al., 2018)—signals that have direct relevance for downstream applications like ASD screening, where clinicians specifically assess such social gaze behaviors (Edition, 2013).

While the previous works have focused on third-person gaze analysis, first-person or egocentric gaze estimation focuses on predicting the gaze target of the individual wearing an AR/VR device. With the growing ubiquity of such AR and VR devices (Ray-Ban, 2025; Apple, 2025), egocentric gaze estimation is increasingly valuable for enabling immersive and intuitive interaction. This is typically addressed using eye-tracking, which infers gaze direction from close-up views of the eyes and maps it onto the observed scene to identify the gaze target. However, it comes with limitations such as additional hardware, calibration requirements and increased power consumption (Hansen and Ji, 2009). Recent efforts (Lai et al., 2023; Tavakoli et al., 2019; Huang et al., 2018) aim to address these challenges by predicting egocentric gaze targets directly from outward-facing RGB video.

In this thesis, we build upon these prior works in gaze estimation, and propose new methods, datasets, and evaluation protocols for gaze following, social gaze prediction and egocentric gaze estimation. We additionally apply these models in naturalistic child-centric settings.

The rest of this chapter proceeds as follows: we first present an overview of the different tasks and their associated challenges, we then outline our research objectives and finally summarize the main contributions of this thesis.

1.1 Addressed Tasks

In this section, we formally introduce the three vision tasks tackled in this thesis—gaze following, social gaze prediction, and egocentric gaze estimation—and summarize the specific challenges each of them poses.

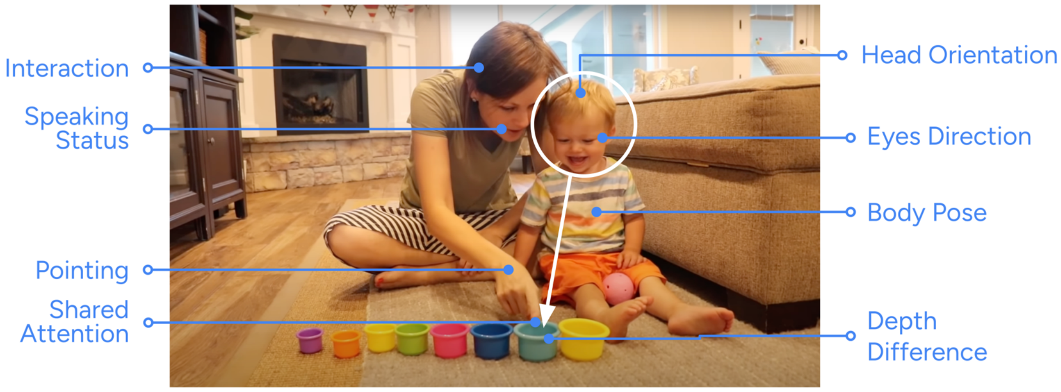


Figure 1.2: An illustration of the gaze following task, defined as predicting the 2D gaze target of a person in the scene. Accurate prediction requires estimating the person's visual field-of-view, identifying salient elements in the scene, and making a final decision about the gaze target—potentially incorporating additional context such as social or interactive cues.

1.1.1 Gaze Following

The computer vision task of gaze following aims to predict the 2D location of a person's gaze target in an image or video. An example is shown in Figure 1.2, where the child's gaze is directed toward the toy in front of him. Accurate gaze following requires the model to integrate multiple levels of visual and contextual reasoning. Broadly, the task can be decomposed into the following three components:

- **Visual Field-of-View:** When the person's eyes are visible, the model can directly estimate gaze direction. In more challenging scenarios—such as when the eyes are occluded or low-resolution—the model must rely on indirect cues like head orientation and body pose. This information along with scene geometry helps establish the person's visual field-of-view (FoV), *i.e.*, the region of the scene that is likely visible to them.
- **Scene saliency:** The model must identify salient elements in the environment—such as toys, furniture, or people—and determine which of these fall within the estimated FoV, and are thus visible targets.
- **Decision making:** With the FoV and salient scene elements identified, the model must then infer the gaze target. Most methods formulate this as a heatmap prediction, allowing for multimodal outcomes. In ambiguous cases with multiple plausible targets, the model may leverage additional context such as social cues (e.g., shared gaze, speaking status) or interactive signals (e.g., gestures, pointing) to disambiguate the target.

An important sub-task within gaze following involves predicting a binary **in vs. out-of-frame** gaze label. This label indicates whether the gaze target is visible within the image frame and helps exclude gaze following predictions when the target lies outside the scene.

It is worth noting that, while the gaze following task is general in its formulation—thereby enhancing its applicability—it also comes with certain limitations. First, the predicted gaze

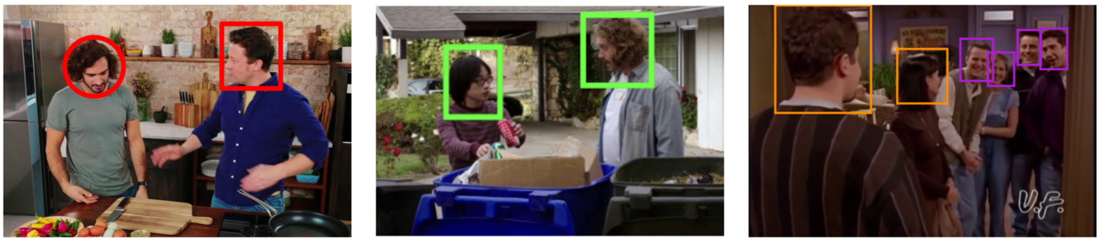


Figure 1.3: Social gaze prediction enables semantic interpretation of gaze behavior, which is often more valuable for downstream applications. We focus on three social gaze prediction tasks - looking at heads or *LAH* (left), looking at each other or *LAEO* (middle) and shared attention or *SA* (right).

point lacks semantic grounding, which can limit its utility in downstream applications where categorical or relational understanding (e.g., object of interest, interaction intent) is more informative. Second, the annotation process for gaze following is labor-intensive, as it requires precisely marking the gaze target within the image, often necessitating frame-by-frame effort.

1.1.2 Social Gaze Prediction

As discussed above, gaze following lacks semantic labels which are often more valuable for downstream tasks. When considering human interactions or human-robot interactions, a first level of semantic understanding is related to the analysis of social gaze patterns between people. For instance, in autism spectrum disorder (ASD) screening, behaviors like lack of eye-contact serve as vital diagnostic cues (Edition, 2013). Similarly, in human–robot interaction, gaze information helps identify who is speaking to whom and guides appropriate conversation behaviour (Admoni and Scassellati, 2017). In this thesis, we focus on three frame-level social gaze prediction tasks (illustrated in Figure 1.3):

- **Looking at Heads (LAH):** Indicates whether one person is looking at another person's head. Notably, the gaze does not need to be reciprocated.
- **Looking at Each Other (LAEO):** Captures instances of mutual gaze or eye-contact, where two individuals are simultaneously looking at each other.
- **Shared Attention (SA):** Refers to the scenario in which two or more individuals are attending to the same object or person in the scene.

These tasks are defined as atomic and frame-specific; they do not rely on a sequence of sub-events, distinguishing them from more temporally extended constructs such as joint attention (Fan et al., 2019). Specifically, we focus on predicting social gaze labels at the level of *person pairs*. For shared attention, pairwise predictions can subsequently be aggregated to identify the group of individuals attending to the same target.

Similar to gaze following, accurate prediction of social gaze labels requires an understanding of both scene-level and person-level cues—with particular emphasis on social interactions. This

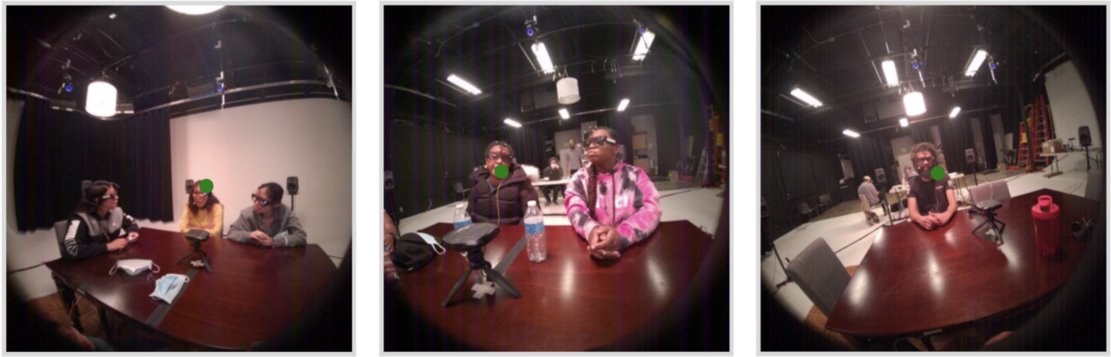


Figure 1.4: Egocentric gaze estimation focuses on predicted the 2D gaze target of the person wearing an AR/VR device, in the scene captured from the first-person perspective. In these samples from RLR-CHAT (Murdock et al., 2024; Yun et al., 2024), the egocentric gaze target marked with a green dot.

includes interpreting mutual orientation, gaze alignment, and contextual relationships among individuals. A key advantage of social gaze annotation is that it is generally less labor-intensive than gaze following, as it involves assigning discrete categorical labels to pairs or groups of individuals, rather than precisely marking gaze points within the image.

1.1.3 Egocentric Gaze Estimation

Egocentric gaze estimation focuses on predicting the 2D gaze target of the AR/VR device wearer within the observed scene (Figure 1.4). Compared to third-person tasks like gaze following and social gaze prediction, this setting introduces two major challenges:

- **Lack of direct visual gaze cues of the person of interest:** Since the camera wearer is not visible in the video, the model cannot observe the eyes or facial orientation to directly estimate gaze direction.
- **Camera motion:** Egocentric videos exhibit significant self-motion, in contrast to third-person videos that are typically captured from a static viewpoint. This introduces additional complexity, as the model must learn to compensate for the wearer's head and body movements.

To address the second challenge, models must instead rely on other signals. As for gaze following, scene-saliency and contextual cues from other people in the environment can offer valuable information about potential gaze targets. Dynamics signals, such as estimated camera motion or inertial measurements from IMUs, can also offer cues related to gaze shifts or head gestures. Additionally, due to natural oculomotor biases (Pelisson et al., 1988), gaze targets in egocentric views tend to be strongly concentrated near the image center, which can be exploited as a spatial prior.

A key advantage of this task formulation is the potential for large-scale automatic annotation using glasses with eye-tracking such as the Aria (Engel et al., 2023). This enables efficient data collection at scale—unlike gaze following or social gaze prediction, which typically require manual labeling of gaze targets or social interactions.

1.2 Research Objectives

Our research objectives build directly on the tasks outlined in the previous section, and target key limitations identified in existing methods, datasets, and evaluation protocols at the start of this PhD. Beyond advancing the state-of-the-art, we also explore the applicability of the developed models in real-world settings.

1.2.1 Leverage Multimodal Information for Improving Gaze Prediction

Contemporary gaze methods rely predominantly on RGB input, hence, the model must implicitly learn all the relevant cues necessary to solve the task effectively. However, this is challenging given the limited size of existing datasets.

Our first objective is therefore to *explicitly* incorporate relevant cues by leveraging additional modalities that can be obtained at little extra cost—monocular depth, body-pose skeletons from strong pre-trained models, or spatial audio from microphones—so as to resolve difficult cases while remaining deployable “in the wild.” We will investigate:

- fusion strategies that integrate standard RGB input with modalities such as pose, depth, and audio within a single gaze-prediction architecture;
- geometrically consistent 3D scene reasoning (depth, point clouds, explicit gaze cones) that lets the network discard targets outside the viewer’s 3D field of view;
- the use of recent vision–language models (VLMs) as a source of gaze relevant contextual cues such as people’s pose, gestures and interactions, and a model allowing their exploitation towards gaze tasks;
- privacy-preserving variants that rely only on anonymised modalities (pose and depth) when identifiable imagery cannot be used.

1.2.2 Develop a Unified System for Modelling Gaze-Based Social Interactions

Traditionally, gaze following, LAH, LAEO, and SA are tackled by separate networks, each trained with its own loss function. This siloed approach overlooks the strong dependencies among these tasks and incurs unnecessary computational overhead when multiple models must be run in parallel.

Our second objective is therefore to devise a *single* framework that jointly reasons about gaze targets and social relations in third-person views. Specifically, the system should:

- process all people in the scene in one forward pass, avoiding repeated per-person inference;
- capture person–person interactions and their temporal evolution, enabling the network to reason about mutual gaze, shared attention, and conversational dynamics;
- jointly train across all tasks to leverage complementary information and enable unified prediction within a single multi-task framework;
- seamlessly integrate the multimodal streams introduced in the previous objective so that the same architecture can exploit whichever modalities are available at deployment.

1.2.3 Learn Ego-Exo Gaze Representations

Egocentric (ego) gaze estimation and third-person (exo) gaze following are typically studied in isolation, yet the two views are naturally linked in multi-party interactions. When the AR/VR device wearer's view contains several plausible targets, knowing where other people in the scene are looking can resolve the ambiguity. For instance, shared attention from exocentric observers can be a strong cue.

This objective explores whether such exocentric information can be captured without extra manual labels. Specifically, we will investigate:

- self-supervised approaches for learning exocentric gaze representations, with the goal of improving egocentric gaze estimation in social settings;
- novel probing tasks to evaluate the learned representations.

1.2.4 Create New Datasets and Evaluation Protocols

Most publicly available datasets tackle only one facet of gaze—for example, gaze following or LAH—making them ill-suited for training and assessing unified models that predict multiple gaze behaviours at once. Additionally, they often lack annotations for contextual signals that can inform gaze prediction. Our next objective is therefore to develop new datasets to support these objectives. Concretely, we will:

- merge and expand current datasets by adding missing labels (e.g. LAH, LAEO, SA), thereby increasing both scale and scene diversity;
- enrich the data with complementary signals such as speaking status so that multimodal models can be trained and evaluated.

Alongside the data effort, we will design evaluation protocols that better reflect downstream needs. Traditional gaze following metrics judge heatmap or pixel-level accuracy but ignore *what* or *whom* is being looked at. Conversely, some existing metrics such as those for shared attention suffer from drawbacks as they overlook which specific people share attention. We will introduce metrics that capture these semantic aspects, as well as provide a more complete picture of model performance.

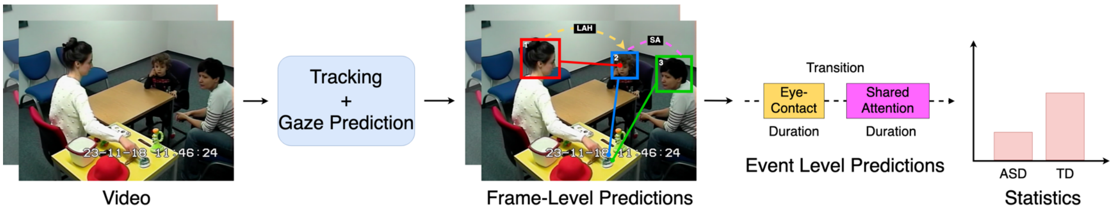


Figure 1.5: A potential pipeline for applying gaze prediction models in ASD settings. We can first extract frame-level behaviors (e.g., gaze points, social gaze), followed by event-level patterns (e.g., durations, transitions). These predicted behaviors can be compared with clinician annotations to compute relevant behavioral statistics.

1.2.5 Study Applications of Gaze Models in Naturalistic Settings

The value of more accurate gaze modeling ultimately lies in its real-world applications. We therefore turn to child-centred domains such as autism-diagnostic sessions and early language learning where gaze is a key behavioural marker. Notably, this PhD was supported by the SNSF Sinergia project *AI4Autism* (Tafasca et al., 2023a), which aims to leverage automated tools—including computer vision methods—for early screening and detailed phenotyping of autism.

Existing studies in these areas often rely on controlled laboratory protocols or specialised hardware, restricting investigations to small cohorts. By contrast, the methods developed in this thesis operate directly on unconstrained video and audio, allowing researchers to analyse much larger, more varied datasets collected in the wild. In doing so, they open the door to ecologically valid studies of joint attention, eye-contact patterns, and other gaze-based behaviours that matter for both diagnosis and developmental science.

In Figure 1.5, we illustrate a potential pipeline for analyzing gaze behaviors of children during autism diagnostic sessions. The automatically predicted gaze patterns can be correlated with clinician annotations to compute relevant behavioral statistics.

1.3 Contributions

The contributions of this thesis are organised around the research objectives set out in the preceding section.

(O1) Multimodal gaze following (§1.2.1).

- *Multimodal models.* We introduce two gaze following architectures that fuse RGB with complementary streams. The first, published in Gupta et al. (2022), combines predicted depth and body pose with RGB and can even operate on depth and pose alone for privacy-sensitive use cases. The second, published in Tafasca et al. (2023b), performs explicit geometric reasoning: it combines the predicted scene point cloud with the estimated 3D gaze vector to construct a 3D field-of-view cone that highlights scene regions physically visible to the person. Both achieve state-of-the-art results and are detailed in Chapter 5.

- *Vision–language cues.* In Gupta et al. (2024c) we show that large vision–language models can supply rich contextual cues—pose, gestures, social interactions—that, when injected into a gaze following network, boost cross-dataset generalisation. This work is detailed in Chapter 7.

(O2) Unifying gaze following and social gaze prediction (§1.2.2). We design transformer-based systems that, in a single pass and for any number of people, predict both the gaze target and three frame-level social labels—LAH, LAEO, and SA. The first model, published in Gupta et al. (2024a), starts from a frozen gaze following backbone and adds a graph module that reasons over person interactions to infer social gaze. Building on these insights, the second model, published in Gupta et al. (2024b), models person interactions and their temporal evolution at multiple levels of the architecture. It is trained end-to-end to predict both gaze targets and social labels, and can accept several of the multimodal cues introduced in (O1). Both architectures are presented in Chapter 6.

(O3) Ego–exo representation learning (§1.2.3). We introduce novel techniques that align egocentric (ego) and third-person (exo) gaze features captured during multi-party conversations. The exo features are learned via self-supervision and help inform the prediction of the egocentric gaze target by providing information about where others are looking, enabling the model to resolve ambiguous, multi-target scenes and improve single-frame egocentric gaze prediction. This work is currently under submission and is detailed in Chapter 8.

(O4) Datasets and protocols (§1.2.4).

- *VSGaze.* We merge and extend multiple public datasets to create VSGaze, the largest corpus to date with unified annotations for gaze following, LAH, LAEO, and SA. This dataset supports multi-task learning and evaluation for unified models. The work was presented in Gupta et al. (2024a,b) and detailed in Chapter 3.
- *ChildPlay-audio.* We extend the ChildPlay dataset (Tafasca et al., 2023b) with frame-level annotations for speaking status. These labels enable the use of speaker information as an additional modality in the multimodal models described in Objectives O1 and O2. This contribution is also described in Chapter 3.
- *New metrics.* We propose evaluation protocols and metrics for assessing semantic gaze following, particularly for the social gaze tasks LAH, LAEO, and SA. In particular, we introduce a new pairwise evaluation protocol for SA that explicitly identifies the individuals involved in shared attention, addressing a key limitation of prior work. These contributions were presented in Gupta et al. (2024a,b) and discussed in Chapter 4.

(O5) Applications in naturalistic child studies (§1.2.5). We apply our developed models to child language learning settings to evaluate their ability to automatically capture joint attention behaviours in naturalistic interactions. This work demonstrates the potential of gaze-based models in supporting large-scale, ecologically valid behavioural analysis. It was published in Dickerman et al. (2025) and is detailed in Chapter 9.

1.4 Thesis Plan

In summary, this thesis is organized as follows:

- **Chapter 2** provides an overview of related work.
- **Chapter 3** details the datasets used throughout the thesis.
- **Chapter 4** introduces the metrics and evaluation protocols.
- **Chapter 5** presents our proposed multimodal gaze following models that leverage depth and pose information.
- **Chapter 6** introduces our unified architectures that jointly address gaze following and social gaze prediction.
- **Chapter 7** explores vision–language models for extracting gaze-relevant contextual cues and integrating them into a gaze following model.
- **Chapter 8** explores self-supervised approaches for learning exocentric features to improve egocentric gaze estimation performance.
- **Chapter 9** investigates the application of our developed models in naturalistic child language acquisition settings.
- **Chapter 10** concludes the thesis by summarizing contributions, and discussing limitations and promising future directions.

2 Related Work

In Chapter 1, we introduced the gaze tasks addressed in this thesis: gaze following (Section 1.1.1), which predicts the 2D gaze target of a person in an image or video; social gaze prediction (Section 1.1.2), which infers frame-level social interaction labels such as eye contact and shared attention; and egocentric gaze estimation (Section 1.1.3), which aims to estimate the 2D gaze target of an AR/VR device wearer from their first-person view.

In this chapter, we survey prior work across these domains. We begin with early methods for predicting a person's gaze target, referred to as their visual focus of attention (VFOA), that relied on controlled lab setups (Section 2.1). We then review modern deep learning approaches for gaze following (Section 2.2) that allow gaze target prediction "in the wild". We then discuss models for social gaze prediction (Section 2.3), egocentric gaze estimation (Section 2.4), and approaches that bridge first- and third-person views through ego–exo representation learning (Section 2.5).

We also discuss the recent emergence of vision–language models (VLMs) and prompting strategies that offer new ways to extract contextual cues relevant to gaze understanding (Section 2.6). Finally, we highlight applications of gaze modeling in child-centered domains, particularly autism diagnosis and early language development (Section 2.7).

2.1 VFOA Estimation

These works typically formulated the task as a two-step process: (1) estimate the person's gaze direction, and (2) map it to a predefined set of candidate targets to determine their visual focus of attention (VFOA).

Since accurately estimating gaze direction is difficult, particularly when the eyes are occluded, initial approaches used head pose as a proxy. Probabilistic inference techniques such as Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), and Dynamical Bayesian Networks (Stiefelhagen et al., 1999; Otsuka et al., 2006; Ba and Odobez, 2008) were commonly used to estimate VFOA from sequences of head pose. These models often incorporated

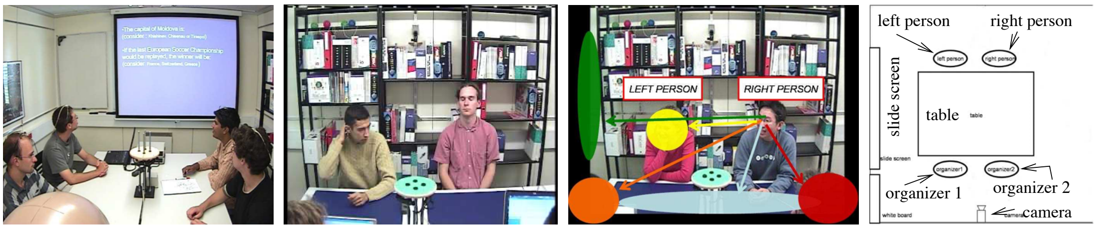


Figure 2.1: Early approaches for VFOA estimation relied on specialized hardware, such as calibrated camera setups or pre-specified scene layouts. In this example from Ba and Odobez (2008), both object and person locations are predefined within the environment.

additional contextual cues (Gorga and Otsuka, 2010) that served as priors for VFOA estimation, including speaking status, speech semantics (Sheikhi and Odobez, 2015; Otsuka et al., 2018), and inter-personal dynamics such as joint VFOA modeling across participants (Ba and Odobez, 2011; Masse et al., 2018).

With advances in gaze estimation, even simple frame-based geometric approaches have proven effective for VFOA prediction (Yücel et al., 2013). More recently, deep learning architectures—including convolutional neural networks (CNNs) and recurrent neural networks (RNNs)—have been introduced, leading to further improvements in performance (Siegfried and Odobez, 2021).

Despite these developments, most existing methods depend on controlled setups—either requiring calibrated hardware or prior knowledge of the scene structure, such as object locations (Figure 2.1). This reliance limits their applicability in unconstrained, real-world environments.

2.2 Gaze Following

As discussed in Section 1.1.1, gaze following can be decomposed into three steps: 1) estimating the visual field-of-view (FoV), 2) identifying salient elements within the scene, and 3) making a decision about the likely gaze target.

Building on this framework, Recasens et al. (2015) proposed the first deep learning-based model for gaze following. Their architecture uses a CNN with two branches (Figure 2.2): one branch processes the overall scene to identify salient elements, while the other analyzes a cropped image of the target person's head to predict a gaze mask that approximates the FoV. The output is a heatmap over the scene indicating probable gaze target locations. Similar to works in pose estimation (Tompson et al., 2015), such a heatmap representation enables multimodal predictions, which is essential for capturing the inherent ambiguity of gaze target inference in many situations.

Subsequently, many studies adopted a similar two-branch architecture. For instance, Chong et al. (2018) extended the model to also determine whether a person's gaze lies within or

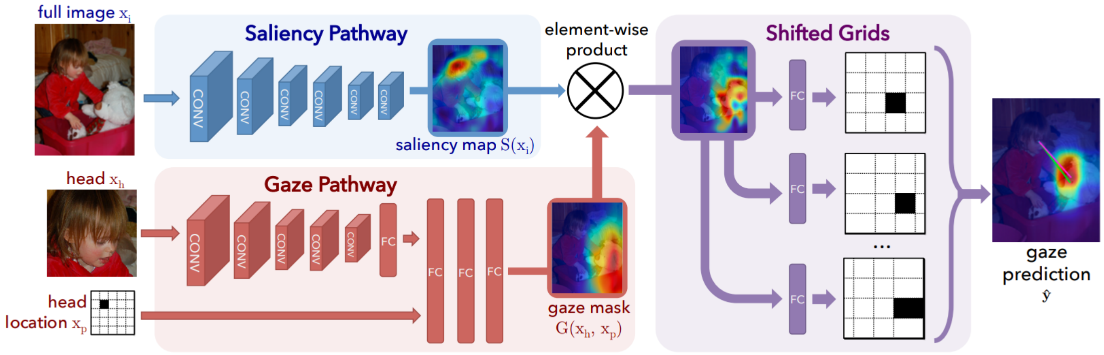


Figure 2.2: The two-branch architecture for gaze following proposed by (Recasens et al., 2015) which was subsequently adopted by many works. The top branch analyzes the entire scene for salient elements, while the bottom branch processes the head crop to model gaze direction and head orientation. The output is a heatmap over the scene, highlighting potential gaze targets.

beyond the image boundaries, and Lian et al. (2018) predicted a 2D gaze vector from the head-branch features to generate explicit “gaze cones” that capture the 2D field-of-view, which were then fused with the original image for inference. Drawing on human pose estimation methods, Zhao et al. (2020) proposed predicting the line of sight along with an attention “landmark”. More recent works expanded on these ideas by incorporating multimodal inputs and modern transformer-based architectures to enhance accuracy.

2.2.1 Multimodal Approaches

As discussed in Research Objective 1.2.1, it is difficult for gaze following models to capture all relevent cues for accurate prediction from RGB input alone given the limited size of existing datasets. For this reason, a number of studies have examined the use of multimodal information obtained from pre-trained models for improving gaze following.

Guan et al. (2020) employed a person’s body pose to strengthen the gaze branch in cases where facial information is unavailable. Likewise, Nan et al. (2021) combined task-driven top-down attention with bottom-up signals (optical flow and pose) to infer the gaze target. Several methods (Fang et al., 2021; Jin et al., 2022; Bao et al., 2022) leveraged predicted disparity maps for identifying whether particular objects lie within the person’s 3D line of sight, thereby disambiguating potential gaze targets. However, disparity maps produce geometrically inconsistent point clouds, limiting the extent of 3D reasoning. These methods also do not match the predicted 3D gaze with the inferred point cloud to construct explicit 3D Field-of-View maps. Furthermore, the combined use of depth and body pose has not been investigated. In a related direction, Hu et al. (2022a) performed 3D gaze target prediction using RGB-D images, but this thesis focuses on standard RGB-only inputs.

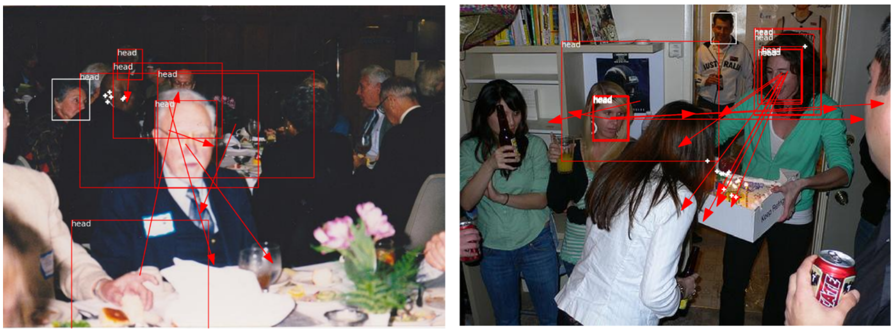


Figure 2.3: Sample predictions from (Tonini et al., 2023). Ground-truth annotations are in white, and model predictions are in red. In several cases, the model either misses certain heads or generates multiple bounding boxes for the same individual, each with a different gaze target.

2.2.2 Transformer and Multi-Person Approaches

The works discussed above focused on single-person gaze following. However, in multi-person scenes, this approach requires reprocessing the entire scene separately for each individual, leading to significant computational overhead. Moreover, such models cannot take advantage of person–person interactions, which can provide valuable context for accurate gaze prediction.

Jin et al. (2021) proposed the first multi-person gaze following architecture that uses a CNN-based scene encoder to extract a scene-level representation, which is subsequently combined with features from each individual's head crop. More recent work (Ryan et al., 2025; Song et al., 2024) has followed a similar approach while replacing the CNN encoder with a stronger vision transformer (ViT) pre-trained with DINOv2 (Oquab et al., 2023). Notably, Ryan et al. (2025) demonstrated state-of-the-art performance by using a frozen encoder and injecting head position embeddings, followed by a lightweight decoder to predict the final gaze heatmap. While these methods lower computational cost, they do not explicitly model interactions among different people, since each head is still processed independently.

Another line of research (Tu et al., 2022; Tonini et al., 2023) employs DETR-like transformer architectures (Carion et al., 2020) for multi-person gaze following. These approaches take a single image as input and jointly estimate the bounding boxes and gaze targets of all individuals in the scene. Although this design can implicitly capture person-person interactions, it also introduces certain limitations. First, their models do not reach state-of-the-art head detection performance and can miss heads, especially for individuals who are partially occluded or facing away. Second, as noted in Ryan et al. (2025), these models typically generate a fixed number of bounding box predictions, which may exceed the actual number of heads. Consequently, multiple predictions can arise for the same person, each associated with a distinct gaze target. However, no solution is provided for resolving these duplicate outputs, which is problematic when applying the model in real-world settings. Figure 2.3 shows examples of

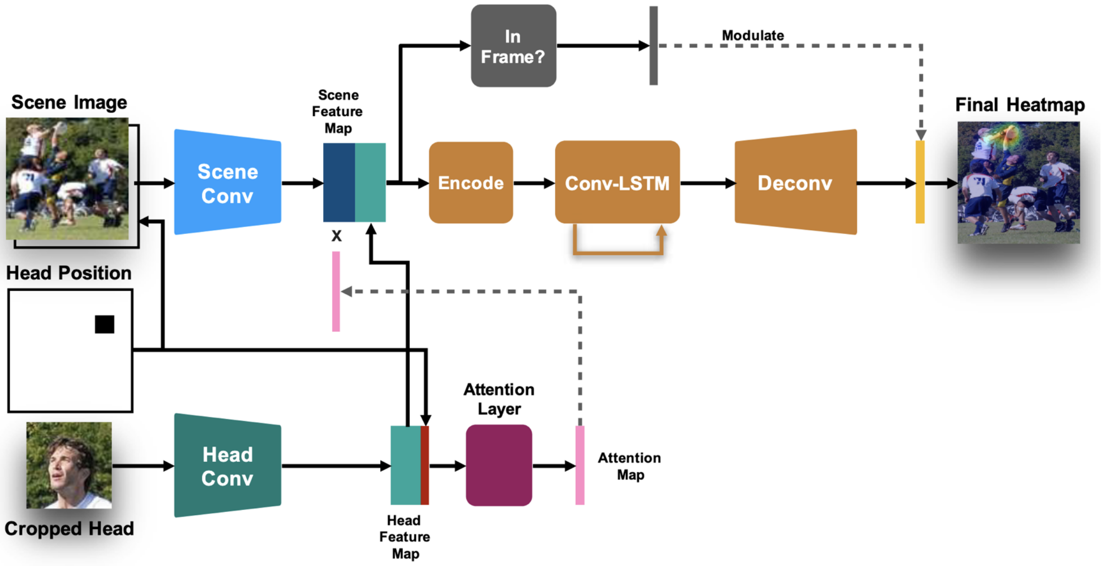


Figure 2.4: Chong et al. (2020b) were the first to explore the use of temporal information for enhancing gaze following by applying an LSTM over bottleneck features from the scene and head branches. However, this approach yielded only marginal improvements compared to their static version.

these issues in Tonini et al. (2023).

2.2.3 Temporal Approaches

Temporal information has the potential to significantly enhance gaze following. At the scene level, moving objects or people often serve as strong indicators of saliency. Additionally, due to the close coordination between eye and hand movements (Johansson et al., 2001), recognizing gestures such as object manipulation can provide informative cues about gaze targets. At the head level, prior work on 3D gaze direction estimation (Kellnhofer et al., 2019; Vuillecard and Odobez, 2025) has demonstrated that modeling head and gaze dynamics over time leads to more accurate predictions.

However, exploiting temporal information for gaze following remains underexplored. To our knowledge, only two previous works have investigated its benefit. (Chong et al., 2020b) introduced a convolutional LSTM at the bottleneck layer of their network (Figure 2.4), and (Miao et al., 2023) leveraged a temporal attention mechanism to aggregate frame-level features. However, both approaches only achieve marginal gains over their static counterparts, underscoring the difficulties in leveraging temporal dynamics for this task. Moreover, neither method explicitly models 2D gaze direction over time, nor do they address multi-person scenarios.

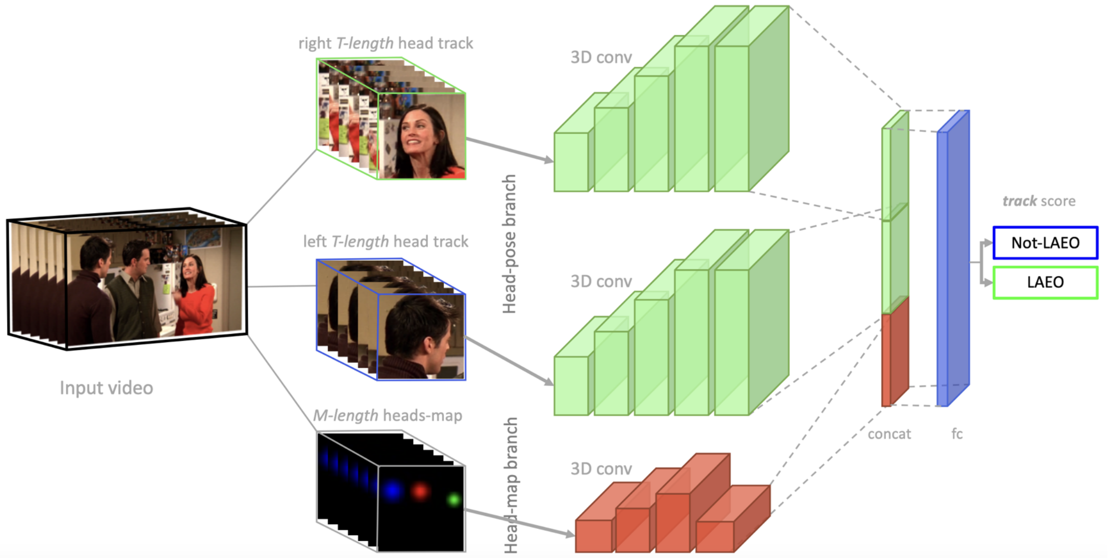


Figure 2.5: Several prior works for social gaze estimation proposed task-specific architectures, such as this one from Marín-Jiménez et al. (2021) for LAEO. It also only supports processing pairs of people, requiring multiple forward passes for all pairs of people in the scene.

2.3 Social Gaze Prediction

As discussed in Section 1.1.2, we focus on three frame-level social gaze tasks - looking at heads (LAH), looking at each other (LAEO) and shared attention (SA). A number of works have tackled social gaze prediction, particularly LAEO and SA. These typically follow one of two approaches:

- *Task-specific methods:* These target individual tasks like LAEO or SA (Marin-Jimenez et al., 2019; Sumer et al., 2020). While effective for their respective tasks, they cannot generalize to other gaze related tasks. As a result, a suite of task-specific models would be needed to extract different social gaze labels.
- *Post-processing methods:* These operate on gaze point, gaze vector, or head pose predictions (Chong et al., 2020b; Cantarini et al., 2021). While the post-processing protocols can be adapted to target a range of social gaze tasks, they typically achieve lower performance than task-specific models as they are not explicitly trained for social gaze prediction.

In the following sections, we review representative works from both categories, as well as well as the few works that attempt multi-task modeling of social gaze.

2.3.1 Looking at Each Other

For LAEO, most approaches rely on processing head crops to extract gaze directional information, which is subsequently combined with 2D or inferred 3D geometry to classify LAEO

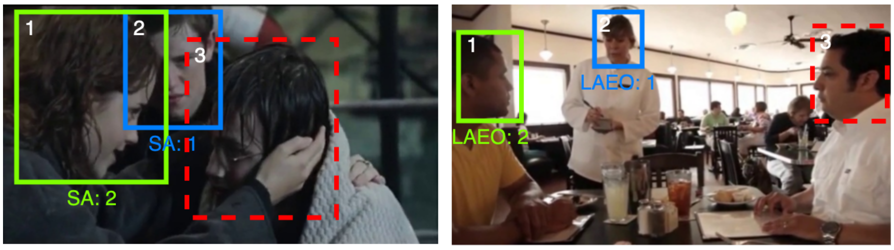


Figure 2.6: Samples from the VACATION dataset (Fan et al., 2019). In both cases, Person 3 is missed because the associated individual is already annotated with a social gaze 'state'. Similarly, the method proposed in the paper predicts only a single social gaze state per person, preventing the modeling of simultaneous social gaze behaviours.

(Marin-Jimenez et al., 2019; Doosti et al., 2021; Marín-Jiménez et al., 2021; Cantarini et al., 2021) (see Figure 2.5). However, these methods typically treat each pair of individuals independently and focus exclusively on head crops, ignoring global scene context and limiting extensibility to social gaze tasks such as SA. A recent paper (Guo et al., 2022) introduced an encoder-decoder transformer to jointly predict heads and LAEO labels with good results, yet encounters similar drawbacks.

2.3.2 Shared Attention

For shared attention, the first method to address it in unconstrained environments was Fan et al. (2018), which framed the problem as two sub-tasks: (i) classifying whether SA occurs in a frame, and (ii) localizing the target of attention, often assuming only one such target exists. Their method combined 2D gaze cones of each person with a heatmap of object region proposals, while other works (Sumer et al., 2020) directly inferred SA from raw images. More recent approaches have leveraged gaze following heatmap predictions for all people in the scene (Chong et al., 2020b; Tu et al., 2022), thereby enhancing performance.

Nevertheless, the task formulation proposed by Fan et al. (2018) faces two key issues: (i) it cannot differentiate multiple SA instances within the same frame, and (ii) it cannot identify which specific individuals share attention.

2.3.3 Multi-Task Prediction

As discussed in Research Objective 1.2.2, developing multi-task frameworks for social gaze prediction can provide not only computational advantages but also potentially lead to better overall performance by leveraging complementary information across tasks. However, while many studies address social gaze prediction, few have tackled multiple social gaze tasks simultaneously.

One notable exception is Fan et al. (2019), which examined gaze communication activities (including LAEO and SA) in a graph-based framework. However, their method infers only a

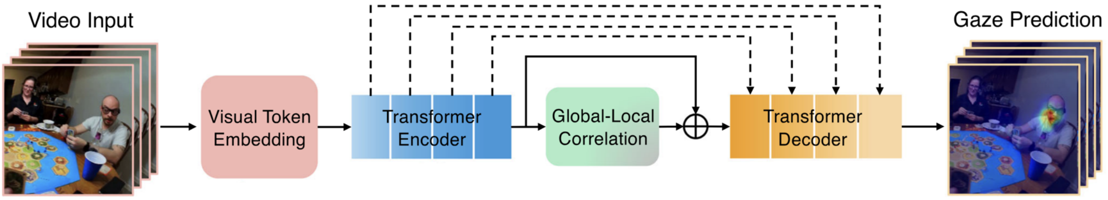


Figure 2.7: Typical architectures for egocentric gaze estimation process a sequence of video frames to predict a heatmap over the scene highlighting potential gaze targets. Illustrated here is the transformer-based architecture from Lai et al. (2023).

single gaze "state" per person, disallowing simultaneous LAEO and SA (examples in Figure 2.6). Also, it does not specify which other person is involved in the social gaze interaction. Another approach (Chang et al., 2023), extends a gaze following style two-branch architecture to analyze dyadic communication, but it requires $\frac{N_p!}{(N_p-2)!}$ forward passes to process all pairwise interactions for N_p people, leading to high computational cost. Moreover, neither method addresses the gaze following task as part of an integrated framework.

2.4 Egocentric Gaze Estimation

There have been several works on egocentric gaze estimation using deep learning (Huang et al., 2018, 2020; Tavakoli et al., 2019; Lai et al., 2023; Li et al., 2021; Thakur et al., 2021). Similar to approaches in gaze following, these methods typically generate a heatmap over the scene to enable multimodal predictions. Furthermore, like recent advances in gaze following, they recognize the value of incorporating multimodal information to improve performance. For instance, motivated by the strong coordination between gaze and hand movements (Johansson et al., 2001), Huang et al. (2020) demonstrated the benefit of jointly modeling egocentric gaze with hand actions. On the other hand, Thakur et al. (2021) incorporated IMU measurements to improve gaze estimation. As discussed in Section 1.1.3, even in the absence of direct visual cues for gaze direction, proxies such as head movement and orientation from such sensors can be informative. More recently, Lai et al. (2023) introduced the first transformer-based model for this task (Figure 2.7), achieving state-of-the-art results by leveraging more robust pre-trained representations and learning global-local scene relations. Notably, all of these works process a sequence of frames. However, hardware limitations of existing device may not allow storing a frame buffer.

Another interesting line of work investigated social gaze from egocentric views. In particular, Soo Park and Shi (2015) leveraged data from multiple egocentric views to reconstruct the 3D scene and predict joint attention based on social formation. However, it assumes access to the egocentric streams of all individuals in the scene during inference, which may not always be feasible.

A related task is that of egocentric gaze anticipation first explored by Zhang et al. (2017). This

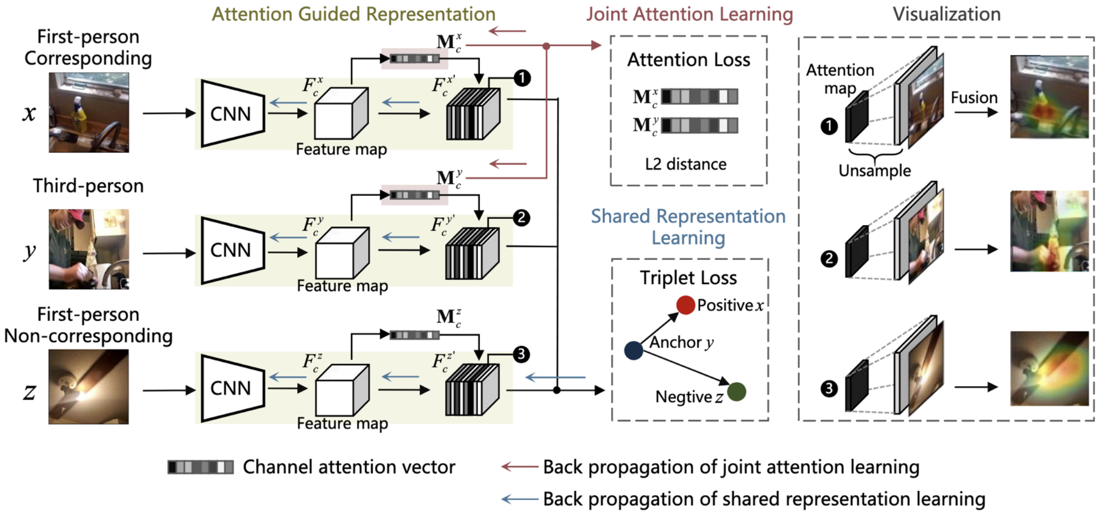


Figure 2.8: The architecture from Yu et al. (2020) for learning view-invariant features. It uses a triplet loss to maximize similarity between representations from corresponding ego-exo frames, while minimizing similarity between representations from non-corresponding frames.

task aims to predict the egocentric gaze target in future frames. More recent work by Lai et al. (2025) performed autoregressive egocentric gaze anticipation, leveraging single-channel audio information for improved performance. Exploring the benefit of multi-channel audio, which can provide spatial localization of the sound source, remains unexplored.

2.5 Ego-Exo Representation Learning

Despite significant progress in egocentric gaze estimation, models can still fail in scenes with several plausible targets (e.g., multiple people clustered near the image centre). As discussed in Research Objective 1.2.3, such ambiguity can potentially be resolved if third-person (exo) gaze following information from surrounding actors is also available. For instance, when several bystanders are simultaneously looking at the same external object.

However, research in bridging ego and exo views is relatively nascent, with existing studies falling into two broad categories. The first centres on identity matching: models are trained with contrastive objectives to align the embedding of the camera-wearer in the egocentric view with the corresponding actor in a third-person view (Fan et al., 2017; Xu et al., 2018; Wen et al., 2021). These methods essentially answer the question “Who in the room is holding the camera?” and do not tackle gaze or other behavioural cues.

The second category aims to learn view-invariant features. Here egocentric and exocentric clips that depict the same action are mapped to a shared latent space, encouraging cross-view consistency (Sigurdsson et al., 2018; Yu et al., 2020; Xue and Grauman, 2023). The network of Yu et al. (2020) (Figure 2.8), for example, shows that such alignment can benefit several

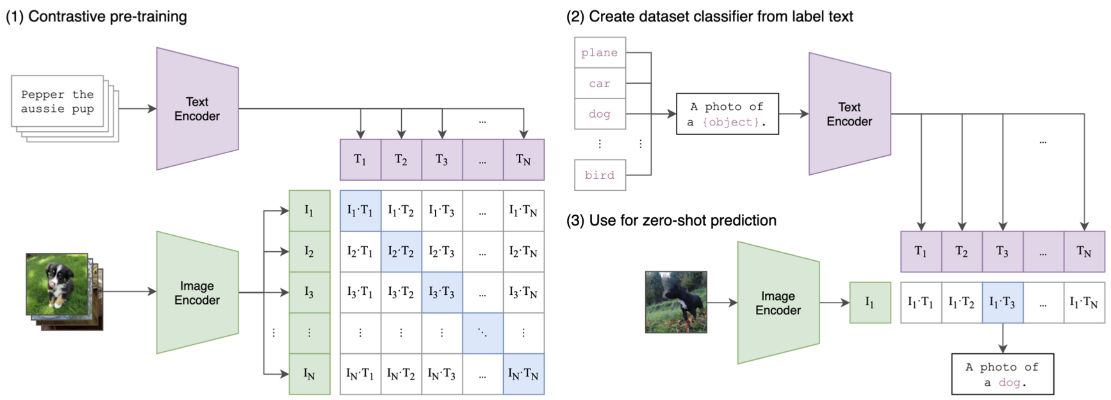

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Figure 2.9: CLIP (Radford et al., 2021) uses contrastive training to align image and text embeddings (1). At inference time, zero-shot classification is enabled by constructing a set of text prompts (2), computing their embeddings, and selecting the one that has the highest similarity with the image embedding (3).

downstream tasks, including coarse 3D gaze-angle prediction. However, they only considered a single visible actor, so neither multi-person gaze following nor precise egocentric target localisation is addressed.

Moreover, in all of these prior works the exocentric stream comes from a static third-person camera; aligning an egocentric feed with another moving egocentric view—arguably the more natural configuration for social interaction—remains largely unexplored and presents a significantly harder problem.

2.6 Vision–Language Models

In Research Objective 1.2.1, we discussed exploiting recent Vision–Language Models (VLMs) as a source of gaze relevant contextual cues for improving gaze estimation performance. VLMs gained significant attention with the introduction of CLIP (Figure 2.9) (Radford et al., 2021), which learns multi-modal representations from image–text pairs. CLIP demonstrated impressive zero-shot performance across standard image classification tasks. Subsequent models such as BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023a) introduced improvements including additional training objectives, curated datasets, and better caption generation via caption filtering. BLIP-2, in particular, combines a frozen image encoder with a Querying Transformer (Q-Former), allowing for more nuanced visual representation learning.

These models have shown strong results even on video tasks like action recognition (Radford et al., 2021) and text-to-video retrieval (Li et al., 2022), despite not explicitly modelling temporal information. However, their ability to localise gaze-relevant actions or cues in images remains largely unexplored.

Although some efforts have adapted image-based VLMs for video tasks, they face challenges.

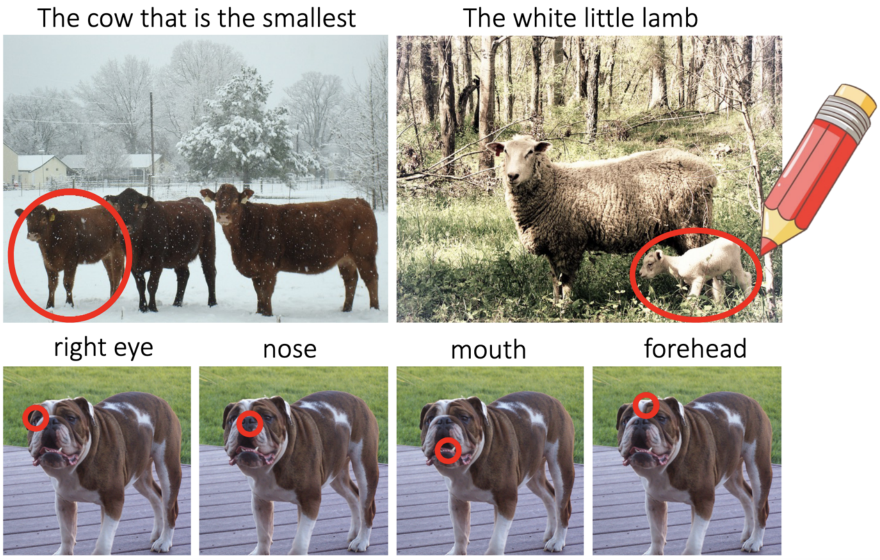


Figure 2.10: Shtedritski et al. (2023) explored visual prompting methods for VLMs, demonstrating that a red circle drawn over the area of interest provided the best results.

The limited availability of large-scale video–text datasets restricts generalisation (Xu et al., 2021), and adapting image VLMs to video often reduces zero-shot performance (Wang et al., 2021a; Ju et al., 2022). These models are also computationally demanding and less suited for static gaze following datasets.

More recently, VLMs have integrated Large Language Models (LLMs) to generate textual outputs. LLMs demonstrate strong world-modelling abilities, including basic reasoning about agents, beliefs, and actions (Andreas, 2022), as well as commonsense and mathematical reasoning (Chowdhery et al., 2023). This suggests they could potentially capture complex human–object and human–human relationships relevant to gaze estimation. VLMs have also benefited from in-context learning (ICL) (Brown et al., 2020), where a few demonstrations are provided at inference without modifying model weights. Recent VLMs leveraging LLMs (Alayrac et al., 2022; Tsimpoukelli et al., 2021) show improved performance with ICL, although whether these benefits extend to capturing gaze-related contextual cues remains an open research question.

2.6.1 Visual and Textual Prompting

Extracting gaze relevent contextual cues using VLMs would require visual prompting to direct the model to focus on the person of interest, as well as textual prompting to better describe the specific cue to be extracted.

Shtedritski et al. (Shtedritski et al., 2023) explored visual prompting methods for CLIP, comparing cropping the area of interest to drawing a red ellipse around it. The latter performed better for keypoint localisation and referring expression comprehension, with further gains when the background was blurred or greyed out. However, such approaches have not yet been tested for action or cue recognition tasks.

On the textual side, the original CLIP paper (Radford et al., 2021) showed that prompt engineering, such as using the format "a photo of a {label}", improved performance over using simple label text alone (Russakovsky et al., 2015). Combining multiple prompts (ensembling) further boosted results. More recent works (Zhou et al., 2022b,a) introduced learnable prompts fine-tuned for specific tasks. While effective when adapted to task-specific data, this approach does not support true zero-shot transfer and still lags behind manual prompt engineering for unseen classes (Zhou et al., 2022a). Applying visual and textual prompting techniques to gaze-related tasks remains a promising direction for future work.

2.7 Gaze Prediction Applications

In Section 1, we discussed the various applications of gaze estimation. However, as outlined in Research Objective 1.2.5, this thesis focuses specifically on applications towards child-centric settings, where understanding gaze behaviour is essential for studying social development and supporting diagnostic processes. In particular, we focus on two areas where automatic gaze estimation can provide valuable insights at scale: analyzing the components of joint attention in naturalistic interactions and detecting social gaze signals relevant to Autism Spectrum Disorder (ASD) diagnosis.

2.7.1 Analyzing the Components of Joint-Attention

Joint attention is an interactional framework defined by shared focus and interaction about a third object or activity of interest (Bakeman and Adamson, 1984). Gaze behaviour is a central component in joint attention, and variations in gaze patterns have been linked to differences in learning outcomes (Abney et al., 2020).

Existing research on joint attention in child-centric settings has largely relied on specialised eye-tracking equipment or manual coding, typically conducted in controlled environments. Most studies constrain some aspect of the interaction—whether the identities of participants, the assigned tasks, or the physical setting. For example, Abney et al. (2020) studied mother–infant dyads in a laboratory setting with a predefined set of toys and instructed free play, a common design described as “naturalistic free-flowing interaction.” Similar dyadic setups, typically involving mothers and infants, dominate both early studies (Bakeman and Adamson, 1984; Tomasello and Farrar, 1986) and recent work (Suarez-Rivera et al., 2022), despite findings that joint attention with fathers may contribute even more to vocabulary growth (Ataman-Devrim et al., 2023).

While some studies vary the physical setting, including home environments (Tomasello and

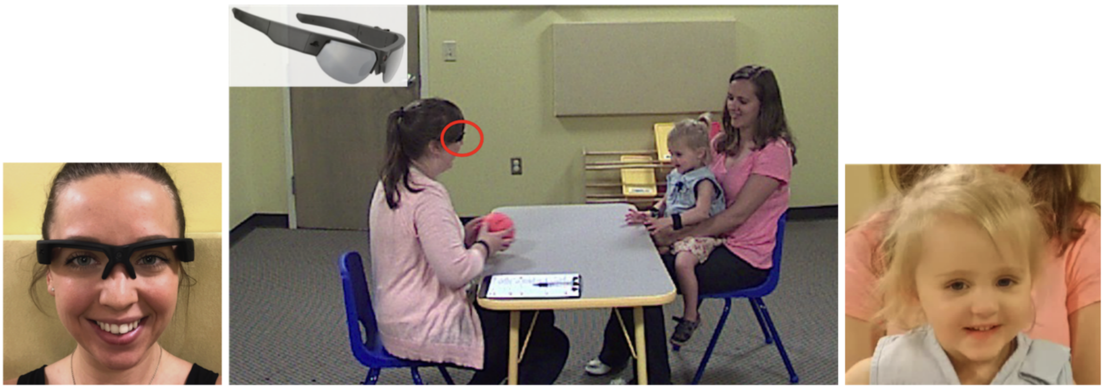


Figure 2.11: Several prior works on gaze applications have relied on specialized hardware like eye-trackers or head-mounted cameras. Besides limiting scaling to wider populations, this setup does not capture the point of regard and thus limits gaze analysis. This figure illustrates the data capture setup from Chong et al. (2017). The examiner wears glasses with an embedded camera that records the child.

Todd, 1983; Bakeman and Adamson, 1984; Suarez-Rivera et al., 2022), and a few extend to diverse cultural and socio-economic contexts (Callaghan et al., 2011; Mastin and Vogt, 2016; Abels, 2020; Bard et al., 2021), the field remains centred on prompted dyadic interactions, often with mothers, during semi-structured free play. Research outcomes are known to be sensitive to context (Tamis-LeMonda et al., 2017), and this narrow focus may limit the generalisability of findings.

2.7.2 Detecting Signals for ASD Diagnosis

Gaze is a key behavioural signal in ASD diagnosis, and behaviours such as eye contact are routinely evaluated by clinicians during diagnostic assessments (Edition, 2013).

As with joint attention studies, early work explored the use of eye-trackers to capture gaze behaviours (Magrelli et al., 2013; Noris et al., 2012). However, eye tracking provides only the point of regard within the observed image and does not directly reveal the semantics of the gaze target—an essential piece of information for understanding higher-level social gaze behaviours.

More recent studies have applied machine learning-based gaze prediction to ASD-related tasks. Chong et al. (2017) developed a CNN-based method for detecting eye contact towards a clinician from egocentric views, tested on data from autistic children. Their follow-up study (Chong et al., 2020a) further correlated predicted eye contact in videos from BOSCC diagnostic sessions (Grzadzinski et al., 2016) with social affect severity scores recorded during the same BOSCC sessions, as well as ADOS-2 (Lord et al., 2012) sessions, finding a negative correlation. However, both studies relied on specialised head-mounted cameras (see Figure 2.11) which limits applicability.

More recent work by Guo et al. (2023) predicted eye contact from third-person videos of autism therapy sessions, with predictions improving the estimation of social visual behaviour scores. Ahn et al. (2024) evaluated eye contact during ADOS-2 sessions and found that predicted eye contact between child and parent negatively correlated with ADOS-2 social affect severity scores, but this approach also required head-mounted cameras. Also, none of the above works have attempted to consider gaze behaviours beyond eye-contact such as shared attention.

2.8 Conclusion

Across the areas reviewed in this chapter, gaze estimation has advanced considerably—from early VFOA models requiring explicit scene knowledge to modern deep learning approaches for gaze following, social gaze prediction, and egocentric gaze estimation. Yet, several key limitations remain.

Existing methods often treat related tasks separately, missing opportunities for joint modelling and contextual reasoning. Multimodal information, which could help disambiguate difficult cases, is underutilised. Social gaze tasks have lacked unified frameworks that can scale to multi-person, temporally dynamic scenes. In Chapters 5 and 6, we propose new architectures that leverage multimodal information to improve gaze estimation performance and unify gaze following and social gaze tasks into a single framework while incorporating temporal information.

Moreover, while promising steps have been made toward ego–exo alignment, prior efforts have not explored leveraging exocentric cues to improve egocentric gaze prediction. In Chapter 8 we propose novel self-supervised approaches for learning exocentric gaze representations to improve egocentric gaze estimation.

In parallel, while recent advances in vision–language models (VLMs) and prompting strategies offer promising tools for extracting contextual cues—such as pose, gestures, and social interactions—their potential for improving gaze understanding remains largely unexplored. We investigate different VLMs and prompting strategies for extracting these cues in Chapter 7, and demonstrate improvements in cross-dataset gaze following performance.

Finally, although gaze-based tools show potential for real-world applications—particularly in child-centric domains such as joint attention analysis and ASD diagnosis—most studies to date rely on controlled environments or specialised equipment, limiting their scalability and ecological validity. Our gaze models can be applied to arbitrary scenes, and we demonstrate this in Chapter 9 by applying them to child-language acquisition settings.

3 Datasets

This chapter presents the datasets used for training and evaluating the models developed in this thesis, spanning third-person gaze following, social gaze prediction, and egocentric gaze estimation. In addition to presenting existing benchmarks, this chapter introduces several key dataset contributions developed as part of this work:

- **ChildPlay-audio**, which extends the ChildPlay dataset (Tafasca et al., 2023b) with frame-level speaking status annotations to enable multimodal gaze prediction;
- **VSGaze**, which unifies and extends annotations across multiple video datasets to support joint training and evaluation of gaze following and social gaze tasks (LAH, LAEO, and SA);
- **RLR-CHAT extensions**, which augment the original dataset (Murdock et al., 2024; Yun et al., 2024) with head identity matching and LAH annotations, facilitating the study of ego–exo gaze alignment.

We provide a summary of the datasets in Table 3.1. The remainder of this chapter details the datasets, the annotations developed, and the rationale for their use in the experiments described in subsequent chapters.

3.1 Gaze Following

These datasets address the task of gaze following (Section 1.1.1) and are used for training and evaluating the models presented in Chapters 5, 6, and 7.

3.1.1 GazeFollow

The GazeFollow dataset, introduced by Recasens et al. (2015), was the first large-scale dataset created specifically for the gaze following task. It contains approximately 122,000 static images,

Name	Type	Shows (Clips)	Frames	Samples	Origin	Annotations
Gaze Following Datasets						
GazeFollow (Recasens et al., 2015)	image	-	122K	130K	SUN, MS-COCO, ImageNet, ...	eye location · gaze point · inside/outside
VAT (Chong et al., 2020b)	video	50 (606)	71K	164K	TV shows	gaze point · inside/outside
DL Gaze (Lian et al., 2018)	video	4 (86)	95K	6K	Manual collection	gaze point
GOO (Tomas et al., 2021)	image	-	201K	172K	Manual collection Synthetic	gaze point · gaze object · object bbox
ChildPlay (Tafasca et al., 2023b; Farkhondeh et al., 2024)	video	95 (401)	120K	257K	YouTube	gaze point · gaze class · speaking status · HOI labels
Social Gaze Datasets						
VideoCoAtt (Fan et al., 2018)	video	20 (400)	493K	138K	TV shows	shared gaze object bbox
UCO-LAEO (Marin-Jimenez et al., 2019)	video	4 (129)	18K	36K	TV shows	LAEO labels
AVA-LAEO (Marin-Jimenez et al., 2019)	frames	-	50K	172K	Movies	LAEO labels
Gaze Following and Social Gaze Datasets						
VACATION (Fan et al., 2019)	video	50	96K	164K	TV Shows	gaze object bbox · gaze communication label
VSGaze (Gupta et al., 2024b)	video	169 (1536)	700K	6M	VAT, ChildPlay, VideoCoAtt, UCO-LAEO	gaze point · inside/outside · LAH, LAEO, SA labels
Egocentric Gaze Datasets						
Ego4D gaze (Lai et al., 2023)	video	27	~3M	-	Ego4D	gaze point
RLR-CHAT (Murdock et al., 2024; Yun et al., 2024)	video	170	2.6M	-	Manual collection	gaze point · multi-view · head identity · LAH labels

Table 3.1: Summary of gaze estimation datasets. All datasets other than Ego4D also provide head bounding boxes.

most of which are annotated with a single person's head bounding box and a corresponding 2D gaze target point. The images are sourced from a variety of public datasets, including ImageNet (Russakovsky et al., 2015), COCO (Lin et al., 2014), Places (Zhou et al., 2014), Actions 40 (Yao et al., 2011), PASCAL (Everingham et al., 2010) and SUN (Xiao et al., 2010), offering wide diversity in scenes and contexts.

The test set includes gaze point annotations provided by multiple annotators, offering a measure of label variability. Later, Chong et al. (2018) extended the train split of the dataset by adding annotations indicating whether a person's gaze target lies inside or outside the image frame.

Overall, GazeFollow features simpler scenes, where the annotated person is typically the one whose gaze target is more clearly visible. Additionally, both the images and annotations are of relatively lower quality compared to more recent datasets. Yet, it remains widely used due to its diversity in scenes and settings. Its scale and variability make it especially valuable for pre-training gaze following models before fine-tuning on more specific or higher-quality datasets. We provide samples from the dataset in Figure 3.1.

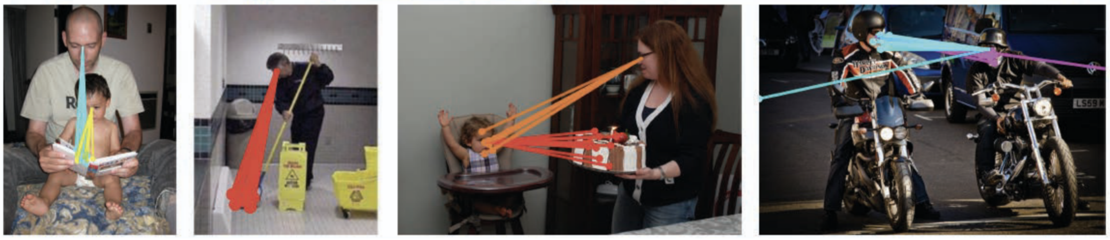


Figure 3.1: Samples from the GazeFollow test set (Recasens et al., 2015). GazeFollow is a large-scale static dataset for gaze following. Despite lower quality images and annotations, it's rich diversity makes it a valuable dataset for pre-training.

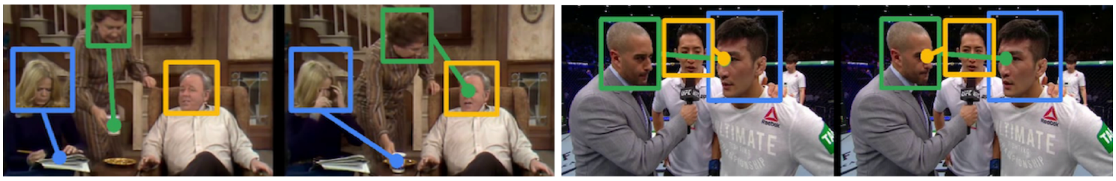


Figure 3.2: Samples from the VideoAttentionTarget dataset (Chong et al., 2020b). It is a video dataset for gaze following, featuring scenes from conversational settings such as TV shows and movies. For this reason, faces commonly appear as gaze targets.

3.1.2 VideoAttentionTarget (VAT)

The VideoAttentionTarget dataset (Chong et al., 2020b) was the first video dataset created for the gaze following task. It consists of 1,331 video clips collected from 50 television shows and YouTube series, covering a broad range of everyday conversational settings.

Each clip is annotated with head bounding boxes, gaze points, and inside-versus-outside frame gaze labels for a subset of people in the scene. The primary motivation for creating this dataset was to extend gaze following beyond static images, enabling the study of gaze behaviour in dynamic social interactions where temporal information and motion cues are present.

Compared to GazeFollow, the resolution and overall quality of both the videos and annotations are generally higher. Moreover, as the dataset focuses on conversational settings such as talk shows, faces frequently appear as gaze targets (see Table 3.3). We provide samples from the dataset in Figure 3.2.

3.1.3 ChildPlay

The ChildPlay dataset, introduced in our work Tafasca et al. (2023b), is the first publicly available gaze following dataset focused on children interacting with other children and adults in naturalistic settings. Existing child-centric gaze datasets are typically private (de Belen et al., 2020) or limited to anonymized pose data, making it difficult to study gaze behaviour directly. Others offer only coarse labels or are restricted to fixed lab environments (Billing et al., 2020;

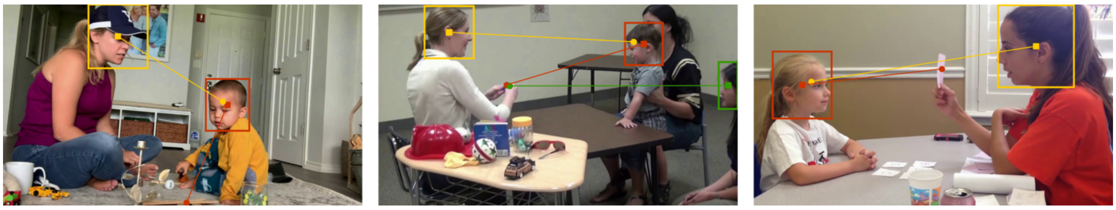


Figure 3.3: Samples from the ChildPlay dataset (Tafasca et al., 2023b). It features high quality videos from YouTube of children playing and interacting with other children and adults. Given the setting, children tend to look at nearby objects such their toys, while adults tend to observe the children and their activities.

Rehg et al., 2013).

Other gaze benchmarks, such as GazeFollow and VideoAttentionTarget, predominantly feature adults, and models trained on them can underperform when applied to children—as seen in literature for body landmark estimation (Sciortino et al., 2017). Given the importance of pose information in gaze prediction (Gupta et al., 2022; Belkada et al., 2021) and the known differences in gaze behaviour between adults and children (Franchak et al., 2016), there is a clear need for gaze-annotated datasets featuring younger age groups in general, everyday settings.

ChildPlay contains 401 high quality clips from 95 YouTube videos, with annotations including head bounding boxes, gaze points, and seven gaze classes covering inside/outside frame gaze, gaze shifts, and related behaviours. Later (Farkhondeh et al., 2024) extended the dataset with Human-Object Interaction (HOI) labels. Since the dataset focuses on child-play scenarios, children tend to gaze at nearby objects such as toys, whereas adults are more likely to observe the children and their activities. We provide samples from the dataset in Figure 3.3

3.1.4 Quality of Annotations

Annotating for gaze following can be challenging, especially when the eyes are not visible or when there are several possible targets in the scene. In such cases, different annotators may disagree on the correct gaze target. To account for this, GazeFollow provides annotations from multiple people in its test set, while ChildPlay introduces a special "uncertain" gaze class that annotators can choose for ambiguous cases.

In addition, GazeFollow, VideoAttentionTarget, and ChildPlay measure annotator agreement by evaluating the performance of one annotator (used as a prediction) against another (used as ground truth). This additionally provides an upper bound on human performance for each dataset. The same metrics are used later for model evaluation, as discussed in Chapter 4.

On GazeFollow, human performance reaches an average distance of 0.096 and a minimum distance of 0.040. Recent models are approaching these numbers, again highlighting the simpler nature of this dataset. For VideoAttentionTarget, human performance yields a distance of 0.051 and an in-vs-out of frame AP of 0.925. On ChildPlay, humans achieve a distance of 0.048 and an in-vs-out of frame AP of 0.993. For these datasets, human performance remains much better, especially for the distance metric. This reflects their higher complexity and shows that there is still room for model improvement.

3.2 ChildPlay-audio

Previous studies on analyzing conversations during meetings have shown that people usually look at the other speaking participants (Stiefelhagen et al., 2002), and such cues can be exploited for gaze target selection (Otsuka et al., 2005). Hence, we expect that identifying speaking persons can provide better scene understanding for gaze following, and help recognize attentiveness towards people, especially speakers. The latter is especially important in autism diagnosis, as eye contact is closely monitored by the clinician when they call out to the tested child (Lord et al., 2012).

Given the importance of having joint speaking and gaze information especially regarding children, we extended the ChildPlay gaze dataset (Tafasca et al., 2023b) with speaking status annotations. We leverage this dataset in Chapter 6 for training and evaluating models.

3.2.1 Annotation Protocol

We mark the speaking status for every gaze annotated person in ChildPlay. It is defined by a set of 5 non-overlapping labels:

- *Speaking*: the person is very likely speaking;
- *Not-speaking*: the person is very likely not speaking;
- *Vocalizing*: the person is very likely making a sound with their mouth (ex. to draw attention);
- *Laughing*: the person is very likely laughing (this does not include smiling);
- *Not-annotated*: the status cannot be inferred.

Note that in some sequences, the person's face may be occluded or the audio may not matching with the video (ex. in the case of voice-overs). Nevertheless, it is still often possible to estimate the speaking status based on head dynamics and and gaze (of others), similar to when annotating for gaze following. Hence, we ask annotators to mark the speaking status when it very likely a particular label, otherwise the 'Not-annotated' label has to be selected.

3.2.2 Annotation Statistics

They are given in Figure 3.4. We observe that children mostly do not speak, while adults are more balanced between speaking and not speaking. This makes sense as ChildPlay clips

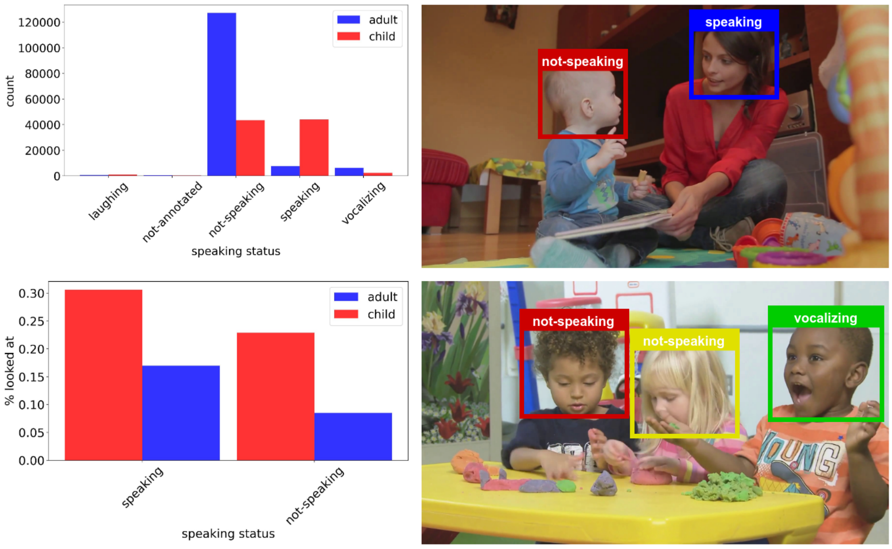


Figure 3.4: Annotation statistics and samples for ChildPlay-audio.

Method	**AVA-activespeaker** (Roth et al., 2020)	**ChildPlay-audio**
random	25.06	26.66
SPELL (Min et al., 2022)	85.30	56.81

Table 3.2: mAP for active speaker detection on AVA-activespeaker and ChildPlay-audio using a SoTA method (Min et al., 2022).

mainly contain children focused on their play activities, with adults supervising them. We also compute the percentage of cases when a speaking person[1] is looked at by any of the annotated people. The statistics indicate that both children and adults are more likely to be looked at when speaking. However, it is worth noting that adults are not looked at much overall, as clips mostly contain a single adult and the children are focused on their play activities. This is in contrast to meeting situations where the speaking person is looked at most of the time (Stiefelhagen et al., 2002).

3.2.3 Speaking Status Prediction

We re-trained a state of the art model for active speaker detection (Min et al., 2022) using the visual-only modality on the large-scale AVA-activespeaker benchmark (Roth et al., 2020). Table 3.2 provides prediction results of the model on the AVA and ChildPlay test sets. We see that while the model performs well on AVA, it does not generalize well to ChildPlay, which indicates the challenge of detecting speech in natural scenes with children, and the potential of ChildPlay-audio as a challenging new benchmark for speaking status prediction, as it contains significantly different settings in terms of people (children), poses (sitting) and environments

[1] Here onward, ’speaking’ includes the speaking, laughing and vocalizing labels.

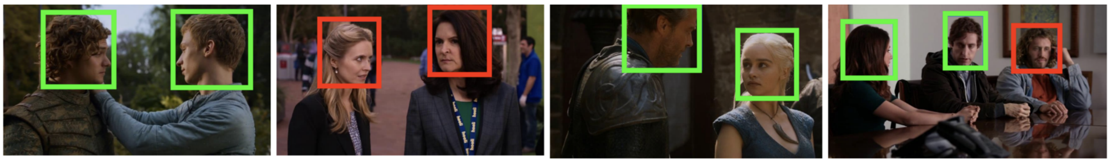


Figure 3.5: Samples from the UCO-LAEO dataset (Marin-Jimenez et al., 2019). The dataset features high-quality frames from four TV shows with pairwise LAEO (Looking At Each Other) annotations. As most scenes involve only two people, it does not require complex social reasoning.

(schools, daycare centers).

3.3 Social Gaze Prediction

These datasets address the tasks of Looking at Each Other (LAEO) and Shared Attention (SA) (see Section 1.1.2), and are used for training and evaluating models in Chapter 6. While no dataset exists for LAH, we extend existing gaze datasets with LAH labels using the protocol described in Section 3.4.

3.3.1 UCO-LAEO

The UCO-LAEO dataset (Marin-Jimenez et al., 2019) was developed for the task of detecting whether two people in a video frame are LAEO. The dataset contains 22,398 high-quality frames sourced from four TV shows, covering a variety of conversational and social scenarios.

Unlike prior datasets (Marin-Jimenez et al., 2014), UCO-LAEO provides *frame-level* annotations including head bounding boxes and binary LAEO labels for all possible head pairs. We additionally used a head detector (Jocher et al., 2022) to predict the head bounding boxes of people in "negative" videos (*i.e.* without positive LAEO instances) that are provided without head annotations.

The dataset primarily focuses on scenes involving two people, which simplifies the task and reduces the need for social interaction reasoning. We provide samples from the dataset in Figure 3.5.

3.3.2 VideoCoAtt

VideoCoAtt (Fan et al., 2018) was the first publicly available dataset to provide shared attention annotations in unconstrained, real-world video data. It contains 380 videos (approximately 492,000 frames) collected from TV shows, covering diverse group interaction settings.

When shared attention occurs (approximately 140,000 frames), the relevant frames are annotated with the bounding box of the shared target object and the head bounding boxes of the

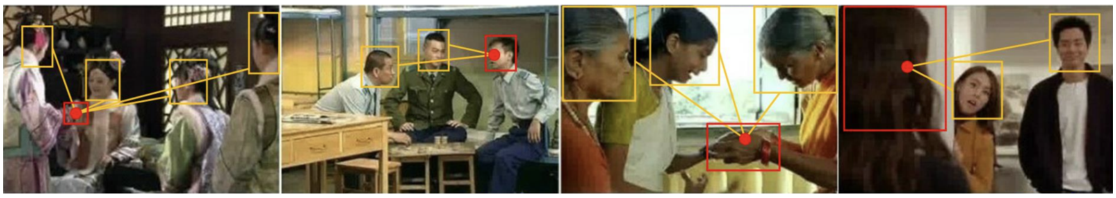


Figure 3.6: Samples from the VideoCoAtt dataset (Fan et al., 2018). This large-scale dataset comprises frames from TV shows annotated for shared attention. The frames are typically lower in resolution.

individuals involved in the SA event. Since we also need negative instances, we run a head detector (Jocher et al., 2022) to identify other people in the scene that are not sharing attention. Any pair containing at least one of such people is automatically labeled as a negative instance.

Unlike UCO-LAEO, VideoCoAtt includes a wider variety of group sizes but is generally lower in resolution. We provide samples from the dataset in Figure 3.6.

3.4 VSGaze

A limitation of the above datasets is that they only contain annotations for either gaze following or specific social gaze tasks, but not both. To address this, we introduced the **V**ideo dataset with **S**ocial gaze and **Gaze** following annotations, or **VSGaze**, in our paper (Gupta et al., 2024b). Samples from the dataset are provided in Figure 3.7 and experiments are detailed in Chapter 6.

VSGaze extends head track annotations and unifies annotation types across VideoAttentionTarget, ChildPlay, VideoCoAtt, and UCO-LAEO. This enables joint training of temporal gaze following and social gaze models and supports the development of new tasks and metrics for evaluating performance across the component datasets. Note that we also extend GazeFollow with LAH annotations as described below, but it is not considered part of VSGaze due to being a static dataset.

3.4.1 Extending Head Track Annotations

Since each dataset contains annotations for only a subset of people in the scene, we detect all missing heads using a pre-trained Yolov5 head detection model (Jocher et al., 2022), and track them using the ByteTrack algorithm (Zhang et al., 2022). We manually verified the accuracy of the obtained tracks. This step is essential for identifying both positive and negative social gaze pairs. For example, consider a scene with three people, i, j, and k, where i is looking at j. If only i is annotated, the positive LAH pair $i \rightarrow j$ would be missed, as would the negative LAH pair $i \rightarrow k$.

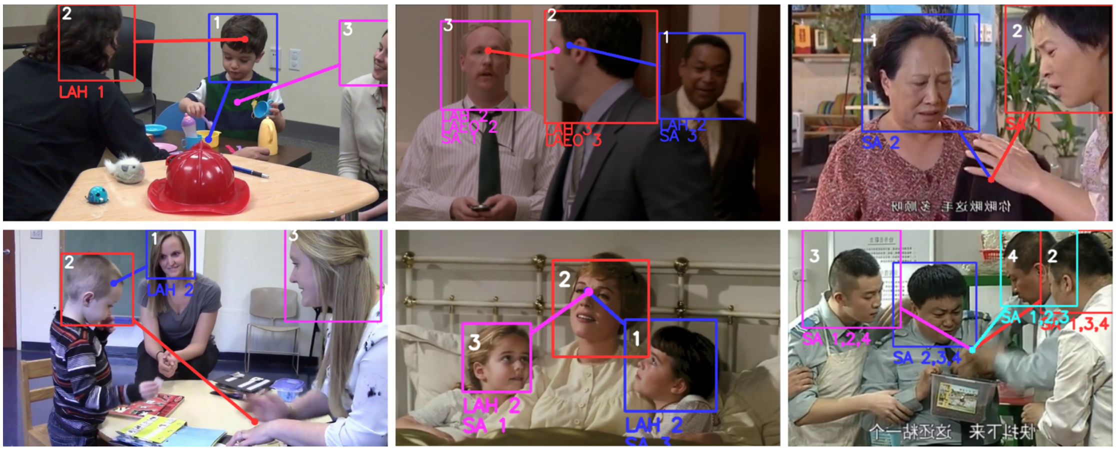


Figure 3.7: Samples from the VSGaze dataset. VSGaze unifies and extends annotation types across multiple gaze datasets, resulting in the largest and most diverse resource of its kind. It spans a wide range of scenes—from TV shows and movies to childcare environments—and includes annotations for both gaze following and social gaze behaviors (LAH, LAEO, SA). For each individual, the predicted social gaze relations are listed alongside the corresponding person IDs (*e.g.*, in the middle column, Person 1 shares attention with Person 3).

3.4.2 Unifying Gaze Following and Social Gaze Annotations

Given the extended set of head bounding box annotations, along with existing gaze following and social gaze annotations, we process these labels to generate unified gaze following and social gaze labels for the tasks of Looking at Heads (LAH), Looking at Each Other (LAEO), and Shared Attention (SA), as described in Section 1.1, across all datasets.

Gaze target points. For VideoAttentionTarget and ChildPlay, we use the annotated gaze points. For people sharing attention in VideoCoAtt, we compute their gaze points as the centre of the SA object's bounding box. Similarly, for person pairs labelled LAEO in UCO-LAEO, we compute their gaze points as the centre of the other person's head bounding box.

LAH. We generate LAH annotations for all datasets. We check whether the gaze point for an annotated person falls inside another person's head bounding box. For the GazeFollow test set, at least two of the annotated gaze points must fall inside another person's head bounding box.

LAEO. We use the provided annotations for UCO-LAEO. For VideoAttentionTarget and ChildPlay, we generate LAEO annotations using the LAH annotations by checking whether the LAH target for a pair of people corresponds to the other person. LAEO annotations cannot be generated for GazeFollow, since most images are annotated for only a single person, or for VideoCoAtt, where individuals acting as SA targets are typically not annotated with their own gaze targets.

SA. We use the provided annotations for VideoCoAtt. For VideoAttentionTarget and ChildPlay, we generate new SA annotations from the LAH annotations by checking whether two people

Dataset	Gaze Points	LAH	LAEO	SA
GazeFollow (Recasens et al., 2017)	118k	27k/493k	0	0
VideoAttentionTarget (Chong et al., 2020b)	109k	74k/729k	13k/461k	16k/94k
ChildPlay (Tafasca et al., 2023b)	217k	59k/682k	7k/351k	4k/55k
VideoCoAtt (Fan et al., 2018)	367k	290k/1551k	0	400k/918k
UCO-LAEO (Marin-Jimenez et al., 2019)	21k	21k/36k	10k/54k	0
VSGaze	714k	444k/2998k	30k/866k	420k/1067k

Table 3.3: Person-wise gaze point and pair-wise social gaze annotation (positive/negative) statistics for our datasets. VSGaze unifies annotation types across VAT, ChildPlay, VideoCoatt and UCO-LAEO.

share attention to the same third person. SA annotations cannot be obtained for GazeFollow, where most images are annotated for only a single person, or for UCO-LAEO, where pairs annotated with LAEO cannot also be sharing attention.

3.4.3 Annotation statistics

Annotation statistics are summarised in Table 3.3. Overall, VideoCoAtt provides the largest number of annotations, except for LAEO. As expected, the pairwise annotations are skewed towards negative cases. The statistics also offer insight into the content of the datasets. Since VideoAttentionTarget, VideoCoAtt, and UCO-LAEO contain clips from TV shows, they include many more instances of people looking at each other or at others. In contrast, for ChildPlay, LAH primarily occurs when the supervising adult looks at a child, and there are relatively few instances of LAEO.

3.5 Egocentric Gaze Estimation

These datasets address the task of egocentric gaze estimation (Section 1.1.3) and are used for training and evaluating the models in Chapter 8.

3.5.1 RLR-CHAT

The Reality Labs Research Conversations for Hearing Augmentation Technology (RLR-CHAT) dataset (Murdock et al., 2024; Yun et al., 2024) is a large-scale collection of egocentric multi-sensory recordings captured from individuals engaging in natural conversations. Each conversation session is approximately one hour in duration and is recorded using Aria glasses (Engel et al., 2023), which capture RGB frames at 5Hz, 7-channel spatial audio at 48kHz, and eye-tracking data at 30Hz, among other modalities. To maximize visual diversity, we subsample RGB frames by selecting every third frame and align them with the nearest eye-tracking annotations in time. The distribution of session sizes by number of participants is illustrated in Figure 3.8. The dataset contains a total of 170 sessions, the majority of which involve two participants.

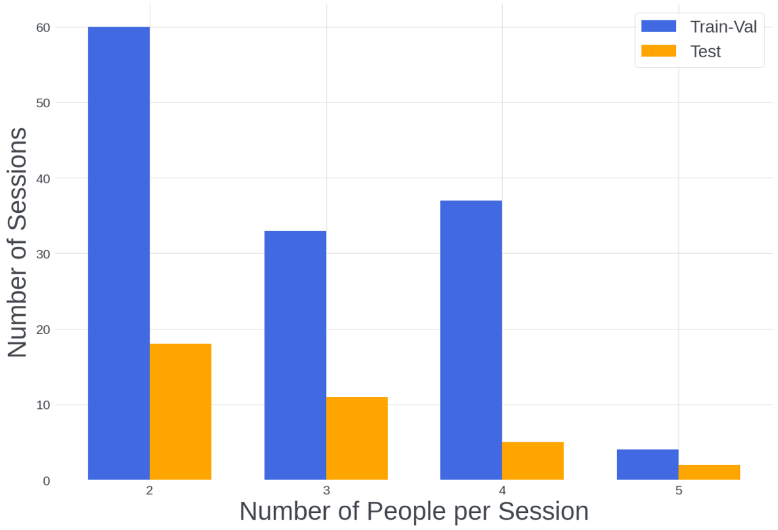


Figure 3.8: RLR-CHAT session distribution by number of people.

A distinctive feature of RLR-CHAT is the synchronized availability of modalities from all participants within the conversation. This synchronization uniquely enables the exploration of ego-exo alignment techniques to learn richer gaze representations. To our knowledge, the only comparable accessible dataset is the Aria Everyday Activities dataset (Lv et al., 2024), which is significantly smaller and primarily focuses on single-person activities.

We augment RLR-CHAT by incorporating head bounding box detections and automatically assigning identities to these boxes using spatial audio cues, as detailed in the next section. The test set includes manually corrected head bounding boxes and high-quality, OptiTrack-based head identity matching.

Obtaining Head Box Identities. To reliably identify the head bounding boxes of individuals visible in a participant's field of view (FoV), we leverage a pre-trained MAV-ASL model (Jiang et al., 2022) to obtain active speaker heatmaps for each egocentric image frame. The MAV-ASL model produces two types of heatmaps: one that indicates the direction of the active speaker over a full 360-degree span, and another that provides the 2D location of the active speaker when present in the FoV. Both heatmaps are initially computed in a head-locked coordinate system.

We begin by utilizing the directional heatmap and converting it into world-locked coordinates with the help of SLAM data. By processing 5000 frames per participant, we compute the average world-locked location of the other participants in the conversation. For each frame, these average locations are then transformed back into the head-locked FoV coordinate system.

Subsequently, we match the detected head bounding boxes to these averaged locations by selecting the nearest match within a threshold of 200 pixels. Although this thresholding

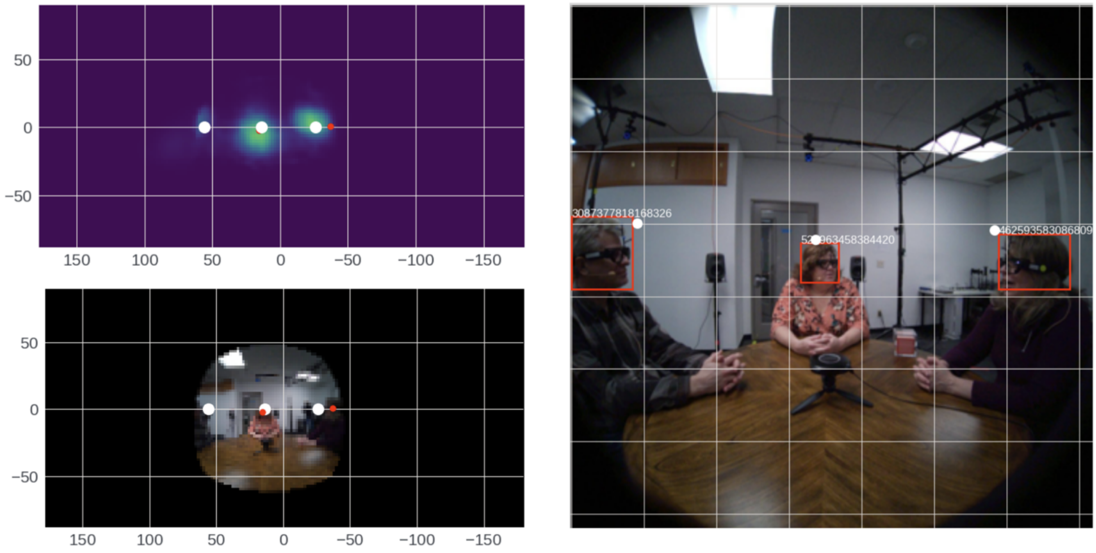


Figure 3.9: Overall pipeline for identifying head bounding boxes. We first obtain the average locations of other participants (white dots) in the conversation in 360 degrees world-locked coordinates. These are then mapped to the head-locked FoV coordinates and matched to the nearest head bounding box within a threshold.

Split	Number of Frames	LAH Pairs	
		Positive	Negative
Train	1848555	273464	1107699
Validation	448173	77914	309426
Test	385039	36361	351216

Table 3.4: RLR-CHAT number of frames and LAH statistics.

process means that not all head bounding boxes are assigned an identity, the matches that are made have been verified to be of high quality. The overall pipeline is illustrated in Figure 3.9.

Obtaining LAH Annotations. By leveraging these identity-aligned head bounding boxes alongside eye-tracking annotations, we first determine if person A is looking at person B from A's egocentric perspective. This information is then mapped to another person's viewpoint (e.g., person C) to obtain exocentric annotations indicating whether person A is looking at the head of person B. This annotation process is applied to all pairs of individuals present in the scene. The resulting LAH statistics are summarized in Table 3.4.

3.5.2 Ego4D Gaze

Ego4D (Grauman et al., 2022) is a large-scale, publicly available egocentric dataset that captures individuals performing daily life activities. We use the subset with gaze annotations introduced by Lai et al. (Lai et al., 2023), which consists of 27 approximately hour-long videos

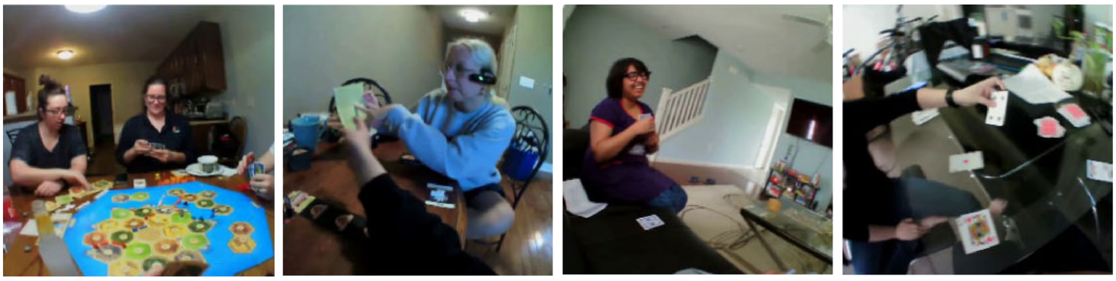


Figure 3.10: Samples from the Ego4D gaze (Lai et al., 2023) dataset. It features videos of people engaged in social interactions like playing board games. Despite the focus on social settings like RLR-CHAT, key differences include a narrower field of view, greater environmental diversity, and fewer gaze instances directed toward faces.

featuring 80 participants engaged in social interactions such as playing board games.

A key distinction of Ego4D compared to other datasets, such as EGTEA Gaze (Li et al., 2018) and Aria (Lv et al., 2024), is that those datasets primarily focus on single-person activities. Since Ego4D emphasizes social settings, it is better suited for evaluating improvements derived from learning exocentric gaze representations.

However, despite the focus on social settings in both RLR-CHAT and Ego4D, a significant domain gap remains between the datasets. Ego4D videos have a smaller FoV and contain much more diverse environments. Furthermore, as participants are often engaged in playing games rather than conversation, gaze points tend to fall less frequently on faces. We provide samples from the dataset in Figure 3.10.

3.6 Conclusion

This chapter has presented the datasets used throughout the thesis, covering third-person gaze following, social gaze prediction, and egocentric gaze estimation. In addition to leveraging existing benchmarks, we contributed several new resources: ChildPlay-audio, which provides multimodal annotations for child-centric gaze prediction; VSGaze, which unifies gaze following and social gaze annotations across multiple datasets; and extended annotations for RLR-CHAT, enabling identity assignment and ego–exo alignment.

Together, these datasets support the unified, multimodal, and self-supervised models developed in the following chapters and enable semantic evaluation.

4 Metrics

This chapter presents the evaluation metrics and protocols used to assess model performance across the three core tasks addressed in this thesis: gaze following, social gaze prediction, and egocentric gaze estimation. These metrics are computed at the frame-level. Our contributions in this context include:

- **Semantic metrics for gaze following:** To complement standard distance and heatmap-based metrics, we introduce social gaze metrics that measure whether gaze predictions align with meaningful social targets and interactions. These metrics provide a semantically grounded assessment of gaze following performance.
- **New protocols for social gaze prediction:** We define evaluation procedures for our addressed social gaze tasks using both post-processed gaze following predictions and direct decoder outputs. In particular, our updated protocol for *Shared Attention* resolves key issues in prior work (see Section 2.3.2), such as identifying multiple shared attention groups and specifying which individuals are involved.
- **Gaze following inspired metrics for egocentric gaze:** We extend gaze point and social gaze based evaluation to egocentric settings, providing interpretable, socially relevant alternatives to conventional heatmap-based metrics.

Together, these metrics enable fine-grained evaluation of model performance, capturing not only spatial accuracy but also socially and semantically grounded behaviour in both third-person and egocentric gaze estimation tasks.

4.1 Gaze Following

This section provides an overview of the metrics and protocols used to evaluate gaze following performance. As described in Section 2.2, typical gaze following architectures—including those proposed in this thesis—predict a heatmap over the scene that highlights potential gaze targets. This predicted heatmap is then processed to compute evaluation metrics.

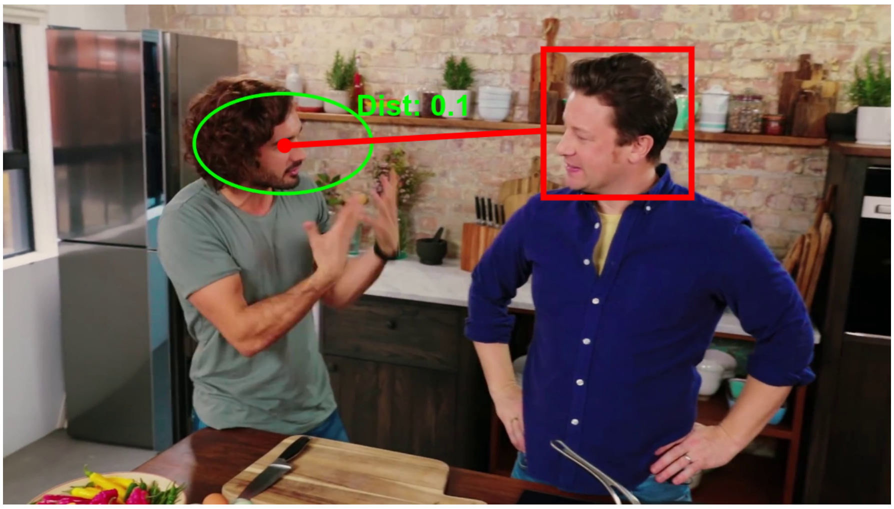


Figure 4.1: Illustration of the distance metric defined in Section 4.1.1. The green ellipse represents the points that lie within an distance of 0.1 from the ground truth gaze point (red).

4.1.1 Gaze Point

Distance. Following the protocol of Recasens et al. (2015), we compute the distance between the predicted and ground truth gaze points. The predicted gaze point is obtained by taking the *arg max* of the predicted gaze heatmap. We then calculate the L2 distance between this predicted point and the ground truth gaze point on a 1×1 normalized square, resulting in values ranging from 0 to $\sqrt{2}$. Figure 4.1 illustrates the region corresponding to a distance of 0.1 from the ground truth gaze point.

Since GazeFollow (Recasens et al., 2015) includes multiple annotations per person in the test set, we additionally report:

- *Minimum Distance:* the distance of the predicted gaze point to the closest annotated gaze point, reflecting whether the model aligns with *any* plausible target.
- *Average Distance:* the distance of the predicted gaze point to the average of all annotated gaze points, reflecting agreement with the overall annotator consensus.

It is worth noting that the Average Distance metric may be unreliable in scenes where the annotated gaze points are widely dispersed across the image, as the computed average can then fall at a location that does not correspond to any plausible gaze target.

4.1.2 Heatmap

AUC (per Recasens et al. (2015)). Distance-based metrics rely on the *arg max* of the heatmap and may overlook cases where the predicted heatmap is multimodal—often a reflection of the model's uncertainty about the target. To account for this, and to compare against annotator uncertainty, Recasens et al. (2015) introduced an AUC (Area Under the Curve) metric. Here,

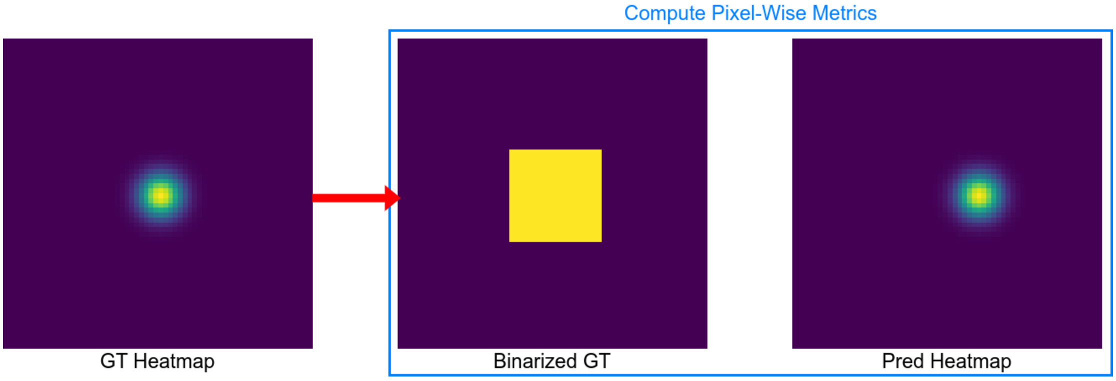


Figure 4.2: The heatmap metric for gaze following proposed by Chong et al. (2020b), and by Li et al. (2018) for egocentric gaze estimation, compares the predicted heatmap to a binarized ground truth heatmap at the pixel level. However, the binarization threshold is arbitrary and, in the case of Chong et al. (2020b), results in an overly large positive region. This setup also penalizes multimodal predictions when they fall outside this region. Moreover, the metric is difficult to interpret and lacks a clear correspondence to meaningful spatial or semantic error, which limits its utility. For these reasons, we do not compute this metric in our later works.

the predicted heatmap is compared against a binary ground truth map, which assigns a value of 1 to the annotated gaze locations and 0 elsewhere. The AUC is then computed as the area under the resulting ROC curve.

AUC (per Chong et al. (2020b)). The VideoAttentionTarget dataset (Chong et al., 2020b) contains only a single gaze point annotation per person, even in the test set. Therefore, the original AUC protocol from Recasens et al. (2015) cannot be applied. Instead, Chong et al. (2020b) proposed an alternative AUC protocol. In this version, the predicted heatmap is compared against a binarized ground truth heatmap (created by thresholding the ground truth heatmap), and the area under the ROC curve is computed. We provide an illustration of the protocol in Figure 4.2.

However, this variant has several limitations. First, the choice of threshold for binarizing the ground truth heatmap is arbitrary. In practice, the originally proposed threshold creates a large positive region, which penalizes more precise heatmap predictions. Second, while this AUC variant can account for non-maximal heatmaps (which distance metrics cannot), it would still penalize multimodal predictions outside the positive region, and hence does not effectively capture gaze target uncertainty unlike the original formulation. For these reasons, we do not report this metric in our later works.

More generally, both AUC formulations can be difficult to interpret compared to the distance metric, which directly translates to an average pixel error. Additionally, given the class imbalance between positive and negative regions in the heatmap, AUC scores tend to overestimate performance and may saturate—a trend also observed in egocentric gaze estimation research (Lai et al., 2023).

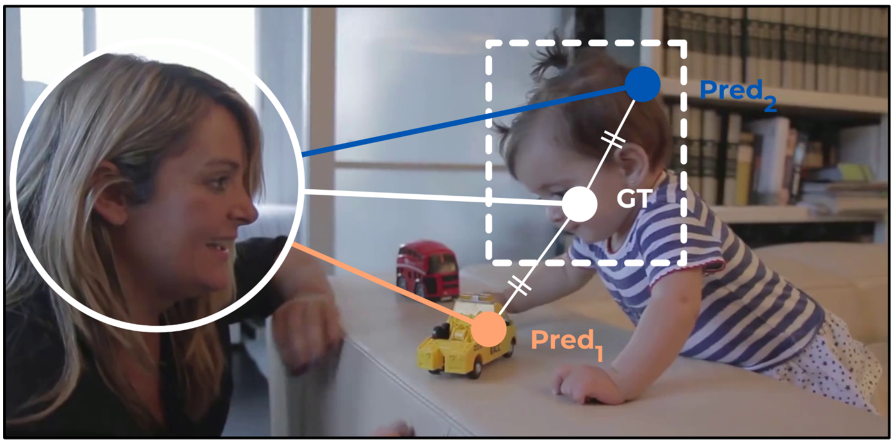


Figure 4.3: Standard gaze following metrics, such as the distance metric, assign the same error to both Pred$_1$ and Pred$_2$. However, Pred$_2$ is a better prediction, as it falls on the same semantic target (the child's head). In contrast, our proposed social gaze metrics account for head-related semantics.

4.1.3 In-Out

The In-Out metric, proposed by Chong et al. (2020b), evaluates whether the model correctly predicts whether the gaze target lies inside or outside the image frame. Performance is measured using Average Precision (AP), comparing the predicted in-out score with the corresponding ground truth binary label.

4.2 Social Gaze Prediction

This section describes the protocols and metrics used to evaluate performance on the social gaze tasks addressed in this thesis: looking at each other (LAEO), shared attention (SA), and looking at heads (LAH). These metrics apply both to post-processed gaze following predictions and to explicit social gaze labels produced by task-specific decoders.

4.2.1 Post-processing

The gaze following metrics discussed previously capture the spatial location of the predicted gaze point but do not account for its semantic meaning. As illustrated in Figure 4.3, two predicted gaze points, Pred$_1$ and Pred$_2$, would receive the same error according to the distance metric. However, Pred$_2$ is clearly a better prediction, as it lies on the same semantic target (the child's head). Our social gaze metrics incorporate head-related semantics and represent a step toward more meaningful, semantic evaluation. As we demonstrate in later chapters (see chapters 5, 6), these metrics do not always correlate with standard gaze following metrics, highlighting their value as complementary indicators of model performance.

LAH. We determine whether the predicted and ground truth gaze points fall inside a head bounding box. To compute performance, we have employed different protocols:

- **P.Head; Tafasca et al. (2023b)**: We assessed whether predictions generally captured looks towards faces. A true positive was recorded when both the predicted and ground truth gaze points fell inside *any* head bounding box. A false positive occurred when the ground truth gaze point fell inside a head box, but the prediction did not. Precision was then computed from these values.

- **LAH; Gupta et al. (2024a,b)**: We refined the protocol to determine whether the prediction captured gaze toward the *same* head as the ground truth, using the following definitions:
  - *True Positive*: Predicted and ground truth gaze points fall in the same head box.
  - *False Positive*: Predicted gaze point falls in a head box, but ground truth does not.
  - *False Negative*: Ground truth gaze point falls in a head box, but predicted gaze point does not, or falls on a different head or object.
  - *True Negative*: Neither predicted nor ground truth gaze points fall in a head box.

  From these values, we compute precision, recall, and F1 scores.

LAEO. For each pair of people, we check whether their gaze points fall within each other's head bounding box. If so, it is counted as a positive instance; otherwise, it is considered negative. This process is applied to both the ground truth and predicted gaze points, and precision, recall, and F1 scores are computed accordingly.

SA. For each pair of people, we evaluate whether their gaze points fall within a specified distance threshold of each other. If so, the instance is considered positive; otherwise, it is negative. We repeat this process across a range of thresholds and compare the predictions to the ground truth to compute an AP score.

For SA, we additionally report two metrics that quantify shared attention target localization performance:

- *Distance:* Similar to the gaze following distance metric, this is the L2 distance (computed on a 1×1 normalized square) between the average of the predicted gaze points of all individuals involved in a shared attention instance (as defined by the ground truth), and the center of the ground truth bounding box of the shared attention target.
- *Accuracy:* The proportion of shared attention instances where the average predicted gaze point falls within the ground truth target bounding box. To account for near-boundary predictions, we expand the bounding box by 5% on each side.

It is important to note that, unlike previous works (Fan et al., 2018; Chong et al., 2020b; Tu et al., 2022), which compute these metrics on a per-frame basis, we evaluate them per shared

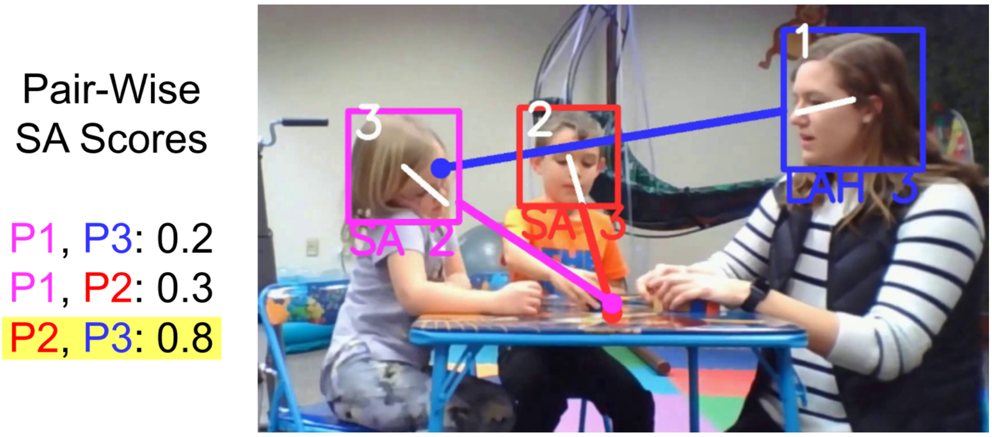


Figure 4.4: An example of social gaze inference. For each individual (right), the predicted social gaze relations are listed alongside the corresponding person IDs (*e.g.*, Person 2 shares attention with Person 3). To obtain LAH, we post-process the predicted gaze point to determine whether it falls within another person's head bounding box; for LAEO, we perform the same check in both directions. For SA, we use the pairwise scores predicted by the decoder (left) and classify pairs as positive instances if the score exceeds a predefined threshold (here 0.5).

attention instance. Additionally, our distance metric is normalized assuming an image size of 1×1.

4.2.2 Task-Specific Decoders

In Chapter 6, we introduce methods that *directly* predict pairwise scores for each social gaze task using task-specific decoders. These predicted scores are denoted as:

- **LAH:** $\mathbf{e}_{i\to j}$, representing whether person i is looking at person j.
- **LAEO:** $\mathbf{e}_{i\leftrightarrow j}$, representing mutual gaze between persons i and j.
- **SA:** $\mathbf{c}_{i,j}$, representing whether persons i and j share attention.

We provide an example of these pairwise predictions in Figure 4.4. To evaluate these predictions, we adopt the following protocols, introduced in Gupta et al. (2024a,b).

LAH. Each sample corresponds to an individual person i. At inference, for each possible pair (i, j), we compute the predicted LAH score $\mathbf{e}_{i\to j}$ and select the person $\hat{j}$ with the highest predicted score:

$$\hat{j} = \arg\max_{j}, \mathbf{e}_{i\to j}. \tag{4.1}$$

For a ground truth positive case, the prediction is counted as a true positive if $\hat{j}$ matches the ground truth target and $\mathbf{e}_{i\to\hat{j}}$ exceeds the specified threshold. Otherwise, it is a false negative. For ground truth negatives, the prediction is a true negative if $\mathbf{e}_{i\to\hat{j}}$ is below the threshold; otherwise, it is a false positive. Using these values, we compute either AP and AUC scores across different thresholds, or F1 scores at a fixed threshold.

LAEO. Each sample corresponds to a pair of people (i, j). For each person i, we select $\hat{j}$ with the highest predicted LAEO score:

$$\hat{j} = \arg\max_{j} \mathbf{e}_{i \leftrightarrow j}. \tag{4.2}$$

We then set $\mathbf{e}_{i \leftrightarrow j} = 0$ for all $j \neq \hat{j}$ before computing performance metrics. This procedure is applied symmetrically for both i and j. Using these values, we compute either AP and AUC scores across different thresholds, or F1 scores at a fixed threshold.

SA. Each sample corresponds to a pair of people (i, j). We compute AP and AUC scores by thresholding the predicted scores $\mathbf{c}_{i,j}$ at varying levels.

4.3 Egocentric Gaze Estimation

This section describes the protocols and metrics used to evaluate egocentric gaze estimation performance. Like gaze following methods, typical architectures for egocentric gaze estimation—including those proposed in this thesis—predict a heatmap over the scene that highlights potential gaze targets. This predicted heatmap is then processed to compute evaluation metrics.

4.3.1 Heatmap

The heatmap-based metric, introduced by Li et al. (2018), directly compares the predicted and ground truth heatmaps at the pixel level. As with the AUC metric proposed by Chong et al. (2020b) (see Section 4.1.2), the ground truth heatmap is first binarized using a fixed threshold. However, unlike AUC, where the continuous predicted heatmap is compared directly to the binary ground truth, the predicted heatmap is also binarized in this case.

Following the implementation of Lai et al. (2023), a range of thresholds is applied to the predicted heatmap. The binarized predicted and ground truth heatmaps are then compared to compute precision, recall, and F1 scores for each threshold. The threshold yielding the best F1 score across the test samples is selected for final reporting.

As with the AUC metric discussed earlier, this heatmap metric is highly sensitive to the choice of binarization thresholds. It is also difficult to interpret and, importantly, does not account for semantic information about the gaze target.

4.3.2 Gaze Point

We propose two evaluation protocols inspired by gaze following—drawing from both prior literature and our own contributions.

Distance. We follow the protocol described above in Section 4.1.1 to compute the L2 distance between the predicted and ground truth gaze points on a normalized 1 × 1 square.

LAH. This semantic metric evaluates whether the predicted and ground truth gaze points fall within the head bounding boxes of other people in the scene. It is particularly valuable for egocentric gaze estimation in social settings—such as conversations—where gaze directed at other people's heads is both a common behaviour and a meaningful indicator of attention and engagement.

We follow the protocol introduced in Gupta et al. (2024a,b), as described in Section 4.2.1, to compute LAH performance. A true positive is recorded when both the predicted and ground truth gaze points fall on the *same* head bounding box. Precision, recall, and F1 scores are then computed across all data points.

4.4 Conclusion

In this chapter, we presented the evaluation metrics and protocols used for gaze following, social gaze prediction, and egocentric gaze estimation. We introduced social gaze metrics that extend standard gaze following evaluation by providing semantic interpretation of predicted gaze targets. These include LAH, LAEO, and SA, applied as post-processing metrics to assess whether gaze predictions correspond to socially meaningful targets and interactions.

We also proposed new evaluation protocols for social gaze prediction tasks, including an updated SA protocol that addresses key limitations of prior methods by enabling pairwise evaluation and handling multiple SA groups within a frame. For egocentric gaze estimation, we introduced gaze following inspired metrics, including distance and LAH, to complement conventional heatmap comparisons with semantically meaningful evaluation.

Together, these metrics offer deeper insight into model behaviour and error patterns, and were instrumental in guiding the development of the methods presented in the following chapters.

5 Gaze Following

Gaze following—the task of predicting where a person is looking within a scene—is a fundamental component of social perception and visual understanding. Accurately predicting someone's gaze target requires reasoning about a multitude of contextual cues (see Section 1.1.1). As early methods relied exclusively on RGB input (Recasens et al., 2015; Chong et al., 2020b; Lian et al., 2018), the limited size of available datasets may restrict the model's ability to implicitly learn all the relevant cues necessary to solve the task effectively.

In this chapter, we investigate how additional modalities such as depth and body pose can be explicitly leveraged to improve gaze prediction. We propose two models: the first, published in Gupta et al. (2022), introduces a modular multimodal framework that integrates RGB, depth, and pose cues, and can optionally operate in privacy-sensitive settings using only anonymized modalities. The second, published in Tafasca et al. (2023b), builds on this work by introducing a geometrically grounded method that constructs a 3D Field of View (3DFoV) from predicted gaze direction and a geometrically-consistent point cloud, enabling improved reasoning about visibility in 3D space.

The main contributions are:

- A multimodal architecture that fuses RGB, depth, and pose signals using an attention mechanism, achieving state-of-the-art performance on public benchmarks and supporting privacy-aware inference using only depth and pose (Section 5.1).
- A 3DFoV-based model that combines predicted gaze direction with geometry-preserving depth to highlight visible regions of the scene in 3D, also achieving state-of-the-art results and improving cross-dataset generalization (Section 5.2).
- A comprehensive evaluation across multiple datasets, including GazeFollow, VideoAttentionTarget, and the newly introduced ChildPlay dataset, demonstrating the importance of multimodal cues for improved performance.
- An analysis of semantic gaze behaviour, more specifically gaze toward other people's

heads, showing that it can capture complementary information to standard metrics.

5.1 Multimodal Gaze Target Prediction

In Gupta et al. (2022), we study the use of explicit depth and pose information to improve gaze target inference, as illustrated in Figure 5.1.

Depth provides the model with cues about object shapes and overall 3D scene structure. It helps disambiguate whether a person is looking at a foreground or background object and resolves ambiguities along the line of sight when the predicted gaze and scene geometry are inconsistent. Pose, meanwhile, offers accurate localization of body parts related to attention—such as the head and hands—which are common gaze targets in interactive and manipulation contexts. Additionally, pose encodes physical state and activity, providing further context to constrain likely gaze targets.

In privacy-sensitive scenarios, using raw RGB images can be problematic due to the presence of identifiable facial information—at both training and inference time. This is especially true for applications in health and surveillance. For instance, reduced eye contact and shared attention have been identified as early indicators of autism in children (Zwaigenbaum et al., 2019), but due to privacy constraints, raw videos from diagnostic sessions are typically not publicly available—even with facial blurring. In contrast, pose and depth data are inherently anonymized and more amenable to sharing. To address this, we develop models that operate solely on depth and pose and evaluate their performance against multimodal and RGB-only baselines.

Beyond modality fusion, we also study key technical aspects of model architecture. First, we examine the role of feature resolution. Gaze target localization is a dense prediction task, similar to pose estimation, where the output is a heatmap indicating the likelihood of each pixel being the gaze target. Many existing models (Lian et al., 2018; Chong et al., 2020b; Fang et al., 2021) rely on ResNet-style backbones that downsample spatial resolution considerably before upsampling, potentially losing important spatial detail. We instead adopt a Feature Pyramid Network (FPN) (Lin et al., 2017) architecture, which incorporates skip connections during upsampling to preserve spatial precision and improve localization accuracy.

Second, we examine fusion strategies. In many gaze models, gaze-related features are fused with the image early in the pipeline, requiring the full image to be processed for each person. We evaluate whether a late fusion approach—where gaze information is merged with intermediate feature maps—can achieve similar performance. Our findings indicate that early fusion remains more effective in this setting.

Approach and Contributions. We address the task of gaze target localization and make the following contributions:

- We propose a modular, multimodal gaze prediction architecture that integrates RGB,

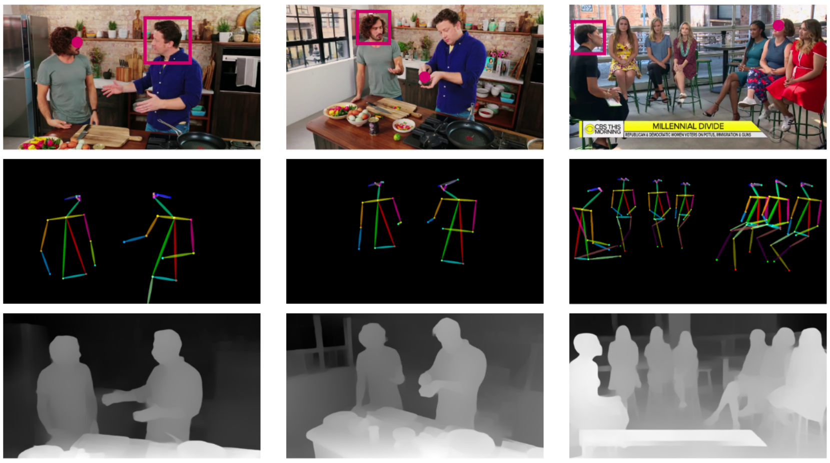


Figure 5.1: Sample cases where depth and pose information can improve gaze target inference. Left: pose cues indicate interaction, while depth helps ignore salient background distractors. Middle: manipulation activities benefit from hand localization. Right: depth allows disambiguation between potential face targets in the background.

depth, and pose via an attention-based fusion scheme and supports end-to-end training.

- We evaluate the use of only pose and depth inputs, demonstrating competitive performance in privacy-sensitive scenarios where RGB is unavailable or unsuitable.
- We adopt a Feature Pyramid Network for dense regression and show that preserving spatial resolution is important for accurate gaze heatmap prediction.

We conduct experiments on the GazeFollow (Recasens et al., 2017) and VideoAttentionTarget (Chong et al., 2020b) benchmarks. Our model achieves state-of-the-art results using all modalities and demonstrates strong performance even when trained on anonymized inputs, supporting its use in privacy-sensitive applications.

5.1.1 Method

An overview of our system is illustrated in Figure 5.2. It takes as input an image or a video frame, a set of derived modality images, and the head bounding box of a target person. The output is a gaze heatmap **H** where the location of the maximum value corresponds to the desired gaze point prediction.

Our network architecture consists of 3 modules. The first one is a *Human-Centric module* whose goal is, given the head crop of a person, to predict a gaze cone representing their visual field of view, i.e., the set of pixel locations where the person might be looking. The second one is a *Scene-Centric module* which is fed the image, the person's location (head mask), and the gaze cone in order to generate a feature saliency map **F** highlighting possible gaze target locations. The last one is a *Prediction module* comprising two heads: one for inferring the

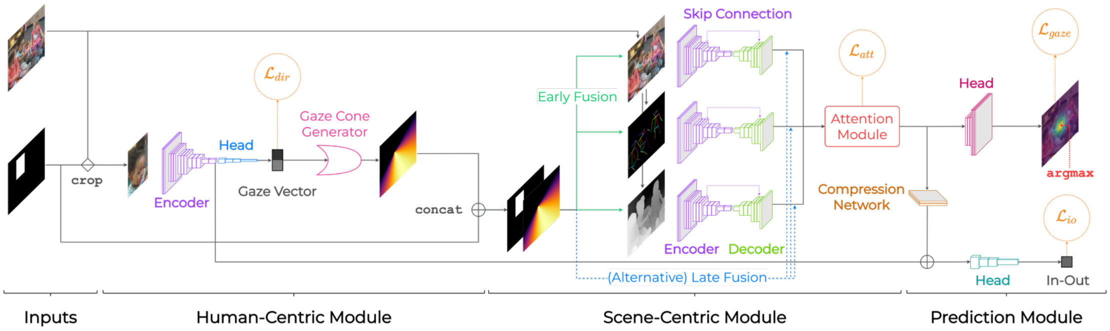


Figure 5.2: Given an input image and a target subject's head location, we first extract depth and pose maps from the image using off-the-shelf pre-trained models. Next, the Human-Centric module takes the person's head crop as input and predicts a 2D gaze vector which is used to generate a gaze cone image. Then, the Scene-Centric module processes the original image, the depth image and the pose map in order to produce modality saliency feature maps (using modality-specific feature extractors) which are fused by an Attention module. The resulting saliency map is used by the Prediction module to regress a gaze heatmap, and optionally predict an in-vs-out gaze classification score.

gaze heatmap, and the second one for predicting whether the gaze target is located within the frame. These components are detailed below.

Human-Centric Module

In this module, a sub-network $\mathcal{G}$ takes the head crop image $\mathbf{I}^{head}$ of the target person as input and predicts a normalized 2D gaze vector $\mathbf{g}^p_{2D} = \mathcal{G}(\mathbf{I}^{head})$. This gaze vector is used by a *gaze cone generator* to produce a gaze cone image $\mathbf{I}^{co}$. Finally, the gaze cone image is concatenated with the binary head mask of the target person $\mathbf{I}^{mask}$ and passed to the Scene-Centric branch for further processing.

Gaze Cone Generator. The gaze cone is a way to modulate the image information appearing in the gaze direction of the person. It is encoded as an image in order to be consistent with the rest of the architecture. The gaze cone generator produces $\mathbf{I}^{co}$ by drawing a cone from the subject's eyes location $\mathbf{p}_{eye}$ (i.e. eye mid-point if available from the pose modality; otherwise, using a prototypal location in the head bounding box) along the direction of $\mathbf{g}^p_{2D}$. To account for uncertainties in gaze prediction, the cone has an aperture of α_{co} (set to π in practice), and the intensity decays the farther we are from the gaze direction angle-wise. Specifically, the value at each pixel location $\mathbf{p}$ is scored according to the cosine similarity between the predicted gaze vector and the eye-to-target direction (see example in Fig. 5.2).

Formally:

$$\begin{gathered} \forall \mathbf{p} = (i, j) \text{ where } (i, j) \in [1,\ w] \times [1,\ h], \\ \mathbf{I}^{co}(\mathbf{p}) = \max\left(0, \cos(\mathbf{g}^p_{2D}, \mathbf{p} - \mathbf{p}_{eye})\right) \end{gathered} \tag{5.1}$$

Note that given this definition, the gaze cone generator is differentiable, allowing us to train our architecture end-to-end.

Scene-Centric Module

In the Scene-Centric module, the input image **I** is first transformed using different networks (see implementation details) into a set of modality images $\mathbf{I^m}$, where $m \in \{raw, pose, depth\}$ and $\mathbf{I^{raw}} = \mathbf{I}$ by definition. These modalities are passed through feature extractors to produce feature maps, which are then fused using an attention mechanism to create a single combined feature map.

Feature Extractors. A set of modality-specific feature extractors $\mathscr{F}_m$ are used to compute feature maps $\mathbf{F}_m$ to encode the person-specific salient regions of the scene according to the input modality. Thus, each feature extractor $\mathscr{F}_m$ processes its corresponding modality $\mathbf{I^m}$ concatenated with the output of the Human-Centric module, so that we have:

$$\mathbf{F}_m = \mathscr{F}_m(\mathbf{I^m}, \mathbf{I}^{co}, \mathbf{I}^{mask}) \tag{5.2}$$

Note that the concatenation can be seen as an early fusion scheme, whereas an alternative (so far less successful) approach consists in fusing the Human-Centric module information later at the feature level (see late fusion experiments). While multiplication is a more straightforward way to fuse the modality image and the gaze cone, in practice it produced worse results, probably because it performs a hard decision based on potentially inaccurate gaze direction predictions. This is particularly the case when the subject's head is facing backwards and the gaze vector is more difficult to estimate. Concatenation, on the other hand, allows the model to make that decision later in the processing.

Regarding the network, we use a typical image-to-image approach, relying on an encoder-decoder architecture. However, in contrast to previous works which simply upsample the lowest resolution representation produced by the encoder (Lian et al., 2018; Chong et al., 2020b), we use skip connections from different intermediate representations (at different resolutions) to their corresponding decoder representations in the style of a Feature Pyramid Network (Lin et al., 2017). This architectural choice aims to retain information from higher resolution representations, which is important in dense prediction tasks, and is further evidenced by our experiments.

Attention Module. Its goal is to perform a soft-selection of the most appropriate input modality given the scene. It takes as input the set of feature maps $\mathbf{F}_m \in \mathbb{R}^{w \times h \times d_m}$ and produces a single combined feature map **F**, which we use to predict the outputs.

Concretely, it performs four steps:

1. Each feature map $\mathbf{F}_m$ is passed through a modality-specific convolution layer to produce

a transformed feature map $\mathbf{T}_m$.

2. Each map $\mathbf{T}_m$ is passed through a network $\mathscr{C}_m$ consisting of three strided convolution layers followed by a global max pooling to generate an embedding vector $\mathbf{e}_m$. All embeddings are then concatenated to form the global embedding $\mathbf{e}$.
3. The global embedding is passed through a projection layer P followed by a softmax operation to get the attention weights: $\{w_m\} = \text{softmax}(P(\mathbf{e}))$.
4. Finally, the output is computed as the weighted sum of the transformed feature maps: $\mathbf{F} = \sum_m w_m \mathbf{T}_m$.

This loosely resembles the self-attention mechanism in a transformer (Vaswani et al., 2017): the transformed feature maps $\mathbf{T}_m$ act as the values, whereas the attention weights w_m simulate a dot product between an implicit query and a set of keys.

In addition, this attention mechanism allows us to use a variable number of modalities during inference because the model can simply assign a weight of 0 when a modality is absent. To encourage this behaviour, we perform modality dropout during training, i.e., we randomly provide a white noise image instead of the dropped modality, and use an attention loss for supervision (see Section 5.1.1).

Prediction Module

This module uses the feature map to predict the quantities of interest: a gaze heatmap $\mathbf{H}$, and a binary In-Out flag o indicating whether the gaze target is inside or outside the image. It comprises two parts, which are explained below.

Gaze prediction head. The gaze target heatmap $\mathbf{H}$ is regressed from the combined feature map $\mathbf{F}$ using a prediction decoder $\mathscr{R}$ that consists of an analytic upsampling followed by a set of convolution layers:

$$\mathbf{H} = \mathscr{R}(\mathbf{F}) \tag{5.3}$$

The location where the heatmap is maximal is then used as the gaze target prediction.

In-Out prediction head. In general, we want to predict whether the person is looking at a scene location which is visible in the image or not. This is important as we do not want to use the gaze target prediction when a person is looking outside the frame. To accomplish this, we attach an In-Out network prediction head $\mathscr{O}$ which takes as input the feature map $\mathbf{F}$ resulting from the attention step as well as a gaze embedding $\mathbf{e}_g$ coming from the human-centric module (see Fig. 5.2):

$$o = \mathscr{O}(\mathbf{F}, \mathbf{e}_g) \tag{5.4}$$

More precisely: first, the map $\mathbf{F}$ is passed through a network having the same architecture as $\mathscr{C}_m$ to produce a scene embedding $\mathbf{e}_s$ which is concatenated with the gaze embedding $\mathbf{e}_g$ and fed into an In-Out predictor consisting of 2 linear layers followed by a sigmoid activation.

Loss

The complete model is trained end-to-end using a combination of four losses:

1. **Gaze loss** $\mathscr{L}_{hm}$. It measures the error in gaze location prediction, which is done by computing the pixel-wise L2 loss between the predicted heatmap $\mathbf{H}^{pred}$ and the ground truth gaze target heatmap $\mathbf{H}^{gt}$, defined as a Gaussian blob centered on the ground-truth location.
2. **Gaze direction loss** $\mathscr{L}_{dir}$. This loss constrains the Human-Centric module by maximizing the cosine similarity between the predicted 2D gaze vector $\mathbf{g}_{2D}^{p}$ and the ground truth vector $\mathbf{g}_{2D}^{gt}$, derived from the ground-truth gaze point.
3. **In-Out loss** $\mathscr{L}_{io}$. We use a standard binary cross-entropy loss to measure whether a person is looking inside or outside the image frame.
4. **Attention loss (modality drop)** $\mathscr{L}_{att}$. This loss supervises the Attention module (Section 5.1.1) by pushing the attention weight w_m of a dropped modality toward 0:

$$\mathscr{L}_{att} = \sum_{m} w_m \cdot \mathbb{1}_{m \in \text{dropped}}$$

The final loss is a linear combination of the four terms:

$$\mathscr{L} = \lambda_{hm}\mathscr{L}_{hm} + \lambda_{dir}\mathscr{L}_{dir} + \lambda_{io}\mathscr{L}_{io} + \lambda_{att}\mathscr{L}_{att} \tag{5.5}$$

Implementation Details

Modality extraction. The pose maps are extracted using HRFormer (Yuan et al., 2021), a hybrid of HRNet (Wang et al., 2020) and transformer architectures. Depth maps are generated using MiDaS (Ranftl et al., 2020). Pose maps are represented as RGB skeleton renderings, with color-coded limbs and keypoints.

Feature extraction networks $\mathscr{F}_m$**.** The feature extractors use EfficientNet backbones (Tan and Le, 2019). We use EfficientNet-B1 for RGB (7.8M parameters) and EfficientNet-B0 for depth and pose (5.3M parameters). Input modalities are resized to 224 × 224 and processed through encoder-decoder networks with intermediate resolutions between 56 × 56 and 7 × 7. These intermediate representations are passed through skip connections in a Feature Pyramid Network decoder to produce final saliency maps at 56 × 56 resolution.

Gaze subnetwork $\mathscr{G}$**.** The Human-Centric branch uses a ResNet-18 (11M parameters) backbone, with a custom head to predict the 2D gaze direction from a resized 224 × 224 head crop.

Prediction module. In the prediction module (Figure 5.2), the feature maps $\mathbf{F}_m$, $\mathbf{T}_m$, and $\mathbf{F}$ are maintained at 56 × 56 resolution. Before predicting the gaze heatmap, $\mathbf{F}$ is upsampled to 64 × 64 and passed through convolutional layers. Embedding vectors $\mathbf{e}_g$, $\mathbf{e}_m$, and $\mathbf{e}_s$ all have dimensionality 512.

5.1.2 Experiments

We evaluate our proposed models using two benchmark datasets and follow established evaluation protocols and metrics.

Datasets. The first dataset is the *GazeFollow* dataset (Recasens et al., 2015), introduced in Section 3.1.1.

The second is the *VideoAttentionTarget* dataset (Chong et al., 2020b), described in Section 3.1.2. Compared to GazeFollow, it features higher-resolution frames and more close-up, front-facing views of people. It also contains a greater number of instances involving gaze directed at other people's heads (see Section 3.4).

Training Protocol. Ground-truth gaze heatmaps $\mathbf{H}^{gt}$ are constructed by placing a Gaussian blob centered at the ground-truth gaze location, using $\sigma = 3$ at a 64×64 resolution following previous works (Chong et al., 2020b).

The backbone of the Human-Centric sub-network is pre-trained on the Gaze360 dataset (Kellnhofer et al., 2019) for 3D gaze estimation. The scene branch encoders are initialized with ImageNet-pretrained weights (Russakovsky et al., 2015). To train our multimodal models, we first pre-train each modality-specific model independently (see Section 5.1.2) and initialize the multimodal model using their learned weights.

Following the protocol of Fang et al. (2021), all VideoAttentionTarget models are initialized with weights trained on GazeFollow. To reduce redundancy, we subsample the VideoAttentionTarget training frames by selecting every third frame.

All models are trained end-to-end using the AdamW optimizer (Loshchilov and Hutter, 2018). For GazeFollow, we use a learning rate of 1×10^{-4}, and for VideoAttentionTarget, 1×10^{-5}. Loss weights are set to: $\lambda_{gaze} = 100$, $\lambda_{dir} = 0.1$, $\lambda_{io} = 1$, and $\lambda_{att} = 1$. We train for 35 epochs on GazeFollow and 20 epochs on VideoAttentionTarget (40 for the multimodal model).

Performance Metrics. We evaluate performance using standard metrics for gaze target prediction: *AUC; Recasens et al. (2015)*(Section 4.1.2) and *Distance* (Section 4.1.1) for spatial localization, and *In-Out Average Precision (AP)* (Section 4.1.3) for binary classification of whether the gaze target lies inside or outside the image frame.

In-Out AP is computed over the full test set, whereas AUC and Distance are evaluated only on the subset of samples where the ground truth gaze point is annotated and falls within the image bounds.

Tested Models

Individual Modalities. To isolate the contribution of each modality, we train single-modality models by disabling the attention fusion mechanism and using only the corresponding modal-

Model	**AUC↑**	**AvgDist↓**	**MinDist↓**
Lian et al. (2018)	0.906	0.145	0.081
Chong et al. (2020b)	0.921	0.137	0.077
Jin et al. (2021)	0.919	0.126	0.076
Fang et al. (2021)	0.922	0.124	0.067
Image	0.933	0.134	0.071
Depth	0.921	0.141	0.080
Pose	0.902	0.164	0.100
Multimodal	0.943	0.114	0.056
Depth-privacy	0.920	0.152	0.088
Pose-privacy	0.893	0.175	0.109
Multimodal-privacy	0.928	0.136	0.075
Image-NoSkip	0.932	0.133	0.073
Multimodal-NoMoDrop	0.941	0.115	0.057
Multimodal-Late	0.931	0.128	0.068

Table 5.1: Results for our models on the GazeFollow dataset. Best scores are given in red and second best scores are given in blue.

Model	**AUC ↑**	**Dist ↓**	**AP ↑**
Chong et al. (2020b)-static	0.854	0.147	0.848
Chong et al. (2020b)	0.860	0.134	0.853
Jin et al. (2021)	0.870	0.127	0.882
Fang et al. (2021)	0.905	0.108	0.896
Image	0.918	0.122	0.864
Depth	0.899	0.134	0.852
Pose	0.904	0.131	0.866
Multimodal	0.913	0.110	0.879
Depth-privacy	0.891	0.156	0.831
Pose-privacy	0.881	0.150	0.823
Multimodal-privacy	0.895	0.140	0.826
Image-NoSkip	0.906	0.133	0.857
Multimodal-NoMoDrop	0.905	0.118	0.874
Multimodal-Late	0.905	0.113	0.863

Table 5.2: Results on the VideoAttentionTarget dataset. Best scores are given in red and second best scores are given in blue.

ity's feature map **F** as input to the prediction module.

Privacy Approach. To support privacy-sensitive settings, we modify the Human-Centric module to operate on the head crop from the pose image (i.e., a rendered skeleton) instead of the original RGB image. Further, the Feature Extractors use only processed and anonymized input data. The rest of the architecture and training pipeline remain unchanged.

Late Fusion. We test a late fusion strategy where the gaze cone image g_{cone} and binary head mask h_{loc} are downsampled and concatenated with the feature map $\mathbf{F}_m$ of each modality later in the architecture, instead of fusing earlier. This variant is illustrated by the dashed blue line in Figure 5.2.

Skip Connections. To assess the impact of spatial detail retention, we train an image-only model without skip connections in the decoder of the Scene-Centric branch.

Modality Dropout. We train a multimodal model without modality dropout to evaluate the effect of this regularization strategy on the attention mechanism.

State-of-the-Art. We compare our models with leading gaze prediction methods, including those of Chong et al. (Chong et al., 2020b), Lian et al. (Lian et al., 2018), Jin et al. (Jin et al., 2021), and Fang et al. (Fang et al., 2021). For VideoAttentionTarget, we focus on static variants when available to ensure fair comparisons with our single-frame models.

5.1.3 Results

Our results on the GazeFollow and VideoAttentionTarget datasets are summarized in Table 5.1 and Table 5.2 respectively.

Individual Modalities. As expected, the image modality performs best overall due to its

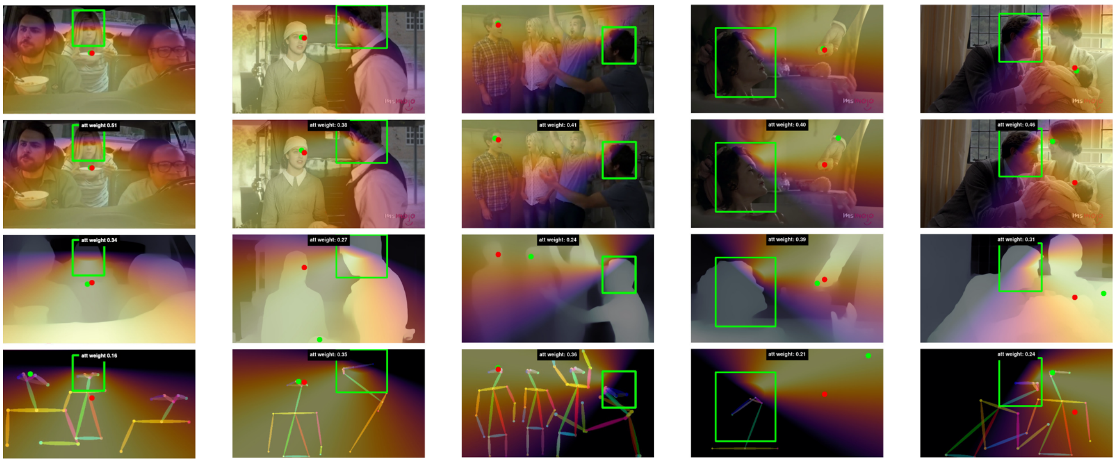


Figure 5.3: Qualitative results of our models (from top to bottom: multimodal, image, depth and pose). The image (or modality) is superimposed with the predicted gaze cone, the predicted gaze target (in **green**) and the ground truth target (in **red**). We observe that the attention scores reflect the reliability of the respective modalities for a particular sample (pose in 2nd and 3rd column; depth in 4th column), and that the fusion is able to ignore wrong information (pose in 1st and 4th columns; depth in 3rd column), and improve predictions of the image modality (4th and 5th column).

rich content. However, depth and pose still yield competitive results—often comparable to several state-of-the-art baselines. Notably, on GazeFollow, depth outperforms pose across all metrics, whereas on VideoAttentionTarget, the two perform similarly. This likely reflects the higher prevalence of face-directed gaze in VideoAttentionTarget, where pose information (e.g., skeletal landmarks) becomes more informative.

Multimodal. The multimodal model consistently outperforms all single-modality variants across both datasets, particularly in distance metrics, where we observe up to a 21% error reduction on GazeFollow. This improvement stems from the complementary nature of pose and depth cues—pose improves localization near people and hands, while depth resolves spatial ambiguities along the line of sight.

Compared to existing state-of-the-art methods, our multimodal approach achieves the best performance on GazeFollow across all metrics. On VideoAttentionTarget, our results match those of Fang et al. (Fang et al., 2021), despite the fact that their method uses explicit eye-region processing—a strong cue in high-resolution frontal faces. We hypothesize that incorporating such a component into our framework could yield further gains.

Qualitative Results. Figure 5.3 presents qualitative examples and attention scores for various models. The attention weights highlight the relative reliability of each modality per example: for instance, pose receives higher weights in person-focused scenarios, while depth is emphasized when background clutter must be resolved.

On average, attention weights on GazeFollow are 0.41 (image), 0.36 (depth), and 0.23 (pose), while on VideoAttentionTarget they are more balanced: 0.37, 0.31, and 0.32 respectively. This shift suggests that pose plays a more prominent role in datasets with more face-directed attention.

Privacy-Sensitive Models. Our privacy-preserving models, using only depth and pose, perform competitively—especially on GazeFollow, where the head and eye cues in the original image are less visible. Performance drops are more pronounced on VideoAttentionTarget, likely due to its high-resolution, front-facing faces, where RGB features offer additional gaze cues. Still, the multimodal privacy model achieves performance comparable to many baselines, demonstrating its utility in sensitive applications.

Skip Connections. The use of skip connections in our decoder contributes notably to improved performance, especially on the VideoAttentionTarget dataset. Removing them (NoSkip model) leads to a clear drop in distance accuracy, highlighting the importance of preserving spatial detail in high-resolution scenes.

Late Fusion. While late fusion outperforms single-modality models, it underperforms compared to early fusion in the multimodal setting. This likely occurs because late fusion delays the integration of gaze direction, requiring the model to identify targets for all people in the scene, rather than conditioning early on the specific subject's gaze.

Modality Dropout. Disabling modality dropout leads to reduced performance on VideoAttentionTarget, with less balanced attention weights (image: 0.32, depth: 0.29, pose: 0.39). When dropout is enabled, the learned weights (0.37, 0.31, 0.32) better reflect modality relevance and contribute to more robust predictions. This demonstrates that modality dropout improves generalization by regularizing the fusion process.

5.1.4 Conclusion

In this work, we proposed a modular multimodal architecture that explicitly leverages pose and depth information to enhance gaze target prediction. Our approach achieves state-of-the-art performance on two public benchmarks and improves gaze localization through more accurate spatial reasoning. We also investigated a late fusion scheme, which allows the scene to be parsed in a person-agnostic manner before incorporating subject-specific information. In addition, we demonstrated that our model can be adapted for privacy-sensitive applications by using only anonymized input in the form of pose skeletons and depth maps, while still achieving competitive results.

Our architecture is inherently modular and can be extended to incorporate additional modalities such as optical flow in video settings, which may further enhance prediction quality. Another promising direction is the integration of temporal information directly within the model, enabling more robust predictions in dynamic scenes. Furthermore, while depth appears useful for validating the geometric compatibility between the gaze direction and scene layout, future work is needed to systematically understand how this cue is utilized by the

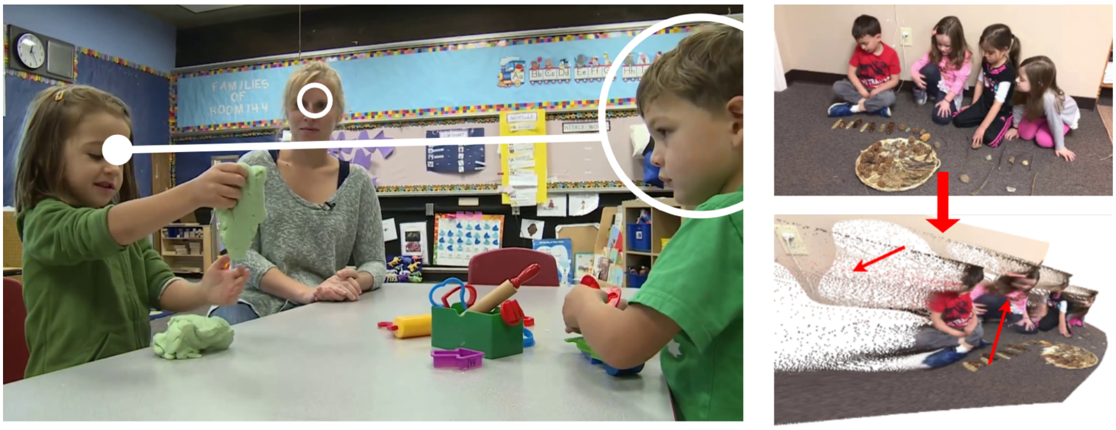


Figure 5.4: Depth is crucial for gaze following, as it enables reasoning about the 3D structure of the scene and helps rule out salient objects that lie along the 2D line of sight but are not visible in 3D (left). However, typical methods rely on disparity maps (*e.g.* (Ranftl et al., 2020)), which often produce distorted point clouds (right) and thus limit 3D reasoning.

model. Lastly, our current attention mechanism assumes that a single dominant modality is selected, which implicitly treats all modalities as equivalent. Future research could explore more expressive fusion strategies that capture complementary information across modalities.

5.2 Geometry Preserving Gaze Target Prediction

As discussed in the previous section, reasoning about depth is crucial for accurate gaze target prediction. Depth provides information about the 3D structure of the scene, enabling geometric reasoning that helps determine whether objects along a 2D line of sight are actually visible to the observer. For example, salient objects may fall along the projected 2D gaze vector but be occluded or outside the viewer's true field of view in 3D space.

Given that most gaze datasets lack ground-truth depth, several prior works have used depth estimated from pre-trained monocular networks to augment gaze prediction models (Fang et al., 2021; Bao et al., 2022; Gupta et al., 2022). However, these depth estimators (Ranftl et al., 2020; Yin et al., 2020) often produce outputs that are only accurate up to an unknown scale and shift, which can lead to distorted point clouds and limit their usefulness for 3D geometric analysis (see Figure 5.4 and Section 5.4).

In this work published in Tafasca et al. (2023b), we propose a more geometrically grounded formulation by leveraging a recent depth estimation method (Patakin et al., 2022) that corrects for these shift and scale ambiguities. This results in geometry-preserving depth maps suitable for reliable 3D reconstruction. We use these maps to construct point clouds and explicitly intersect them with the predicted 3D gaze vector to compute the person's 3D field-of-view (3DFoV).

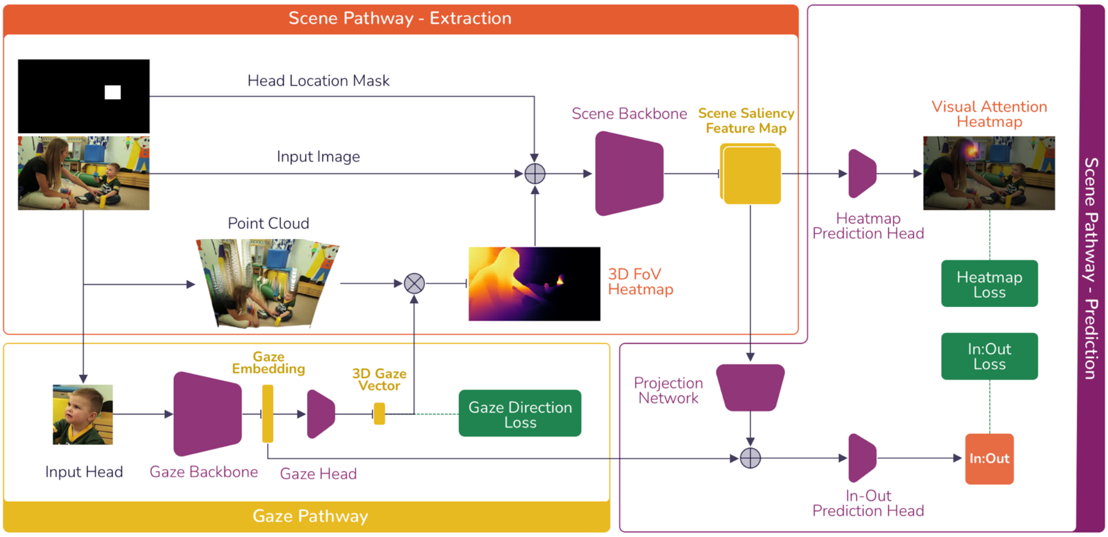


Figure 5.5: The Gaze Pathway processes the head crop to predict a 3D gaze vector, which is then used with the inferred point cloud to generate a heatmap of the 3D Field-of-View. The Scene Pathway further combines this map with the image and a head location mask to predict a feature map highlighting salient items in the scene. This map is further used to predict, on one hand, the visual attention map $\mathbf{H}^p$, and on the other hand, with the gaze embedding $\mathbf{e}_g$, the in-vs-out gaze label.

In experiments, we show that our approach leads to improved performance, especially in cross-dataset generalization settings, validating the benefit of maintaining geometric fidelity in depth-based gaze estimation.

5.2.1 Method

Our network architecture is illustrated in Figure 5.5. It builds on the structure of our previous design, with key modifications to the gaze and scene pathways to explicitly incorporate 3D scene geometry.

On one hand, the Gaze Pathway aims at predicting the scene elements which are in the 3D Field of View (3DFoV) of the person, represented by the heatmap $\mathbf{V}$. To do so, it takes as input the image crop $\mathbf{I}^{head}$ of the person's head and predicts their 3D gaze direction $\mathbf{g}^p_{3D}$, as well as a gaze embedding $\mathbf{e}_g$. The gaze $\mathbf{g}^p_{3D}$ is then combined with the scene-consistent 3D point cloud $\mathbf{P}$ inferred from the image to generate the 3DFoV heatmap.

On the other hand, the scene pathway combines the image with the Gaze Pathway information (the head location mask $\mathbf{I}^{mask}$, the $\mathbf{V}$, and the gaze embedding $\mathbf{e}_g$) to infer the in-out label o^p (i.e., whether the person is looking inside the frame or outside) and the attention heatmap $\mathbf{H}^p$. We detail these elements below. However, given the importance of scene structure representation, we first describe how the scene point cloud is obtained.

Point cloud generation

To obtain our point cloud in the camera coordinate system $\mathbf{P^c} = \{\mathbf{P}_i^\mathbf{c} = (X_i^\mathbf{c}, Y_i^\mathbf{c}, Z_i^\mathbf{c})\}$ associated with the 2D pixels defined in the image plane $\mathbf{p}_i = (x_i, y_i)$, we need to know the scene depth as well as the intrinsic parameters of the camera. As these are not available, we infer them from the data and make standard assumptions.

Regarding depth, we leverage the pre-trained model of (Patakin et al., 2022) to predict the depth $Z_i^\mathbf{c}$ of each pixel. We chose this model because it generates geometrically consistent depth maps, which are crucial for performing reliable 3D analysis of the scene.

As to camera parameters, we make standard assumptions: square pixels, no skew, and the principal point at the image center. The more important parameter is the focal length, which is required to avoid scene stretching. In this paper, we estimate it using the pre-trained model of (Yin et al., 2021). As a result, denoting by W and H the image width and height, we obtain the simplified projection equation:

$$\begin{bmatrix} X_i^\mathbf{c} \\ Y_i^\mathbf{c} \\ Z_i^\mathbf{c} \end{bmatrix} = \begin{bmatrix} f & 0 & W/2 \\ 0 & f & H/2 \\ 0 & 0 & 1 \end{bmatrix}^{-1} . \begin{bmatrix} x_i.Z_i^\mathbf{c} \\ y_i.Z_i^\mathbf{c} \\ Z_i^\mathbf{c} \end{bmatrix} \tag{5.6}$$

enabling us to build our point cloud $\mathbf{P^c}$.

Note that $\mathbf{P^c}$ is defined in the camera coordinate system. However, as our aim is to evaluate the scene elements visible from the person's viewpoint, we transform it into the local eye coordinate system C^{eye} in which the gaze vector is predicted (see next section), resulting in $\mathbf{P^e}$. Following (Kellnhofer et al., 2019), the origin of C^{eye} is defined by the eye location $\mathbf{P}^\mathbf{c}_{eye}$, and the basis vectors (E_x, E_y, E_z) are such that E_z is the unit vector from the camera to the eye, and E_x and E_y lie in the plane perpendicular to E_z.

$$\mathbf{P^e} = \begin{bmatrix} E_x^\mathrm{T} \\ E_y^\mathrm{T} \\ E_z^\mathrm{T} \end{bmatrix} .(\mathbf{P^c} - \mathbf{P}^\mathbf{c}_{eye}) \tag{5.7}$$

We provide an analysis of the quality of the generated point clouds in the appendix.

Gaze pathway

This pathway comprises several steps to generate the 3DFoV heatmap $\mathbf{V}$, as described below.

Gaze Prediction Network. Its aim is to predict the gaze direction $\mathbf{g}_{3D}^p$ defined in the local coordinate system C^{eye} associated with the head crop $\mathbf{I}^{head}$ (Kellnhofer et al., 2019). C^{eye} is used rather than the camera coordinate system, as the gaze depends primarily on appearance (head pose and eyes) and not on the head location within the image. This network is composed of a gaze prediction backbone $\mathscr{G}_b$ and a gaze prediction head $\mathscr{G}_h$. The first one, $\mathscr{G}_b$, is a ResNet-18 (He et al., 2016) network that predicts the gaze embedding $\mathbf{e}_g$ from the head crop $\mathbf{I}^{head}$, while the second is an MLP with 2 layers followed by a tanh activation which transforms this gaze embedding into the unit 3D gaze vector prediction $\mathbf{g}^p_{3D}$:

$$\mathbf{e}_g = \mathscr{G}_b(\mathbf{I}^{head}) \quad \text{and} \quad \mathbf{g}^p_{3D} = \mathscr{G}_h(\mathbf{e}_g) \tag{5.8}$$

3DFoV heatmap V generation. Its goal is to highlight the scene parts lying in the gaze direction of the person. To do so, given the point cloud $\mathbf{P^e}$ and the gaze prediction $\mathbf{g}^p_{3D}$, we compute the cosine similarity $\mathbf{c}$ between $\mathbf{g}^p_{3D}$ and every point $\mathbf{P}^\mathbf{e}_i$ in $\mathbf{P^e}$, and apply an exponential decay function for values with lower similarity to enhance the scene parts that are more in the focus of gaze:

$$\mathbf{V}_i = \begin{cases} \mathbf{c}_i, & \text{if } \mathbf{c}_i > 0.9 \\ 0.9 \cdot \frac{\exp(5 \cdot \mathbf{c}_i)}{\exp(5 \cdot 0.9)}, & \text{otherwise} \end{cases} \tag{5.9}$$

where $\mathbf{c}_i = \mathbf{g}^p_{3D} \cdot \frac{\mathbf{P}^\mathbf{e}_i}{\|\mathbf{P}^\mathbf{e}_i\|}$. This formulation of the 3DFoV is differentiable, allowing for end-to-end training.

Scene Pathway

The scene pathway combines the scene information (image $\mathbf{I}$) with the 3DFoV heatmap $\mathbf{V}$ of the person (along with the head location mask $\mathbf{I}^{mask}$ to better characterize the person's position and scale in the scene) and the gaze embedding $\mathbf{e}_g$ to infer their attention: the in-out indicator o_p and the visual attention heatmap $\mathbf{H}_p$, according to:

$$\mathbf{F} = \mathscr{F}([\mathbf{I}, \mathbf{V}, \mathbf{I}^{mask}]) \tag{5.10}$$

$$\mathbf{H}^p = \mathscr{R}(\mathbf{F}), \quad o^p = \mathscr{O}([\mathscr{C}(\mathbf{F}), \mathbf{e}_g]) \tag{5.11}$$

Saliency feature extraction. The scene backbone network $\mathscr{F}$ is an encoder-decoder architecture that produces a set of gaze saliency feature maps $\mathbf{F}$. The encoder is an EfficientNet-B1 (Tan and Le, 2019) network, and the decoder is a Feature Pyramid Network (FPN) (Lin et al., 2017). The FPN includes skip connections that help retain high-resolution spatial information, which improves gaze target localization. The concatenation of inputs in Eq. 5.10 represents an early fusion scheme of scene and gaze information, which as we showed in our previous work (Gupta et al., 2022) above, was more effective than late fusion.

Visual attention prediction. Summarized in Eq. 5.11, this comprises two parts: the attention prediction head $\mathscr{R}$ processes the feature maps $\mathbf{F}$ to predict the gaze target heatmap $\mathbf{H}^p$, whose maximum gives the predicted gaze target location. It is implemented as a CNN block with six dilated convolution layers and a final 1 × 1 regression layer.

The in-out prediction head $\mathscr{O}$ determines whether the gaze target lies within the image frame. It is implemented as an MLP with two layers followed by a sigmoid activation, and takes as input the scene embedding $\mathbf{e}_s$ (from the compression network $\mathscr{C}$) and the gaze embedding $\mathbf{e}_g$. $\mathscr{C}$ is a 3-layer CNN with strided convolutions and max pooling.

Ground Truth and Loss Definition

Heatmap Ground Truth. As in the protocol described in Section 5.1.1, we generate the ground truth heatmap by placing a Gaussian at the annotated 2D gaze location. However, instead of using a fixed standard deviation, we define it relative to the heatmap resolution:

$$\sigma = \frac{(W_{hm} + H_{hm})}{2} \cdot \frac{3}{64},$$

where W_{hm} and H_{hm} denote the width and height of the heatmap. For a 64 × 64 heatmap, this yields $\sigma = 3$, consistent with prior work (Chong et al., 2020b).

3D Gaze Vector Pseudo-Ground Truth. Unlike prior methods that supervise gaze prediction only in 2D, we leverage our geometry-preserving point cloud to derive a pseudo-ground truth 3D gaze vector. Given the 2D annotated gaze point, we retrieve its corresponding 3D location $\mathbf{P}^{\mathbf{e}}_{gaze}$ from the point cloud $\mathbf{P}^{\mathbf{e}}$ defined in the eye coordinate system (see Section 5.2.1). The pseudo-ground truth 3D gaze direction is then defined as:

$$\mathbf{g}^{gt}_{3D} = \frac{\mathbf{P}^{\mathbf{e}}_{gaze}}{\|\mathbf{P}^{\mathbf{e}}_{gaze}\|}.$$

This allows us to directly supervise the predicted 3D direction against a geometrically meaningful reference.

Loss Definition. The model is trained using a weighted combination of three loss components:

$$\mathscr{L} = \lambda_{hm}\mathscr{L}_{hm} + \lambda_{dir}\mathscr{L}_{dir} + \lambda_{io}\mathscr{L}_{io}.$$

1. **Heatmap Loss $\mathscr{L}_{hm}$:** As described in Section 5.1.1, it measures the pixel-wise mean squared error between the predicted heatmap $\mathbf{H}^p$ and the ground truth heatmap $\mathbf{H}^{gt}$:

2. **In-Out Loss $\mathscr{L}_{io}$:** Following the definition in Section 5.1.1, this term applies binary cross-entropy to supervise the predicted in-vs-out label o^p against the ground truth label o^{gt}.

3. **3D Gaze Direction Loss $\mathscr{L}_{dir}$:** To encourage accurate 3D gaze direction prediction, we introduce a direction loss that maximizes the cosine similarity between the predicted gaze vector $\mathbf{g}^p_{3D}$ and the pseudo-ground truth vector $\mathbf{g}^{gt}_{3D}$:

$$\mathscr{L}_{dir} = 1 - \langle \mathbf{g}^p_{3D}, \mathbf{g}^{gt}_{3D} \rangle,$$

where $\langle a, b\rangle$ denotes the inner product between vectors a and b.

5.2.2 Experiments

Implementation Details. The gaze network head $\mathcal{G}_h$ is pre-trained on the Gaze360 dataset (Kellnhofer et al., 2019) and processes the head crop at a resolution of 224 × 224. The Scene Pathway encoder is pre-trained on ImageNet (Russakovsky et al., 2015) and operates on scene images resized to 512 × 288. During testing, we maintain the original aspect ratio and scale the longer side of the scene image to 512 pixels.

Datasets. We train our models on three datasets—GazeFollow, ChildPlay, and VideoAttentionTarget—as described in Section 3.1. Notably, GazeFollow and VideoAttentionTarget include more instances of gaze directed toward heads, whereas children in ChildPlay more often gaze at nearby objects such as the toys they are interacting with (Tafasca et al., 2023b).

Training. We train the model for 40 epochs on GazeFollow. Following the protocol of Chong et al. (2020b), we fine-tune the GazeFollow-trained model for 20 additional epochs on VideoAttentionTarget. We adopt the same fine-tuning protocol for ChildPlay. Training is conducted using the AdamW optimizer (Loshchilov and Hutter, 2018), with a learning rate of 2.5×10^{-4} for GazeFollow and 2.5×10^{-5} for both VideoAttentionTarget and ChildPlay. The loss coefficients are set to $\lambda_{hm} = 100$, $\lambda_{dir} = 0.1$, and $\lambda_{io} = 1$.

Validation. Since GazeFollow and VideoAttentionTarget do not provide validation splits, we create them by holding out a subset of the training data. Our GazeFollow validation set contains 4499 samples, while the VideoAttentionTarget validation set contains 6726 samples from 3 TV shows. We select the best-performing checkpoint based on the distance metric evaluated on the validation set.

Metrics. We evaluate model performance using standard metrics for gaze target prediction: *AUC* (Recasens et al., 2015) (Section 4.1.2) and *Distance* (Section 4.1.1) for spatial localization, and *In-Out Average Precision (AP)* (Section 4.1.3) for binary classification of whether the gaze lies inside or outside the frame.

Additionally, we report performance using the *P.Head* metric introduced in Tafasca et al. (2023b) (see Section 4.2.1), which captures performance for looking towards heads, providing a more semantic evaluation of gaze predictions.

Tested Models

ChildPlay. In addition to our proposed model, we train the Image-only model from Gupta et al. (2022) on ChildPlay. For reference, we also evaluate models trained solely on GazeFollow, without fine-tuning on ChildPlay.

GazeFollow and VideoAttentionTarget. We train our model from scratch on GazeFollow

	Children				Adults				Full data			
Model	**AUC↑**	**Dist↓**	**AP↑**	**P.Head↑**	**AUC↑**	**Dist↓**	**AP↑**	**P.Head↑**	**AUC↑**	**Dist↓**	**AP↑**	**P.Head↑**
Gupta et al. (2022)†	0.926	0.136	-	0.435	0.919	0.151	-	0.621	0.923	0.142	-	0.518
Ours - 2D cone†	0.929	0.125	-	0.472	0.934	0.131	-	0.664	0.931	0.127	-	0.567
Ours†	0.934	0.112	-	0.509	0.930	0.119	-	0.681	0.932	0.115	-	0.602
Gupta et al. (2022)	0.923	0.106	0.980	0.648	0.914	0.123	0.987	0.731	0.919	0.113	0.983	0.694
Ours - 2D cone	0.925	0.118	0.937	0.564	0.927	0.125	0.955	0.717	0.926	0.121	0.944	0.644
Ours	0.939	0.098	0.989	0.604	0.928	0.121	0.983	0.704	0.935	0.107	0.986	0.663
Human	-	-	-	-	-	-	-	-	0.911	0.048	0.993	-

Table 5.3: Results on the ChildPlay dataset. The best results are given in red and the second best results are given in blue. † indicates that the model was not fine-tuned on ChildPlay.

Model	**AUC↑**	**Avg.Dist↓**	**Min.Dist↓**	**P.Head↑**
Fang et al. (2021)	0.922	0.124	0.067	-
Hu et al. (2022a)	-	0.135	0.075	-
Bao et al. (2022)	0.928	0.122	-	-
Jin et al. (2022)	0.920	0.118	0.063	-
Chong et al. (2020b)	0.921	0.137	0.077	0.708
Gupta et al. (2022)	0.933	0.134	0.071	0.750
Ours - 2D cone*	0.939	0.122	0.062	0.762
Ours*	0.936	0.125	0.064	0.760
Human	0.924	0.096	0.040	-

Model	**AUC↑**	**Dist↓**	**AP↑**	**P.Head↑**
Gupta et al. (2022)†	0.907	0.137	-	0.887
Ours - 2D cone†	0.915	0.128	-	0.894
Ours†	0.911	0.123	-	0.900
Fang et al. (2021)	0.878	0.124	0.872	-
Bao et al. (2022)	0.885	0.120	0.869	-
Jin et al. (2022)	0.898	0.109	0.897	-
Chong et al. (2020b)	0.854	0.147	0.848	-
Gupta et al. (2022)*	0.897	0.134	0.864	0.903
Ours - 2D cone*	0.909	0.120	0.856	0.892
Ours*	0.914	0.109	0.834	0.902
Human	0.921	0.051	0.925	-

Table 5.4: Results on GazeFollow (left) and VideoAttentionTarget (right) with the best results in red and second best results in blue. * indicates that the model follows a proper protocol, using a validation split to select the model. † indicates that the model was not fine-tuned on VAT.

and fine-tune it on VideoAttentionTarget. For comparison, we also retrain the Image-only model from Gupta et al. (2022) using our updated training and validation splits. For state-of-the-art baselines, we include the static model from Chong et al. (2020b) and other models incorporating depth and head crop features, including those by Fang et al. (2021), Hu et al. (2022a), Gupta et al. (2022), Bao et al. (2022), and Jin et al. (2022). We exclude comparisons with methods that use additional inputs such as eye crops or pose, to ensure fairness.

Ablation: 3DFoV vs 2D Cone. To assess the benefit of explicitly modeling the 3D Field of View (3DFoV), we compare our approach against a baseline using a standard 2D gaze cone, similar to that of Gupta et al. (2022). In this baseline, the 2D gaze cone is generated by projecting the 3D gaze vector into the image plane and computing cosine similarity with each pixel location—without applying any decay function.

5.2.3 Results

GazeFollow and VideoAttentionTarget (VAT). Our results on GazeFollow and VideoAttentionTarget are shown in Table 5.4. Overall, our model achieves strong performance across all metrics. Compared to the current state-of-the-art, it is on par with the best-performing method (Jin et al., 2022), which also uses depth information but does not model an explicit 3DFoV and may not rely on a validation set for evaluation. On GazeFollow, although our

AvgDist is slightly worse, the MinDist is identical, and on VAT both methods perform equally well in terms of distance (0.109). Compared to the image-only model of (Gupta et al., 2022), which follows the same training protocol, our approach yields significantly better results.

Looking at the P.Head metric, we observe generally high performance, especially on the VAT dataset, which is known to have a strong bias toward gaze directed at heads (see Table 3.3). This leads to a high number of true positives relative to false positives.

ChildPlay. Results on the ChildPlay dataset are presented in Table 5.3. Our model demonstrates significantly better cross-dataset generalization compared to both the 2D gaze cone baseline and the image-only model from Gupta et al. (2022). Fine-tuning improves performance across all models, with the (Gupta et al., 2022) model benefitting more—potentially due to overfitting to dataset-specific priors. Despite this, our method remains the best-performing across all metrics except for P.Head.

Interestingly, the performance gap with respect to human-level accuracy suggests that substantial room for improvement still exists.

Children vs. Adults. Closer inspection of the ChildPlay results reveals slightly better distance scores for children than for adults. However, this is largely explained by the fact that gaze targets tend to be closer to children on average (Tafasca et al., 2023b). This finding is consistent with observations from Tu et al. (2022), who noted that prediction models struggle more with distant gaze targets.

In contrast, the P.Head metric tells a different story: performance is significantly lower for children—by 18.7%—compared to adults. This supports our argument that a single metric is insufficient for fully evaluating model behavior, and that performance can vary significantly across subgroups. Moreover, all models exhibit larger performance gains after fine-tuning on child data, particularly for children and across both distance and P.Head metrics, highlighting the importance of domain-specific training data.

3DFoV vs. 2D Cone Saliency. Results indicate that our model (Ours), which incorporates an explicit 3D Field of View (3DFoV), performs on par with—or outperforms—the 2D cone baseline (Ours-2D cone). On GazeFollow, the performance of both models is similar, likely due to the dataset's relatively simple scenes and bias toward foreground subjects—gaze annotations are typically only provided for the person with the largest face in the scene.

However, on VideoAttentionTarget and ChildPlay, our 3DFoV model exhibits superior generalization and continues to outperform the 2D cone baseline even after fine-tuning. This emphasizes the value of incorporating depth cues and the advantage of using a geometrically consistent 3D representation of the visual field.

Qualitative Results. Qualitative examples are provided in Figure 5.6. We observe that the 3DFoV allows the model to selectively highlight only the parts of the scene that are physically visible from the person's viewpoint, thereby improving predictions in cluttered scenes with

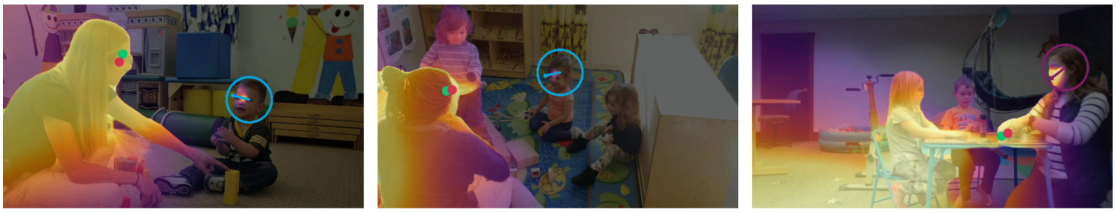


Figure 5.6: Qualitative results of our geometrically grounded model on ChildPlay. Our 3D Field of View (3DFoV) highlights potential gaze targets, excluding objects where the depth does not match. Gaze target predictions are given in green and GT ones in red.

multiple salient distractors along the line of sight.

5.2.4 Conclusion

In this work, we introduced a geometrically grounded model for gaze prediction that explicitly constructs a 3D Field of View (3DFoV) using estimated gaze direction and a predicted geometrically-consistent point cloud. By integrating this 3DFoV representation into the prediction pipeline, our approach provides a principled way to rule out occluded or implausible targets along the 2D line of sight and better localize the true gaze target.

Our model achieves state-of-the-art results on public gaze following benchmarks, and demonstrates strong cross-dataset generalization to the VideoAttentionTarget and ChildPlay datasets. These results highlight the importance of geometrically consistent depth cues for improving robustness across diverse settings.

In addition, we showed that training and evaluating on ChildPlay improves performance for child gaze prediction—particularly when assessed using semantic metrics such as the likelihood of looking at other people's heads. These findings reinforce the value of including child-centric data and semantically meaningful evaluation protocols for advancing gaze understanding in real-world social contexts.

5.3 Conclusion

In this chapter, we explored how multimodal cues—specifically depth and body pose—can be explicitly leveraged to enhance gaze following. We introduced two models: a modular attention-based framework that combines RGB, depth, and pose inputs, and a geometrically grounded model that constructs a 3D Field of View (3DFoV) using predicted gaze vectors and geometrically-consistent depth maps. Together, these approaches achieve state-of-the-art results on multiple public benchmarks and demonstrate strong cross-dataset generalization, including under privacy-sensitive constraints.

While these models advance the accuracy and robustness of gaze target prediction, they remain fundamentally limited to spatial localization and do not capture the semantic or social meaning of gaze behaviours. As we highlighted in earlier chapters, such semantic cues—like whether a person is looking at another person or engaging in shared attention—are crucial for understanding interactions in social scenes. Rather than treating them as separate tasks, given their conceptual overlap, they may also benefit from being jointly modeled.

In the next chapter, we address this gap by proposing unified models that jointly predict gaze targets and social gaze behaviours, enabling a richer, multi-task understanding of visual attention in complex, multi-person environments.

5.4 Appendix

Point Cloud Comparison

Monocular Depth Estimation. Depth datasets can be put under three categories:

- *Absolute Depth:* These datasets provide the absolute depth of the scene. The data is recorded using sensors such as LiDARS, time of flight cameras etc. ex. KITTI (Geiger et al., 2013)
- *Up to Scale (UTS) Depth:* These datasets provide the depth of the scene up to an unknown scale C_1. The absolute depth $d*$ can be recovered from UTS depth d as $d*^{-1} = C_1.d^{-1}$. ex. Megadepth (Li and Snavely, 2018)
- *Up to Shift and Scale (UTSS) Depth:* These datasets provide the disparity of scene. They are obtained from stereo movies and photos by computing the optical flow. The absolute depth can be recovered from the disparity D as $d*^{-1} = C_1.(D + C_2)$. C_2, also known as shift, depends on the camera parameters and is crucial for reconstructing geometry preserving point clouds. However, the shift is typically unknown. ex. MiDaS (Ranftl et al., 2020)

Recent methods for monocular depth estimation (Ranftl et al., 2020)(Yin et al., 2020) have leveraged UTSS depth data due to it's high diversity, and shown better generalization when tested on unseen datasets. However, they can only predict UTSS depth so the reconstructed point clouds are not geometry preserving. Hence, methods for gaze target prediction that use these algorithms rely on course matching (Fang et al., 2021) or attempt to correct the point cloud based on prior assumptions (Bao et al., 2022).

We study two recent methods for monocular depth estimation that aim to generate geometry-preserving point clouds while still leveraging UTSS data. Wei et al. (Yin et al., 2021) predict UTSS depth and use it to construct a (distorted) point cloud. A point cloud module then recovers the shift factor from the distorted point cloud. On the other hand, Patakin et al. (Patakin et al., 2022) train on a mix of absolute, UTS and UTSS depth data. The absolute and UTS depth data provide supervision such that the algorithm predicts UTS depth.

Qualitative Results. We provide a qualitative comparison of point clouds generated using the depth maps from Ranftl et al. (Ranftl et al., 2020), Wei et al (Yin et al., 2021) and Patakin et al. (Patakin et al., 2022) in Figure 5.7. We observe that the point clouds generated using the depth maps from Wei et al. and Patakin et al. generally have less distortion of scene elements, and better maintain the depth between objects. The point clouds from Patakin et al. in particular seem to preserve the geometry of the scene best.

Gaze Vector Stability. To quantitatively compare the methods of Wei et al. (Yin et al., 2021) and Patakin et al. (Patakin et al., 2022), we investigate which algorithm generates more stable gaze vectors. This is crucial as we rely on their generated gaze vectors as ground truth. The test is based on the fact that the gaze vector for a person (camera coordinate system) should be the same irrespective of their distance from the camera. We perform the test as follows:

- We take 5 random crops of an image
- For each crop, we compute the depth (Wei et al. or Patakin et al.) and focal length
- We then reconstruct the point cloud $\mathbf{P}^{\mathbf{c}}$ following the protocol defined in Section 4.2, and obtain the gaze vector for each crop as $\mathbf{g}_{3D}^{gt}{}^{\mathbf{c}} = \frac{\mathbf{P}^{\mathbf{c}}_{gaze}-\mathbf{P}^{\mathbf{c}}_{eye}}{||\mathbf{P}^{\mathbf{c}}_{gaze}-\mathbf{P}^{\mathbf{c}}_{eye}||}$
- The stability is given by the standard deviation of the gaze vector across the crops

For a more robust estimate, we perform this procedure for the first frame of every clip in the ChildPlay training set, and compute the median standard deviation. The values for the method of Wei et al. are [0.041, 0.032, 0.095] while the values for the method of Patakin et al. are [0.026, 0.019, 0.075]. The median standard deviation for Patakin et al. is lower, especially for the z component, indicating that it generates more stable gaze vectors.

Training Details

Head Bounding Boxes. The provided head box annotations for GazeFollow are not consistent and sometimes include the whole head, and at other times just the face of the person. Hence, we re-extract the head boxes using a pre-trained Yolov5 model (Jocher et al., 2022) and use these for all our experiments.

Eye Location. For GazeFollow, we use the annotated eye location, and for the VideoAttention-Target and ChildPlay datasets we use the center of the annotated head bounding box as the eye location.

Input Aspect Ratio. Previous methods (Chong et al., 2020b)(Gupta et al., 2022) distort the scene and head images to the model input size. To avoid this, we expand the head bounding box to a square to match the Human-Centric module's input aspect ratio. We also carefully crop and pad scene images to the Scene-Centric module's input aspect ratio during training and validation so that there is no distortion. During the test phase, we do not perform any cropping/padding and instead scale the longer side of the scene image to the Scene-Centric module's input width.

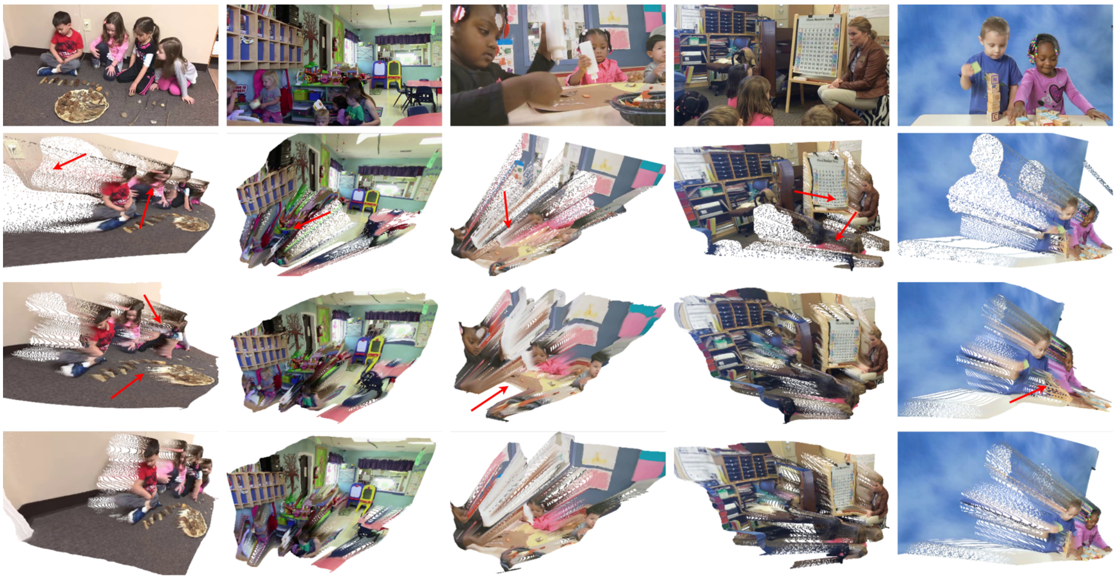


Figure 5.7: Comparison of point clouds generated using the depth maps from Ranftl et al. (Ranftl et al., 2020) (row 2), Wei et al. (Yin et al., 2021) (row 3) and Patakin et al. (Patakin et al., 2022) (row 4) on ChildPlay images. The point clouds generated using Patakin et al. appear to best preserve the geometry of the scene.

6 Unifying Gaze Following and Social Gaze Prediction

As discussed in Chapters 1 and 2, interpreting gaze—particularly in the context of communicative cues such as eye contact or shared attention—is crucial for understanding social interactions, with important applications in developmental assessment (Senju and Johnson, 2009), human–robot interaction (Sheikhi and Odobez, 2015), and behavioral science (Otsuka et al., 2018).

In this chapter, we propose unified models that jointly predict the *gaze target and social gaze labels*—including Looking at Heads (LAH), Looking at Each Other (LAEO), and Shared Attention (SA) as introduced in Section 1.1—in a single stage for *all individuals* in the scene. This requires new architectures capable of handling multiple people, reasoning about social interaction dynamics, and integrating both spatial and temporal cues.

Previous work in social gaze prediction generally adopts one of two approaches (detailed in Section 2.3). The first approach focuses on designing task-specific networks that process pairs of head crops and potentially additional scene information (Marin-Jimenez et al., 2019; Doosti et al., 2021; Cantarini et al., 2021; Fan et al., 2018; Sumer et al., 2020). While these specialized models are effective for their respective tasks, they offer limited generalization to other gaze-related tasks. The second approach first addresses the gaze following task, then applies ad-hoc post-processing to derive social gaze attributes from predicted gaze points. For instance, inferring shared attention by combining gaze heatmaps from multiple individuals (Chong et al., 2020b).

However, gaze following is itself a challenging task, and these methods are not explicitly trained for social gaze prediction. Moreover, standard gaze following metrics, such as distance, do not capture the semantic quality of predictions, such as whether a predicted gaze point falls on a person's head.

We address these limitations by proposing transformer-based models that unify gaze following and social gaze prediction. These models are trained end-to-end and evaluated using semantic metrics based on the same social gaze tasks (see Chapter 4.2). Our main contributions are:

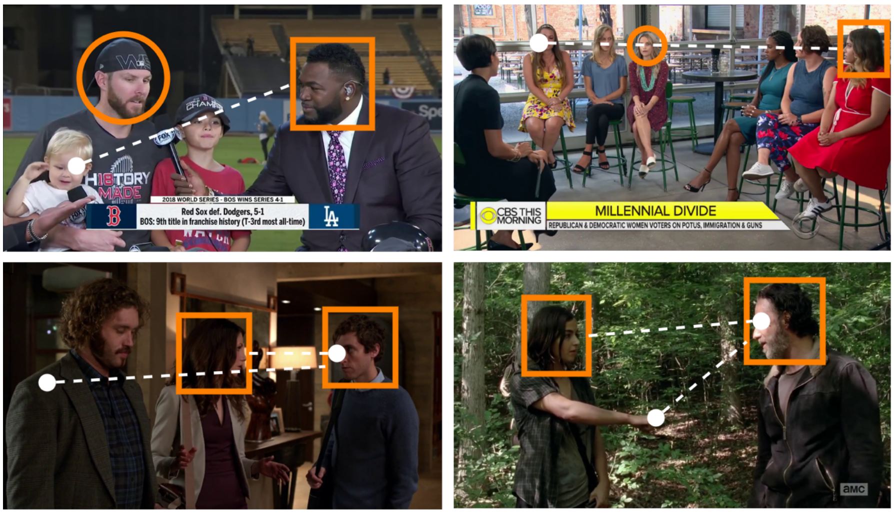

HISTORY MADE
2018 WORLD SERIES - BOS WINS SERIES 4-1
Red Sox def. Dodgers, 5-1
BOS: 9th title in franchise history (T-3rd most all-time)
CBS THIS MORNING
MILLENNIAL DIVIDE
REPUBLICAN & DEMOCRATIC WOMEN VOTERS ON POTUS, IMMIGRATION & GUNS
amc

Figure 6.1: Our method accurately predicts social gaze in cases where state-of-the-art gaze following methods fail. Row 1 shows examples for LAH, and Row 2 for LAEO. Orange highlights true positive pairs predicted by our model, while white indicates gaze predictions from (Tonini et al., 2022), which when post-processed, result in incorrect social gaze predictions.

- We propose a graph-based transformer model that extends a frozen gaze following backbone (Tafasca et al., 2024) with a graph reasoning module for social gaze prediction. This model performs multi-person inference and supports unified prediction of gaze and social gaze labels. This work was published in Gupta et al. (2024a) and is detailed in Section 6.1).
- We introduce a unified temporal transformer model that jointly *trains* for gaze following and social gaze prediction, and supports optional auxiliary signals such as speaking status. This model enables temporal modeling of social dynamics and improves upon some limitations of the first work. This work was published in Gupta et al. (2024b) and is detailed in Section 6.2).
- We evaluate both models across multiple tasks and datasets, and show that they improve over task-specific and post-processing based baselines. We also demonstrate the benefit of semantic metrics for interpreting and evaluating model predictions.

6.1 Sharingan Social

Given the general nature of the gaze following task, trained models capture rich representations of visual attention and social context that can be repurposed for higher-level reasoning. In particular, the Sharingan model that we proposed in Tafasca et al. (2024) supports efficient multi-person gaze following, making it a natural candidate for extension to social gaze prediction tasks.

At the time of this work, a unified dataset annotated for all social gaze tasks was not available

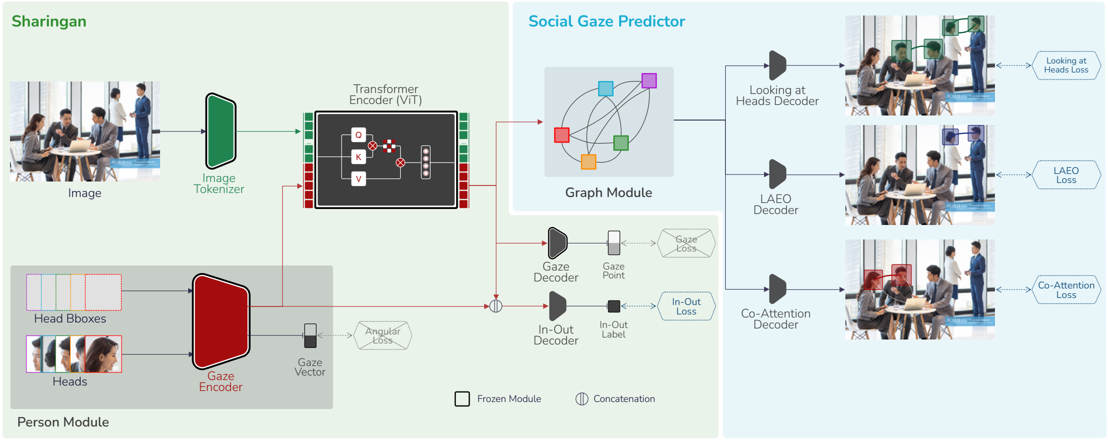


Figure 6.2: Our architecture comprises two main components—a Sharingan encoder and the Social Gaze Predictor. The Sharingan encoder first processes the scene and head crops to produce image and gaze tokens. These tokens are then passed to a ViT encoder to generate person tokens encoding gaze and attention information, which are fed to the gaze target regression decoder and the Social Gaze Predictor. The latter module includes a Graph Attention Network with task-specific decoders, which jointly model person-person multimodal interactions and predict different social gaze behaviors. The modules outlined in black are frozen during training.

(prior to the release of VSGaze). To address this, we leverage a frozen, pre-trained Sharingan model and adopt a modular strategy in which separate decoders are trained for each task using available datasets.

Concretely, our model treats the token-based person representations returned by Sharingan as nodes in a fully connected graph. These node embeddings are either directly passed to task-specific decoders to predict social gaze behavior for each edge (i.e., pair of people), or first updated via a Graph Attention Network (GAT) to explicitly model inter-person interactions. We evaluate this approach against post-processing baselines that infer social gaze from gaze following predictions and show consistent improvements across tasks (Figure 6.1).

6.1.1 Method

Our architecture is illustrated in Figure 6.2. It consists of two main components—a Sharingan (Tafasca et al., 2024) encoder and the Social Gaze Predictor. First, Sharingan processes the scene and head crops to produce both image and person tokens. These tokens are then used as input to a ViT encoder (Dosovitskiy et al., 2020), which produces a set of person tokens encoding gaze and attention information. The resulting tokens are passed to both the gaze target regression decoder and the Social Gaze Predictor. The latter consists of an optional Graph Attention Network and task-specific decoders to predict different social gaze behaviors. More precisely, the graph updates person node features (along with speaking status, when available) to jointly model person-person (or multimodal) interactions, while the task-specific decoders operate on pairs of the resulting node embeddings. We detail each component below.

Sharingan

Sharingan is a hybrid CNN–Transformer that extends the standard ViT (Dosovitskiy et al., 2020) for gaze following. The input to the transformer is a set of image tokens and person gaze tokens. The image tokens are computed using standard patchification and linear projection, with added positional encodings. The person gaze tokens are obtained by feeding input heads and their bounding boxes to a Gaze Encoder. In this module, a CNN backbone processes the head crops to produce gaze embeddings. These embeddings are used to predict a 2D gaze vector and are also linearly projected into gaze tokens. Simultaneously, the head bounding boxes are linearly projected to match the transformer token dimensions. This location information is added to the gaze tokens to form the final person tokens.

All tokens (image and person) are fed to a ViT encoder composed of standard transformer blocks. This allows the image and person tokens to interact via self-attention, updating the person tokens with gaze-relevant scene context. The output tokens $\mathbf{x}^{\text{out}}$corresponding to the input person tokens are passed to the Social Gaze Predictor and other decoders to predict pairwise social gaze labels, individual 2D gaze points, and in-vs-out-of-frame classification. We refer readers to the original paper (Tafasca et al., 2024) for full architectural details.

Social Gaze Predictor

We explore two approaches for social gaze prediction. In the first, we explicitly model interactions between people by passing Sharingan's output tokens through a Graph Module, followed by task-specific decoders. In the second, we bypass the graph and directly pass pairs of person tokens to the task-specific decoders.

Graph Module. Our goal is to model all person-to-person interactions, including multimodal cues such as speaking status. This allows the model to reason over socially consistent configurations (e.g., a person cannot be in mutual gaze with two others at the same time). We use a Graph Attention Network (GAT) (Veličković et al., 2018; Brody et al., 2022) where the nodes are the person tokens $\mathbf{x}^{\text{out}}$from Sharingan. The graph is fully connected, and each node i aggregates information from all others j using attention weights α_{ij}:

$$\mathbf{x}_i^{\text{gr}} = \alpha_{ii} W_s + \sum_{j=1, j\neq i}^{N} \alpha_{ij} (W_{\text{gr}} \mathbf{x}_j^{\text{out}}), \tag{6.1}$$

$$e_{ij} = s^{\text{T}} \text{LeakyReLU}(W_s \mathbf{x}_i^{\text{out}} + W_{\text{gr}} \mathbf{x}_j^{\text{out}}), \tag{6.2}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{N} \exp(e_{ik})}. \tag{6.3}$$

Here, s, W_s, and W_{gr} are learnable parameters. Although the graph is fully connected and could be implemented as a Transformer, we found the GAT to generalize better due to its lower parameter count and reduced tendency to overfit.

Speaking Status. In long videos, speaking status can be estimated from head trajectories (Min et al., 2022). We incorporate this into the model by transforming the score into an embedding $\mathbf{s}$, which is added to the person token before graph processing: $\mathbf{x}^{\text{out}} + \mathbf{s}$.

Task-Specific Decoders. While the graph models interactions globally, the final predictions are made by task-specific decoders, each processing concatenated node pairs. These decoders are implemented as MLPs that operate on $\mathbf{x}_i^{\text{gr}} \oplus \mathbf{x}_j^{\text{gr}}$.

$$\mathbf{e}_{i\to j} = E_{LAH}(\mathbf{x}_i^{\text{gr}} \oplus \mathbf{x}_j^{\text{gr}}), \tag{6.4}$$

$$\mathbf{e}_{i\leftrightarrow j} = E_{LAEO}(\mathbf{x}_i^{\text{gr}} \oplus \mathbf{x}_j^{\text{gr}}), \tag{6.5}$$

$$\mathbf{c}_{i,j} = C(\mathbf{x}_i^{\text{gr}} \oplus \mathbf{x}_j^{\text{gr}}), \tag{6.6}$$

where $\mathbf{e}_{i\to j}$ indicates if i is looking at j, $\mathbf{e}_{i\leftrightarrow j}$ indicates mutual gaze, and $\mathbf{c}_{i,j}$ denotes shared attention.

Loss Definition and Model Details

Loss. We train the model using a weighted sum of the following task-specific binary cross-entropy losses:

$$\mathscr{L} = \lambda_{LAH}\mathscr{L}_{LAH} + \lambda_{LAEO}\mathscr{L}_{LAEO} + \lambda_{SA}\mathscr{L}_{SA} + \lambda_{io}\mathscr{L}_{io}$$

where $\mathscr{L}_{LAH}$, $\mathscr{L}_{LAEO}$, and $\mathscr{L}_{SA}$ correspond to the LAH, LAEO, and SA tasks, respectively, and $\mathscr{L}_{io}$ is the in/out-of-frame classification loss. Each loss is computed over all pairs of people and then averaged.

While we could also train the gaze target location prediction loss, we found that doing so provided no performance gain on either gaze or social gaze tasks—likely due to the small size of social gaze datasets compared to GazeFollow. As a result, we freeze Sharingan and train only the Social Gaze Predictor.

6.1.2 Experiments

Implementation Details. The Sharingan encoder is pre-trained on GazeFollow (Recasens et al., 2017), following the protocol described in (Tafasca et al., 2024). For all experiments, we keep its weights frozen. Note that we use the version that predicts 2D points and not heatmaps as it was not available at the time of writing the paper. The Graph Module is implemented as a GAT (Brody et al., 2022) with two message passing layers, and the Task-Specific Decoders are implemented as three-layer MLPs with residual connections.

Inputs. Both the scene image and head crops are provided at a resolution of 224 × 224.

Speaking Status. We re-train a state-of-the-art speaker detection model (Min et al., 2022) using the AVA-ActiveSpeaker dataset (Roth et al., 2020), relying solely on the visual modality. The trained model is then used to generate speaking status predictions on our datasets.

Datasets. At the time of writing, no single dataset included annotations for all social gaze tasks. We therefore train using separate datasets for each task:

- *VideoAttentionTarget* (Chong et al., 2020b): LAH labels are generated following the protocol in Section 3.4.
- *UCO-LAEO* (Marin-Jimenez et al., 2014): We use the LAEO annotations as described in (Section 3.3.1).
- *VideoCoAtt* (Fan et al., 2018): We use the SA annotations as described in Section 3.3.2.

Two notable exceptions, VACATION (Fan et al., 2019) and GP-static (Chang et al., 2023), contain annotations for multiple social gaze behaviours. However, VACATION assigns only a single gaze "state" to each person per frame, even when multiple behaviours (e.g., LAEO and SA) occur simultaneously (see Figure 2.6). This limitation reflects model design choices discussed in Section 2.3. In contrast, GP-static only considers dyadic communication and was not available at the time of our study.

Training. During training, we randomly sample up to six people per scene to enable batch processing. At test time, we process all detected individuals with a batch size of one. The optimization details for each task are given below:

VideoAttentionTarget. We train for 12 epochs using the AdamW optimizer (Loshchilov and Hutter, 2018) with a learning rate of $3e^{-4}$. The loss coefficients are set as $\lambda_{LAH} = 10$, $\lambda_{io} = 2$, and $\lambda_{LAEO} = \lambda_{SA} = 0$.

UCO-LAEO & VideoCoAtt. We train for 20 epochs with the AdamW optimizer using learning rates of $3e^{-5}$ (UCO-LAEO) and $1e^{-3}$ (VideoCoAtt). Only the relevant binary cross-entropy loss for the task is used; other losses are set to zero.

Metrics. We report Area Under the ROC Curve (AUC) and Average Precision (AP) based on the Precision-Recall curve for all tasks, as described in Section 4.2.2.

Tested Models

Post-processing Baselines. We evaluate three Gaze Following models—Chong *et al.* (Chong et al., 2020b), Tonini *et al.* (Tonini et al., 2022), and Sharingan (Tafasca et al., 2024)—as baselines for social gaze prediction. Chong *et al.*is a strong and widely used baseline; Tonini *et al.*represents a recent state-of-the-art approach; and Sharingan is the current state-of-the-art and serves as our model's backbone. We post-process their gaze predictions following the protocol described in Section 4.2.1.

Model	AUC$_{LAH}$ ↑	AP$_{LAH}$ ↑
Ours-MLP	0.724	0.896
Ours-MLP$_{spk}$	**0.738**	**0.906**
Ours-Graph	0.733	0.902
Ours-Graph$_{spk}$	0.722	0.893

Table 6.1: Results for LAH on VAT. Best results are given in bold.

Model	AUC$_{SA}$ ↑	AP$_{SA}$ ↑
Chong et al. (2020b)	0.695	0.297
Tonini et al. (2022)	0.715	0.300
Sharingan (Tafasca et al., 2024)	0.760	0.301
Ours-MLP	**0.896**	0.594
Ours-Graph	0.890	**0.604**

Table 6.2: Results for SA on VideoCoAtt. Best results are given in bold.

Model	AUC$_{LAEO}$ ↑	AP$_{LAEO}$ ↑
LAEO-Net (Marin-Jimenez et al., 2014)	-	0.795
Doosti et al. (2021)	-	0.651
Chang et al. (2023)	-	0.803
Ours-MLP	**0.986**	**0.957**
Ours-Graph	0.981	0.946

Table 6.3: Results for LAEO on UCO-LAEO. Best results are given in bold.

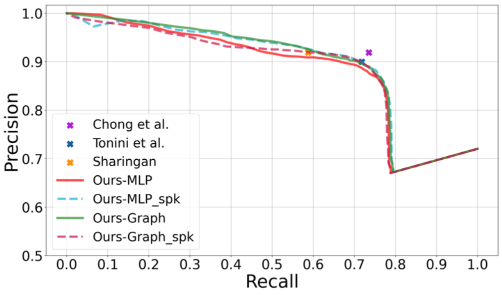


Figure 6.3: Prec-Recall curve for LAH on VAT.

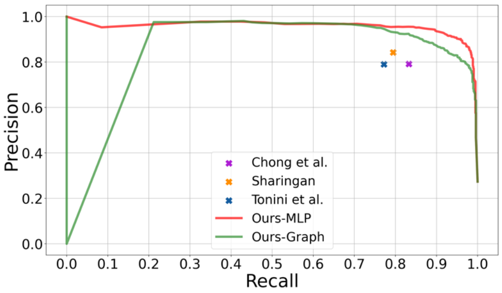


Figure 6.4: Prec-Recall curve for LAEO on UCO-LAEO.

Ours-MLP. This variant feeds the output person tokens from Sharingan directly to the task-specific decoders, without updating them via a Graph Module.

Ours-Graph. Here, the Sharingan output tokens are first updated using the Graph Module before being passed to the task-specific decoders. To ensure fair comparison, we match the number of parameters between the MLP and Graph variants.

Speaking. We evaluate the impact of augmenting both MLP and Graph-based models with predicted speaking status information. These variants are denoted by the subscript *spk*.

6.1.3 Results

VideoAttentionTarget. Our results for LAH performance on VideoAttentionTarget are summarized in Table 6.1. We observe that *Ours-Graph* provides modest gains over *Ours-MLP*, with an improvement of approximately 1 point in AUC and 0.6 in AP. It also matches or exceeds the performance of the post-processing baselines Tonini (Tonini et al., 2022) (Prec.=0.900, Recall=0.718) and Sharingan (Tafasca et al., 2024) (Prec.=0.919, Recall=0.587), as seen in the precision-recall curve in Fig. 6.3. However, the Chong baseline (Chong et al., 2020b) (Prec.=0.919, Recall=0.735) slightly outperforms both of our models.

VideoCoAtt. Results for SA performance on the VideoCoAtt dataset are presented in Table 6.2. *Ours-Graph* achieves the highest AP, outperforming all post-processing baselines. However, it slightly underperforms *Ours-MLP* in terms of AUC. For the shared attention target localization metrics, Sharingan (Tafasca et al., 2024) demonstrates superior performance (Dist: 0.116, Acc: 0.665), followed by Chong (Chong et al., 2020b) (Dist: 0.139, Acc: 0.579) and Tonini (Dist: 0.127, Acc: 0.452). Since the gaze decoder is frozen during SA training, the distance and accuracy scores for both *Ours-MLP* and *Ours-Graph* remain fixed at Dist=0.116 and Acc=0.665.

UCO-LAEO. Table 6.3 summarizes our results on the UCO-LAEO dataset. *Ours-MLP* achieves the best overall performance, establishing a new state of the art, while *Ours-Graph* performs slightly below it. Given that most frames in UCO-LAEO contain only two people, the graph-based modeling offers limited additional benefit. The precision-recall curve in Fig. 6.4 confirms that both of our models outperform all baselines, including Chong (Chong et al., 2020b) (Prec.=0.791, Recall=0.832), Tonini (Tonini et al., 2022) (Prec.=0.790, Recall=0.771), and Sharingan (Tafasca et al., 2024) (Prec.=0.842, Recall=0.794).

Speaking Status. Preliminary results for including speaking status in LAH prediction are reported in Table 6.1. For *Ours-MLP*, we observe an improvement of roughly 1 point in both AUC and AP, indicating that speaking cues can enhance performance. Conversely, for *Ours-Graph*, incorporating this information slightly degrades performance by about 1 point. These findings suggest that while speaking status is a promising signal for social gaze inference, more work is needed to model and integrate it effectively.

Gaze Following Metrics vs Social Gaze Metrics. A key observation is that better performance on standard gaze following metrics does not necessarily translate to improved social gaze prediction. For instance, Tonini (Tonini et al., 2022) outperforms Chong (Chong et al., 2020b) on gaze following, but underperforms it for both LAH and LAEO (see Figures 6.3 and 6.4). Similarly, while Sharingan (Tafasca et al., 2024) achieves the best gaze following results overall, it underperforms Chong for LAH. This discrepancy is often due to models predicting salient nearby objects or individuals instead of the correct gaze target, an issue not captured by standard gaze metrics but revealed by semantic social gaze metrics.

Qualitative Results.

Figure 6.5 shows qualitative predictions from *Ours-Graph*. The model successfully captures social gaze across diverse scenarios, including cases where the face is not visible. However, it occasionally fails to detect subtle gaze cues from the eyes, resulting in missed or incorrect predictions.

6.1.4 Conclusion

In this work, we proposed a unified model for Gaze Following and social gaze prediction, focusing on the tasks of looking at heads (LAH), mutual gaze (LAEO), and shared attention (SA). Unlike previous approaches that rely on task-specific architectures or post-processing of

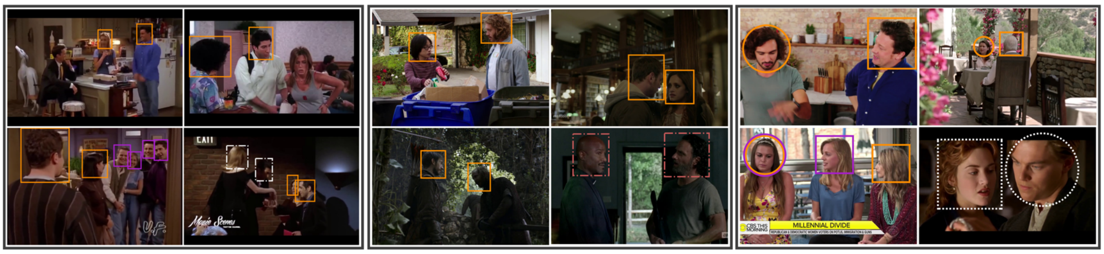


Figure 6.5: Sample qualitative results from the test sets. Left: SA (VideoCoAtt). Middle: LAEO (UCO-LAEO). Right: LAH (VideoAttentionTarget). Shapes of the same color indicate a positive instance (e.g., shared attention). Orange and Pink denote different true positives. Dashed white lines indicate false negatives, and Dashed red lines indicate false positives. For LAH, the source person is shown with a bounding box and the predicted target with a circle.

gaze predictions, our model leverages rich person-level representations from the Sharingan encoder—pretrained on gaze following—and is explicitly trained for social gaze tasks.

Our experiments demonstrate improved performance across multiple public benchmarks compared to post-processing baselines. Interestingly, results indicate that updating the person tokens with a Graph Module does not consistently outperform the MLP variant when matched for parameter count. This may be due to the strong self-attention capabilities of Sharingan, which already models scene-level and inter-person interactions. We also observed oversmoothing effects in the graph network, which necessitated careful architectural design such as incorporating skip connections to mitigate the issue (effectively resembling a transformer). However, these measures may not have been sufficient to fully prevent the degradation.

Finally, we explored the integration of speaking status as a multimodal signal and observed modest performance improvements for LAH in the MLP variant. This suggests promise for future work on incorporating auxiliary cues into social gaze estimation. In the next work, we address some of these limitations by introducing a new transformer architecture capable of jointly modelling all tasks while also incorporating temporal information.

6.2 MTGS

The Sharingan Social model demonstrated that gaze following representations can serve as an effective foundation for training social gaze prediction models. A natural next step is to ask whether joint training across all tasks—gaze following, LAH, LAEO, and SA—can provide mutual benefits and lead to improved performance. Another key limitation of most existing methods is their reliance on static images. Temporal dynamics, such as coordinated head and gaze movements (Sheikhi and Odobez, 2015), can offer critical cues for inferring gaze direction, particularly when the eyes are occluded or poorly visible (Nonaka et al., 2022). However, the absence of appropriate datasets-temporal and with annotations covering all

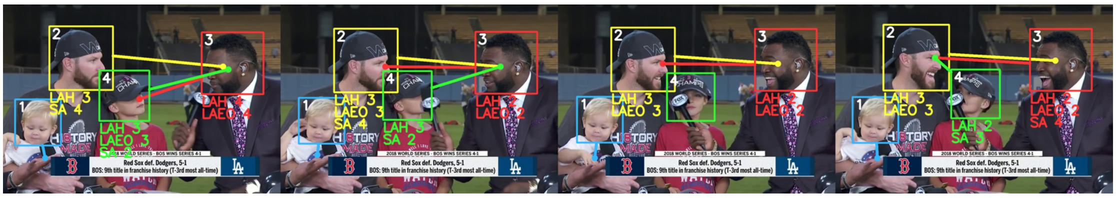


Figure 6.6: MTGS is a unified framework for jointly modeling gaze following and social gaze behaviors, including Looking at Humans (LAH), Looking at Each Other (LAEO), and Shared Attention (SA). For each individual, the predicted social gaze relations are listed alongside the corresponding person IDs (e.g., in Frame 1, Person 2 shares attention with Person 4).

social gaze tasks-has hindered progress in this direction.

The Sharingan Social model also had several architectural limitations. All person and scene tokens are processed uniformly by the ViT encoder, which may bias attention toward scene content and underrepresent person–person interactions. Moreover, since scenes typically remain static over short video segments, temporal modeling may be more effective when applied to person-level representations. Finally, while Sharingan Social employed a GAT module to capture pairwise relationships, like many GNNs, it was prone to oversmoothing.

In this work, published in Gupta et al. (2024b), we propose *MTGS* (Multi-person, Temporal Gaze Following and Social Gaze Prediction), a new architecture that addresses these limitations to jointly model all tasks (Figure 6.6). Our contributions are as follows:

- We introduce a novel dual-branch transformer architecture that separately encodes person and scene tokens, allowing them to interact via cross-attention. This separation enables: 1) explicit modelling of person-person interactions, 2) temporal modeling at multiple stages within the person branch—from gaze direction to target prediction.
- We leverage the VSGaze dataset (Section 3.4), introduced in the same paper, to jointly train for both gaze following and social gaze prediction. Importantly, we show that these tasks are complementary: training for one improves performance on the other.
- In an unpublished extension, we show that incorporating a recent foundation model for scene encoding (DINOv2 (Oquab et al., 2023)) further enhances performance. Our final architecture achieves state-of-the-art results for both gaze following and social gaze prediction (LAH, LAEO, and SA).

6.2.1 Method

Our approach is illustrated in Fig. 6.7. It takes as input a sequence of $t = 1 \ldots T$ frames, as well as the head bounding box tracks $\mathbf{B}_{i,1:T}$ and corresponding head crops $\mathbf{I}^{head}_{i,1:T}$ which are assumed to have been extracted for each of the $i = 1 \ldots N_p$ persons. The outputs are the sequence of gaze heatmaps $\mathbf{H}_{i,1:T}$ and in-out gaze labels $o_{i,1:T}$ for each person i, as well as the sequence of

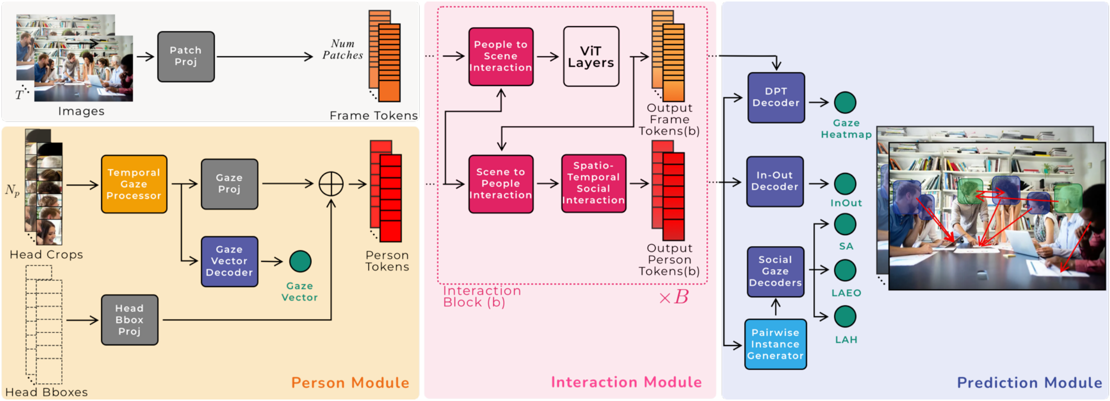


Figure 6.7: Proposed architecture for multi-person temporal gaze following and social gaze prediction. See approach overview in Section 6.2.1.

per-frame pair-wise social gaze labels for each task and pair $i, j \in \{1 \dots N_p\}$: $\mathbf{e}_{i\rightarrow j,1:T}$ for LAH, $\mathbf{e}_{i\leftrightarrow j,1:T}$ for LAEO, and $\mathbf{c}_{i,j,1:T}$ for SA.

The model proceeds as follows. First, each frame t is processed by a standard ViT tokenizer to produce the set of patch-wise frame tokens $\mathbf{f}_t$, resulting in a sequence of frame tokens $\mathbf{f}_{1:T}$. In parallel, the Person Module processes the sequence of head crops from each person i using the Temporal Gaze Processor, and the resulting sequence outputs are then tokenized at each frame along with the bounding box locations to produce the sequence of person token $\mathbf{p}_{i,1:T}$. Secondly, the Interaction Module jointly processes the frame and person tokens, iteratively updating them at each time step through person-scene cross-attention interaction components and scene ViT self-attention, and in time through person spatio-temporal social interaction components. Finally, the Prediction Module processes at each time step the resulting frame and person tokens (from multiple blocks) to infer the sequence of gaze heatmaps and in-out gaze labels for each person, as well as pair-wise social gaze labels. We detail the three modules in the next sections.

Person Module

This module aims to model person-specific information relating to gaze and head location.

Temporal Gaze Processor. It aims to capture all gaze-related information (direction, dynamics). First, individual head crops $\mathbf{I}_{i,t}^{head}$ are processed by a Gaze Backbone $\mathcal{G}_b$ to produce gaze embeddings according to $\mathbf{g}_{i,t}^{\text{stat}} = \mathcal{G}_b(\mathbf{I}_{i,t}^{head})$. Then, to model the gaze dynamics of a person, we rely on a Temporal Gaze Encoder $\mathcal{G}_{\text{temp}}$ to process the sequence $\mathbf{g}_{i,1:T}^{\text{stat}}$ of gaze embeddings plus learnable temporal position embeddings $\mathbf{z}_{1:T}$ and obtain their temporal counterparts: $\mathbf{g}_{i,1:T}^{\text{temp}} = \mathcal{G}_{\text{temp}}(\mathbf{g}_{i,1:T}^{\text{stat}} + \mathbf{z}_{1:T})$. $\mathcal{G}_{\text{temp}}$ is implemented as a single Transformer layer with self-attention. Finally, to supervise the learning of relevant gaze embeddings, we attach a Gaze Vector Decoder that predicts a person's 2D gaze vector at each time step, $\mathbf{g}_{2D}^{p}{}_{i,t} = \mathcal{G}_h(\mathbf{g}_{i,t}^{\text{temp}})$, where $\mathcal{G}_h$ is implemented as a 2-layer MLP.

Person tokenization. The person tokens are obtained by projecting the temporal gaze embeddings and normalized 4d head box locations using learnable linear layers ($\mathscr{P}_{\text{gaze}}$ and $\mathscr{P}_{\text{box}}$ respectively) to tokens of same dimension than frame token, and adding them together:

$$\mathbf{p}_{i,t} = \mathscr{P}_{\text{gaze}}(\mathbf{g}^{\text{temp}}_{i,t}) + \mathscr{P}_{\text{box}}(\mathbf{B}_{i,t}). \tag{6.7}$$

Interaction Module

The Interaction module aims at modeling the exchange of information between persons and the scene at each time step, as well as the spatio-temporal social interactions between people. One important goal of this process is to align the person and frame token representations so that (i) *person-specific* gaze heatmaps can be predicted from the set of output frame tokens and each person output token; (ii) in-out gaze and social gaze prediction can be made from the person tokens.

To do so, we designed the module to consist of B blocks, each comprising Person-Scene Interaction and Spatio-Temporal Social Interaction components. The input to the first block is the set of person tokens $\mathbf{p}_{1:N_p,1:T}$ from the Person Module, and the frame tokens $\mathbf{f}_{1:T}$. Each block then processes the set of output[1] person tokens $\mathbf{p}^{\text{o,b-1}}_{1:N_p,1:T}$ and output frame tokens $\mathbf{f}^{\text{o,b-1}}_{1:T}$ from the previous block, and returns updated tokens after a series of self/cross-attention layers through the components.

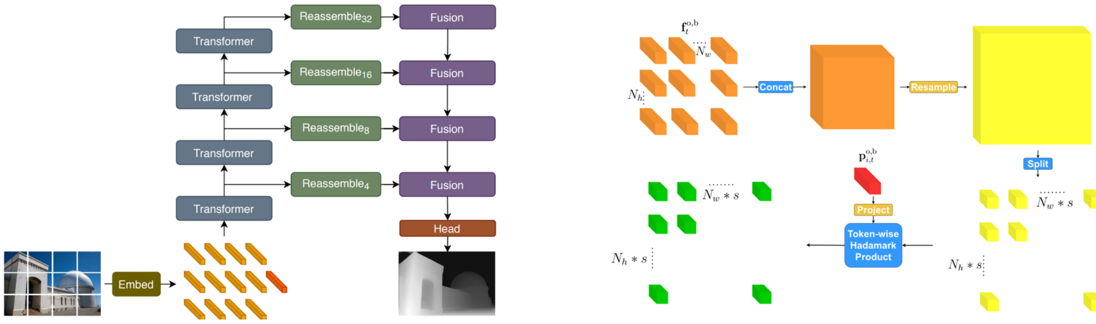
Person-Scene Interaction. This component models the interactions between people and the scene and can capture inferring gaze to scene objects or body parts like hands or exploit some global context. It is inspired by ViT-Adaptor (Chen et al., 2022) which has shown good performance for dense prediction tasks when relying on pretrained models and small amounts of data for the target task. It proceeds in 3 steps:

1. People-to-Scene Encoder $\mathscr{I}^{\text{b}}_{\text{ps}}$: it updates the frame tokens with person information relevant to gaze by processing the frame tokens $\mathbf{f}^{\text{o,b-1}}_t$ and frame-level person tokens $\mathbf{p}^{\text{o,b-1}}_{1:N_p,t}$ according to $\mathbf{f}^{\text{p,b}}_t = \mathscr{I}^{\text{b}}_{\text{ps}}(\mathbf{f}^{\text{o,b-1}}_t, \mathbf{p}^{\text{o,b-1}}_{1:N_p,t})$. It is implemented as a single Transformer layer with cross-attention, where $\mathbf{f}^{\text{o,b-1}}_t$ generate the queries and $\mathbf{p}^{\text{o,b-1}}_{1:N_p,t}$ generate the keys and values.

2. The updated frame tokens $\mathbf{f}^{\text{p,b}}_t$ pass through the standard set of of ViT layers $\mathscr{V}_b$ to process the scene information, resulting in the output frame tokens for the block b: $\mathbf{f}^{\text{o,b}}_t = \mathscr{V}_b(\mathbf{f}^{\text{p,b}}_t)$.

3. Scene-to-People Encoder $\mathscr{I}^{\text{b}}_{\text{sp}}$: it updates the person tokens so that they capture location information related to the salient items they are probably looking at. It works by processing the frame-level person tokens $\mathbf{p}^{\text{o,b-1}}_{1:N_p,t}$ and obtained frame tokens $\mathbf{f}^{\text{o,b}}_t$ according to: $\mathbf{p}^{\text{s,b}}_{1:N_p,t} = \mathscr{I}^{\text{b}}_{\text{sp}}(\mathbf{p}^{\text{o,b-1}}_{1:N_p,t}, \mathbf{f}^{\text{o,b}}_t)$.

[1] Note that the superscript {o, p, s} do not represent indices, but intermediate token updates within a block b.



(a) (b)

Figure 6.8: The standard DPT (a, taken from (Ranftl et al., 2021)) and the person-conditioned re-assemble stage from Tafasca et al. (2024) (b). This transformed DPT is used for predicting gaze heatmaps for each person in the scene.

It is also implemented as a single Transformer layer with cross-attention, where the set $\mathbf{p}^{\text{o,b-1}}_{1:N_p,t}$ generates the queries and $\mathbf{f}^{\text{o,b}}_{t}$ generates the keys and values.

Spatio-temporal Social Interaction. This component allows the sharing of information between people and of the alignment of their representations for social gaze prediction. This also include modeling the temporal evolution of individual tokens. To achieve this, a Social Encoder $\mathscr{I}^{\text{b}}_{\text{pp}}$ first processes and updates the frame-level person tokens $\mathbf{p}^{\text{s,b}}_{1:N_p,t}$ to capture interactions between people at each frame, according to: $\mathbf{p}^{\text{p,b}}_{1:N_p,t} = \mathscr{I}^{\text{b}}_{\text{pp}}(\mathbf{p}^{\text{s,b}}_{1:N_p,t})$. It is followed by a Temporal Person Encoder $\mathscr{I}^{\text{b}}_{\text{pt}}$ that processes the updated person token sequences $\mathbf{p}^{\text{p,b}}_{i,1:T}$ of each person i and updates them to capture temporal patterns of attention, resulting in the output person tokens for the block: $\mathbf{p}^{\text{o,b}}_{i,1:T} = \mathscr{I}^{\text{b}}_{\text{pt}}(\mathbf{p}^{\text{p,b}}_{i,1:T})$. Both $\mathscr{I}^{\text{b}}_{\text{pp}}$ and $\mathscr{I}^{\text{b}}_{\text{pt}}$ are implemented as a single Transformer layer with self-attention.

Prediction Module

The Prediction Module processes the set of output person $\{\mathbf{p}^{\text{o,b}}_{1:N_p,1:T}\}$ and frame $\{\mathbf{f}^{\text{o,b}}_{1:T}\}$ tokens from all Interaction Module blocks to predict the person-specific gaze heatmaps and in-out labels, as well as the pair-wise social gaze labels.

Gaze Heatmap Prediction.

Here, we follow the model introduced in (Tafasca et al., 2024) which takes inspiration from the DPT decoder (Ranftl et al., 2021) for dense prediction tasks, and adapts it to handle multiple heatmap predictions from the same ViT outputs. This is performed *by conditioning the decoding on each person's token.* The standard DPT decodes the image features from multiple layers of a ViT in a Feature Pyramid Network (Lin et al., 2017) style. It works by fusing at block level b the feature maps from level b+1 after an upsampling stage, and the feature maps computed by a reassemble stage from the ViT output of block b. We aim to apply this approach to the frame tokens $\{\mathbf{f}^{\mathrm{o,b}}_{1:T}, \mathrm{b} = 1 : B\}$, but conditioned on a specific person. In our model, this is achieved through a modification in the reassemble stage, in which the image feature maps produced by the standard reassemble stage are multiplied at every location (using a Hadamard product) with the projected person token $\mathbf{p}^{\mathrm{o,b}}_{i,t}$ of that same block level. The gaze heatmap $\mathbf{H}_{i,t}$ for each person at each frame is thus obtained as:

$$\mathbf{H}_{i,t} = \mathscr{D}(\{(\mathbf{f}^{\mathrm{o,b}}_{t}, \mathbf{p}^{\mathrm{o,b}}_{i,t}), \mathrm{b} = 1 : B\}) \tag{6.8}$$

where $\mathscr{D}$ denotes this conditional DPT. See Figure 6.8 and (Tafasca et al., 2024) for details.

Social Gaze Prediction. This decoder processes the person tokens from all B Interaction Module blocks to predict the social gaze label for every pair of people in every frame. In practice, the B tokens $\{\mathbf{p}^{\mathrm{o},1}_{i,t} \dots \mathbf{p}^{\mathrm{o},B}_{i,t}\}$ corresponding to a single person in a frame are linearly projected and concatenated to produce a multi-scale person token $\mathbf{p}^{\mathrm{ms}}_{i,t}$. Then, to predict a social gaze label, pairs of these tokens are concatenated and processed by the decoders E for LAH and C for SA (illustrated through the Pairwise Instance Generator in Fig. 6.7). Their outputs are the predicted LAH score $\mathbf{e}_{i\to j,t}$ for person i looking at j, and the predicted SA score $\mathbf{c}_{i,j,t}$ for i, j.

$$\mathbf{e}_{i\to j,t} = E_{LAH}(\mathbf{p}^{\mathrm{ms}}_{i,t}, \mathbf{p}^{\mathrm{ms}}_{j,t}) \text{ and } \mathbf{c}_{i,j,t} = C(\mathbf{p}^{\mathrm{ms}}_{i,t}, \mathbf{p}^{\mathrm{ms}}_{j,t}). \tag{6.9}$$

E_{LAH} and C are implemented as 3-layer MLPs with residual connections. For LAEO, both people i, j need to be looking at each other for a positive label, and either one can be looking away for a negative label. Hence, we simply compute the LAEO score $\mathbf{e}_{i\leftrightarrow j,t}$ as $\min(\mathbf{e}_{i\to j,t}, \mathbf{e}_{j\to i,t})$.

In-Out Prediction. This decoder $\mathscr{O}$ processes the multi-scale person tokens $\mathbf{p}^{\mathrm{ms}}_{i,t}$ to predict at every frame whether people are looking inside the frame or outside the frame, as $o_{i,t} = \mathscr{O}(\mathbf{p}^{\mathrm{ms}}_{i,t})$, where $\mathscr{O}$ is implemented as a 5-layer MLP with residual connections.

Losses

The total loss $\mathscr{L}$ is a linear combination of the gaze heatmap loss $\mathscr{L}_{hm}$, gaze vector loss $\mathscr{L}_{dir}$, social gaze losses $\mathscr{L}_{LAH}, \mathscr{L}_{SA}$ and the in-out loss $\mathscr{L}_{io}$:

$$\mathscr{L} = \lambda_{hm}\mathscr{L}_{hm} + \lambda_{dir}\mathscr{L}_{dir} + \lambda_{LAH}\mathscr{L}_{LAH} + \lambda_{SA}\mathscr{L}_{SA} + \lambda_{io}\mathscr{L}_{io} \tag{6.10}$$

$\mathscr{L}$ is applied at each time step per person for $\mathscr{L}_{hm}, \mathscr{L}_{dir}, \mathscr{L}_{io}$, and per pair for $\mathscr{L}_{LAH}, \mathscr{L}_{SA}$. All losses are standard: $\mathscr{L}_{hm}$ is defined as the pixel-wise MSE loss between the GT and predicted heatmaps, $\mathscr{L}_{dir}$ as the cosine loss, and the social gaze and in-out losses as binary cross-entropy losses. Since LAEO is inferred from LAH predictions (Sec 6.2.1), we do not have any LAEO loss.

6.2.2 Experiments

Datasets. We leverage the GazeFollow dataset, extended with LAH annotations using the protocol detailed in Section 3.4, for pre-training. Its diversity makes it a strong candidate for pre-training, though it is a static dataset and lacks annotations for LAEO and SA.

We then fine-tune our models on VSGaze (Section 3.4) to enable temporal modeling and joint training across all gaze following and social gaze tasks. Additionally, we perform fine-tuning and evaluation on its constituent datasets: VideoAttentionTarget (Section 3.1.2), ChildPlay (Section 3.1.3), UCO-LAEO (Section 3.3.1), and VideoCoAtt (Section 3.3.2).

Training and Validation. Following standard practice (Chong et al., 2020b; Fang et al., 2021; Gupta et al., 2022), we first train a static version of our model—i.e., without temporal attention subnetworks $\mathscr{G}_{\text{temp}}$ and $\mathscr{I}_{\text{pt}}^{\text{b}}$ —on GazeFollow. This model is trained for 20 epochs with a learning rate of 1×10^{-4}, and the resulting weights are used to initialize our temporal model.

The temporal model is then trained on VSGaze for 20 additional epochs with a learning rate of 3×10^{-6}, while freezing the ViT $\mathscr{V}$. For validation, we use the official splits for UCO-LAEO, VideoCoAtt, and ChildPlay, and the splits proposed by Tafasca *et al.* (Tafasca et al., 2023b) for GazeFollow and VAT.

For all experiments, we use a temporal window of $T = 5$ frames with a stride of 3. To enable batch training, we randomly sample up to $N_p = 4$ people per scene (padding if necessary). At test time, we evaluate with a batch size of 1 and consider all visible people. The Interaction Module consists of $B = 4$ blocks, interacting with the ViT at layers $\{2, 5, 8, 11\}$.

Implementation Details. The Gaze Backbone $\mathscr{G}_b$ is a ResNet-18 pretrained on Gaze360 (Kellnhofer et al., 2019), and processes head crops at a resolution of 224×224. The scene is processed using a ViT-base model $\mathscr{V}$ (Arnab et al., 2021) initialized with MultiMAE weights (Bachmann et al., 2022), also at a resolution of 224×224. The temporal position embedding $\mathbf{z}$ is zero-initialized, and all other new layers are randomly initialized.

We train using the AdamW optimizer (Loshchilov and Hutter, 2018) with warm-up and cosine annealing. The loss coefficients are set to: $\lambda_{hm} = 1000$, $\lambda_{dir} = 3$, $\lambda_{io} = 2$, and $\lambda_{LAH} = \lambda_{SA} = 1$. To address class imbalance between positive and negative social gaze labels, we upweight positive samples by a factor of 2.

All models are trained on a single NVIDIA RTX 3090 GPU (24GB). Each experiment takes approximately 8 hours to train on GazeFollow or VSGaze, totaling around 140 GPU hours across all experiments.

Metrics. We use standard gaze following metrics: AUC; Recasens et al. (2015), Distance, and In-Out AP as defined in Section 4.1.

For social gaze, we follow the inference protocol described in Section 4.2, computing metrics for both the task-specific decoder outputs and post-processed gaze following predictions. In

Method	**PP**	**Dist.** ↓	$\mathbf{AP_{IO}}$ ↑	$\mathbf{F1_{LAH}}$ ↑	$\mathbf{F1_{LAEO}}$ ↑	$\mathbf{AP_{SA}}$ ↑
Chong_S* (Chong et al., 2020b)	✓	0.121	0.918	0.778	0.562	0.288
Chong_T* (Chong et al., 2020b)	✓	0.130	**0.956**	0.764	0.529	0.331
Gupta* (Gupta et al., 2022)	✓	0.119	0.929	0.784	0.590	0.335
Ours-noGF	✗	-	-	0.738	0.579	<u>0.515</u>
Ours-noSoc	✓	<u>0.111</u>	<u>0.945</u>	<u>0.802</u>	<u>0.598</u>	0.339
Ours	✗	**0.107**	0.940	0.795	0.590	**0.576**
Ours-PP	✓	**0.107**	0.940	**0.812**	**0.603**	0.352

Table 6.4: Comparison against gaze following methods on VSGaze. All models were trained on VSGaze. PP indicates social gaze predictions from post-processing gaze following outputs (✓) vs predictions from decoders (✗). Best results are in bold, second best results are underlined.

particular, we threshold LAH decoder outputs at 0.5 to compute F1 scores, enabling direct comparison with post-processed F1 scores.

6.2.3 Results

Comparison against the State-of-The Art

We compare against recent SoTA methods addressing either social gaze tasks or gaze following. In addition, for fairness and to evaluate the impact of the VSGaze dataset, we also re-trained on this dataset the static image based models of Chong (Chong et al., 2020b) (Chong_S*) and Gupta (Gupta et al., 2022) (Gupta*), as well as the temporal model of (Chong et al., 2020b) (Chong_T*), the only temporal gaze following model with available code.

VSGaze. The results on VSGaze are given in Table 6.4. Note that regarding our approach, for social gaze, we compute the scores by leveraging either the predictions from the respective task decoders (Ours), or by post-processing the gaze following outputs of our model (Ours-PP).

Compared to the baselines, we observe that our model achieves the best performance for all tasks except for in-out gaze prediction. In particular, we achieve significant gains in the distance and AP_{SA} metrics when leveraging the predictions from the SA decoder. The latter highlights the importance of modeling SA as a classification task compared to post-processing gaze following outputs, which struggles to capture whether the gaze points for a pair of people falls on the same semantic item.

In addition, we note that better gaze following performance does not always translate to better social gaze performance. For instance, although Chong_S* has a better distance score compared to Chong_T*, it performs worse for shared attention. This effect is even more pronounced on ChildPlay (appendix Table 6.4), and suggests the benefit of considering social gaze metrics for better characterizing the performance of gaze following models, especially its semantic performance. In appendix Table 6.4, we provide a breakdown of performance on each of the component datasets of VSGaze.

State-of-the-art comparison: fine tuning on individual datasets. Table 6.5 compares our

Method	Multi	AUC↑	Avg.Dist.↓	Min.Dist.↓
Fang et al. (2021)	✗	0.922	0.124	0.067
Tonini et al. (2022)	✗	0.927	0.141	-
Jin et al. (2022)	✗	0.920	0.118	0.063
Bao et al. (2022)	✗	0.928	0.122	-
Hu et al. (2022b)	✗	0.923	0.128	0.069
Tafasca et al. (2023b)	✗	0.936	0.125	0.064
Chong$_S$ (Chong et al., 2020b)	✗	0.921	0.137	0.077
Gupta (Gupta et al., 2022)	✗	0.933	0.134	0.071
Jin et al. (2021)	✓	0.919	0.126	0.076
Ours-static	✓	0.929	**0.116**	**0.059**

(a) Results on GazeFollow (Recasens et al., 2017).

Method	Multi	Dist.↓	AP_{IO} ↑
Tafasca et al. (2023b)	✗	0.107	0.986
Gupta (Gupta et al., 2022)	✗	0.113	0.983
Ours	✓	0.117	**0.994**
Ours†	✓	**0.113**	0.993

(c) Results on ChildPlay (Tafasca et al., 2023b).

Method	Multi	Dist.↓	AP_{IO} ↑
Fang et al. (2021)	✗	0.108	0.896
Tonini et al. (2022)	✗	0.129	-
Jin et al. (2022)	✗	0.109	0.897
Bao et al. (2022)	✗	0.120	0.669
Hu et al. (2022b)	✗	0.118	0.881
Tafasca et al. (2023b)	✗	0.109	0.834
Chong$_T$ (Chong et al., 2020b)	✗	0.134	0.853
Gupta (Gupta et al., 2022)	✗	0.134	0.864
Jin et al. (2021)	✓	0.134	**0.880**
Ours	✓	**0.105**	0.869
Ours†	✓	**0.105**	0.869

(b) Results on VAT (Chong et al., 2020b).

Method	Dist.↓	AP_{LAEO} ↑
Marin-Jimenez et al. (2019)	-	0.795
Doosti et al. (2021)	-	0.762
Marín-Jiménez et al. (2021)	-	0.867
Ours	0.023	0.963
Ours†	**0.019**	**0.974**

(d) Results on UCO-LAEO (Marin-Jimenez et al., 2019).

Table 6.5: Comparison against task specific methods fine-tuned on individual datasets. Best multi-person results are in bold, overall best results are underlined. Multi indicates multi-person (✓) vs single-person (✗) gaze following methods. Ours is initialized from training on GazeFollow, while Ours† is initialized from training on VSGaze.

model against task specific methods. For GazeFollow, we use our static model (Ours-static) that was trained on GazeFollow and used to initialize our model trained on VSGaze. For the video datasets, as SoTA methods were trained (or finetuned) on individual datasets, for fairness we also fine-tune our model on these datasets, investigating two initialization alternatives: either from the model trained on GazeFollow (Ours), or from the model trained on VSGaze (Ours†). Note that we are unable to compare against previous results for VideoCoatt due to our new pair-wise evaluation protocol that better captures SA performance (Sec 4.2).

On GazeFollow and VAT, our model outperforms the only other comparable multi-person gaze following model of Jin (Jin et al., 2021). It also achieves competitive or better results to single-person methods, even those leveraging auxiliary modalities such as depth (Fang et al., 2021; Tonini et al., 2022; Bao et al., 2022; Jin et al., 2022; Hu et al., 2022b; Tafasca et al., 2023b). Importantly, on the social LAEO task, we set the new state of the art on UCO-LAEO, far outperforming methods designed specifically for LAEO (Marin-Jimenez et al., 2019; Marín-Jiménez et al., 2021; Doosti et al., 2021).

We also note that fine-tuning using the VSGaze model initialization can improve results compared to the standard protocol of fine-tuning after training on GazeFollow (ex. distance on ChildPlay and AP_{LAEO} on UCO-LAEO). This suggests that training on VSGaze can leverage the complementary knowledge provided by the different tasks and datasets, which follows observations made in other works addressing multi-task training (Ci et al., 2023).

Analysis

Impact of Architecture. Comparing the performance of our model with no social gaze losses (Ours-noSoc) against the baselines (Table 6.4), we see that it already performs on par or better than them while being much more efficient as it processes the image only once for all people in the scene. It also serves as a strong gaze following baseline to compare performance against.

Impact of Social Gaze Loss. Our architecture can further benefit from the social gaze losses, showing improved gaze following performance and social gaze prediction (Ours and Ours-PP, Table 6.4). In particular, we observe significant gains for the SA compared to Ours-noSoc. Interestingly, the addition of the social gaze losses also better aligns the gaze following outputs for social gaze prediction. Comparing Ours-PP and Ours-noSocial, we see that performance for all social gaze tasks is improved.

Impact of Gaze Following Loss. We additionally train our model without the standard gaze following losses: heatmap, gaze vector and in-out (Ours-noGF, Table 6.4). Across VSGaze, we see that performance for all social gaze tasks drops, which indicates that the gaze following and social gaze losses provide complementary information, and using both can give improved performance.

Impact of VSGaze. When comparing the performance of the models trained on VSGaze (appendix Table 6.7) against their versions fine-tuned on individual datasets (Table 6.5), we see that the fine-tuned models always perform better. This is because fine-tuning allows the models to learn dataset specific priors (ex. more LAH cases in VAT, Table 3.3). For instance, on VAT, Gupta* has a distance score of 0.138 when trained on VSGaze, compared to a score of 0.134 when directly fine-tuned on VAT. Also, our model has a distance score of 0.112 when trained on VSGaze, and a score of 0.105 when fine-tuned on VAT. This highlights the challenge in leveraging multiple datasets: while we may expect better performance by having more data, the different priors and statistics bring additional difficulties. Nevertheless, our model trained on VSGaze is able to achieve strong performance across all datasets.

Impact of temporal information. Comparing the static and temporal versions of our model trained on VSGaze, we observe improvements in performance for shared attention (0.555 vs 0.576, Tab. 6.8 in the appendix), and similar or slightly improved performance for other metrics. This is in contrast with Chong$_T$, which often has lower performance than Chong$_S$ (for distance, LAH and LAEO metrics in Table 6.4). These results follow observations from prior work (Section 2.2) and highlight the challenge in leveraging temporal information for gaze following. We provide a detailed analysis in appendix Section 6.4.

Method	Res.	AUC↑	Avg.Dist.↓	Min.Dist.↓
Gaze-LLE (Ryan et al., 2025)	448	**0.956**	0.104	0.045
MTGS-static	224	0.929	0.116	0.059
MTGS-DINO-static	224	0.944	0.101	0.045
MTGS-DINO-static	448	0.949	**0.098**	**0.043**

(a) Results on GazeFollow (Recasens et al., 2015).

Method	Res.	Dist. ↓	AP$_{\text{IO}}$ ↑
Gaze-LLE (Ryan et al., 2025)	448	0.107	**0.897**
MTGS-static	224	0.114	0.843
MTGS-DINO-static	224	0.102	0.852
MTGS-DINO-static	448	**0.096**	0.878

(b) Results on VAT (Chong et al., 2020b). MTGS models were trained on VSGaze and not further finetuned on VAT.

Method	Res.	Dist. ↓	AP$_{\text{IO}}$ ↑	F1$_{\text{LAH}}$ (PP)↑	F1$_{\text{LAEO}}$ (PP)↑	AP$_{\text{SA}}$ ↑
MTGS-static	224	0.108	0.946	0.806	0.599	0.555
MTGS-DINO-static	224	0.096	0.951	**0.833**	**0.624**	0.551
MTGS-DINO-static	448	**0.089**	**0.960**	0.832	0.621	**0.595**

(c) Results on VSGaze. We provide results for LAH and LAEO using the post-processing approach.

Table 6.6: Performance of MTGS with a DINOv2 (Oquab et al., 2023) backbone for scene encoding. We observe consistent improvements across all metrics, surpassing the previous state of the art, Gaze-LLE (Ryan et al., 2025). Results are reported for different scene input resolutions (Res.).

6.2.4 Leveraging DINOv2 Representations

In this unpublished extension, we investigate the benefits of leveraging recent foundation models, specifically DINOv2, to improve the performance of MTGS. A recent gaze following method, Gaze-LLE (Ryan et al., 2025), demonstrated state-of-the-art performance by using a frozen DINOv2 scene encoder. We evaluate whether a similar approach can enhance our own model.

To this end, we initialize the ViT encoder ($\mathcal{V}$) with DINOv2 weights instead of the MultiMAE weights (Bachmann et al., 2022) used in earlier experiments. We follow the same training protocol described in Section 6.2.2, first pretraining on GazeFollow and then fine-tuning on VSGaze. Throughout this process, the ViT encoder remains frozen.

Results are presented in Table 6.6. We observe consistent improvements across all metrics, allowing us to outperform Gaze-LLE and set a new state of the art. We also experiment with different scene input resolutions. We see that increasing the resolution further improves performance on the distance and in–out metrics. The observed improvement for SA performance in Table 6.6c is likely due to using a lower learning rate for the SA decoder in the 224 × 224 experiment compared to the 448 × 448 setting. Note that all results in the tables are using the static version of MTGS. We also trained a temporal version but did not see much improvement compared to the static version, potentially due to hyperparameter tuning, but also due to the challenges detailed in Section 6.4.

6.2.5 Conclusion

In this work, we proposed a new framework for multi-person, temporal gaze following and social gaze prediction. Building on limitations identified in Sharingan Social, our approach introduces a dual-branch transformer architecture for person and scene encoding, explicitly incorporating person interactions and enabling targeted temporal modeling within the person branch.

A key strength of our method lies in its ability to perform joint training across gaze following and multiple social gaze tasks (LAH, LAEO, SA), leveraging the VSGaze dataset to learn from diverse supervision sources in a unified manner. Through extensive experiments, we demonstrated that our model can achieve strong gaze following and social gaze prediction performance across component datasets of VSGaze. Notably, joint training across tasks leads to mutual benefits—improving social gaze performance through gaze following and vice versa. The model can also be further fine-tuned on specific datasets to optimize for targeted deployment scenarios. Finally, we show that integrating DINOv2 scene representations further improves performance, establishing a new state of the art for gaze following.

Our approach thus represents a step forward in unified gaze modeling, bridging spatial, temporal, and social dimensions in a cohesive framework.

6.3 Conclusion

In this chapter, we advanced the task of gaze prediction through a series of methods that progressively unify gaze following and social gaze understanding. We began by demonstrating how representations learned for gaze following can be leveraged to support higher-level social gaze tasks. Building on this, we proposed a novel, temporal, multi-person transformer architecture (MTGS) capable of jointly modeling gaze following and social gaze behaviors—namely, looking at heads (LAH), mutual gaze (LAEO), and shared attention (SA). Our design supports targeted temporal modeling, interaction-aware person encoding, and unified training across tasks using the VSGaze dataset.

Experimental results confirmed that the tasks benefit from being trained jointly, with improvements in both spatial and semantic gaze performance. Our model generalizes well across datasets with diverse visual and interaction statistics, while still allowing targeted fine-tuning for specific domains.

In the following chapters, we extend this framework in two directions. First, we explore how gaze-relevant contextual cues can be extracted using vision-language models (VLMs) and integrated into MTGS to increase robustness (Chapter 7). Second, we demonstrate the application of MTGS to naturalistic child–adult interaction data to (Chapter 9).

6.4 Appendix

In this section, we provide some additional analysis and ablations of MTGS. Unless specified, all experiments and visualizations here use LAH and LAEO predictions obtained via the post-processing strategy, and SA predictions from the corresponding decoder. This is because post-processing gaze following outputs for LAH and LAEO ensures that outputs for these three tasks align. We also observe slightly better performance for LAH and LAEO using the post-processing approach compared to using the predictions from their respective decoders (Table 6.4, Ours-PP vs Ours).

Breakdown of Results on VSGaze

We provide a breakdown of results on VSGaze by component dataset in Table 6.7. Note that following the results in the main paper, LAH and LAEO results for Ours-noGF and Ours are obtained from their respective decoders. We can observe that performance trends on individual datasets can differ from the aggregated results on VSGaze. For instance, although we have a small improvement for LAH compared to the baselines across VSGaze, we perform significantly better on ChildPlay. In general, as VideoCoAtt represents the highest number of samples in VSGaze, it also has the highest impact.

Also on ChildPlay, once again we see that better gaze following performance does not translate to better social gaze performance. Although Gupta* has a better distance score compared to Chong$_S$*, it performs significantly worse for all social gaze tasks.

Ablations

Temporal Window Length

We compare performance of our model for different temporal window lengths on VSGaze in Table 6.8. Note that $T = 1$ corresponds to a static model. We observe that incorporating temporal information can improve performance, especially in the case of shared attention. For the other metrics performance remains comparable. As a temporal window of 9 does not necessarily give better performance than a temporal window of 5, we use $T = 5$ for our experiments.

We note that these observations are in contrast to those from the static (Chong$_S$) and temporal models (Chong$_T$) of (Chong et al., 2020b). As seen in Table 6.4, Chong$_T$ often performs worse than Chong$_S$, with especially lower scores for the distance and LAEO metrics. This follows prior observations from the state of the art regarding temporal modelling for gaze following (Section 2.2), and illustrates the challenge in leveraging temporal information.

While architecture design may be a reason for the lack of greater improvement in performance, the data itself is an important factor. Firstly, despite the larger number of samples in VSGaze

Dataset	Method	PP	Dist. ↓	AP$_{\mathrm{IO}}$ ↑	F1$_{\mathrm{LAH}}$ ↑	F1$_{\mathrm{LAEO}}$ ↑	AP$_{\mathrm{SA}}$ ↑
VAT (Chong et al., 2020b)	Chong$_S$* (Chong et al., 2020b)	✓	0.132	0.798	0.785	0.486	0.288
	Chong$_T$* (Chong et al., 2020b)	✓	0.137	0.843	0.783	0.479	0.332
	Gupta* (Gupta et al., 2022)	✓	0.138	0.795	0.766	0.518	0.300
	Ours-noGF	✗	-	-	0.766	0.503	0.435
	Ours-noSoc	✓	0.121	**0.847**	0.812	**0.557**	0.440
	Ours	✗	**0.112**	0.845	0.791	0.526	**0.521**
	Ours-PP	✓	**0.112**	0.845	**0.825**	0.548	0.497
ChildPlay (Tafasca et al., 2023b)	Chong$_S$* (Chong et al., 2020b)	✓	0.123	0.973	0.597	0.470	0.154
	Chong$_T$* (Chong et al., 2020b)	✓	0.137	0.985	0.572	0.416	0.165
	Gupta* (Gupta et al., 2022)	✓	0.119	0.979	0.571	0.428	0.132
	Ours-noGF	✗	-	-	0.609	0.404	0.207
	Ours-noSoc	✓	0.118	**0.994**	0.620	0.412	0.188
	Ours	✗	**0.113**	0.993	**0.682**	0.426	0.179
	Ours-PP	✓	**0.113**	0.993	0.651	**0.436**	**0.216**
VideoCoAtt (Fan et al., 2018)	Chong$_S$* (Chong et al., 2020b)	✓	0.120	-	0.793	-	0.290
	Chong$_T$* (Chong et al., 2020b)	✓	0.126	-	0.790	-	0.337
	Gupta* (Gupta et al., 2022)	✓	0.115	-	0.815	-	0.347
	Ours-noGF	✗	-	-	0.733	-	0.524
	Ours-noSoc	✓	0.107	-	0.822	-	0.335
	Ours	✗	**0.106**	-	0.804	-	**0.601**
	Ours-PP	✓	**0.106**	-	**0.825**	-	0.345
UCO-LAEO (Marin-Jimenez et al., 2019)	Chong$_S$* (Chong et al., 2020b)	✓	0.031	-	0.986	0.811	-
	Chong$_T$* (Chong et al., 2020b)	✓	0.064	-	0.941	0.774	-
	Gupta* (Gupta et al., 2022)	✓	0.031	-	0.989	0.859	
	Ours-noGF	✗	-	-	0.989	**0.939**	-
	Ours-noSoc	✓	0.043	-	0.978	0.840	-
	Ours	✗	**0.027**	-	0.990	0.888	-
	Ours-PP	✓	**0.027**	-	**0.994**	0.870	-
VSGaze	Chong$_S$* (Chong et al., 2020b)	✓	0.121	0.918	0.778	0.562	0.288
	Chong$_T$* (Chong et al., 2020b)	✓	0.130	**0.956**	0.764	0.529	0.331
	Gupta* (Gupta et al., 2022)	✓	0.119	0.929	0.784	0.590	0.335
	Ours-noGF	✗	-	-	0.738	0.579	0.515
	Ours-noSoc	✓	0.111	0.945	0.802	0.598	0.339
	Ours	✗	**0.107**	0.940	0.795	0.590	**0.576**
	Ours-PP	✓	**0.107**	0.940	**0.812**	**0.603**	0.352

Table 6.7: Comparison against gaze following methods on VSGaze and its component datasets: VAT (Chong et al., 2020b), ChildPlay (Tafasca et al., 2023b), VideoCoAtt (Fan et al., 2018) and UCO-LAEO (Marin-Jimenez et al., 2019). All models were trained on VSGaze. PP indicates social gaze predictions from post-processing gaze following outputs (✓) vs predictions from decoders (✗). Best results are in bold, second best results are underlined.

compared to standard video based gaze datasets, there is high redundancy between frames so data diversity is not comparable to that of GazeFollow. Secondly, the moments where temporal information is important, such as during gaze shifts, only form a small percentage of total instances. Hence, improvements for predictions in these moments are not reflected in overall metrics. For instance, gaze shifts form less than 10% of total instances in ChildPlay (Tafasca et al., 2023b). However, in our qualitative analysis (Section 6.4) we can see situations where temporal information helps. In future work we plan to investigate new metrics for evaluating the performance of temporal models.

T	**Dist. ↓**	**AP_{IO} ↑**	**F1_{LAH} ↑**	**F1_{LAEO} ↑**	**AP_{SA} ↑**
1	0.108	**0.946**	0.806	0.599	0.555
5	0.107	0.940	**0.812**	**0.603**	**0.576**
9	**0.106**	0.943	0.811	0.590	0.563

Table 6.8: Ablations for different temporal window lengths T on VSGaze. Best results are in bold, second best results are underlined.

Method	**Dist. ↓**	**AP_{IO} ↑**	**F1_{LAH} ↑**	**F1_{LAEO} ↑**	**AP_{SA} ↑**
Ours-noI$_{ppt}$	0.108	0.937	0.806	**0.612**	0.547
Ours-noI$_{sp}$	**0.105**	0.936	**0.813**	0.581	0.545
Ours-noI$_{ps}$	0.112	0.939	0.799	0.605	0.538
Ours-noDPT	0.111	**0.941**	0.810	0.587	0.528
Ours	0.107	0.940	0.812	0.603	**0.576**

Table 6.9: Ablations on different novel components of our architecture on VSGaze. I$_{ppt}$ refers to the Spatio-Temporal Social Interaction component, I$_{sp}$ refers to the Scene-to-Person encoder, I$_{ps}$ refers to the Person-to-Scene encoder and DPT refers to the gaze heatmap decoder. Best results are in bold, second best results are underlined.

Architecture Components

We systematically remove different novel components of our architecture to analyse their impact and provide results in Table 6.9.

Interaction Module. Removing the Person-to-Scene Interaction encoder $\mathscr{I}^{\text{b}}_{\text{ps}}$ (Ours-noI$_{ps}$) has the largest impact on performance, especially for distance, LAH and SA. Without this encoder, the frame tokens cannot access the person tokens, so encoding their gaze relevant salient items is much harder. Removing the Scene-to-Person Interaction encoder $\mathscr{I}^{\text{b}}_{\text{sp}}$ (Ours-noI$_{sp}$) decreases LAEO and SA performance. Without this encoder, the person tokens cannot access the frame tokens, so they cannot capture the locations of gazed at salient items. As the frame tokens may be able to adapt to this change, gaze following performance is not impacted negatively. Finally, removing the Spatio-Temporal Social Interaction component $\mathscr{I}^{\text{b}}_{\text{pp}}$, $\mathscr{I}^{\text{b}}_{\text{pt}}$ (Ours-noI$_{ppt}$) decreases LAH and SA performance. Without this component, there is no interaction between person tokens, so identification of social dynamics is hindered. Interestingly, we see a boost in LAEO performance. However, a major portion of LAEO positives come from the UCO-LAEO dataset (see Table 3.3) which consists mainly of two person scenes, so capturing social interactions may be less important.

DPT Decoder. We replace our proposed modified DPT decoder for gaze heatmap prediction (Tafasca et al., 2024) with a simpler decoder (Ours-noDPT). This decoder projects the frame and person tokens from the last Interaction block, performs a dot product between them, and then upscales the output to the heatmap resolution. Using the simpler decoder results in drops in performance for the distance, LAEO and SA metrics. Unlike the DPT, it lacks multi-scale representations which impacts heatmap prediction and supervision of tokens.

Dataset	Method	Dist. ↓	$\mathrm{AP_{IO}}$ ↑	$\mathrm{F1_{LAH}}$ ↑	$\mathrm{F1_{LAEO}}$ ↑	$\mathrm{AP_{SA}}$ ↑
ChildPlay	Ours-spk	0.114	0.993	0.657	0.448	0.314
	Ours-spk‡	**0.113**	**0.994**	**0.659**	**0.449**	**0.321**
	Ours	**0.113**	0.993	0.658	0.436	0.319
VSGaze	Ours-spk	**0.107**	0.938	0.810	0.588	**0.590**
	Ours	**0.107**	**0.940**	**0.812**	**0.603**	0.576

Table 6.10: Performance for incorporating people's speaking information in our model. Best results are in bold. ‡ indicates ground truth speaking information.

Overall, we observe that removing the components tends to impact shared attention performance. This is similar to the observation in the previous section regarding temporal information. Unlike LAH and LAEO where the target is always another person, SA is more challenging as the shared attention target can be any person or object/point. Hence, this task may be benefitting more from additional information or architecture components.

Incorporating Auxiliary Information

Previous studies on analyzing conversations during meetings have shown that people usually look at the other speaking participants (Stiefelhagen et al., 2002), and such cues can be exploited for gaze target selection (Otsuka et al., 2005). Hence, we expect that identifying speaking persons can provide better scene understanding for gaze following, and help recognize attentiveness towards people, especially speakers. The latter is especially important in autism diagnosis, as eye contact is closely monitored by the clinician when they call out to the tested child (Edition, 2013).

Experiments and Results

To incorporate speaking information in our model, we adapt the Person Module (Section 6.2.1). Specifically, we first obtain speaking scores for each person using an active speaker detection model (Min et al., 2022). The model is retrained to detect speakers using the video alone and with no audio. The obtained speaking scores $\mathbf{s}_{i,t}$ for each person are linearly projected to the token dimension using $\mathscr{P}_{\text{spk}}$, and added to the person token. Hence, the new person token is obtained as:

$$\mathbf{p}_{i,t} = \mathscr{P}_{\text{gaze}}(\mathbf{g}_{i,t}^{\text{temp}}) + \mathscr{P}_{\text{spk}}(\mathbf{s}_{i,t}) + \mathscr{P}_{\text{box}}(\mathbf{B}_{i,t}). \tag{6.11}$$

We note that this formulation can easily be extended to incorporate other kinds of person-specific auxiliary information such as gestures. We explore this aspect in a follow up work where we extract gestures and other cues using vision-language models (Gupta et al., 2024c) (see Chapter 7).

We provide results for incorporating speaking information in Table 6.10. On VSGaze, once again we see improvements for SA while other metrics remain similar. On ChildPlay, we observe some improvements for LAEO. In particular, using ground truth speaking information

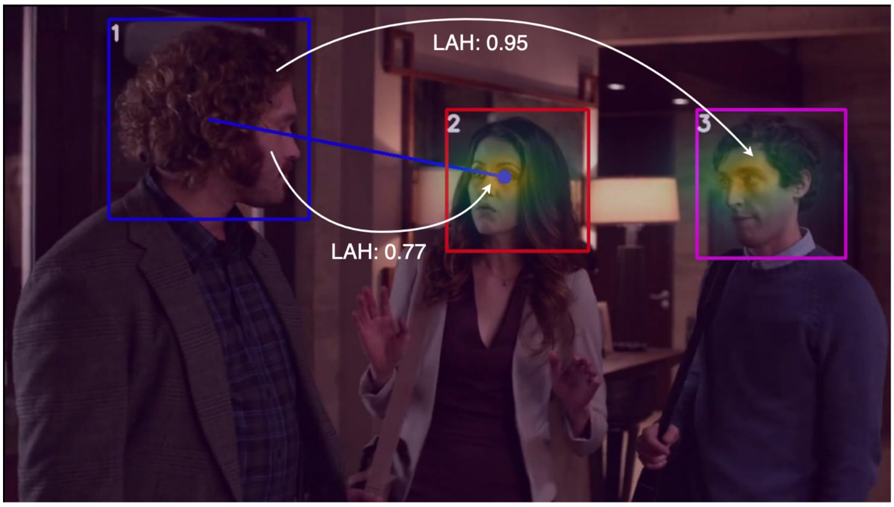


Figure 6.9: An illustration of the few cases where the predicted gaze point does not match with the predicted LAH label. The uncertainty in the gaze target is reflected in the heatmap, while the uncertainty in the LAH target is reflected in the LAH scores.

from ChildPlay-Audio (Section 3.2) gives us similar or better performance for all metrics.

As the speaker detection model tends to fail in cases with side-view faces or for children in the case of ChildPlay (see Table 3.2), we may further benefit from a better speaker detection model. In addition, investigating other ways to better capture and incorporate such auxiliary information in our model is an interesting direction for future research.

Qualitative Analysis

Qualitative Results and Comparisons

We provide qualitative results from our models in Figure 6.10. We observe that all models perform well, accurately capturing people's gaze targets and social gaze behaviour. We also see that incorporating temporal and speaking information can help improve predictions.

In the first sequence, the static model occasionally mispredicts person 1's gaze target as person 2's hands, as it cannot distinguish blinking from lowered gaze. In contrast, the temporal models recognize blinking and correctly maintain the target as person 2's face. In the second sequence, the static model misses shared attention between persons 1, 2, and 3 in the first frame, and between persons 2 and 3 in the third frame. This sequence is challenging due to subtle head motions, which the temporal models better capture. In the third sequence, both static and temporal models mispredict person 3's gaze target in frame 2. However, our model with speaking information correctly identifies person 1 as the target due to their high speaking score.

In addition, we provide qualitative comparisons of our model against other methods in

Figure 6.11. We see that our model performs better overall, accurately inferring people's gaze target and social gaze behaviour despite the complexity of the scenes with multiple salient targets, obscured eyes, varied settings and age groups.

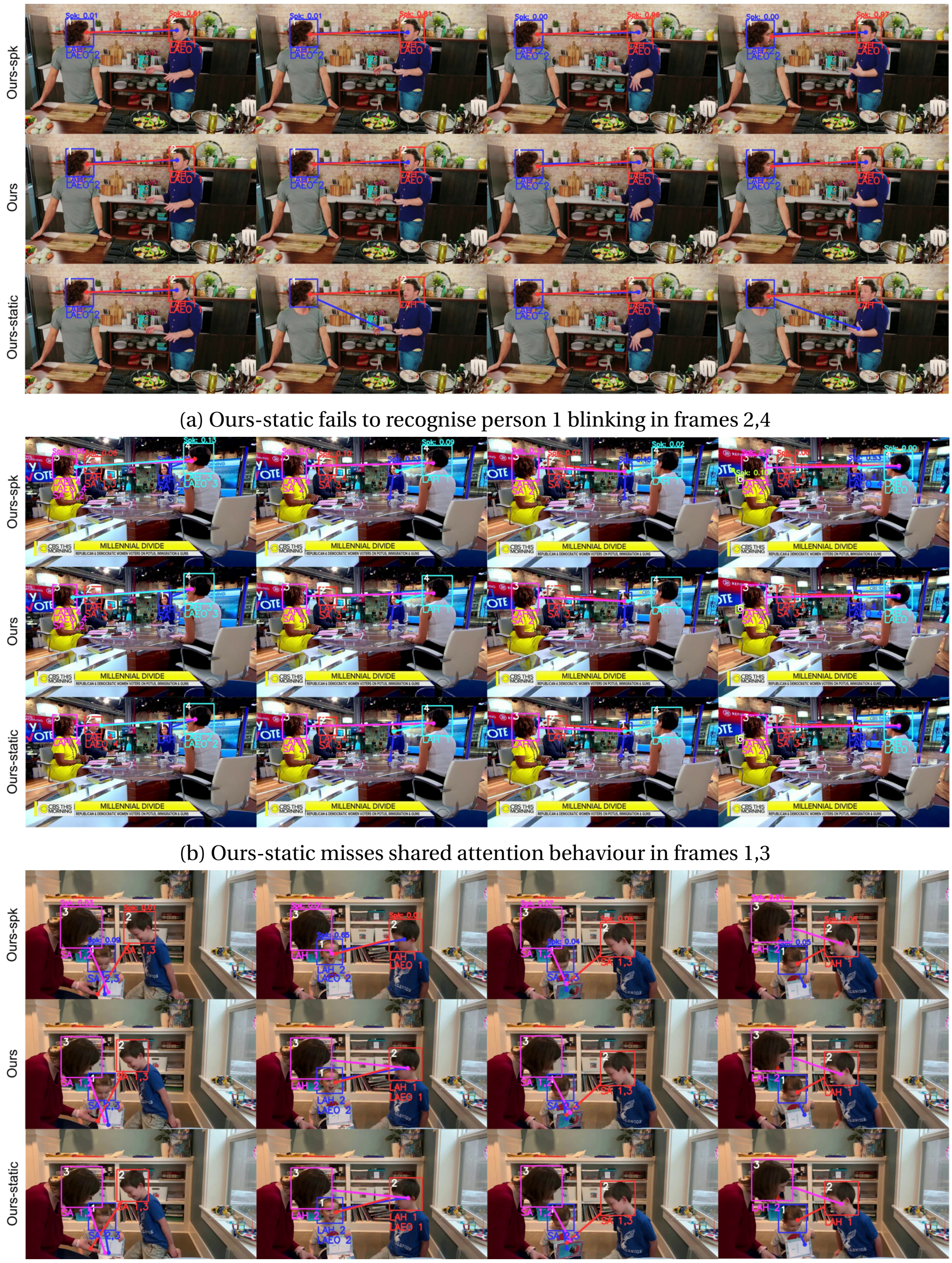
Alignment Between Gaze Following Outputs and Social Gaze Decoders

We analyze the difference in performance between post-processing gaze following outputs and using predictions from task-specific decoders. For LAH, predictions from both schemes align 87% of the time. When they diverge, it is typically due to the model being uncertain between two potential targets. The gaze heatmap arg max selects one, while the LAH score arg max selects the other. This confusion is illustrated for person 1 in Figure 6.9, where the predicted heatmap highlights both persons 2 and 3, and both corresponding LAH scores are high.

Limitations

As discussed in the section on qualitative results, gaze following outputs and predictions from the social gaze decoders do not always align (13% cases). One possible solution is to post-process gaze following outputs for LAH and LAEO as done in the case of Ours-PP. This ensures that outputs for gaze following, LAH and LAEO align. However, further aligning the outputs with SA is more challenging as post-processing for SA does not account for whether the gaze points fall on the same semantic item. Ensuring that outputs for all social gaze tasks and gaze following align is a challenging and interesting direction for future research. Potential directions include more refined post-processing techniques, or the addition of task consistency losses and regularization for the social gaze decoders. The latter is particularly interesting as it may further increase the benefit of the social gaze losses.

Secondly, as discussed in the ablations, temporal information does not seem to bring large improvements in performance, with improvements observed mainly for SA. Investigating new architectures, datasets and metrics to improve training and evaluation of temporal information is another important and interesting direction for future research.



(a) Ours-static fails to recognise person 1 blinking in frames 2,4

(b) Ours-static misses shared attention behaviour in frames 1,3

(c) Ours-spk captures the right target for person 3 in frame 2

Figure 6.10: Qualitative results of our proposed model (Ours), our model with speaking information (Ours-spk) and our model without temporal information (Ours-static). When the target is predicted to be inside the frame, we display the predicted gaze point and the social gaze tasks with the associated person id(s).

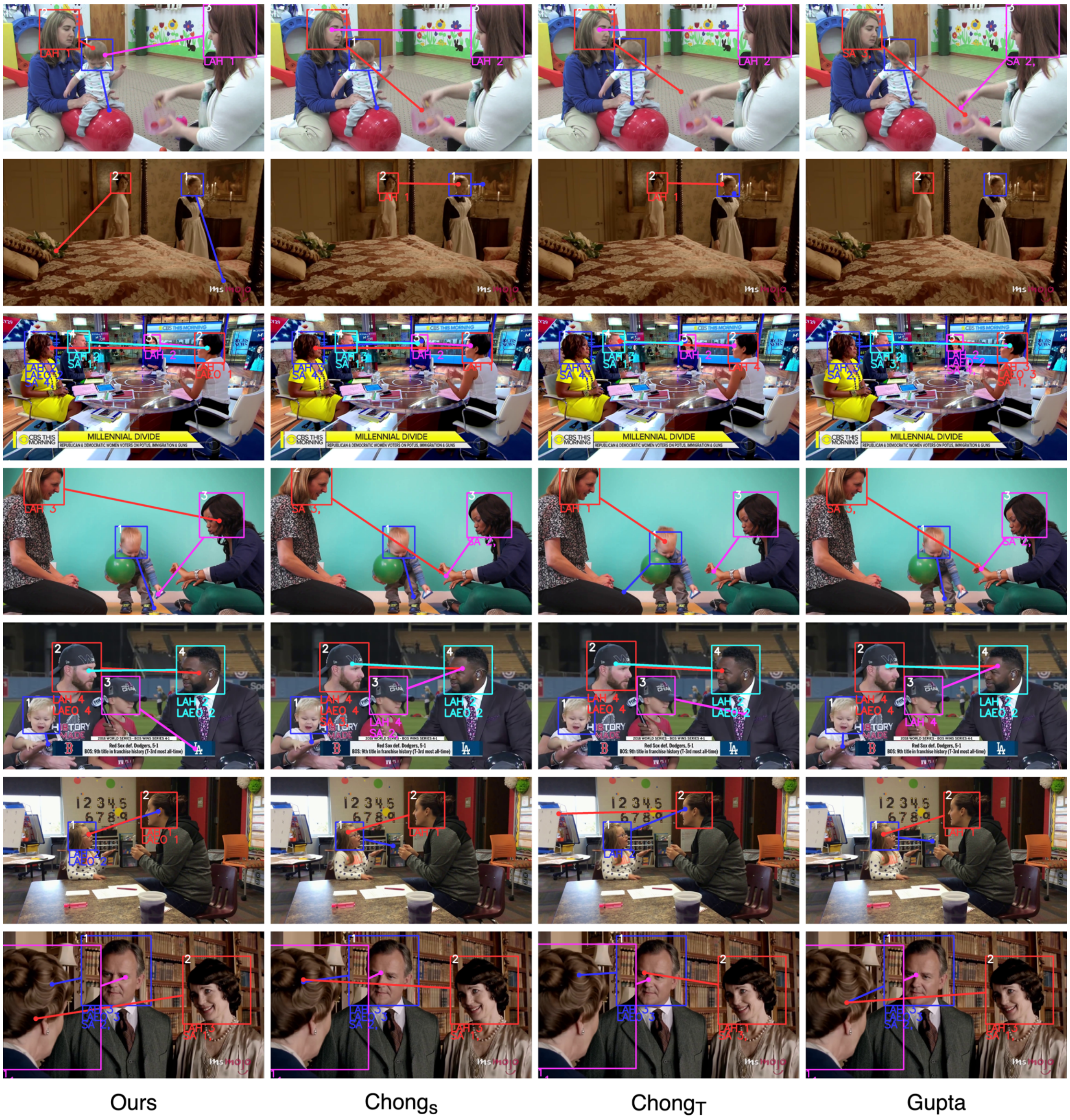


Figure 6.11: Qualitative comparison of our model against other methods Chong_S, Chong_T (Chong et al., 2020b), Gupta (Gupta et al., 2022). Our model performs better overall, outperforming other methods in complex scenes with obscured eyes, multiple salient targets, varied settings and age groups.

7 VLMs for Contextual Cues

As highlighted in Section 1.1.1, gaze following is a challenging task, demanding a model to interpret a large spectrum of contextual cues such as the person's interactions with objects and other people in the scene. For instance, it has been shown that eye and hand movements are coordinated during manipulation activities (Johansson et al., 2001). Or that during conversations, people usually look at the person talking (Stiefelhagen et al., 1999) and that leveraging this information can help in gaze target selection in meeting settings (Otsuka et al., 2005).

In order for a model to capture these cues and learn their impact on gaze target selection, it would need to be trained on a high-quality and large-scale labeled dataset. However, existing gaze following datasets (Chong et al., 2020b; Tafasca et al., 2023b) are small-scale, hindering the effective utilization of these cues. To address these challenges, prior works have relied on dedicated cue extraction methods such as inferred body pose (Gupta et al., 2022; Lian et al., 2018), and supplied them to models for improved performance. However, these approaches focus on specific cues and do not supply the larger array of contextual cues which could be needed for accurate gaze target prediction. Traditional methods to address this limitation include: (1) manually annotating for relevant cues; but it is cost-intensive and not always available during inference, or (2) pseudo-labelling with expert models; however this requires access to a multitude of task specific models for each cue or a subset of cues. Hence, it is evident that novel solutions are required.

Given these challenges, in Gupta et al. (2024c), we investigated the potential of Visual Language Models (VLMs) to extract valuable contextual cues for gaze following, aiming to overcome the constraints of traditional labeling approaches. VLMs have shown promising zero-shot performance for a variety of tasks (Radford et al., 2021; Zhang et al., 2023), owing to their ability to learn visual-text associations at scale. Hence, a single model may be capable of extracting all relevant contextual cues. At the same time, given the zero-shot setting, the set of cues to be considered can be adjusted based on the domain, further increasing this approach's applicability.

In this work, we consider cues related to pose, person-person interactions, and person-object

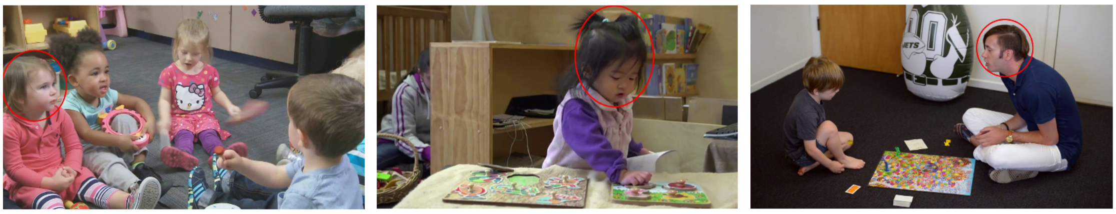


Figure 7.1: Accurately predicting a person's gaze target requires interpreting a variety of contextual cues such as body pose (left), object interactions (middle), and social engagement (right). In these samples from the ChildPlay dataset (Tafasca et al., 2023b), the child on the left is seated, the child in the middle is reaching for an object, and the man on the right is speaking.

interactions (see Figure 7.1). We first evaluate the zero-shot performance of different VLMs for recognising these cues (Section 7.1), and leverage the best performing approach to extract them. We then investigate whether incorporating these extracted cues can improve gaze following performance (Section 7.2).

Challenges. While VLMs have shown impressive zero-shot performance for a variety of tasks, these tasks (ex. image classification) usually involve processing the entire image. However, for accurate gaze following we also need to capture contextual cues related to each person in the scene. Hence, we need to consider an appropriate visual prompt to allow the VLM to focus on the person of interest. At the same time, it is important to consider the choice of text prompt as VLMs have been shown to benefit from prompt engineering (Radford et al., 2021). Finally, given the extracted cues from the VLMs, we need to consider how to incorporate such information into a gaze following model. Following these research questions, we make the following contributions:

- *VLMs for contextual cues extraction*: We explore 4 state of the art VLMs (Radford et al., 2021; Li et al., 2022, 2023a) for this task. We also investigate different visual prompts to focus on the person of interest, and different text prompts to describe the cue of interest. We show that VLMs can indeed capture contextual cues although the choice of VLM, visual prompt and text prompt impacts performance.
- *Text improved Gaze Following*: We incorporate the extracted contextual cues into MTGS (Gupta et al., 2024b). As shown in Section 6.4, MTGS's token based representation allows for the seamless integration of auxiliary information. Our results indicate that incorporating these cues can result in better generalization performance, especially when considering larger sets of cues.

7.1 Contextual Cues Extraction

In the first stage, we evaluate the zero-shot performance of different VLMs and prompting approaches for recognition of cues. Note that we interchangeably refer to a cue as a class.

7.1.1 Method

We investigate different visual and textual prompting strategies, as well as two different variants of VLMs for zero-shot contextual cues extraction, namely image-text matching (ITM) and visual question-answering (VQA).

Visual Prompting. In a complex scene involving multiple people, ITM becomes challenging as our task requires conditioning on a specific target person. To address this, we investigate various visual prompting techniques that enable ITM to focus on a chosen individual. We employ several approaches, including no prompting, drawing a red ellipse around the person following (Shtedritski et al., 2023), blurring or graying the background. These techniques are applied either to the entire image (image-based) or to the cropped target person (person-based), resulting in a total of eight distinct visual prompting approaches. We provide an example of the different visual prompts in Figure 7.8 in the appendix.

Text Prompting. In our approach to text prompting, we employed a structured method for generating prompts systematically based on templates. This method allows us to meticulously examine the impact of each textual component within the prompt. A template, in this context, is a fixed sentence where only specific parts can be altered. For examples, "a {photo} of a {person} {class}", and "a {person} is {class}", are two instances of templates. Beyond the varied sentence structures, the placeholders {photo}, {person}, and {class} can be substituted with semantically related components. For instance, {photo} could be replaced with "picture" or "snapshot," {person} might be substituted with "individual" or "human," and {class} can refer to class synonyms such as "talking" or "narrating" if the original class is "speaking". In this work, synonyms refer to changes in class synonyms otherwise mentioned explicitly. In the appendix, Figure 7.9 presents all the different templates and synonyms used in this section.

Image-Text Matching (ITM). In ITM, the objective is to compute a cosine similarity between the visual and textual embedding. A high similarity suggests that the image contains the textual description. Formally, given an image I of size $H \times W \times 3$ and a set of K class names, we use the visual and text encoders of a pre-trained VLM (e.g., CLIP) to get a visual embedding $e_I \in \mathbb{R}^d$ and K text embeddings $e_T \in \mathbb{R}^{K \times d}$. We perform the following matching:

$$S = dot(e_I \cdot e_T^{\mathrm{T}}) \tag{7.1}$$

where, $S \in \mathbb{R}^K$ are the resulting similarity scores. When multiple textual prompts refer to the same class name k, i.e. $e_{T_k} \in \mathbb{R}^{P \times d}$, we can perform an *Ensemble* to get the score.

$$S_k = dot(e_I \cdot \frac{1}{P} \sum_{e \in e_{T_k}} e^{\mathrm{T}}) \tag{7.2}$$

The *Ensemble* approach utilizes the mean embedding, acting as a centroid for a given class and thus is expected to be more robust. The scores for each class are then normalized across samples to have a zero mean and a standard deviation of one. In this work, we investigate three different pre-trained VLMs such as CLIP (Radford et al., 2021), BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023a). For more details regarding these models, we refer the readers to the original papers and details in Section 2.6.

Visual Question Answering (VQA). In order to explore the potential of LLMs for our task, we investigate a recent VQA model, BLIP-2 VQA (Li et al., 2023a), that leverages a LLM called FlanT5 (Raffel et al., 2020). In VQA models, a textual question is jointly input with an image to the model, and the model outputs a textual answer. We convert the text prompts described previously into a set of questions that result in simple *yes* or *no* answers, which we then convert into a binary score. Examples of prompts are displayed in appendix Fig 7.10. To further explore the benefits of ICL, we provide additional textual context in the form of a generated caption from the same model. Thus, the text input to the model is of the form *{generated caption} {text prompt}*. It is worth noting that the BLIP-2 VQA model is much slower to run than the ITM models as (1) the model is much larger due to the LLM, and (2) the answer is conditioned on the image *and* question, so we need to run a forward pass for each image-prompt pair. This is unlike the ITM models where the images and prompts can be processed separately, with a similarity score computed afterward.

7.1.2 Experiments

Datasets. We employ two datasets to shed light on the VLMs' ability to extract meaningful cues.

ChildPlay: We manually annotated 6 cues from the ChildPlay (Tafasca et al., 2023b) dataset, which is a recently proposed dataset for gaze following. For each class, we selected around 50 clear positives and 50 clear negatives. The classes and statistics are presented in the appendix Table 7.5.

AVA-Actions: Then, to scale our evaluation we used the validation split of the AVA dataset (Gu et al., 2018), which is a human action localization dataset. This dataset is much more challenging since it is heavily unbalanced and large scale containing around 41000 images. A subset of the classes of interest was selected. In Table 7.1, a summary of the dataset classes and distribution is shown.

Metrics. We leverage two metrics:

- *AP*: To evaluate the performance of different VLMs and prompting approaches, we use Average Precision (AP). It is computed per class between the ground truth and the scores obtained from the VLMs. We also consider the mean of the AP scores across all classes or mean Average Precision (mAP).
- *Accuracy*: Since the output of the VQA variants is a binary decision, we cannot compute AP; instead, we compute accuracy. To compare with ITMs, we binarize their output by

Selected Classes - AVA	**Support**
Pose (P)	
stand	23424
sit	16660
bend/bow (at the waist)	1512
Person-Person Interaction (P-P)	
talk to (e.g., self, a person, a group)	25985
hug (a person)	340
hand clap	330
give/serve (an object) to (a person)	313
Person-Object Interaction (P-O)	
carry/hold (an object)	17199
touch (an object)	5099
read	658
write	273
lift/pick up	118
text on/look at a cellphone	112
work on a computer	111

Table 7.1: Selected classes from the AVA Dataset (validation set) categorized as Pose (P), Person-Person Interaction (P-P), and Person-Object Interaction (P-O), including the number of samples (support) for each.

applying a threshold of zero since the scores are normalized with a zero mean (however they may benefit from optimizing the threshold).

7.1.3 ITM Results

Visual prompting. We compare the performance of the different visual prompts described in Section 7.1.1 in Fig. 7.2. The results are aggregated across VLMs and different text prompts, and categorized by the type of visual input, i.e. image-based versus person-based. We see that image-based approaches outperform person-based variants. This suggests that a broader visual input provides additional context, enhancing the zero-shot recognition for the target person in the image. Furthermore, among the visual prompts, the red ellipse approach outperforms others, aligned with findings in (Shtedritski et al., 2023). Therefore, in subsequent experiments, we employ the image-based red ellipse as the visual prompt.

VLMs. We compare the performance of three VLMs, namely CLIP, BLIP, and BLIP-2. In Fig. 7.3, we present a class-wise comparison of the three VLMs on AVA. Note that, for each VLM, we aggregate the results from different textual prompts. Firstly, we observe that no single model always outperforms the others. However, BLIP and BLIP-2 surpass CLIP in pose and person-to-person classes, while CLIP performs well when the class refers to a clear object, such as *work on a computer* or *text on a cellphone*. This may be related to differences in training data, and is a direction for investigation. On average, BLIP-2 is the top performing model. In the subsequent analysis, we continue focusing on BLIP-2 while varying the text prompting aspects.

Text prompting. We investigate the impact of the text prompts described in Section 7.1.1 at

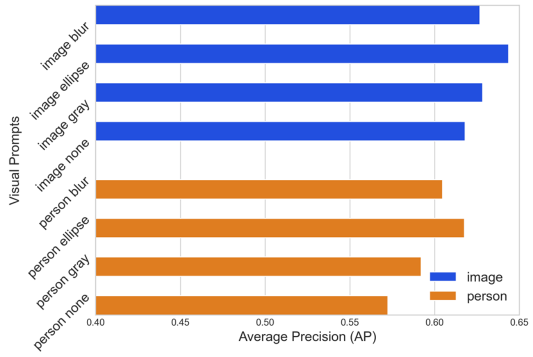


Figure 7.2: Results of different visual prompting approach on Childplay. *Image* corresponds to input the full image whereas *person* refers to the use of person crop as input.

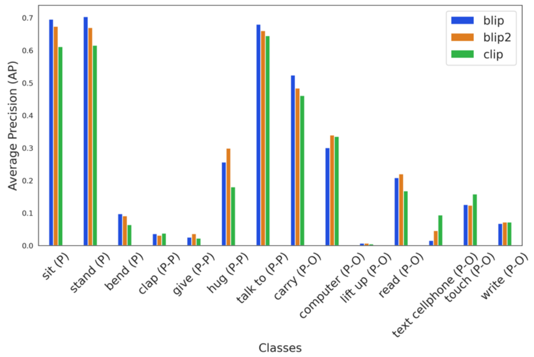


Figure 7.3: Results of different VLMs following the ITM approach on AVA. Three VLMs are compared across different classes categorized as Pose (P), Person-Person Interaction (P-P), and Person-Object Interaction (P-O).

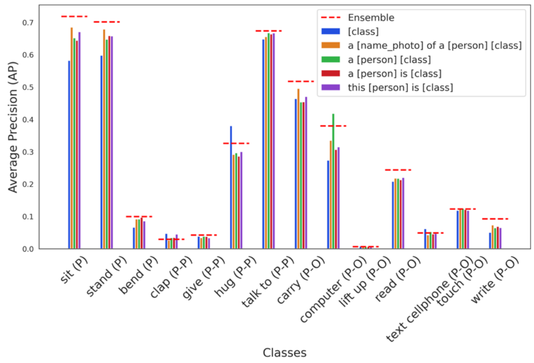


Figure 7.4: Results of different templates using BLIP-2 on AVA. Six templates are compared across different classes.

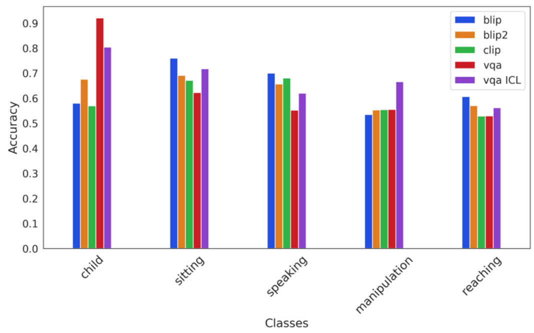


Figure 7.5: Results of BLIP-2 vqa with and without in-context learning, vqa ICL and vqa respectively, on ChildPlay. It is compared with the VLMs CLIP, BLIP and BLIP-2.

two different levels, at the template level and synonym level. When evaluating the template, we aggregate results over the other text prompt component variations. Similarly for when we evaluate the class synonym. In Figure 7.4, performance for different templates are shown on AVA per class. Firstly, there is no best template overall, which correlates to the finding of (Radford et al., 2021) that VLMs are prompt sensitive. However, using the *Ensemble* approach described in Section 7.1.1 provides more robust performance, often outperforming the best template, and always outperforming the worst template. In addition, the wording in textual prompts matters, as can be seen in the appendix Figure 7.11, where different class synonyms can change the performance by a large margin. However, we notice that for most of the classes, including *{person}* in the prompt improves performance. This suggests that conditioning the prompt to an individual helps to extract person-centric information.

7.1.4 VQA Results

To investigate the potential of LLMs and in-context learning for contextual cues extraction, we evaluate the BLIP-2 VQA model on the ChildPlay dataset, and compare it against ITM based VLM models (Figure 7.5). Note that the results are aggregated across all text prompts. As mentioned in Section 7.1.1, the BLIP-2 VQA model is much slower to run compared to the ITM based models which is why we use the smaller ChildPlay dataset. We also use a smaller set of templates and synonyms in the text prompt (Fig. 7.10 in appendix) to reduce computation time.

Benefit of LLM. Comparing the performance of BLIP-2 against BLIP-2 VQA (BLIP-2 and vqa in the figure), we see that BLIP-2 VQA does much better for the 'child' class, but on par of worse for the other classes. This suggests that the LLM in the BLIP-2 VQA model is not necessarily providing better results. However, as mentioned previously, this model uses a smaller set of templates and synonyms in the text prompt for computational reasons so may benefit from using a larger set.

In-Context Learning. We see that the BLIP-2 vqa model with ICL improves for all classes except the 'child' class compared to no ICL. This is in contrast to the observations in the original paper where the architecture is introduced (Li et al., 2023a), and suggests the potential of leveraging ICL for contextual cues extraction.

7.2 Text-Improved Gaze Following

In the second stage, we apply insights from Section 7.1 and leverage BLIP-2 along with the red ellipse visual prompting approach, and the *Ensemble* text prompting approach to extract contextual cues. We then evaluate the impact of incorporating these cues into a gaze following model.

7.2.1 Method

We employ the static version of the MTGS (Gupta et al., 2024b) model (see Section 6.2). This model is a transformer-based architecture designed for multi-person gaze following and social gaze prediction. Given an input image and head crops of people in the scene, it first produces two types of tokens: image tokens ($\mathbf{f} \in \mathbb{R}^{N \times D}$), similar to those in a standard Vision Transformer (ViT) architecture (Dosovitskiy et al., 2020), and person gaze tokens ($\mathbf{p} \in \mathbb{R}^{N_p \times D}$), where N_p represents the number of people in the scene. Person tokens are generated using head crops, a gaze backbone, and a subsequent linear projection layer. This formulation naturally supports incorporating contextual cues for each person, as the information can be fused with the corresponding person token.

Given the success of additive fusion in the case of position embeddings for transformers (Dosovitskiy et al., 2020), and early fusion of body pose and depth information for gaze following

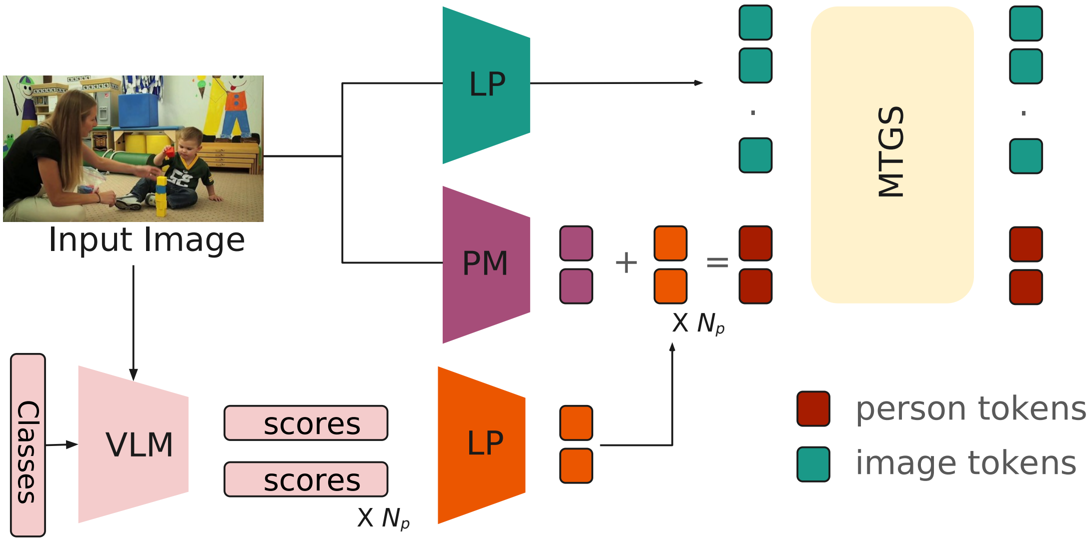


Figure 7.6: An overview of Text-Improved Gaze Following: Given an image containing N_p persons, image tokens and person tokens are generated via a Linear Projection (LP) and a person module (PM) respectively. To incorporate VLM contextual information, we use a VLM to obtain N_p score vectors, each with the dimension as the number of classes (K). We then linearly project these vectors and perform early fusion by adding them to the corresponding person tokens. Scene and updated person tokens are subsequently passed to MTGS (Gupta et al., 2024b) to model person and scene interactions using self and cross-attention modules across multiple blocks.

models (Gupta et al., 2022), we aim to incorporate contextual information derived from VLMs in an early fusion and additive manner. To this end, as illustrated in Fig. 7.6, we use a linear projection layer (Φ) to project the vector of predicted scores ($S_{vlm} \in \mathbb{R}^{N_p \times K}$, K is the number of classes) and generate person context tokens matching the dimensions of the person gaze tokens ($\Phi(S_{vlm}) \in \mathbb{R}^{N_p \times D}$). We then apply the add operation to combine the person context tokens from the VLMs with the corresponding person gaze tokens. Following this, the enriched person gaze tokens, now with added contextual cues, and the image tokens are fed into MTGS, where, people and scene tokens interact through self and cross-attention modules across multiple blocks.

$$\mathbf{p}^{\mathrm{o}}, \mathbf{f}^{\mathrm{o}} = \mathrm{MTGS}\left(\left[\mathbf{p} + \Phi\left(\mathrm{S}_{\mathrm{vlm}}\right), \mathbf{f}\right]\right) \tag{7.3}$$

Finally, a prediction module takes the updated tokens ($\mathbf{p}^{\mathrm{o}}, \mathbf{f}^{\mathrm{o}}$) and predicts the visual attention heatmap for each person, as well as pair-wise social gaze labels.

7.2.2 Experiments

Datasets. We leverage two gaze following datasets:

- *GazeFollow* (Recasens et al., 2015): Detailed in Section 3.1.1. In addition to the original annotations, we use the processed LAH labels generated following the protocol described in Section 3.4. Despite the relatively lower quality of images and annotations, GazeFollow serves as a valuable resource for pre-training due to its scale and diversity.

- *ChildPlay* (Tafasca et al., 2023b): Detailed in Section 3.1.3. Alongside the original gaze annotations, we also use the processed LAH and LAEO labels derived using the protocol in Section 3.4.

Contextual Cues. We define three sets of contextual cues:

- *AVA+CP:* These are the set of 24 cues defined in Section 7.1 for AVA and ChildPlay that were used for evaluating different VLMs and prompting strategies.
- *HICO:* The HICO dataset (Chao et al., 2015) is human-object interaction dataset that defines a list of 117 interaction verbs. We leverage these verbs as contextual cues.
- *SWIG:* The SWIG-HOI dataset (Wang et al., 2021b) is a large-scale human-object interaction dataset that defines 406 verbs. We leverage these verbs as contextual cues.

We provide the manually curated synonyms and templates used for generating different text prompts for AVA+CP in Figures 7.9,7.10 of the appendix. For HICO and SWIG, we use the same set of templates, but generate 4 synonyms for each cue using ChatGPT (OpenAI, 2024).

Training and Validation. Following (Gupta et al., 2024b), we train the model for 20 epochs on GazeFollow using a learning rate of 1e-4 and the AdamW (Loshchilov and Hutter, 2018) optimizer. We supervise using the standard MSE loss for gaze heatmap prediction, and binary cross entropy loss for LAH prediction. For validation, we use the split proposed in (Tafasca et al., 2023b).

Metrics. We use the standard gaze following metrics AUC; Recasens et al. (2015) and distance (see Section 4.1). In addition, we post-process the predicted points to obtain F1 scores for LAH and LAEO following the protocol in Section 4.2.1. Note that we can also use the LAH decoder outputs from MTGS, but it tends to have slightly lower scores than the post-processing approach (Section 6.2.2) and decoder outputs do not always align with gaze following outputs (Section 6.4).

7.2.3 Results

GazeFollow. We provide results for incorporating VLM context on the GazeFollow dataset in Table 7.2. We observe that performance does not change much for the distance score. In contrast, for LAH, we observe a slight improvement with the addition of AVA+CP cues, and a degradation with the addition of SWIG cues. However, the GazeFollow test set is very small (approx. 5k instances), and often contains simple scenes with a single salient target such as the held object. Also, annotations on GazeFollow are not always reliable as mentioned in Section 7.2.2. Hence, analyzing results on GazeFollow alone is not sufficient.

ChildPlay. To further investigate the properties of our models, we perform cross-dataset

Model	AUC↑	Avg.Dist↓	Min.Dist↓	$F1_{LAH}$ ↑
Fang et al. (2021)	0.922	0.124	0.067	-
Tonini et al. (2022)	0.927	0.141	-	-
Jin et al. (2022)	0.920	0.118	0.063	-
Bao et al. (2022)	0.928	0.122	-	-
Hu et al. (2022b)	0.923	0.128	0.069	-
Tafasca et al. (2023b)	0.936	0.125	0.064	-
Chong et al. (2020b)	0.921	0.137	0.077	-
Gupta et al. (2022)	0.933	0.134	0.071	-
Jin et al. (2021)	0.919	0.126	0.076	-
MTGS (Gupta et al., 2024b)	0.929	0.118	0.062	0.639
MTGS + AVA + CP	0.936	0.118	0.061	**0.643**
MTGS + HICO	0.934	**0.116**	**0.060**	0.639
MTGS + SWIG	0.933	0.119	0.061	0.619

Table 7.2: Results for incorporating VLM context with different sets of classes on the GazeFollow dataset. AVA+CP has 24 classes, HICO has 117 classes and SWIG has 406 classes. Best results are given in bold, second best results are underlined.

Method	Dist.↓	$F1_{LAH}$ ↑	$F1_{LAEO}$ ↑
Tafasca et al. (2023b)	**0.115**	-	-
Gupta et al. (2022)	0.142	-	-
MTGS (Gupta et al., 2024b)	0.122	0.588	0.376
MTGS + AVA + CP	0.129	0.586	0.371
MTGS + HICO	0.119	**0.601**	0.407
MTGS + SWIG	0.117	0.600	**0.426**

Table 7.3: Cross-dataset results for the models trained on GazeFollow and evaluated on the ChildPlay dataset. Best results are given in bold, second best results are underlined.

Method	AUC↑	Avg.Dist↓	Min.Dist↓	$F1_{LAH}$ ↑
Multi Fusion	0.932	0.119	0.062	0.633
Early Fusion	0.936	**0.118**	**0.061**	**0.643**

Table 7.4: Ablation on early vs multi-stage fusion of VLM context using the AVA+ChildPlay classes on the GazeFollow dataset. Best results are given in bold.

evaluation on ChildPlay in Table 7.3. The ChildPlay test set has a large number of instances (approx. 20k), and contains challenging scenes with multiple salient targets (ex. toys, other children/adults), making it an interesting benchmark. We observe that incorporating the AVA+CP classes results in a drop in performance for the distance score. However, with the larger set of HICO and SWIG classes, there is an improvement in performance for distance, LAH and LAEO. In particular, incorporating the SWIG classes gives the most improvements, with gaze following results comparable to the state of the art (Tafasca et al., 2023b) and contrasts with our observations on GazeFollow. This suggests that incorporating gaze contextual cues can result in more robust performance with better generalization.

Ablation: Early Fusion vs Multi-Stage Fusion. We perform an ablation with two different fusion mechanisms for incorporating VLM contextual information in MTGS in Table 7.4. The first is early fusion, and follows the approach described in Section 7.2.1. The second is a multi-stage fusion approach, where the VLM context is fused with the person tokens at every block of the architecture (4 times). We observe that the early fusion approach slightly outperforms the multi-stage fusion approach, especially for LAH, so we followed the early fusion approach for all our experiments.

Qualitative results. We provide qualitative results for MTGS, with and without the use of contextual cues in Figure 7.7. We observe that incorporating contextual cues can improve performance, helping identify the gaze target in challenging situations with multiple salient people and objects. For instance, in row 1, person 2 has a high score for *carrying*, which might indicate that this person is looking towards their hand. In row 2, person 3 has a high score for *talking on*, which suggests social interaction such as LAH.

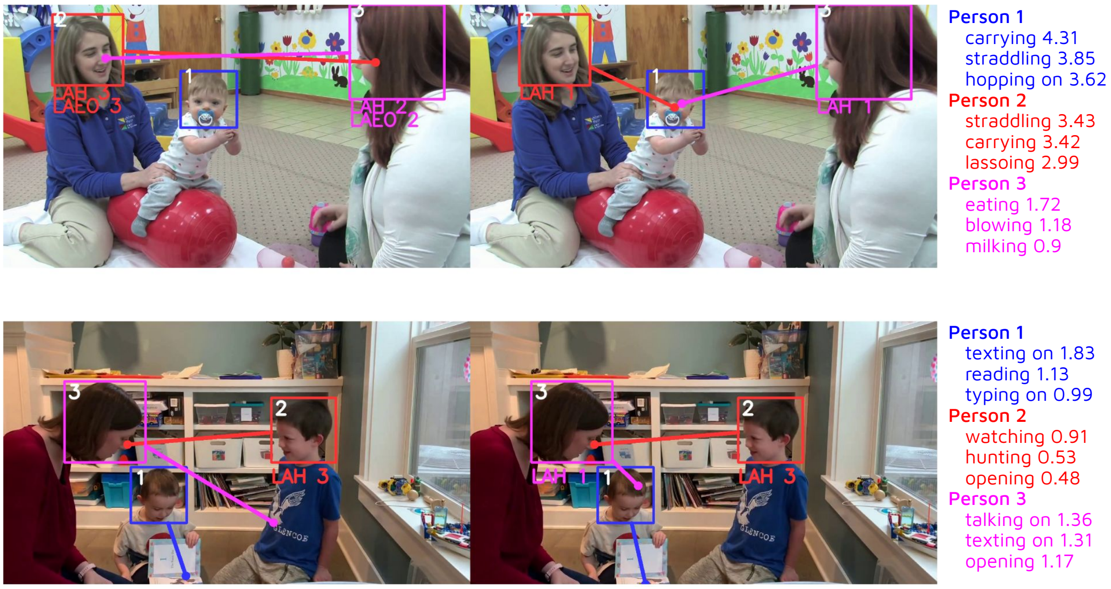


Figure 7.7: Qualitative results of MTGS (Gupta et al., 2024b) trained on GazeFollow and evaluated on ChildPlay. For each person, we display the predicted gaze point as well as social gaze task along with the associated person id. We provide results without contextual cues (left) and with contextual cues from the HICO classes (right). We also display the top three classes with the highest normalized score for each person.

7.3 Discussion

Our observations in Section 7.2.3 suggest that incorporating a larger set of contextual cues can improve generalization performance for gaze following. As the set of cues becomes larger, it can capture more specific situations (ex. unlocking, sewing in SWIG) which are usually associated with certain gaze targets. It is worth noting that increasing the number of classes has a negligible impact on computation time. As mentioned in Section 7.1.1, the ITM approach processes the text prompts and images independently to obtain text and image embeddings. The final score is then a dot product of the two. Hence, all the text embeddings can be computed and saved at the start, and then used with any new image.

We also note that the set of HICO and SWIG classes utilized in our study are obtained from HOI datasets, hence, scores for the different cues could alternatively be obtained from HOI models. This is another interesting direction of investigation, but its main drawback is that the set of cues that can be considered is fixed depending on the chosen model. On the other hand, leveraging VLMs in a zero-shot manner allows us to consider any set of cues, including larger sets than the ones we considered (with a negligible impact on computation time), or more domain specific cues tailored for specific applications.

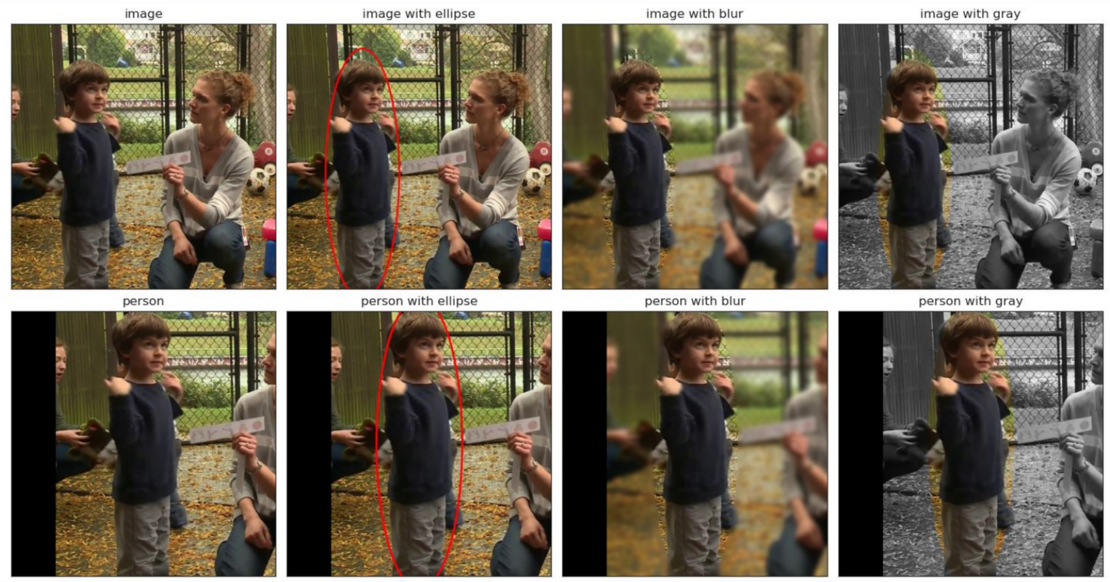


Figure 7.8: Different visual prompts are used to focus on the person of interest. Row-wise, the image-based and person cropped-based variants are displayed. Column-wise, various visual prompts such as ellipse, blur, and gray are presented.

7.4 Conclusion

In this work, we explored the zero-shot capabilities of VLMs for extracting contextual cues related to a person's pose or interactions with objects and other people, and evaluated the impact of incorporating these cues into a gaze following model. We learned that VLMs can indeed extract contextual cues, and that considering the entire image with a red-circle drawn over the person of interest serves as the best visual prompt, and that ensembling scores from different textual prompts serves as the best text prompting strategy. We also observed that BLIP-2 is the overall best performing VLM, and that ICL can potentially bring further benefits. In the second part, we observed that incorporating the extracted cues into a gaze following model can provide better generalization performance, especially when considering a larger set of classes. In future work, we plan to investigate other VLMs and further explore prompting strategies such as ICL. We also plan to explore the option of predicting the different cues rather than providing them as input to the model.

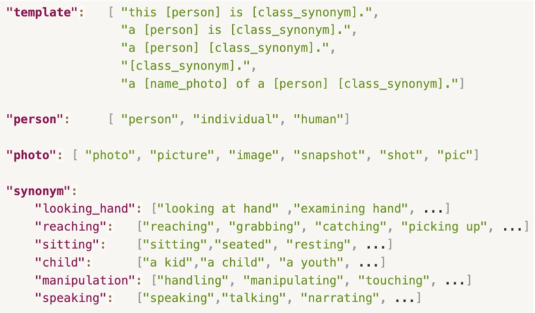
7.5 Appendix

Example of visual prompts

As described in Section 7.1.1, we investigate different visual prompting approaches to focus on a specific individual in the scene. An example of each prompt is provided in Fig. 7.8. These techniques are implemented on either the whole image or specifically on the cropped image of the target person. In total, this leads to eight distinct visual prompting strategies.

Classes	negative	positive
looking at hand	36	35
reaching	36	34
sitting	60	52
child	59	58
manipulation	59	59
speaking	31	30

Table 7.5: Classes and statistics of the ChildPlay dataset annotation.



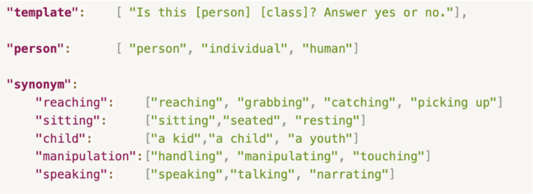
```
"template":   [ "this [person] is [class_synonym].",
                "a [person] is [class_synonym].",
                "a [person] [class_synonym].",
                "[class_synonym].",
                "a [name_photo] of a [person] [class_synonym]."]

"person":     [ "person", "individual", "human"]

"photo": [ "photo", "picture", "image", "snapshot", "shot", "pic"]

"synonym":
    "looking_hand": ["looking at hand" ,"examining hand", ...]
    "reaching":   ["reaching", "grabbing", "catching", "picking up", ...]
    "sitting":    ["sitting","seated", "resting", ...]
    "child":      ["a kid","a child", "a youth", ...]
    "manipulation": ["handling", "manipulating", "touching", ...]
    "speaking":   ["speaking","talking", "narrating", ...]
```

Figure 7.9: List of the different prompts variations used as described in section 7.1.3. A final prompt is a combination of {template},{person}, {photo} and {synonym} such as *"this individual is grabbing"* or *"a snapshot of a human handling"*.



```
"template":    [ "Is this [person] [class]? Answer yes or no."],

"person":      [ "person", "individual", "human"]

"synonym":
    "reaching":    ["reaching", "grabbing", "catching", "picking up"]
    "sitting":     ["sitting","seated", "resting"]
    "child":       ["a kid","a child", "a youth"]
    "manipulation":["handling", "manipulating", "touching"]
    "speaking":    ["speaking","talking", "narrating"]
```

Figure 7.10: List of the different prompt variations used for VQA model. A final prompt is a combination of {template},{person}, and {synonym} such as *"Is this individual grabbing? Answer yes or no."* .

Details of the Childplay dataset

In Table 7.5, we detail the number of annotated negative and positive samples for each class in the ChildPlay dataset.

Details of Text Prompts

ITM. Fig. 7.9, lists different text prompt variations as described in Section 7.1.3 for the ITM approach. A final prompt is a combination of {template},{person}, {photo} and {synonym} such as *"this individual is grabbing"* or *"a snapshot of a human handling"*.

VQA. For the VQA approach, for computational reasons, we consider a single template in the form of a question, and reduce the number of synonyms for the classes. We provide the template and synonyms in Fig. 7.10.

Impact of class synonyms

In Fig. 7.11, we provide the results for varying the class synonym in the text prompt. We observe that performance can change depending on the used synonym by a large margin.

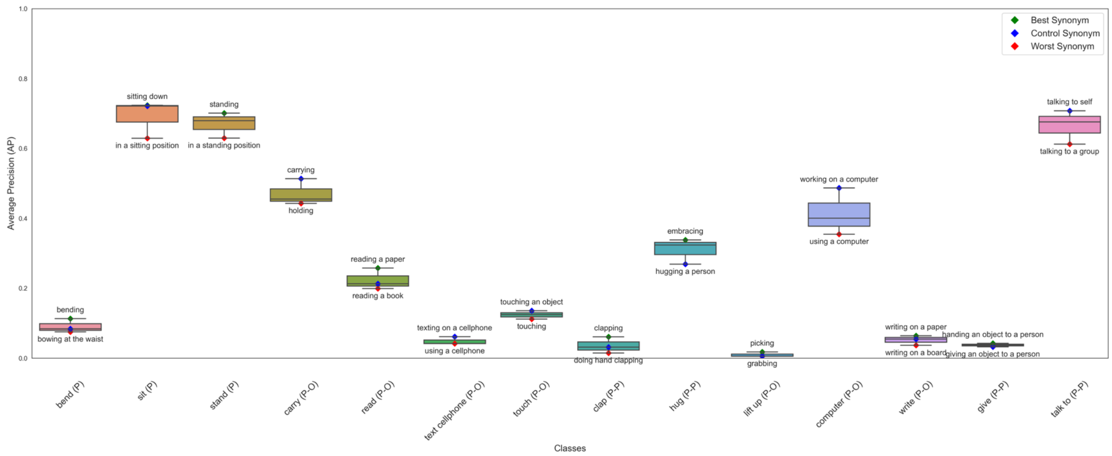


Figure 7.11: Performance when varying the class synonym in the text prompt. We display the mean and variance of results, as well as the best and worst synonym.

8 Ego-Exo Gaze Representation Learning

As discussed in Section 1, egocentric gaze estimation is valuable for enabling immersive and intuitive interaction in AR/VR devices. However, eye-tracking brings limitations such as additional hardware and calibration requirements. Hence a separate line of work aims to directly predict the egocentric gaze target using scene saliency and other contextual cues in the wearers visual environment (Lai et al., 2023; Tavakoli et al., 2019; Huang et al., 2018). These approaches predominantly focus on video-based prediction, leveraging temporal cues to improve accuracy. Yet, the hardware limitations of current platforms often make the storage and processing of a frame buffer impractical.

In this work, we focus on the single-frame egocentric gaze estimation task. We aim to leverage exocentric gaze (also known as gaze following) which involves predicting the gaze targets of other individuals in the scene from a third-person perspective. This additional context can help disambiguate between multiple potential gaze targets. For instance, in Figure 8.1, predicting egocentric gaze is ambiguous due to the presence of multiple individuals in the field of view. However, by incorporating exocentric gaze cues, we can better understand social interactions—such as shared attention towards a person—allowing us to significantly reduce this ambiguity.

Although existing models for egocentric gaze estimation may implicitly learn representations related to gaze following, the complexity of the task and the limitations of current datasets may restrict the extent to which these representations are effectively captured. Indeed, prior research has demonstrated the benefits of explicitly integrating additional contextual information such as hand actions (Huang et al., 2020) for egocentric gaze estimation. Guided by these insights, we explore whether explicitly learning exocentric gaze cues can enhance single-frame egocentric gaze estimation.

Contribution. Given these motivations, we propose novel approaches that leverage simultaneous views from a pair of individuals to jointly learn ego-exo gaze representations. Egocentric representations are learned via supervised training, while exocentric representations are learned through a self-supervised alignment task. Specifically, this task aims to match the ego

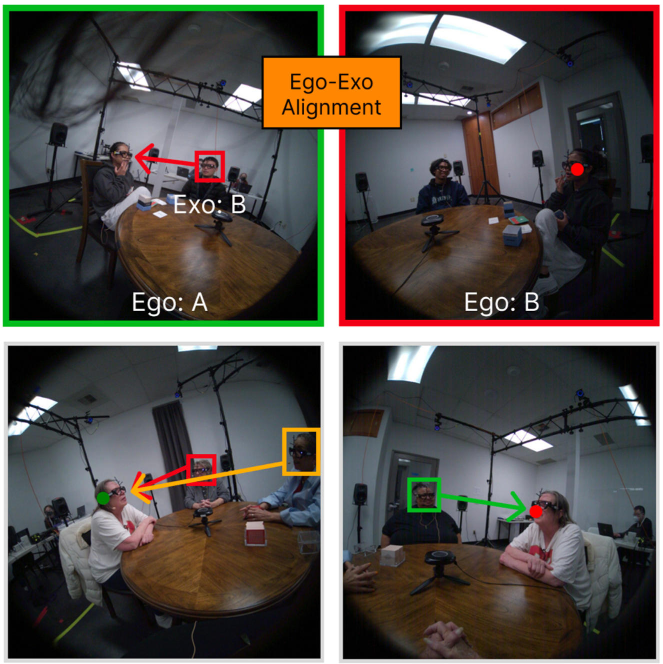


Figure 8.1: Estimating the egocentric gaze target of a person from a single frame is challenging, but can be improved using exocentric gaze cues of other individuals in the scene. During training, we employ a siamese-style architecture: one branch captures a person's ego gaze features (top right), the other captures the *same person's* exo features (top left), which are then aligned. Through symmetric ego-exo alignment and shared weights, the encoder learns to exploit exocentric gaze information from *other individuals* to improve egocentric gaze estimation from a single view (bottom). Images are from the RLR-CHAT dataset.

representation of a person with the exo representation of the *same person* as captured from another person's view (Figure 8.1, top).

We adopt a siamese style architecture, where one branch captures egocentric gaze information, while the other branch captures exocentric gaze information. Through symmetric ego-exo alignment and weight sharing, the same encoder learns to capture not only the egocentric gaze features of an individual, but also the exocentric gaze features of *other individuals* in the scene. As a result, during inference, we can use a single branch to improve egocentric gaze estimation from a single view by leveraging the learned exocentric representations (Figure 8.1, bottom). Our contributions can be summarized as follows:

- *Exploring single-frame egocentric gaze estimation:* We show that single-frame methods can achieve strong performance by leveraging modern CNN and transformer architectures.
- *Learning ego-exo gaze representations:* We propose three approaches that leverage ego-exo alignment to jointly learn egocentric and exocentric gaze representations, using self-supervision for the exocentric features. Our results show that these models improve egocentric gaze estimation, likely by utilizing the learned exocentric representations.

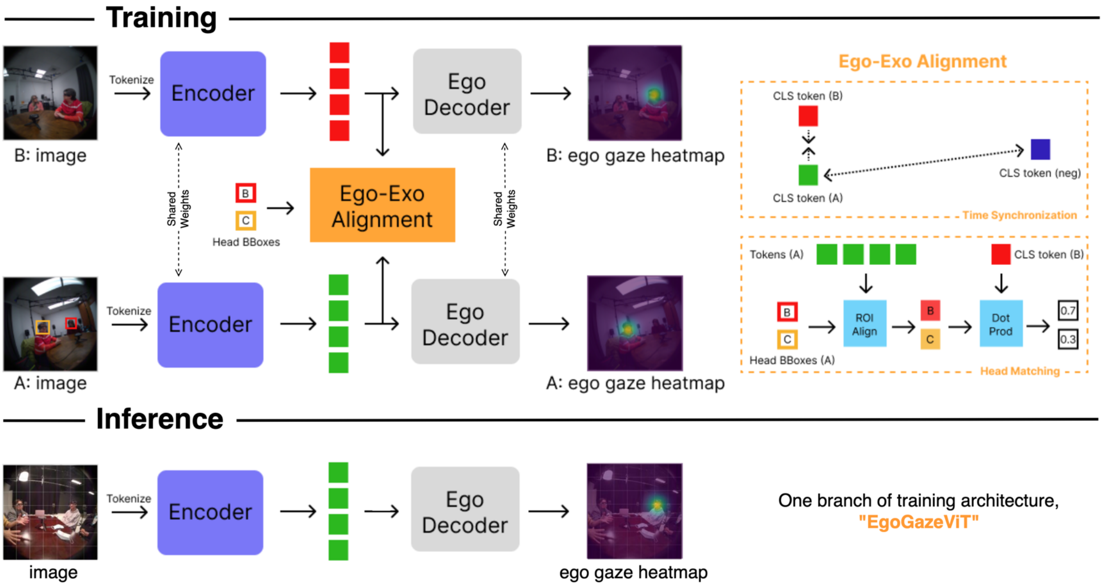


Figure 8.2: During training, we sample two views corresponding to individuals A and B. The Encoder extracts features, which are aligned using one of two Ego-Exo Alignment techniques: (1) time synchronization or (2) head matching. The Ego Decoder then uses these aligned features to predict each person's egocentric gaze heatmap. At inference time, only a single branch of the network—termed EgoGazeViT—is used to predict the egocentric gaze heatmap for any input image.

- *Probing for exocentric gaze:* We further probe the models to assess their ability to capture exocentric gaze representations, confirming that they indeed learn meaningful exocentric gaze features.
- *New metrics for egocentric gaze estimation:* We propose a novel suite of metrics inspired from gaze following literature, to enable more comprehensive performance analysis.

In addition, we perform an initial exploration of spatial audio for improving egocentric gaze performance, with promising preliminary results.

8.1 Method

Our training architecture, illustrated in Figure 8.2, follows a siamese design. The bottom branch predicts person A's egocentric gaze, while the top branch predicts person B's. Each branch processes an egocentric image frame, denoted as $\mathbf{I}^A$ and $\mathbf{I}^B$, extracting features $\mathbf{F}^A$ and $\mathbf{F}^B$ using an encoder $\mathcal{V}$. Simultaneously, we align the ego-exo gaze features of the same individual across views. We explore two approaches: (1) *Time Synchronization*, which encodes exocentric features within a single global representation, and (2) *Head Matching*, which encodes exocentric features in local representations extracted via head bounding boxes. Finally, the aligned features are passed through a decoder D_{ego} to generate each person's egocentric gaze heatmap, $\mathbf{H}^A$ and $\mathbf{H}^B$.

Due to weight sharing and symmetric ego-exo alignment, we can use a single branch of the network—termed EgoGazeViT—at inference. This branch leverages the learned exocentric representations of other individuals in the scene to enhance egocentric gaze estimation, with no added cost compared to other methods.

8.1.1 Feature Extraction

The feature extraction module is responsible for obtaining gaze-relevant features from an input egocentric image frame, denoted as **I**. We employ a Vision Transformer (ViT) encoder, denoted as $\mathcal{V}$, to extract these features. Specifically, we utilize the output from the last layer of the ViT.

$$\mathbf{F} = \mathcal{V}(\mathbf{I}) \tag{8.1}$$

This module is applied independently to the egocentric images of both individuals, $\mathbf{I}^A$ and $\mathbf{I}^B$, yielding corresponding feature representations, $\mathbf{F}^A$ and $\mathbf{F}^B$.

8.1.2 Ego-Exo Alignment

The ego-exo alignment module enables self-supervised learning of exocentric gaze representations by aligning egocentric and exocentric features. The key idea is that a person's egocentric gaze features already encode valuable information about where they are looking, which can be leveraged to supervise learning of their corresponding exocentric representations captured from another person's viewpoint.

Through this alignment, we expect the ego and exo features corresponding to an individual to capture complementary information. For instance, in cases of eye contact, the exocentric features from A's FoV can capture B's head orientation and gaze direction, while B's egocentric features can encode complementary information regarding B's own head pose through cues like body orientation. Or in shared attention scenarios, both egocentric and exocentric features can capture similar visual cues about the attended item.

We explore two alignment approaches:

Time Synchronization. Inspired by prior work (Yu et al., 2020), we align egocentric features of two participants observing the same scene simultaneously. Egocentric features for each person (e.g., person A) are obtained from the CLS token of the ViT output:

$$\mathbf{G}^A_{ego} = \mathrm{CLS}(\mathbf{F}^A) \tag{8.2}$$

Here, the exocentric feature for person A at a given timestamp is directly represented by the egocentric feature from person B observing the same scene simultaneously:

$$\mathbf{G}^A_{exo} = \mathbf{G}^B_{ego} \tag{8.3}$$

We compute similarity between these ego-exo features using the L2 distance:

$$S = \|\mathbf{G}^A_{ego} - \mathbf{G}^A_{exo}\|_2 \tag{8.4}$$

Our triplet loss (Section 8.1.4) encourages high similarity between egocentric features from the same timestamp. Simultaneously, it minimizes similarity against negative samples drawn from the same batch. These negatives may include features from different timestamps of the same session or features from entirely different sessions. Due to symmetric alignment, after training, we expect the same CLS token to capture *both* ego and exo gaze information.

Head Matching. We explore a novel approach that explicitly aligns an individual's global egocentric features with their local exocentric features as observed from another person's view. These local exocentric features can then be aggregated by the decoder to predict egocentric gaze. Specifically, given a participant B, we first extract exocentric features for all people visible in B's FoV (including A) using ROI-Align (He et al., 2017):

$$\mathbf{G}^B_{exo} = \text{ROI Align}(\mathbf{F}^B, \mathbf{B}^B) \tag{8.5}$$

where $\mathbf{B}^B$ represents the head bounding boxes in B's FoV. We again obtain egocentric features from the CLS token:

$$\mathbf{G}^A_{ego} = \text{CLS}(\mathbf{F}^A) \tag{8.6}$$

Similarity between the normalized egocentric and exocentric features is computed using a dot product:

$$\mathbf{S}^A = \mathbf{G}^B_{exo} \cdot \mathbf{G}^A_{ego} \tag{8.7}$$

Our loss function (Section 8.1.4) maximizes the similarity of matched ego-exo pairs ($\mathbf{S}^A(A)$) and minimizes it for unmatched pairs. This alignment is again performed symmetrically, so the same feature map learns to encode egocentric (in the CLS token) and exocentric (in tokens corresponding to head box regions) gaze information.

8.1.3 Prediction

The Prediction Module processes the features from the Feature Extraction module, integrating both egocentric and exocentric information to generate an egocentric gaze heatmap for each person. It consists of four transformer layers, followed by a linear projection layer that maps the processed token representations to the spatial dimensions of the gaze heatmap. Specifically, given the extracted feature representation **F** from the Feature Extraction module, the egocentric gaze heatmap $\mathbf{H}$ is predicted as follows:

$$\mathbf{H} = D_{\text{ego}}(\mathbf{F}) \tag{8.8}$$

where D_{ego} represents the transformer-based decoder.

This operation is applied independently to the feature representations of both person A and person B, producing their respective egocentric gaze heatmaps $\mathbf{H}^A, \mathbf{H}^B$.

8.1.4 Losses

Our training objective combines an egocentric gaze estimation loss ($\mathscr{L}_{hm}$) and an ego-exo alignment loss ($\mathscr{L}_{\text{ego-exo}}$). The egocentric gaze estimation loss is a pixel-wise cross-entropy applied independently to the predicted heatmaps $\mathbf{H}^A$ and $\mathbf{H}^B$, comparing them to the corresponding ground truth heatmaps.

The total loss is given by:

$$\mathscr{L} = \mathscr{L}_{hm}^A + \mathscr{L}_{hm}^B + \mathscr{L}_{\text{ego-exo}} \tag{8.9}$$

We explore two formulations for the ego-exo alignment loss ($\mathscr{L}_{\text{ego-exo}}$):

Time Synchronization Loss. This loss uses a triplet formulation based on the similarity between egocentric features at the same timestamp:

$$\mathscr{L}_{\text{ego-exo}} = \frac{e^{S^+}}{e^{S^+} + e^{S^-}} \tag{8.10}$$

where S^+ is the distance between matched ego features from simultaneous views, and S^- is the distance from $\mathbf{G}_{ego}^A$ to a randomly sampled negative ego feature from the batch.

Head Matching Loss. We explore two variants for the head matching loss: an *implicit* approach that automatically learns the alignment between egocentric and exocentric features, and an *explicit* approach leveraging ground truth head-box identities when available.

- **Explicit matching:** This is also a cross-entropy loss, applied to the similarity scores $\mathbf{S}$. The correct "class" corresponds to the similarity score of the same person viewed from the other perspective. For $\mathbf{S}^A$, the correct class is $\mathbf{S}^A(A)$, and vice versa for $\mathbf{S}^B$.
- **Implicit matching:** We apply an entropy loss on the similarity scores $\mathbf{S}$. This encourages the model to select exactly one exocentric feature with maximum similarity. Additionally, for two-person sessions, where head-box identities are trivially identifiable, we apply the explicit cross-entropy loss described above.

In both variants, the total ego-exo loss $\mathscr{L}_{\text{ego-exo}}$ is the sum of the losses computed independently for persons A and B.

8.2 Experiments

8.2.1 Datasets

We perform experiments on two datasets:

RLR-CHAT. The RLR-CHAT dataset, as described in Section 3.5.1, is divided into training, validation, and test splits. In particular, we refer to the test split as the "golden subset" as it has higher quality annotations. Initially, we train and evaluate various baselines on this golden subset, as training on the entire dataset is computationally expensive. Subsequently, we leverage the best performing approach for training on the full dataset.

Ego4D (Grauman et al., 2022). Specifically, the gaze annotated subset of Ego4D as described in Section 3.5.2. We selected Ego4D over alternative datasets such as EGTEA Gaze (Li et al., 2018) and Aria (Lv et al., 2024), as those primarily focus on single-person activities. Since Ego4D emphasizes social settings, it is more suitable for evaluating improvements derived from learning exocentric gaze representations.

However, it is important to note that there is still a significant domain gap between Ego4D and RLR-CHAT. The images have a smaller FoV and include much more diverse settings. Further, as people are playing games instead of mainly conversing, the gaze points tend to fall less on faces.

Training and Evaluation

During training, we randomly sample two people A and B for each timestamp. Models are trained for 20 epochs using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 2×10^{-5}, and with a batch size of 512. We employ standard augmentations, namely center cropping, flipping, and color jittering. The method was trained on a distributed system with two nodes, each equipped with eight H100 GPUs.

During evaluation, we leverage only one of the branches (since both share weights), referred to as EgoGazeViT, to assess egocentric gaze estimation performance. Specifically, EgoGazeViT can be initialized with weights from either the self-supervised training or standard egocentric gaze estimation training.

8.2.2 Trained Models

We train several egocentric gaze estimation baselines along with our proposed self-supervised methods using the RLR-CHAT dataset.

Egocentric baselines. We compare naïve baselines, as well as CNN and transformer-based approaches:

- **Naïve Baselines:** Simple heuristics, including predicting the image center, using the average gaze point from the training set, and selecting the head closest to the image center as the gaze target.
- **U-Net:** A CNN-based model with a ResNet-18 encoder and an FPN-style decoder. It corresponds to the image branch of the MAV-Gaze baseline described below and operates on single-frame inputs.
- **MAV-Gaze:** An adaptation of MAV-ASL (Jiang et al., 2022), originally designed for active speaker localization. It processes both visual and auditory cues, taking in a single image frame and a 7-channel, 200ms audio clip.
- **EgoGazeViT:** A transformer-based model comprising a ViT encoder and a transformer decoder. This architecture corresponds to one of the branches in our proposed SSL methods and operates on single-frames.

Self-supervised Approaches. We initialize EgoGazeVit with one of our three self-supervised alignment methods described in Section 8.1, or with standard training for egocentric gaze prediction (**Standard Training**):

- **Synchronization:** Aligns egocentric features across simultaneous views using temporal correspondence.
- **Implicit Matching:** Aligns ego-exo features without explicit identity annotations.
- **Explicit Matching:** Aligns ego-exo features using ground truth head-box identities.

8.2.3 Metrics

As discussed in Section 4.3.1, existing heatmap-based metrics are highly sensitive to the choice of standard deviation used to generate the ground truth heatmap, as well as the thresholds applied during binarization. Therefore, we include this metric only for comparison with prior state-of-the-art methods.

Instead, we primarily rely on our proposed gaze-following-inspired metrics introduced in Section 4.3.2. Specifically, we report both distance scores, to assess localization accuracy, and Looking at Heads (LAH) scores, to evaluate semantic performance.

8.3 Results

8.3.1 Egocentric Gaze Estimation Baselines

We trained and evaluated the baselines listed in Section 8.2.2 on the RLR-CHAT golden subset. Specifically, each model is trained on the train split of the golden subset for egocentric gaze estimation and evaluated then evaluated on it's test set. The results are presented in Table 8.1.

Despite operating on a single image frame, both transformer and CNN-based models outperform the naive baselines by a significant margin, highlighting their ability to incorporate hu-

Model	Distance↓		LAH↑		
	Mean	Median	Prec	Recall	F1
Naive Baselines					
Predict center	0.107	0.093	0.633	0.146	0.237
Predict avg of train data	0.105	0.092	**0.638**	0.130	0.216
Predict closest head to center	0.131	0.073	0.396	**0.863**	0.543
CNN Baselines					
U-Net	0.105	0.072	0.520	0.610	0.561
MAV-Gaze	0.098	0.065	0.617	0.724	**0.667**
Transformer Baseline					
EgoGazeViT	**0.096**	**0.057**	0.507	0.798	0.620

Table 8.1: Comparison of egocentric gaze estimation baselines on the RLR-CHAT golden subset test split. Best results are in bold.

Initialization	Distance↓		LAH↑		
	Mean	Median	Prec	Recall	F1
Standard training	0.102	0.057	0.538	0.819	0.650
SSL Approaches					
Synchronization	**0.100**	**0.055**	0.536	**0.843**	0.656
Implicit matching	0.101	0.056	0.533	0.833	0.650
Explicit matching	0.101	**0.055**	**0.545**	0.836	**0.660**

Table 8.2: Results for egocentric gaze estimation on the full RLR-CHAT golden subset. We leverage EgoGazeViT with different initializations. Best results are in bold.

man priors and scene saliency for accurate egocentric gaze estimation. EgoGazeViT achieves the highest performance among the image-only models, and the best overall distance score. Therefore, we select this model for all subsequent experiments.

Interestingly, MAV-Gaze achieves the highest LAH F1-score. The incorporation of spatial audio helps the model identify the speaking person, which can serve as a strong cue for gaze target prediction. Exploring the role of spatial audio in egocentric gaze estimation remains an exciting avenue for future work.

8.3.2 Learning Exo Gaze Representations

We leverage the entire RLR-CHAT dataset by training our models on the designated train split and evaluating them on the full test split (the golden subset). Specifically, we compare the performance of EgoGazeViT when initialized with weights from standard egocentric gaze estimation training versus weights obtained via our proposed ego-exo alignment methods. Results, as shown in Table 8.2, indicate that performance on the Distance metric for all methods is comparable. However, the Synchronization and Explicit Matching methods yield some improvements over Standard Training for the LAH metric, with Explicit Matching having the best performance.

As seen in Table 8.3, the Explicit Matching approach consistently improves over Standard

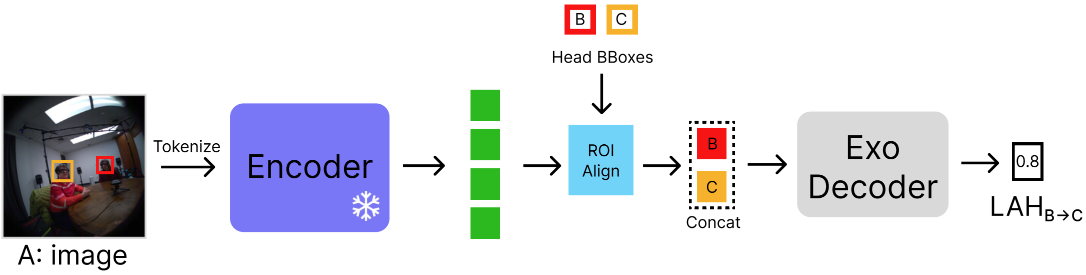


Figure 8.3: Architecture for probing learned exocentric gaze representations. We freeze the encoder, which was initially trained for egocentric gaze estimation, and train a 2-layer MLP probe to predict Looking at Heads (LAH).

Initialization	**Number of People**		
	Full	**≥ 3**	**≥ 4**
Standard Training	0.650	0.630	0.576
SSL Approaches			
Synchronization	0.656	0.634	0.578
Implicit matching	0.650	0.626	0.553
Explicit matching	**0.660**	**0.640**	**0.587**

Table 8.3: LAH F1-scores for different splits of the RLR-CHAT golden subset based on the number of participants in the sessions. Best results are in bold.

Initialization	**LAH AP↑**
Random init	0.178
Standard training	0.262
SSL Approaches	
Synchronization	**0.498**
Implicit Matching	0.371
Explicit Matching	0.304

Table 8.4: Results for exocentric gaze probing on the full RLR-CHAT golden subset. Best results are in bold.

Training across evaluations on different splits of the RLR-CHAT golden subset based on the number of participants in the sessions (full results in supplementary). This suggests that our method captures exocentric gaze behaviors beyond shared attention. Shared attention in RLR-CHAT is predominantly observed when gaze is directed towards other people—a scenario naturally limited in sessions with only two participants, which constitute the majority of the dataset. However, the performance gap implies that the model also learns other behaviors, such as gaze aversion, that help disambiguate the egocentric gaze target.

Probing for Exocentric Gaze. To assess whether training with our proposed method effectively enables the learning of exocentric gaze representations, we probe the trained encoder by evaluating its performance on exocentric gaze prediction. The probing architecture is illustrated in Figure 8.3. Specifically, we freeze the trained encoder and train a new exocentric decoder D_{exo} to predict LAH labels for RLR-CHAT. The decoder is a 2 layer MLP that operates on ROI-aligned features corresponding to individuals visible within the egocentric field of view ($\mathbf{G}_{exo}$). Specifically, it processes their concatenated features and predicts pairwise LAH following the formulation of (Gupta et al., 2024a,b). For instance, to predict whether person B is looking at person C within person A's FoV, the model proceeds as follows:

$$\text{LAH}_{B\to C} = D_{\text{exo}}(\mathbf{G}^{B}_{exo}, \mathbf{G}^{C}_{exo}) \tag{8.11}$$

Initialization/Model	Distance↓		Heatmap↑		
	Mean	Median	Prec	Recall	F1
Cross-Dataset Evaluation					
Standard training	0.174	0.151	26.0	52.0	34.7
Synchronization	**0.169**	**0.144**	**28.6**	50.6	**36.5**
Implicit Matching	0.183	0.154	25.9	**53.7**	34.9
Explicit Matching	0.170	0.147	27.4	49.3	35.2
Within-Dataset Evaluation					
GBVS (Harel et al., 2006)	-	-	11.1	47.2	18.0
Attention Transition (Huang et al., 2018)	-	-	29.5	47.6	36.4
I3D-R50 (Feichtenhofer et al., 2019)	-	-	29.2	52.5	37.5
MViT (Lai et al., 2023)	-	-	31.7	**57.4**	40.9
GLC (Lai et al., 2023)	**0.156**	**0.123**	**34.7**	57.0	**43.1**
EgoGazeViT	0.163	0.131	31.5	56.2	40.4

Table 8.5: Results for egocentric gaze estimation on the Ego4D dataset. Best results are in bold.

Note that the order of individuals supplied to the decoder is crucial, as the LAH prediction is directional.

We present the results in Table 8.4. Unlike egocentric gaze prediction, where discrete LAH labels enable direct precision and recall calculations, the predicted LAH values in this setting are continuous. While applying a threshold can yield discrete values, the precision and recall scores can vary significantly depending on that choice. Therefore, we report the average precision (AP) score, which provides a threshold-independent evaluation.

We find that all self-supervised approaches outperform the baseline, indicating that they successfully capture exocentric gaze information. Notably, the Synchronization approach significantly improves over the other self-supervised approaches. This may be related to the more general nature of the alignment task, which allows the exocentric features to encode global social gaze cues because it does not rely on head crops. The Implicit Matching approach also surpasses Explicit Matching, however, this may be a result of overfitting to scene geometry when learning head-identity correspondences. This interpretation is supported by its lower cross-dataset performance as discussed in the next section.

8.3.3 Evaluation on Ego4D

We provide cross-dataset evaluation results for our RLR-CHAT trained models on Ego4D in Table 8.5. Overall all methods have a marked drop in distance score, highlighting the domain gap between the two datasets.

Despite this gap, all self-supervised approaches except Implicit Matching improve over Standard Training in cross-dataset generalization, illustrating another benefit of ego-exo alignment. The Synchronization approach has the best overall performance, following results from exocentric probing. Interestingly, this trend is not followed for Implicit Matching, which suggests that the method may be overfitting to scene geometry in order to learn head-identity corre-

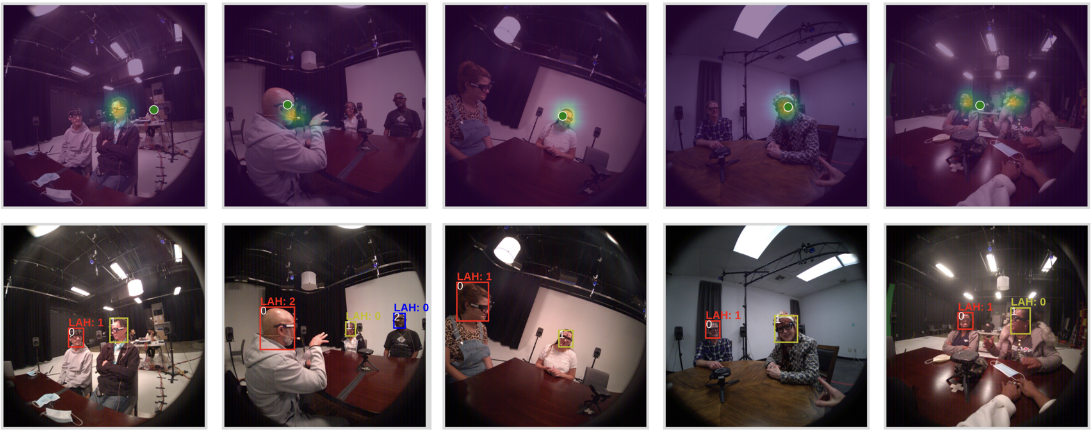


Figure 8.4: Qualitative results on RLR-CHAT for egocentric (top) and exocentric (bottom) gaze prediction using the encoder initialized with our Explicit Matching based self-supervised approach. The predicted egocentric gaze heatmap is overlaid on the image, with the ground truth target marked by a green dot. Predicted exocentric gaze targets are indicated by the person ID following the ’LAH’ prefix.

spondences.

For comparison with state-of-the-art methods, we additionally train our best model—EgoGazeViT with Explicit Matching—on Ego4D. Despite relying solely on single frames, it attains strong performance and even surpasses some temporal models, highlighting the potential of single-frame approaches in this domain. We observe that different initializations of EgoGazeViT do not yield significant performance variations on Ego4D, likely due to the pronounced domain shift between datasets.

8.3.4 Qualitative Results

We present qualitative results for egocentric and exocentric gaze prediction in Figure 8.4, using the encoder initialized with our Explicit Matching-based self-supervised approach. For egocentric predictions, we directly overlay the predicted heatmap onto the image. For exocentric predictions of a given person B, the target is determined by the argmax over LAH pairs $(B, *)$, and is visualized if the corresponding value exceeds a threshold of 0.1.

We observe that the model accurately identifies the egocentric gaze target (columns 2-4). Generally, it tends to focus on salient items such as faces, whereas human gaze can sometimes be directed toward background individuals (column 1) or be in transition during a gaze shift (column 5), which is challenging for a static model to capture.

Additionally, we observe that the model effectively leverages exocentric cues to resolve ambiguities (columns 1–4). However, in scenarios where exocentric gaze information is less informative (column 5), the model exhibits greater uncertainty, as reflected in the multimodal heatmap. Additional qualitative results are provided in the supplementary material.

8.4 Conclusion

In this work, we introduced novel self-supervised learning approaches for egocentric gaze estimation that leverage ego-exo alignment to implicitly learn exocentric gaze representations. Our methods improve egocentric gaze prediction in challenging single-frame setting across RLR-CHAT and Ego4D by likely leveraging the learned exocentric gaze representations. Furthermore, our probing analysis confirms that training with our method enhances the encoder's ability to learn these exocentric gaze representations.

Future research could explore integrating spatial audio cues to further refine gaze estimation, particularly in social settings where auditory information plays a key role in attention and interaction. Additionally, investigating the generalizability of these self-supervised techniques in temporal settings could be another interesting direction of research.

8.5 Appendix

Analysis

Implicit Matching vs. Explicit Matching. In Figure 8.5, we examine how well the Implicit Matching method learns to align egocentric and exocentric gaze features. The model successfully matches features in several cases (top row), demonstrating its ability to capture meaningful ego-exo correspondences. However, it also exhibits failure cases (bottom row), where mismatches occur. In some cases, such as the bottom right example, failure is expected because the corresponding exocentric person is outside the field of view. However, the model also fails in other scenarios (e.g., bottom left), indicating that implicit matching alone may not always be sufficient for robust alignment.

Impact of number of people. Table 8.6 provides a detailed breakdown of egocentric gaze estimation performance on different splits of the RLR-CHAT Golden Subset, based on the number of participants in the included sessions. As expected, performance generally declines in sessions with a higher number of people due to the increased number of potential gaze targets and the resulting complexity of the task. Notably, there is an apparent spike in performance for sessions with 5 participants; however, since this split comprises only 2 sessions, the result is likely subject to high variance and may not be representative.

Implementation Details

The model uses a ViT-B encoder initialized with masked autoencoder (MAE) pretraining (He et al., 2022). The Prediction Module consists of four transformer layers, each with a token dimension of 384—half the dimension of the ViT tokens. Input images are processed at a

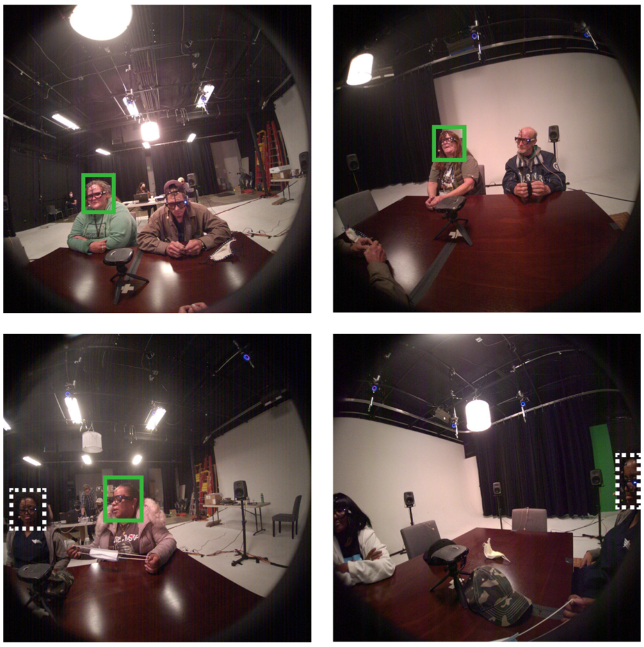


Figure 8.5: Egocentric and exocentric feature alignment results for the Implicit Matching approach. The correct exocentric person is highlighted with a green box. In the top row, these are correctly selected by the model. Incorrect selections made by the model are indicated with dotted white boxes in the bottom row.

resolution of 224 × 224, and the predicted gaze heatmap is generated at the same resolution, following the MAE architecture. The ground truth gaze heatmap $\mathbf{H}_{gt}$ is constructed by placing a Gaussian centered at the gaze point, with a standard deviation of 9.35 pixels. It is converted to two channels $(\mathbf{H}_{gt}, 1 - \mathbf{H}_{gt})$ to facilitate training using the cross-entropy loss detailed in Section 4.4.

Qualitative Comparisons

We provide qualitative comparisons of our models for egocentric gaze prediction in Figures 8.6 and 8.7. In Figure 8.6, all models perform similarly, as these cases involve either a single salient target or a target positioned near the image center, reducing task complexity.

In contrast, Figure 8.7 highlights scenarios where our ego-exo alignment approaches improve egocentric gaze prediction. These cases are more ambiguous, featuring multiple salient targets near the center, making gaze estimation more challenging.

- *Row 1:* The Standard Training and Implicit Matching approaches incorrectly identify the target person.
- *Row 2:* The Synchronization and Implicit Matching approaches struggle to differentiate

Subset	**Initialization**	**Distance**		**LAH**		
		Mean	**Median**	**Precision**	**Recall**	**F1**
Full	Standard Training	0.102	0.057	0.538	0.819	0.650
	Synchronization	**0.100**	**0.055**	0.536	**0.843**	0.656
	Implicit Matching	0.101	0.056	0.533	0.833	0.650
	Explicit Matching	0.101	**0.055**	**0.545**	0.836	**0.660**
≥3 people	Standard Training	0.111	0.067	0.524	0.790	0.630
	Synchronization	**0.110**	**0.064**	0.519	**0.815**	0.634
	Implicit Matching	0.111	0.065	0.512	0.805	0.626
	Explicit Matching	**0.110**	0.065	**0.532**	0.803	**0.640**
≥4 people	Standard Training	0.110	0.074	0.466	0.754	0.576
	Synchronization	0.107	**0.069**	0.461	0.771	0.578
	Implicit Matching	0.111	0.074	0.438	0.750	0.553
	Explicit Matching	**0.106**	**0.069**	**0.473**	**0.773**	**0.587**
≥5 people	Standard Training	0.096	0.054	0.542	0.820	0.653
	Synchronization	**0.090**	**0.049**	0.546	**0.849**	**0.664**
	Implicit Matching	0.101	0.058	0.524	0.796	0.632
	Explicit Matching	0.092	0.052	**0.554**	0.827	**0.664**

Table 8.6: Evaluation results on different splits of the RLR-CHAT Golden Subset based on the number of people in the session. Best results for each split are given in bold.

between two people. The Standard Training approach confidently selects the wrong target, whereas the Explicit Matching approach correctly identifies the target with high confidence.

- *Row 3:* The Standard Training and Synchronization approaches misidentify the target, while the Implicit Matching and Explicit Matching approaches show uncertainty between two possible targets.
- *Row 4:* The Standard Training approach produces a diffused heatmap due to confusion, whereas the ego-exo alignment approaches correctly select the target. The Explicit Matching approach still exhibits some uncertainty.

These results suggest that ego-exo alignment helps disambiguate complex scenarios by leveraging exocentric gaze cues, leading to more precise egocentric gaze predictions.

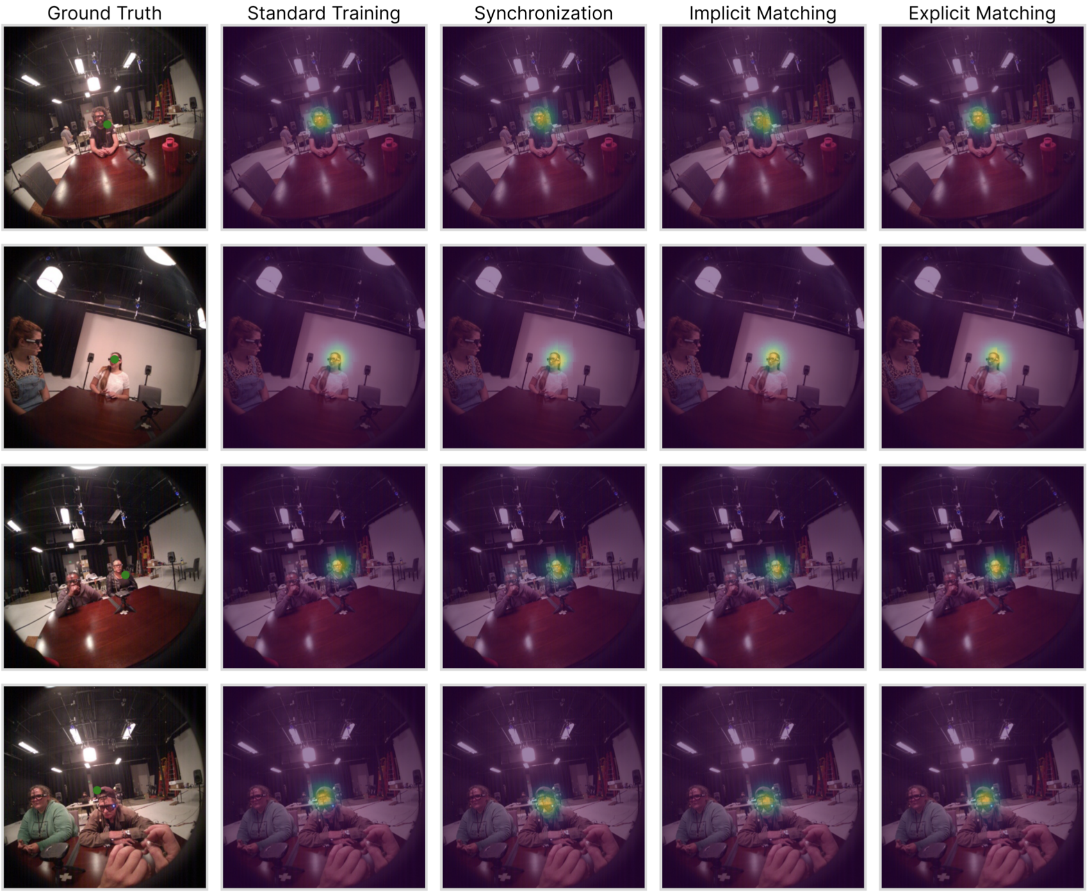


Figure 8.6: Qualitative results on images from RLR-CHAT where all approaches yield similar predictions. In these cases, either a single salient target dominates the scene, making gaze estimation straightforward, or the gaze target is near the image center, reducing ambiguity across models.

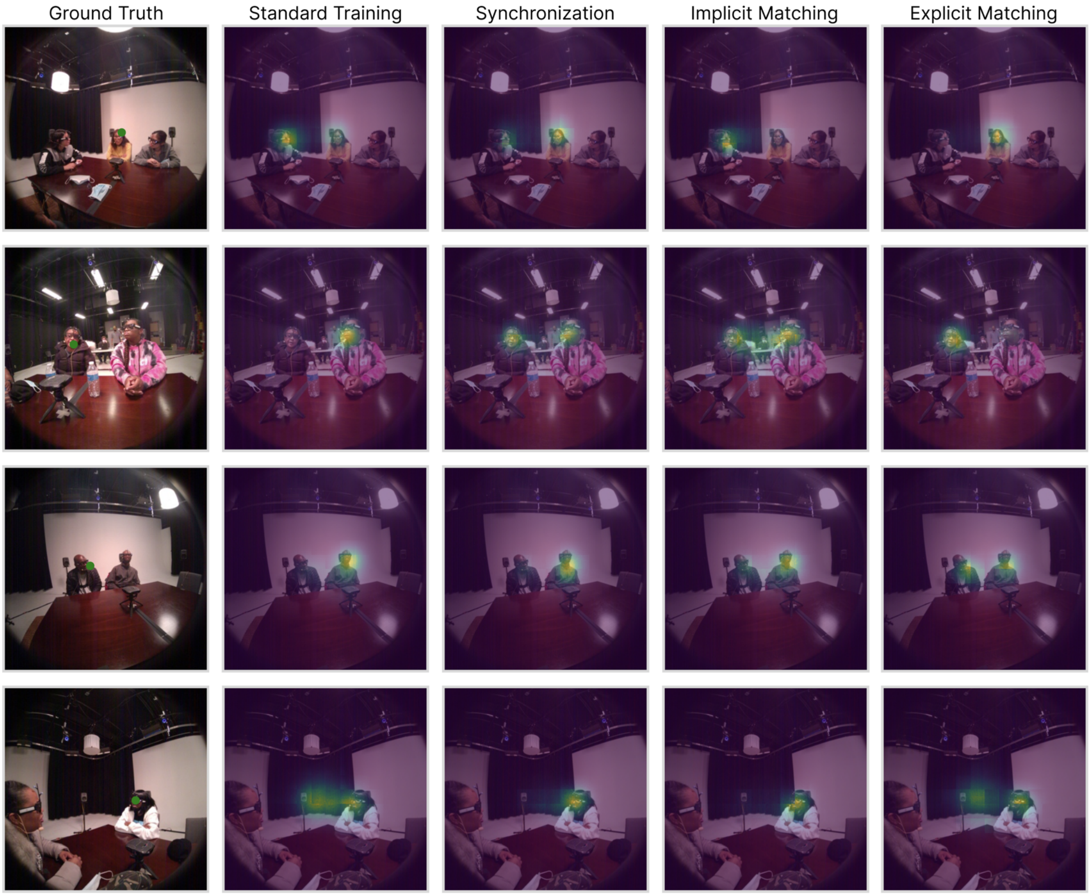


Figure 8.7: Qualitative results on images from RLR-CHAT where our proposed ego-exo alignment approaches improve egocentric gaze prediction compared to standard training. These cases involve greater ambiguity, with multiple salient targets positioned near the center, making gaze estimation more challenging.

9 Gaze Prediction for Automatically Detecting Components of Joint Attention

Early language acquisition is fundamentally social: children learn through interaction, and joint attention—defined as shared focus on an object or activity (Bakeman and Adamson, 1984)—is widely recognized as a foundational mechanism in this process (Baldwin, 1995). Emerging around nine months of age (Trevarthen and Hubley, 1978), joint attention has been repeatedly linked to vocabulary development and broader language skills (Tomasello and Farrar, 1986; Carpenter et al., 1998; Adamson et al., 2019). Its effectiveness, however, depends not only on its occurrence but on its composition—such as who is involved and what modalities are engaged, including gaze, speech, and gesture (Suarez-Rivera et al., 2019; Abney et al., 2020). Among these, gaze behavior plays a central role (Figure 9.1). Studies using eye-tracking have shown that fine-grained gaze behaviors—such as parent gaze shifts or infant sustained attention—can predict future vocabulary outcomes (Yu et al., 2019; Abney et al., 2020).

Despite these insights, as discussed in Section 2.7, much of the empirical research on joint attention has relied on constrained laboratory setups or wearable eye-trackers which may influence participants' natural behavior (Abney et al., 2020). Most studies also tend to feature mother-infant dyads in semi-structured tasks like free play (Akhtar et al., 1991; Suarez-Rivera et al., 2022), limiting generalizability to more diverse or ecologically valid contexts.

An alternative is the manual coding of joint attention related behaviours. However, labor-intensive coding limits dataset size and diversity, contributing to generalizability issues and the broader replication crisis in behavioral science (Shrout and Rodgers, 2018). Recent advances in computer vision and machine learning—particularly in gaze following architectures—offer a promising alternative. Several models, such as the ones proposed in this thesis, now reliably predict gaze targets in static images (Gupta et al., 2022; Chong et al., 2020b; Tafasca et al., 2024). There has now been growing interest in using these models for socially relevant tasks, including shared attention and gaze-to-face prediction (Li et al., 2023b; Gupta et al., 2024b).

Yet, to date, such models have not been applied to naturalistic child language acquisition settings. Challenges include the technical barrier of deploying these models and the lack of

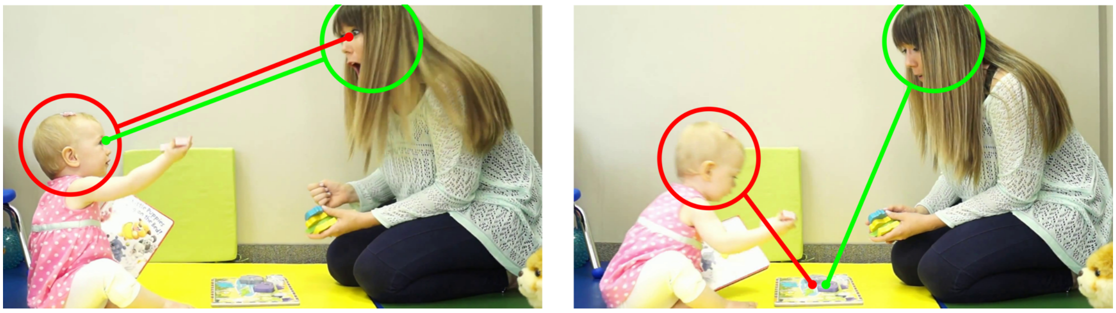


Figure 9.1: Gaze is a central component of joint attention. In this example, we see an instance of a two-step joint attention episode, consisting of shared attention followed by eye-contact. Images are from the ChildPlay dataset (Tafasca et al., 2023b).

datasets featuring children. Most existing gaze following datasets focus on adults, and models trained on these datasets generalize poorly to children (Sciortino et al., 2017). Addressing this gap, Tafasca et al. (2023b) introduced one of the first publicly available datasets focusing on child gaze behavior, but it remains limited in scope for studying developmental language learning in diverse contexts.

In this study published in Dickerman et al. (2025), we present a first step toward adapting automated gaze-target estimation for use in naturalistic child language data. We focus on a longitudinal, observational corpus of child language acquisition in Tuatschin-Romansh, a minority language spoken in Switzerland, featuring children aged 2;0 to 4;0. Our goals are threefold:

- To investigate the feasibility of using pre-trained gaze prediction models in such settings. We leverage two of our models, Sharingan (Tafasca et al., 2024) and MTGS (Gupta et al., 2024b), for this purpose.
- To compare manual and automated annotations on a frame-by-frame basis.
- To assess whether automated gaze point estimates can meaningfully capture joint attention.

9.1 Dataset

Data were extracted from an unpublished corpus of Tuatschin-Romansh (Mazara et al., shed), a dialect of the Sursilvan variety of Romansch spoken in the Grisons region in Switzerland (Maurer-Cecchini, 2021). The corpus consists of longitudinal, naturalistic audiovisual recordings of six children between the ages of 2;0 and 4;3. Each child was recorded for approximately 4.5 hours in monthly intervals for either one or two years. Recordings were carried out by the children's parents. The recordings are unprompted and feature the target children doing typical activities at-home, such as playing alone, with siblings, with their parents, or with grandparents. The corpus is extremely visually diverse, with the camera placed in different locations around the children's homes, capturing interactants both near and far from the camera.

Two datasets were constructed from samples randomly drawn from this corpus – a testing dataset and a training dataset. This split allowed us to first fine-tune a gaze-target estimation model on the training dataset, and then evaluate its performance on the testing dataset. Fine-tuning is particularly important given the domain gap between our dataset and publicly available datasets: while public datasets tend to focus on adults, our dataset is captured with a larger camera field-of-view, and is more centered on household settings, compared to the more diverse contexts in public datasets. Clips were sampled without replacement and always consisted of 10 second long video clips, sampled from the dataset at a frame rate of 10 frames-per-second, for a total of 100 frames per video for annotation.

Testing dataset

The testing dataset comprised 20 video-clips, or 2,000 frames, and was annotated fully by four independent coders, with Coder A treated as ground truth for model evaluations. The data for this dataset was hand-picked from the Tuatschin-Romansch corpus in the following way: clips were randomly sampled from the corpus, and then reviewed for high visual quality. Clips that poorly showed study participants, had camera motion in them, or were otherwise judged to be visually poor were excluded from the testing dataset. The testing dataset was not chosen for ecological validity but rather to examine model estimates given ’good’ use-cases from the naturalistic data.

Training dataset

The training dataset comprised 519 video-clips, or 51,900 frames. Clips were sampled randomly from the corpus. Since some randomly sampled clips lacked participants, additional clips were added to the dataset to compensate. There are a total of 19 clips in which no person is visible at any time, and 500 clips in which at least one person appears at any given time. The dataset was annotated by a team of 14 independent coders.

9.2 Method

We describe the manual annotation protocol, and the automatic annotations obtained using the gaze models below.

9.2.1 Manual Annotation

Annotation was done using the tool Labelbox (Labelbox, 2024). The coding procedure consisted of two ’passes’ through each video. The first pass involved coding frame-wise elements, with coders working on one participant at a time:

- *Head crop:* The head bounding box around the participant’s head in the video frame.

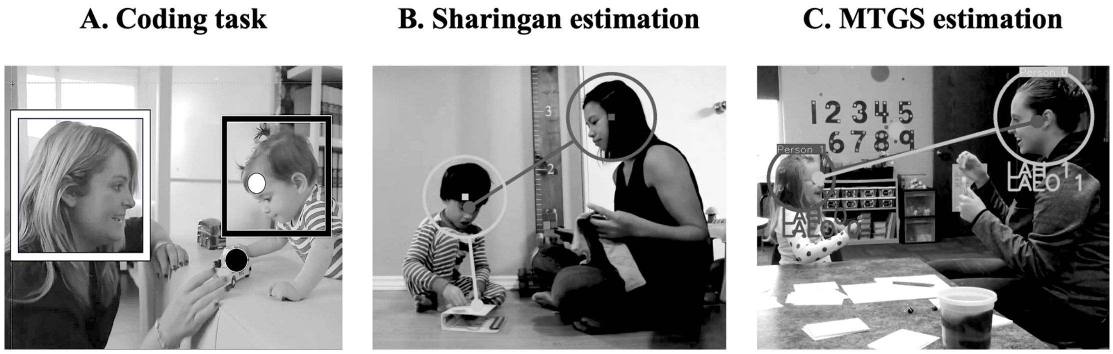


Figure 9.2: Coding task and model outputs. Images are from the ChildPlay dataset (Tafasca et al., 2023b). (A) demonstrates the coding task of drawing a head crop and estimating gaze point within the scene. (B) shows Sharingan model outputs (Tafasca et al., 2024) which estimates a gaze point within the scene. (C) shows MTGS model outputs (Gupta et al., 2024b) which estimates gaze point and social gaze components, including Looking at Head (LAH) and Looking at Each Other (LAEO).

- *Gaze point:* The estimated location in the frame where the participant was looking.
- *Out-of-frame:* Marked when the participant was looking off-screen.
- *Occluded head crop:* Marked when the participant's head was present in the scene but occluded from the camera's view.
- *Occluded gaze point:* Marked when the gaze target was within the camera's field of view but not visible due to occlusion.
- *Eyes closed:* Marked when a gaze target could not be determined because the participant's eyes were closed.
- *Looking-at-each-other:* Annotated when two participants were looking at each other simultaneously.
- *Shared attention:* Annotated when two or more participants were simultaneously looking at the same object or activity.

Joint attention was coded in the second pass. Joint attention was defined to have a minimum duration of 20 frames (2 seconds) following (Gabouer and Bortfeld, 2021), and was coded when two or more people were interacting about some shared focal object, activity, or, in some cases, person.

All participants in all videos were annotated. Coders underwent an in-person training session on the annotation protocol and completed practice annotations. They received feedback on the practice annotations, and corrected them as instructed before continuing to the annotation task. All videos in the training dataset were coded once. 45% of videos in the training dataset were reviewed and corrected by a secondary experienced coder.

9.2.2 Automatic Annotation

We applied two gaze following models to the testing dataset to automatically estimate where individuals in the scene were looking. Visual examples of each model's input and output are shown in Figure 9.2 (B) and (C).

Sharingan (Tafasca et al., 2024) is a transformer-based model designed for multi-person gaze following. It processes both the scene image and individual head crops to predict a per-person gaze heatmap, which represents the spatial distribution of probable gaze targets across the image.

MTGS (Gupta et al., 2024b) builds on Sharingan by incorporating temporal modeling and joint prediction of both gaze following and social gaze behaviors (looking at heads, looking at each other, shared attention), as described in Section 6.2. Like Sharingan, MTGS also produces a gaze heatmap per person, with the addition of pairwise social gaze scores. We use the static version of MTGS for this study.

For both models, we extract the predicted gaze point by selecting the location corresponding to the maximum value in the gaze heatmap.

9.3 Experiments

Dataset. We use the Tuatschin-Romansh dataset described in Section 9.1.

Tested Models. We compare the performance of the following three models on the testing set:

- *Sharingan:* Pre-trained on GazeFollow (Recasens et al., 2015) and fine-tuned on VideoAttentionTarget (Chong et al., 2020b).
- *Publicly-trained MTGS:* Pre-trained on GazeFollow and fine-tuned on VSGaze (Section 3.4). Notably, VSGaze includes the ChildPlay dataset (Tafasca et al., 2023b), which contains child-centric interactions. As a result, this model is trained on data more similar in domain to the Tuatschin-Romansh dataset than Sharingan.
- *Fine-tuned MTGS:* The publicly-trained MTGS model above is further fine-tuned on the Tuatschin-Romansh training set.

Training and Validation. For fine-tuning MTGS, we use 450 clips from the training split, reserving the remaining 50 clips for validation and model selection. The model is trained for 20 epochs using the AdamW optimizer with a base learning rate of 3×10^{-5} and a cosine annealing schedule with warm restarts. Stochastic Weight Averaging (SWA) is applied starting from epoch 12, using a reduced learning rate of 1×10^{-5}.

Metrics. We use two metrics to compute coder reliability between annotators as well as compute model performance:

- *Interclass Correlation Coefficient (ICC):* We calculated the intraclass correlation coefficient (ICC) following (Koo and Li, 2016), separating x- and y- coordinates due to differences in their possible range. Four independent coders annotated all 2,000 frames in the testing dataset. Coder A was treated as the ground-truth. ICC assumes data-point independence; in our testing dataset, temporally adjacent data-points are more similar than non-adjacent data-points. To address interdependency between data-points, one set of gaze point coordinates was sampled per individual in each video clip. This resulted in 48 data-points, above the recommended threshold of 30 (Koo and Li, 2016).

- *Distance:* We compute the standard gaze following distance metric (Section 4.1.1). Coder A was again treated as the ground truth. This metric also faces the issue of interdependent data-points, but in order to follow the standard, distance was calculated by aggregating all data-points. We also calculated a separate distance metric using the random sample extracted for ICC calculation, to allow for comparison between independent data-points.

For both metrics, only data-points coded by all four coders were included in the calculation, as is standard for the ICC metric. Additionally, gaze points coded with the exceptions *eyes closed, obscured head crop, obscured gaze point,* and *out-of-frame* were removed from the pool prior to random sampling.

In addition, for MTGS, we compute precision, recall and F1 scores for looking at heads and looking at each other following the post-processing protocol described in Section 4.2.1. For shared attention, we similarly compute these scores by thresholding decoder outputs at 0.1.

9.4 Results

Manual Annotation Statistics

Of the 51,900 frames in the training dataset, 49,355 contained at least one person. 113,776 head-crops were coded, of which 88,855 had a gaze point estimate. Missing gaze point estimates were largely coded by one or more gaze-exceptions: eyes closed (2,310), out-of-frame (12,308), obscured head-crop (6,605), and obscured gaze point (17,471). Shared attention was coded in 13,378 participants, eye contact in 3,460 participants, and joint attention in 22,487 participants.

In the testing dataset, all 2,000 frames contained at least one person. A total of 4,931 head-crops were coded, of which 4,014 had a gaze point estimate. Gaze exceptions coded were eyes closed (109), out-of-frame (318), obscured head-crop (346), and obscured gaze point (480). Shared attention was coded in 1,418 participants, eye contact in 323 participants, and joint attention in 2,346 participants.

Coder Reliability and Model Performance

Manual coder reliability was assessed with a two-way random effects model for absolute agreement, yielding excellent agreement for both x- (0.91, 95% CI [0.88, 0.95]) and y-coordinates

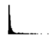
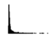
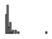
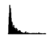
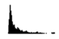
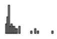
	Distance (n=3,116)		Sampled Distance (n=48)		Intraclass Correlation (n=48)	
	$\mu(\sigma)$	Density Plot of Distance	$\mu(\sigma)$	Histogram of Sample	ICC_x 95% CI	ICC_y 95% CI
Manual estimates						
Coder B	0.047 (0.074)		0.031 (0.031)		0.917 $[0.857, 0.952]$	0.962 $[0.919, 0.981]$
Coder C	0.048 (0.084)		0.05 (0.088)		0.884 $[0.802, 0.933]$	0.964 $[0.936, 0.98]$
Coder D	0.049 (0.081)		0.057 (0.107)		0.94 $[0.896, 0.966]$	0.955 $[0.922, 0.975]$
Model estimates						
Sharingan	0.124 (0.135)		0.134 (0.17)		0.645 $[0.444, 0.784]$	0.684 $[0.498, 0.809]$
MTGS	0.112 (0.112)		0.124 (0.131)		0.832 $[0.719, 0.902]$	0.677 $[0.489, 0.805]$
Fine-tuned MTGS	0.098 (0.111)		0.09 (0.086)		0.876 $[0.79, 0.929]$	0.876 $[0.789, 0.929]$

Table 9.1: Metrics comparing manual and automated gaze point estimates.

	Shared Attention		Looking at Heads		Looking at Each Other	
	Public	**Fine-tuned**	**Public**	**Fine-tuned**	**Public**	**Fine-tuned**
Precision	0.329	**0.407**	0.608	**0.699**	**0.902**	0.852
Recall	**0.793**	0.711	0.432	**0.451**	**0.372**	0.351
F1	0.465	**0.518**	0.505	**0.548**	**0.527**	0.497

Table 9.2: Performance for social gaze for the publicly-trained and fine-tuned MTGS models. Best results are given in bold.

(0.96, 95% CI [0.94, 0.98]). Automated gaze-target estimation reliability was assessed for each model against ground truth (Coder A) using a two-way mixed effects model for absolute agreement. Manual coders were also individually assessed against ground truth with two-way random effects models for absolute agreement. ICC reliability may be interpreted as: < 0.5: *poor*, 0.5 - 0.75: *moderate*, 0.75 - 0.9: *good*, and > 0.9: *excellent*. Both ICC and the distance metric are available in Table 9.1.

Manually coded data generally shows excellent reliability, and model-coded data shows moderate to good reliability. In both metrics, the model which performs best is the fine-tuned MTGS, and human coders out-perform all automatic coding methods. All plots are right-skewed, but manually coded plots show less spread and finer clustering around 0 compared to the model estimates.

We also compare social gaze performance of the publicly-trained and fine-tuned versions of MTGS. Precision, recall, and F1 metrics for both models are available in Table 9.2. The fine-tuned model improves over the publicly-trained version for looking at heads and shared attention, but has a drop in performance for looking at each other.

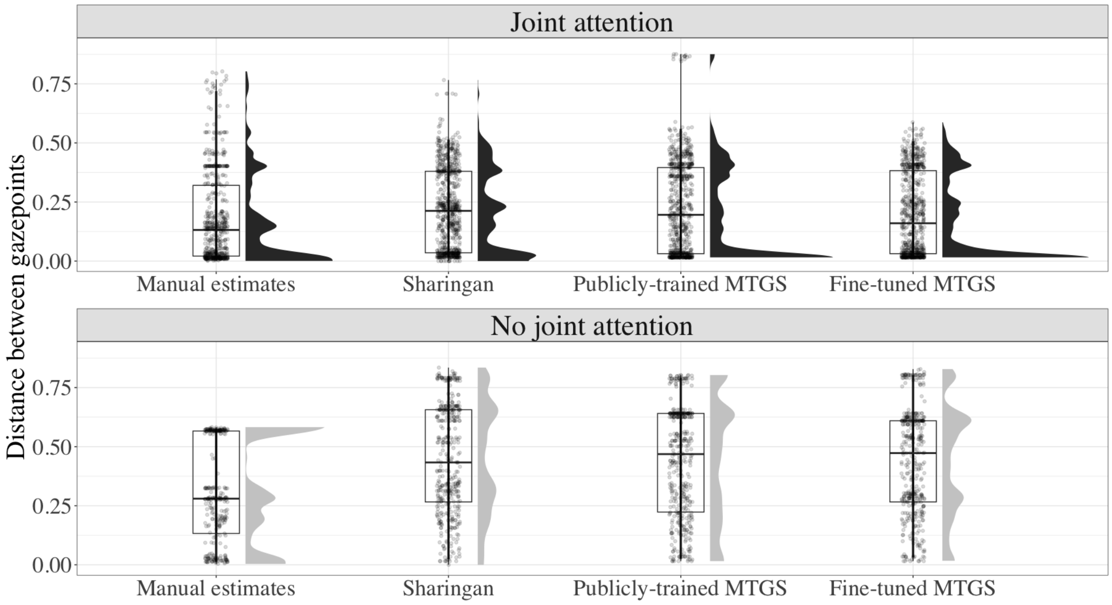


Figure 9.3: Distance between gaze points in frames with only two people in the testing dataset. We compare four coding methods; manual estimates, Sharingan estimates, and both publicly-trained and fine-tuned MTGS estimates.

Distribution of social gaze components relative to joint attention

In the training dataset, we assessed the frame-wise frequency of each social gaze component in participants coded with and without joint attention. Joint attention occurs in 22,487 participants in the dataset. Of these, 2,990 participants (13.3%) were Looking at Heads (LAH), 824 (3.7%) were Looking at Each Other (LAEO), and 4,158 (18.5%) were Sharing Attention(SA). Joint attention does not occur in 91,296 participants in the dataset. Of these, 10,231 participants (11.2%) were Looking at Heads (LAH), 2,636 (2.9%) were Looking at Each Other (LAEO), and 9,220 (10.1%) were Sharing Attention (SA). Significance testing was not carried out due to interrelated data-points.

In the testing dataset, joint attention occurs in 2,346 participants. Of these, 466 (19.9%) were LAH, 219 (9.3%) were LAEO, and 965 (41.1%) were SA. Joint attention does not occur in 2,588 participants. Of these, 368 (14.2%) were LAH, 104 (4.0%) were LAEO, and 453 were (17.5%) SA.

To compare all models, we also calculated the distance between participants' gaze points during and outside of joint attention, as a proxy to shared attention. We restricted the sample to only include frames with two people coded. Distributions are shown in Figure 9.3. We observe similar distributions across all coding methods during and outside of joint attentional frames.

9.5 Conclusion

In this study, we present the first application of automatic gaze-target estimation models to a naturalistic child language acquisition corpus, assessing their ability to approximate human-coded gaze annotations and capture patterns of joint attention. Our results indicate that while human coders still outperform current models in terms of precision and frame-level accuracy, automated estimates—particularly those from models fine-tuned on domain-specific data—fall within accepted inter-rater reliability standards. Notably, the fine-tuned version of MTGS, trained on a subset of the Tuatschin-Romansh corpus, consistently outperforms Sharingan and the publicly trained MTGS in aligning with manual annotations.

The differences between model performances can be partly attributed to the training data: MTGS was trained on a dataset containing both adult and child interactions, while Sharingan was trained mainly on adult data. Our findings further suggest that fine-tuning on visually and contextually similar data improves model reliability, especially for downstream use cases such as inferring joint attention. While exact point-level agreement between human and machine annotations remains imperfect, gaze-point distance distributions produced by the models are still predictive of joint attention states, supporting their utility for broader behavioral analyses over time.

Despite these promising results, we caution that automated estimates may still reflect hidden biases—such as a tendency to predict gaze near salient objects like hands—that warrant further investigation. Moreover, our testing dataset consisted of visually optimal clips chosen to maximize model performance; as such, the results are not fully generalizable to the more visually challenging segments of the corpus. Naturalistic data introduces inherent complexity—children may be in motion, partially occluded, or facing away from the camera—all of which complicate both manual and automatic annotation.

Nevertheless, this work demonstrates the feasibility of applying gaze-target estimation models to naturalistic, third-person video data in developmental research. It also highlights the importance of domain adaptation through fine-tuning, a practical alternative to constructing large task-specific datasets from scratch. As automatic annotation tools continue to evolve, they offer a promising avenue for scaling behavioral research in naturalistic environments—unlocking new opportunities to explore social, cognitive, and linguistic development at scale.

10 Conclusion

This thesis addressed the problem of predicting gaze behaviour in naturalistic settings, with an emphasis on gaze following and higher-level social gaze understanding. Motivated by the need for semantic, scalable, and context-aware models of gaze, we introduced a range of methods and resources that advance the field across multiple dimensions.

10.1 Summary of Contributions

Our contributions span from new datasets and evaluation protocols to multimodal architectures and unified frameworks. Table 10.1 summarizes the overall progression of our gaze following models throughout this PhD, showing consistent improvements over time, with the latest MTGS-DINO model achieving the best overall performance.

More specifically, the work is organized across several key directions.

Datasets. To address the limitations of existing datasets—particularly the lack of contextual signals such as speaking status, limited task coverage, and insufficient semantic annotation—we introduced several new datasets and extensions in Chapter 3. Specifically:

- *ChildPlay-audio* augments the ChildPlay dataset (Tafasca et al., 2023b) with frame-wise speaking annotations, enabling multimodal modeling of social attention.
- *VSGaze* provides unified and extended gaze annotations (including LAH, LAEO, and SA) across multiple gaze datasets, facilitating multi-task learning.
- Semantic gaze annotation extensions to the RLR-CHAT dataset (Murdock et al., 2024; Yun et al., 2024) help support ego-exo modeling and more semantically grounded evaluation.

These resources serve as a foundation for training and evaluating models across a broad range of gaze-related tasks.

Evaluation. Existing gaze evaluation protocols primarily focus on spatial localization, offering

Model	GazeFollow		VideoAttentionTarget	
	Min. Dist.↓	Avg. Dist.↓	Dist.↓	AP IO↑
Gupta et al. (2022)	0.071	0.134	0.122	0.864
Tafasca et al. (2023b)	0.064	0.125	0.109	0.834
Sharingan (Tafasca et al., 2024)	0.057	0.113	0.107	**0.891**
MTGS (Gupta et al., 2024b)	0.059	0.116	0.114	0.843
MTGS-DINO	**0.043**	**0.098**	**0.096**	0.878

Table 10.1: Comparison of gaze following performance of our models on the GazeFollow (Recasens et al., 2015) and VideoAttentionTarget (Chong et al., 2020b) datasets. Best results are given in bold. We see consistent improvements, with our latest MTGS-DINO model providing the best overall performance.

little insight into the social or semantic meaning of gaze behavior. In Chapter 4, we addressed this gap by introducing:

- New metrics and evaluation protocols for assessing social gaze behaviors—looking at heads (LAH), mutual gaze (LAEO), and shared attention (SA)—both as post-processed gaze following outputs and as direct model predictions.
- A revised evaluation protocol for shared attention that resolves key shortcomings in prior work by allowing multiple shared attention instances per frame and explicitly modeling the involved participants.
- Gaze-following-inspired metrics for egocentric gaze estimation, supporting more interpretable and semantically grounded evaluation in first-person settings.

Multimodal Gaze Following. Gaze following is challenging and requires reasoning about multiple contextual cues. Here, additional modalities such as body pose can help signal salient targets and body orientation. Meanwhile, depth information can help filter out potential targets along the 2D line of sight but not visible in 3D. Motivated by this, we introduced two multimodal architectures in Chapter 5:

- The first is a modular architecture that integrates RGB, depth, and body pose using attention-based fusion. It further supports inference using only depth and pose, which are inherently anonymized, allowing applicability in privacy-sensitive settings.
- The second proposes a principled use of depth to construct a 3D Field of View (3DFoV) representation that highlights visible parts of the scene in 3D. This facilitates geometric reasoning and leads to improved generalization across datasets.

In addition, in Chapter 7, we showed that recent vision-language models can be used to extract gaze relevant contextual cues related to people's pose, gestures and interactions, which can then be used to improve gaze performance.

Unified Social Gaze Prediction. While social gaze cues like mutual gaze or shared attention are often more relevant than raw gaze points, prior models have treated these as separate tasks. In Chapter 6, we proposed unified modeling strategies:

- We first extended the Sharingan model (Tafasca et al., 2024) with task-specific decoders trained on disjoint datasets for LAH, LAEO, and SA.
- We then proposed *MTGS*, a unified framework that jointly models gaze following and social gaze tasks. MTGS incorporates architectural improvements and temporal modeling, and demonstrates benefits from multi-task training.
- We further demonstrated that MTGS benefits from recent foundation models such as DINOv2 (Oquab et al., 2023), achieving new state-of-the-art performance.

Learning Ego-Exo Gaze Representations. Egocentric gaze estimation is challenging, but can be improved by leveraging exocentric gaze information of other individuals in the scene. In Chapter 8, we proposed novel approaches for jointly learning ego-exo gaze representations, with the exo representations learned via self-supervision. Our results demonstrated that single frame methods can achieve strong performance, our approaches enable learning of exo representations as validated via a probing task, and that learning these representations leads to improved egocentric gaze predictions.

Application to Naturalistic Child Data. Finally, we demonstrated the feasibility of applying these models to naturalistic, unstructured video in the context of early language development. Gaze is a key component of joint attention, but typical studies have been limited to controlled setups, and the use of eye-trackers or manual coding for annotations. This hinders generalizability of results. In Chapter 9, we applied our models to a corpus of child–adult interactions, aiming to analyze components of joint attention. Our results show that model predictions are reliable, and follow manual annotation distributions across joint attention states—offering a promising step toward scalable behavioral analysis in developmental research.

10.2 Limitations and Future Directions

While this thesis advances the state of gaze prediction across several axes, a number of important questions remain open. We outline several key limitations and corresponding directions for future research.

More Metrics. A diverse set of metrics is essential for capturing the full range of model behaviors and failure modes. Future work could develop metrics that specifically evaluate 3D gaze following performance, or assess model robustness under varying conditions such as when the eyes are visible or occluded. Other possibilities include the development of event-level metrics that can capture temporal behaviours such as joint attention.

Other Modalities. A promising direction for future work is further exploring the integration of additional modalities such as audio and text. We have already seen that speaking status can enhance MTGS, and that spatial audio can improve egocentric gaze estimation. We also used text inputs to vision-language models (VLMs) in order to extract gaze-relevant contextual cues. However, both audio and text carry rich semantic information beyond what has been explored so far—audio may convey conversational dynamics, emotional tone or indicate the interacted object, while text may describe the scene or broader setting. Another potential research direction could involve distilling information from these modalities so that they are not required at inference time.

Leveraging Foundation Models. We showed that MTGS benefits from strong visual representations using DINOv2 (Oquab et al., 2023), and that VLMs like BLIP-2 (Li et al., 2023a) can identify human activities in a zero-shot setting. Going ahead, evaluating the gaze reasoning capabilities of the latest multimodal large language models—such as GPT-4o (Hurst et al., 2024)—offers exciting opportunities to incorporate broader semantic understanding and world knowledge into gaze prediction systems.

Temporal and Interaction Modeling. Although MTGS includes temporal reasoning within the person branch, future models could explore richer temporal modelling including scene dynamics (*e.g.* moving objects) and hand gestures. Incorporating gaze behaviors such as saccades and gaze shifts could further enhance the model's ability to learn fine-grained temporal dependencies. Additionally, longer-term temporal context could aid disambiguation by revealing which objects or people have been interacted with over the temporal window.

Scaling Annotations. The scarcity of large-scale, densely annotated datasets remains a major obstacle. One approach is to generate pseudo-annotations using strong pretrained models, as demonstrated in our VSGaze construction. Another promising avenue is to apply self-supervised or contrastive learning to large unlabelled video corpora, which could enable learning of gaze-relevant representations without relying on costly manual labels.

Expanding Downstream Applications. Beyond joint attention in child–adult interactions, another high-impact application lies in supporting autism diagnosis. As discussed in Chapter 2, gaze is a well-established behavioral marker for autism. We are currently exploring the use of gaze models to analyze child–clinician interactions in diagnostic settings, with the goal of developing automated tools to support clinicians in early screening and intervention. Another promising direction involves applying gaze models in online, interactive settings such as human–robot interaction. These scenarios introduce new challenges, most notably the need for low-latency inferences.

Ethical Considerations. Given the sensitive nature of gaze and attention data—especially in applications involving children or health—future work must prioritize privacy, transparency, and informed consent. It is essential that models be deployed only in contexts that respect participants' autonomy and are aligned with the intended use cases. Additionally, understanding and mitigating bias in pretrained gaze models is an open research priority.

Bibliography

Abels, M. (2020). Triadic interaction and gestural communication: Hierarchical and child-centered interactions of rural and urban gujarati (indian) caregivers and 9-month-old infants. *Developmental Psychology*, 56(10):1817.

Abney, D. H., Suanda, S. H., Smith, L. B., and Yu, C. (2020). What are the building blocks of parent–infant coordinated attention in free-flowing interaction? *Infancy*, 25(6):871–887.

Adamson, L. B., Bakeman, R., Suma, K., and Robins, D. L. (2019). An expanded view of joint attention: Skill, engagement, and language in typical development and autism. *Child Development*, 90(1):e1–e18.

Admoni, H. and Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63.

Ahn, Y. A., Moffitt, J. M., Tao, Y., Custode, S., Parlade, M., Beaumont, A., Cardona, S., Hale, M., Durocher, J., Alessandri, M., et al. (2024). Objective measurement of social gaze and smile behaviors in children with suspected autism spectrum disorder during administration of the autism diagnostic observation schedule. *Journal of autism and developmental disorders*, 54(6):2124–2137.

Akhtar, N., Dunham, F., and Dunham, P. J. (1991). Directive interactions and early vocabulary development: The role of joint attentional focus. *Journal of Child Language*, 18(1):41–49.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Andreas, J. (2022). Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779.

Apple (2025). Apple vision pro. https://www.apple.com/apple-vision-pro/. Accessed: 2025-01-13.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846.

Ataman-Devrim, M., Nixon, E., and Quigley, J. (2023). Joint attention episodes during interactions with fathers but not mothers at age 2 years is associated with expressive language at 3 years. *Journal of Experimental Child Psychology*, 226:105569.

Ba, S. and Odobez, J.-M. (2008). Recognizing visual focus of attention from head pose in natural meetings. *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1).

Ba, S. and Odobez, J.-M. (2011). Multiperson visual focus of attention from head pose and meeting contextual cues. *Transactions on Pattern Analysis and Machine Intelligence*, 33(1).

Bachmann, R., Mizrahi, D., Atanov, A., and Zamir, A. (2022). Multimae: Multi-modal multi-task masked autoencoders. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 348–367. Springer.

Bai, C., Kumar, S., Leskovec, J., Metzger, M., Nunamaker, J., and Subrahmanian, V. S. (2019). Predicting the visual focus of attention in multi-person discussion videos. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 4504–4510. International Joint Conferences on Artificial Intelligence Organization.

Bakeman, R. and Adamson, L. B. (1984). Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Development*, pages 1278–1289.

Baldwin, D. A. (1995). Understanding the link between joint attention and language. In *Joint Attention*, pages 131–158. Psychology Press.

Bao, J., Liu, B., and Yu, J. (2022). Escnet: Gaze target detection with the understanding of 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14126–14135.

Bard, K. A., Keller, H., Ross, K. M., Hewlett, B., Butler, L., Boysen, S. T., and Matsuzawa, T. (2021). Joint attention in human and chimpanzee infants in varied socio-ecological contexts. *Monographs of the Society for Research in Child Development*, 86(4):7–217.

Belkada, Y., Bertoni, L., Caristan, R., Mordan, T., and Alahi, A. (2021). Do pedestrians pay attention? eye contact detection in the wild.

Billing, E., Belpaeme, T., Cai, H., Cao, H.-L., Ciocan, A., Costescu, C., David, D., Homewood, R., Hernandez Garcia, D., Gómez Esteban, P., et al. (2020). The dream dataset: Supporting a data-driven study of autism spectrum disorder and robot enhanced therapy. *PloS one*, 15(8):e0236939.

Brody, S., Alon, U., and Yahav, E. (2022). How attentive are graph attention networks? In *International Conference on Learning Representations*.

Brooks, R. and Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental science*, 8(6):535–543.

Brooks, R. and Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of child language*, 35(1):207–220.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liszkowski, U., Behne, T., Tomasello, M., and Collins, W. A. (2011). Early social cognition in three cultural contexts. *Monographs of the Society for Research in Child Development*, pages i–142.

Cantarini, G., Tomenotti, F. F., Noceti, N., and Odone, F. (2021). Hhp-net: A light heteroscedastic neural network for head pose estimation with uncertainty.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., and Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, pages i–174.

Chang, F., Zeng, J., Liu, Q., and Shan, S. (2023). Gaze pattern recognition in dyadic communication. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, pages 1–7.

Chao, Y.-W., Wang, Z., He, Y., Wang, J., and Deng, J. (2015). Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*, pages 1017–1025.

Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., and Qiao, Y. (2022). Vision transformer adapter for dense predictions. In *The Eleventh International Conference on Learning Representations*.

Chong, E., Chanda, K., Ye, Z., Southerland, A., Ruiz, N., Jones, R. M., Rozga, A., and Rehg, J. M. (2017). Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–20.

Chong, E., Clark-Whitney, E., Southerland, A., Stubbs, E., Miller, C., Ajodan, E. L., Silverman, M. R., Lord, C., Rozga, A., Jones, R. M., et al. (2020a). Detection of eye contact with deep neural networks is as accurate as human experts. *Nature communications*, 11(1):6386.

Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., and Rehg, J. M. (2018). Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 383–398.

Chong, E., Wang, Y., Ruiz, N., and Rehg, J. M. (2020b). Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Ci, Y., Wang, Y., Chen, M., Tang, S., Bai, L., Zhu, F., Zhao, R., Yu, F., Qi, D., and Ouyang, W. (2023). Unihcp: A unified model for human-centric perceptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17840–17852.

de Belen, R. A. J., Bednarz, T., Sowmya, A., and Del Favero, D. (2020). Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Translational psychiatry*, 10(1):1–20.

Dickerman, M., Gupta, A., Tafasca, S., Zhang, X., Odobez, J.-M., and Stoll, S. (2025). Automatic detection of the visual gaze components of joint attention in naturalistic, observational data. In *Proceedings of the 49th annual Boston University Conference on Language Development*. Cascadilla Press.

Doosti, B., Chen, C.-H., Vemulapalli, R., Jia, X., Zhu, Y., and Green, B. (2021). Boosting image-based mutual gaze detection using pseudo 3d gaze. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1273–1281.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Edition, F. (2013). Diagnostic and statistical manual of mental disorders. *American Psychiatric Association*, 21:591–643.

Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al. (2023). Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Fan, C., Lee, J., Xu, M., Kumar Singh, K., Jae Lee, Y., Crandall, D. J., and Ryoo, M. S. (2017). Identifying first-person camera wearers in third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5125–5133.

Fan, L., Chen, Y., Wei, P., Wang, W., and Zhu, S.-C. (2018). Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6460–6468.

Fan, L., Wang, W., Huang, S., Tang, X., and Zhu, S.-C. (2019). Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5724–5733.

Fang, Y., Tang, J., Shen, W., Shen, W., Gu, X., Song, L., and Zhai, G. (2021). Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11390–11399.

Farkhondeh, A., Tafasca, S., and Odobez, J.-M. (2024). Childplay-hand: A dataset of hand manipulations in the wild. In *European Conference on Computer Vision (ECCV) Workshops*.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Franchak, J. M., Heeger, D. J., Hasson, U., and Adolph, K. E. (2016). Free viewing gaze behavior in infants and adults. *Infancy*, 21(3):262–287.

Gabouer, A. and Bortfeld, H. (2021). Revisiting how we operationalize joint attention. *Infant Behavior and Development*, 63:101566.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237.

Gorga, S. and Otsuka, K. (2010). Conversation scene analysis based on dynamic Bayesian network and image-based gaze detection. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction on - ICMI-MLMI '10*. ACM Press.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al. (2022). Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Grzadzinski, R., Carr, T., Colombi, C., McGuire, K., Dufek, S., Pickles, A., and Lord, C. (2016). Measuring changes in social communication behaviors: preliminary development of the brief observation of social communication change (boscc). *Journal of autism and developmental disorders*, 46:2464–2479.

Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., and Malik, J. (2018). Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guan, J., Yin, L., Sun, J., Qi, S., Wang, X., and Liao, Q. (2020). Enhanced gaze following via object detection and human pose estimation. In *International Conference on Multimedia Modeling*, pages 502–513. Springer.

Guo, H., Hu, Z., and Liu, J. (2022). Mgtr: End-to-end mutual gaze detection with transformer. In *Proceedings of the Asian Conference on Computer Vision*, pages 1590–1605.

Guo, Z., Chheang, V., Li, J., Barner, K. E., Bhat, A., and Barmaki, R. L. (2023). Social visual behavior analytics for autism therapy of children based on automated mutual gaze detection. In *Proceedings of the 8th ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, pages 11–21.

Gupta, A., Tafasca, S., Chutisilp, N., and Odobez, J.-M. (2024a). A unified model for gaze following and social gaze prediction. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9. IEEE.

Gupta, A., Tafasca, S., Farkhondeh, A., Vuillecard, P., and Odobez, J.-m. (2024b). Mtgs: A novel framework for multi-person temporal gaze following and social gaze prediction. *Advances in Neural Information Processing Systems*, 37:15646–15673.

Gupta, A., Tafasca, S., and Odobez, J.-M. (2022). A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5041–5050.

Gupta, A., Vuillecard, P., Farkhondeh, A., and Odobez, J.-M. (2024c). Exploring the zero-shot capabilities of vision-language models for improving gaze following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 615–624.

Hansen, D. W. and Ji, Q. (2009). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500.

Harel, J., Koch, C., and Perona, P. (2006). Graph-based visual saliency. *Advances in neural information processing systems*, 19.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hu, Z., Yang, D., Cheng, S., Zhou, L., Wu, S., and Liu, J. (2022a). We know where they are looking at from the rgb-d camera: Gaze following in 3d. *IEEE Transactions on Instrumentation and Measurement*.

Hu, Z., Zhao, K., Zhou, B., Guo, H., Wu, S., Yang, Y., and Liu, J. (2022b). Gaze target estimation inspired by interactive attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8524–8536.

Huang, Y., Cai, M., Li, Z., Lu, F., and Sato, Y. (2020). Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806.

Huang, Y., Cai, M., Li, Z., and Sato, Y. (2018). Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 754–769.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiang, H., Murdock, C., and Ithapu, V. K. (2022). Egocentric deep multi-channel audio-visual active speaker localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10552.

Jin, T., Lin, Z., Zhu, S., Wang, W., and Hu, S. (2021). Multi-person gaze-following with numerical coordinate regression. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE.

Jin, T., Yu, Q., Zhu, S., Lin, Z., Ren, J., Zhou, Y., and Song, W. (2022). Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence*, 113:104924.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., TaoXie, Michael, K., Fang, J., imyhxy, Lorna, Wong, C., Yifu, Z., V, A., Montes, D., Wang, Z., Fati, C., Nadar, J., Laughing, UnglvKitDe, tkianai, yxNONG, Skalski, P., Hogan, A., Strobel, M., Jain, M., Mammana, L., and xylieong (2022). ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations.

Johansson, R., Westling, G., Backstrom, A., and Flanagan, R. (2001). Eye-Hand Coordination in Object Manipulation. *Journal of Neuroscience*, 21(17):6917–6932.

Ju, C., Han, T., Zheng, K., Zhang, Y., and Xie, W. (2022). Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer.

Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., and Torralba, A. (2019). Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Labelbox (2024). Labelbox. Available: Online.

Lai, B., Liu, M., Ryan, F., and Rehg, J. M. (2023). In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision*, pages 1–18.

Lai, B., Ryan, F., Jia, W., Liu, M., and Rehg, J. M. (2025). Listen to look into the future: Audio-visual egocentric gaze anticipation. In *European Conference on Computer Vision*, pages 192–210. Springer.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023a). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Li, P., Lu, H., Poppe, R. W., and Salah, A. A. (2023b). Automated detection of joint attention and mutual gaze in free play parent-child interactions. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 374–382.

Li, Y., Liu, M., and Rehg, J. M. (2018). In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635.

Li, Y., Liu, M., and Rehg, J. M. (2021). In the eye of the beholder: Gaze and actions in first person video. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6731–6747.

Li, Z. and Snavely, N. (2018). Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050.

Lian, D., Yu, Z., and Gao, S. (2018). Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., Bishop, S., et al. (2012). Autism diagnostic observation schedule–2nd edition (ados-2). *Los Angeles, CA: Western Psychological Corporation*, 284:474–478.

Loshchilov, I. and Hutter, F. (2018). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lv, Z., Charron, N., Moulon, P., Gamino, A., Peng, C., Sweeney, C., Miller, E., Tang, H., Meissner, J., Dong, J., Somasundaram, K., Pesqueira, L., Schwesinger, M., Parkhi, O., Gu, Q., Nardi, R. D., Cheng, S., Saarinen, S., Baiyya, V., Zou, Y., Newcombe, R., Engel, J. J., Pan, X., and Ren, C. (2024). Aria everyday activities dataset.

Magrelli, S., Jermann, P., Noris, B., Ansermet, F., Hentsch, F., Nadel, J., and Billard, A. (2013). Social orienting of children with autism to facial expressions and speech: a study with a wearable eye-tracker in naturalistic settings. *Frontiers in psychology*, 4:840.

Marín-Jiménez, M. J., Kalogeiton, V., Medina-Suárez, P., , and Zisserman, A. (2021). LAEO-Net++: revisiting people Looking At Each Other in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Marin-Jimenez, M. J., Kalogeiton, V., Medina-Suarez, P., and Zisserman, A. (2019). Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3477–3485.

Marin-Jimenez, M. J., Zisserman, A., Eichner, M., and Ferrari, V. (2014). Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296.

Masse, B., Ba, S., and Horaud, R. (2018). Tracking gaze and visual focus of attention of people involved in social interaction. *Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2711–2724.

Mastin, J. D. and Vogt, P. (2016). Infant engagement and early vocabulary development: a naturalistic observation study of mozambican infants from 1; 1 to 2; 1. *Journal of Child Language*, 43(2):235–264.

Maurer-Cecchini, P. (2021). *A grammar of Tuatschin: A Sursilvan Romansh dialect (Volume 3)*. Language Science Press.

Mazara, J., Walther, G., Sagot, B., Cathomas, C., Loporcaro, M., and Stoll, S. (Unpublished). Audiovisual longitudinal corpus of 6 children learning romansh tuatschin.

Miao, Q., Hoai, M., and Samaras, D. (2023). Patch-level gaze distribution prediction for gaze following. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 880–889.

Min, K., Roy, S., Tripathi, S., Guha, T., and Majumdar, S. (2022). Learning long-term spatial-temporal graphs for active speaker detection. In *European Conference on Computer Vision*, pages 371–387. Springer.

Mundy, P., Sigman, M., and Kasari, C. (1990). A longitudinal study of joint attention and language development in autistic children. *Journal of Autism and developmental Disorders*, 20(1):115–128.

Murdock, C., Ananthabhotla, I., Lu, H., and Ithapu, V. K. (2024). Self-motion as supervision for egocentric audiovisual localization. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7835–7839. IEEE.

Nan, Z., Jiang, J., Gao, X., Zhou, S., Zuo, W., Wei, P., and Zheng, N. (2021). Predicting task-driven attention via integrating bottom-up stimulus and top-down guidance. *IEEE Transactions on Image Processing*, 30:8293–8305.

Nonaka, S., Nobuhara, S., and Nishino, K. (2022). Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2192–2201.

Noris, B., Nadel, J., Barker, M., Hadjikhani, N., and Billard, A. (2012). Investigating gaze of children with asd in naturalistic settings.

OpenAI (2024). ChatGPT (February 25 version). https://chat.openai.com/chat.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Otsuka, K., Kasuga, K., and Kohler, M. (2018). Estimating visual focus of attention in multi-party meetings using deep convolutional neural networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 191–199.

Otsuka, K., Takemae, Y., Yamato, J., and Murase, H. (2005). A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Inter. Conf. on Multimodal Interfaces*, pages 191–198.

Otsuka, K., Yamato, J., Takemae, Y., and Murase, H. (2006). Conversation scene analysis with dynamic bayesian network basedon visual head tracking. In *2006 IEEE International Conference on Multimedia and Expo*, pages 949–952. IEEE.

Patakin, N., Vorontsova, A., Artemyev, M., and Konushin, A. (2022). Single-stage 3d geometry-preserving depth estimation model training on dataset mixtures with uncalibrated stereo data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1705–1714.

Pelisson, D., Prablanc, C., and Urquizar, C. (1988). Vestibuloocular reflex inhibition and gaze saccade control characteristics during eye-head orientation in humans. *Journal of Neurophysiology*, 59(3):997–1013.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ranftl, R., Bochkovskiy, A., and Koltun, V. (2021). Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., and Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Ray-Ban (2025). Ray-ban meta smart glasses. https://www.ray-ban.com/usa/ray-ban-meta-smart-glasses. Accessed: 2025-01-13.

Recasens, A., Khosla, A., Vondrick, C., and Torralba, A. (2015). Where are they looking? *Advances in Neural Information Processing Systems*, 28.

Recasens, A., Vondrick, C., Khosla, A., and Torralba, A. (2017). Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1435–1443.

Rehg, J., Abowd, G., Rozga, A., Romero, M., Clements, M., Sclaroff, S., Essa, I., Ousley, O., Li, Y., Kim, C., et al. (2013). Decoding children's social behavior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3414–3421.

Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., et al. (2020). Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Ryan, F., Bati, A., Lee, S., Bolya, D., Hoffman, J., and Rehg, J. M. (2025). Gaze-lle: Gaze target estimation via large-scale learned encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28874–28884.

Sciortino, G., Farinella, G. M., Battiato, S., Leo, M., and Distante, C. (2017). On the estimation of children's poses. In *Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part II 19*, pages 410–421. Springer.

Senju, A. and Johnson, M. H. (2009). Atypical eye contact in autism: models, mechanisms and development. *Neuroscience & Biobehavioral Reviews*, 33(8):1204–1214.

Sheikhi, S. and Odobez, J.-M. (2015). Combining dynamic head pose–gaze mapping with the robot conversational state for attention recognition in human–robot interactions. *Pattern Recognition Letters*, 66:81–90.

Shrout, P. E. and Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, 69(1):487–510.

Shtedritski, A., Rupprecht, C., and Vedaldi, A. (2023). What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11987–11997.

Siegfried, R. and Odobez, J.-M. (2021). Visual focus of attention estimation in 3d scene with an arbitrary number of targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3153–3161.

Sigurdsson, G. A., Gupta, A., Schmid, C., Farhadi, A., and Alahari, K. (2018). Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404.

Song, Y., Wang, X., Yao, J., Liu, W., Zhang, J., and Xu, X. (2024). Vitgaze: gaze following with interaction features in vision transformers. *Visual Intelligence*, 2(1):1–15.

Soo Park, H. and Shi, J. (2015). Social saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785.

Stiefelhagen, R., Finke, M., Yang, J., and Waibel, A. (1999). From gaze to focus of attention. In *International Conference on Advances in Visual Information Systems*, pages 765–772. Springer.

Stiefelhagen, R., Yang, J., and Waibel, A. (2002). Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Trans. Neural Networks*, 13(4):928–938.

Suarez-Rivera, C., Schatz, J. L., Herzberg, O., and Tamis-LeMonda, C. S. (2022). Joint engagement in the home environment is frequent, multimodal, timely, and structured. *Infancy*, 27(2):232–254.

Suarez-Rivera, C., Smith, L. B., and Yu, C. (2019). Multimodal parent behaviors within joint attention support sustained attention in infants. *Developmental Psychology*, 55(1):96.

Sumer, O., Gerjets, P., Trautwein, U., and Kasneci, E. (2020). Attention flow: End-to-end joint attention estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3327–3336.

Tafasca, S., Gupta, A., Kojovic, N., Gelsomini, M., Maillart, T., Papandrea, M., Schaer, M., and Odobez, J.-M. (2023a). The ai4autism project: A multimodal and interdisciplinary approach to autism diagnosis and stratification. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 414–425.

Tafasca, S., Gupta, A., and Odobez, J.-M. (2023b). Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20935–20946.

Tafasca, S., Gupta, A., and Odobez, J.-M. (2024). Sharingan: A transformer architecture for multi-person gaze following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2008–2017.

Tamis-LeMonda, C. S., Kuchirko, Y., Luo, R., Escobar, K., and Bornstein, M. H. (2017). Power in methods: Language to infants in structured and naturalistic contexts. *Developmental Science*, 20(6):e12456.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

Tavakoli, H. R., Rahtu, E., Kannala, J., and Borji, A. (2019). Digging deeper into egocentric gaze prediction. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 273–282. IEEE.

Thakur, S. K., Beyan, C., Morerio, P., and Del Bue, A. (2021). Predicting gaze from egocentric social interaction videos and imu data. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 717–722.

Tomas, H., Reyes, M., Dionido, R., Ty, M., Mirando, J., Casimiro, J., Atienza, R., and Guinto, R. (2021). Goo: A dataset for gaze object prediction in retail environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3125–3133.

Tomasello, M. and Farrar, M. J. (1986). Joint attention and early language. *Child Development*, pages 1454–1463.

Tomasello, M. and Todd, J. (1983). Joint attention and lexical acquisition style. *First language*, 4(12):197–211.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., and Bregler, C. (2015). Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656.

Tonini, F., Beyan, C., and Ricci, E. (2022). Multimodal across domains gaze target detection. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 420–431.

Tonini, F., Dall'Asen, N., Beyan, C., and Ricci, E. (2023). Object-aware gaze target detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21860–21869.

Trevarthen, C. and Hubley, P. (1978). Secondary intersubjectivity. *Action, gesture and symbol: The emergence of language*, pages 183–229.

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. (2021). Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., and Shen, W. (2022). End-to-end human-gaze-target detection with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2192–2200. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations*.

Vuillecard, P. and Odobez, J.-M. (2025). Enhancing 3d gaze estimation in the wild using weak supervision with gaze following labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364.

Wang, M., Xing, J., and Liu, Y. (2021a). Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Wang, S., Yap, K.-H., Ding, H., Wu, J., Yuan, J., and Tan, Y.-P. (2021b). Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wen, Y., Singh, K. K., Anderson, M., Jan, W.-P., and Lee, Y. J. (2021). Seeing the unseen: Predicting the first-person camera wearer's location and pose in third-person scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3446–3455.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE.

Xu, H., Ghosh, G., Huang, P.-Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., and Feichtenhofer, C. (2021). Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Xu, M., Fan, C., Wang, Y., Ryoo, M. S., and Crandall, D. J. (2018). Joint person segmentation and identification in synchronized first-and third-person videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–652.

Xue, Z. S. and Grauman, K. (2023). Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36:53688–53710.

Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L., and Fei-Fei, L. (2011). Human action recognition by learning bases of action attributes and parts. In *2011 International conference on computer vision*, pages 1331–1338. IEEE.

Yin, W., Wang, X., Shen, C., Liu, Y., Tian, Z., Xu, S., Sun, C., and Renyin, D. (2020). Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*.

Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., and Shen, C. (2021). Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yu, C., Suanda, S. H., and Smith, L. B. (2019). Infant sustained attention but not joint attention to objects at 9 months predicts vocabulary at 12 and 15 months. *Developmental Science*, 22(1):e12735.

Yu, H., Cai, M., Liu, Y., and Lu, F. (2020). First-and third-person video co-analysis by learning spatial-temporal joint attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6631–6646.

Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., and Wang, J. (2021). Hrformer: High-resolution vision transformer for dense predict. *Advances in Neural Information Processing Systems*, 34.

Yücel, Z., Salah, A. A., Meriçli, Ç., Meriçli, T., Valenti, R., and Gevers, T. (2013). Joint attention by gaze interpolation and saliency. *IEEE Transactions on cybernetics*, 43(3):829–842.

Yun, H., Gao, R., Ananthabhotla, I., Kumar, A., Donley, J., Li, C., Kim, G., Ithapu, V. K., and Murdock, C. (2024). Spherical world-locking for audio-visual localization in egocentric videos. In *European Conference on Computer Vision*, pages 256–274. Springer.

Zhang, J., Huang, J., Jin, S., and Lu, S. (2023). Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*.

Zhang, M., Teck Ma, K., Hwee Lim, J., Zhao, Q., and Feng, J. (2017). Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4372–4381.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., and Wang, X. (2022). Bytetrack: Multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhao, H., Lu, M., Yao, A., Chen, Y., and Zhang, L. (2020). Learning to draw sight lines. *International Journal of Computer Vision*, 128(5):1076–1100.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022a). Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022b). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Zwaigenbaum, L., Brian, J. A., and Ip, A. (2019). Early detection for autism spectrum disorder in young children. *Paediatrics & Child Health*, 24(7):424–432.

Anshul Gupta

✉ *anshul.gupta@epfl.ch*
🔗 *anshul-gupta24.github.io/*
in *anshulgupta24*

Education

2021–Present **Doctorate**, *École polytechnique fédérale de Lausanne (EPFL)*
PhD candidate in the Electrical Engineering Doctoral program

2014–2019 **Bachelors, Masters**, *International Institute of Information Technology, Bangalore*
Bachelors and Masters in Information Technology with specialization in Data Science
Member of the Dean's List

Work Experience

2021–Present **Research Assistant**, *Idiap Research Institute*
Advised by Dr. Jean-Marc Odobez. Researching computer vision and multimodal approaches for human behaviour and social interaction understanding with publications in top-tier venues including ICCV, CVPR and NeurIPS. Key contributions:

- Designed new CNN and transformer-based architectures that (i) unify multiple gaze-based tasks and (ii) support multimodal inputs such as depth, body pose and speaking status;
- Trained and leveraged Vision-Language Models for capturing scene semantics and inferring people's gestures and interactions;
- Created new datasets and metrics for training and evaluating gaze models.

2019–2020 **Machine Learning Engineer**, *Daimler*
Delivered production-level ML solutions for the Mercedes-Benz S-Class cabin experience:

- Built pose estimation models for accurate seat occupancy detection;
- Developed hand-tracking systems enabling touchless control of sunroof and door features;
- Improved pose reliability by integrating automated verification to reduce false positives.

Internships

2024–2025 **Research Scientist Intern**, *Meta Reality Labs Research*, United States
Supervised by Dr. Calvin Murdock in the Audio team. Investigated audio-visual modeling of egocentric conversational videos. Key achievements:

- Curated a large-scale multimodal dataset with time-aligned gaze, speech, and video of conversation participants using distributed pipelines;
- Developed self-supervised models for learning ego-exo gaze representations;
- Explored spatial audio for capturing the locations of speakers.

2023 **Visiting Student**, *Inria Sophia Antipolis*, France
Advised by Dr. François Brémond. Worked on a framework combining gaze following and action detection for automated reports (e.g., eye contact, repetitive behaviours) relevant to autism.

2019 **Masters Thesis**, *Indian Institute of Science*, India
Advised by Prof. Sriram Ganapathy and Prof. Dinesh Babu Jayagopi. Compared human and machine learning of spoken Japanese words via image supervision.

2017 **Summer Intern**, *RIKEN Brain Science Institute*, Japan
Advised by Prof. Hiroyuki Nakahara. Implemented a Reinforcement Learning model for the cognitive mechanism "Theory of Mind" and attended the accompanying summer school.

Summer Schools

2020 **Eastern European Machine Learning Summer School**, *virtual*

2019 **South East Asian Machine Learning Summer School**, *University of Indonesia*, Runner-up for Best Poster

Selected Publications

[1] **Anshul Gupta**, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard, and Jean-Marc Odobez. Mtgs: A novel framework for multi-person temporal gaze following and social gaze prediction. *Proceedings of NeurIPS*, 2024.

[2] Samy Tafasca, **Anshul Gupta**, Victor Bros, and Jean-Marc Odobez. Toward semantic gaze target detection. *Proceedings of NeurIPS*, 2024.

[3] **Anshul Gupta***, Pierre Vuillecard*, Arya Farkhondeh, and Jean-Marc Odobez. Exploring the zero-shot capabilities of vision-language models for improving gaze following. In *Proceedings of the GAZE Workshop at CVPR*, 2024. (* equal contribution) Best Paper.

[4] Samy Tafasca, **Anshul Gupta**, and Jean-Marc Odobez. Sharingan: A transformer-based architecture for multi-person gaze following. In *Proceedings of CVPR*, 2024.

[5] **Anshul Gupta**, Samy Tafasca, and Jean-Marc Odobez. A unified model for gaze following and social gaze prediction. In *Proceedings of IEEE FG*, 2024. Best Student Paper.

[6] Samy Tafasca*, **Anshul Gupta***, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of ICCV*, 2023. (* equal contribution).

[7] Venkat Krishnamohan, Akshara Soman, **Anshul Gupta**, and Sriram Ganapathy. Audiovisual correspondence learning in humans and machines. In *Proceedings of Interspeech*, 2020.

Ongoing Projects

Present **Diffusion and Flow Matching for Heatmap Refinement**
Refining sequences of noisy gaze-heatmaps using diffusion and flow matching strategies.

Organization and Reviewing

2025 **Co-organizer**, *Artificial Social Intelligence Workshop*, ICCV
This workshop explores challenges, collaborations, and ethical safeguards in giving AI human-like social perception, memory and reasoning for richer, safer human–machine interactions.

2023-Present **Reviewer**
Reviewer for the ICMI LBR workshop 2023, ECCV 2024, CVPR 2025.

Recent Talks

01/2025 **A Multimodal and Unified Approach for Gaze Following and Social Gaze Prediction in Everyday Scenes**, *IIIT Bangalore and IISc Bangalore*

04/2024 **Visual Attention: From 3D Gaze to Social Gaze Inference in Everyday Scenes**, *University of Zurich*, LiRi/NCCR Lunchtime Talk

Teaching

2021, 2023 **Teaching Assistant**, *École polytechnique fédérale de Lausanne*
Course: (EE613) Machine Learning for Engineers

2022, 2023 **Teaching Assistant**, *UniDistance Suisse*
Course: (M06) Machine Learning