Learning Ego-Exo Visual Representations for Conversational Gaze Estimation

Anshul Gupta[1,2,3*] Yijun Qian[1] Ruohan Gao[4] Ishwarya Ananthabhotla[1]
Jean-Marc Odobez[2,3] Vamsi Krishna Ithapu[1] Calvin Murdock[1]
[1]Meta Reality Labs Research, USA [2]Idiap Research Institute, Martigny, Switzerland
[3]Ecole Polytechnique Fédérale de Lausanne, Switzerland [4]University of Maryland, USA

Abstract

Egocentric gaze estimation is essential for understanding human attention in first-person scenarios, with applications in augmented reality, social interaction analysis, and assistive technology. While most existing methods rely on a sequence of video frames, real-world hardware constraints often necessitate single-frame inference. In this work, we introduce novel approaches that leverage exocentric gaze information of other individuals in the scene to improve single-frame egocentric gaze estimation. During training, our method leverages simultaneous views from a pair of people to jointly learn ego-exo gaze representations, with the exo representations learned via self-supervision. During inference, the model can leverage the learned exo representations to improve egocentric gaze estimation from a single view. Our experiments demonstrate that single frame models can achieve strong egocentric gaze performance, our approaches enable effective learning of exocentric gaze representations, and that learning these representations leads to improved egocentric gaze predictions.

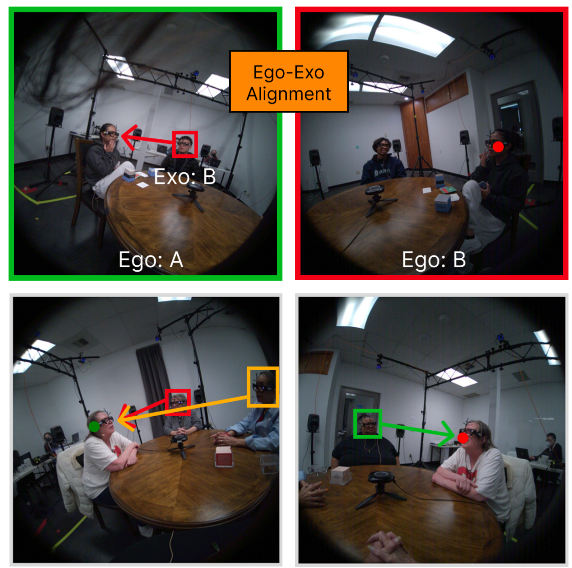

Figure 1. Estimating the egocentric gaze target of a person from a single frame is challenging, but can be improved using exocentric gaze cues of other individuals in the scene. During training, we employ a siamese-style architecture: one branch captures a person's ego gaze features (top right), the 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).

1. Introduction

Augmented Reality and Virtual Reality wearables have seen rapid advancements and widespread adoption in recent years, driven by innovative products such as Ray-Ban Meta glasses [34], Snapchat Spectacles [19], and Apple Vision Pro [18]. These devices promise immersive, gaze-driven interactions for enabling intuitive experiences. In particular, detecting gaze towards faces is crucial for conversation-aware applications. For example, the looked at person is often the auditory focus of attention [1, 36]. In noisy environments, recognizing this can help steer microphones toward the intended speaker and suppress background noise.

However, these devices are often constrained by current hardware limitations. Specifically, integrated eye-tracking systems, though accurate, typically incur complex calibration procedures, substantial power consumption and higher hardware costs, making them infeasible for many lightweight, cost-effective wearable platforms.

As a result, alternative approaches have emerged that estimate egocentric gaze using scene saliency and contextual cues from the wearer's visual environment. Recent work in this domain [16, 24, 43] has demonstrated that such methods can achieve competitive performance compared to traditional eye-tracking systems. However, these approaches predominantly use a sequence of video frames, leveraging temporal cues to improve accuracy. Yet temporal models can take an order of magnitude more compute and memory

*Work completed during an internship at Meta.

than static models, which may limit their practicality (see discussion in supplementary).

Motivation. In this work, we focus on the single-frame egocentric gaze estimation task. In particular, 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 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 (*e.g*. eye-contact, shared attention[1] towards a person) allowing us to significantly reduce 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 [17] 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 representation of a person with the exo representation of the *same person* as captured from another person's view (Figure 1, top).

We adopt a Siamese architecture where one branch captures egocentric gaze information and the other 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 1, bottom). This benefits practical use cases where simultaneous views of other people are often unavailable.

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 ego-exo alignment approaches: time synchronization, implicit matching and explicit matching. These approaches jointly learn ego and exo gaze representations, using self-supervision for the exo features. Our results show that these models improve egocentric gaze estimation, likely by utilizing the learned exocentric representations.
- *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.
- *Additional metrics for egocentric gaze estimation:* We propose a suite of metrics inspired from gaze following literature, to enable more comprehensive analysis of model performance.

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

2. Related Work

Egocentric Gaze Estimation. There have been several works for egocentric gaze estimation using deep learning [16, 17, 24, 27, 43, 44]. In particular, Huang et al. [17] demonstrated the benefit of jointly modelling egocentric gaze with hand actions. Lai et al. [24] proposed the first transformer based model for this task, achieving state of the art results. Meanwhile, Thakur et al. [44] incorporated auxiliary information in the form of IMU measurements for improved performance. In the context of joint attention, Park et al. [40] leveraged data from multiple egocentric views to reconstruct the 3D scene and predict joint attention based on social formation. A related task is that of egocentric gaze anticipation first explored by Zhang et al. [50]. Lai et al. [25] performed autoregressive gaze anticipation, leveraging audio information for improved performance. While these works have pushed the state of the art, they can still fail on cases where gaze following information could have otherwise helped disambiguate the target.

Gaze Following. Recasens et al. [35] first introduced this task, proposing a two-branch architecture for processing the head crop and the scene. This design was then continued by several follow-up works [3, 6, 10, 22, 23, 29, 41], with some leveraging additional inferred modalities such as pose [10], audio [15] and depth [2, 6, 10, 23, 41] for improved performance. More recently, transformer based architectures [9, 37, 39, 42] have achieved state of the art results. In particular, [9, 42] perform multi-person gaze following, with [9] additionally leveraging temporal information and jointly modelling social gaze. On the other hand, [37, 39] leverage higher resolution scene images to allow discarding the head crop branch. A small number of works have also attempted to tackle data limitations by leveraging pseudo labels [31, 33] extracted using pre-trained models. We do not explicitly target the gaze following task, but in-

[1]We provide definitions of key terms in the supplementary

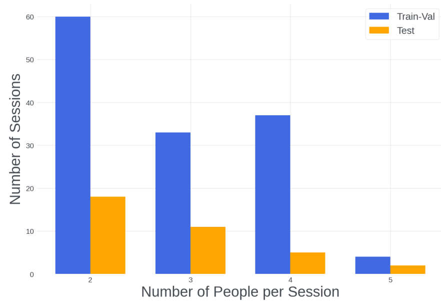

Figure 2. RLR-CHAT session distribution by number of people.

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

Table 1. RLR-CHAT number of frames and exo LAH statistics.

stead aim to learn gaze following representations via self-supervision for improving egocentric gaze estimation.

Ego-Exo Learning. This is a relatively nascent domain with works focusing on either person-matching between ego and exo views [5, 45, 46], learning of view-invariant features [28, 38, 47, 48], and conversational dynamics [20]. Our work shares similarities to [5, 46] as they also leverage contrastive losses to learn to associate the camera wearer in the egocentric view to the person in the exocentric view. However, our method learns to associate not just person identity, but also their gaze between ego and exo views. Another interesting work [48] also learns view invariant features, with the learned representations showing improvements for downstream applications including gaze angle prediction. However, they focus on scenes with a single person, and do not predict egocentric gaze or the gaze following target. Also, unlike all the above works, the exocentric view in our case does not come from a static camera but from another egocentric video which increases complexity.

3. The RLR-CHAT Dataset

The Reality Labs Research Conversations for Hearing Augmentation Technology (RLR-CHAT) dataset [14, 32, 49] is a large-scale collection of egocentric multisensory recordings captured from individuals engaging in natural conversations. Each conversation session is approximately one hour in duration and is recorded using Aria glasses [4], which capture RGB frames at 5Hz, 7-channel spatial audio at 48kHz, and eye-tracking data at 30Hz, among other modalities. Participants were recruited to ensure that no individual appeared in more than one session. Given the conversational setting, faces tend to be a common gaze target ($\sim$40% cases). 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 participant count is illustrated in Figure 2. The dataset contains a total of 170 sessions, the majority of which involve two participants.

A distinctive feature of RLR-CHAT is the synchronized availability of modalities from all participants within the conversation (examples in Figure 1). 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 [30], which is significantly smaller and primarily focuses on single-person activities.

We augment RLR-CHAT by incorporating head box detections and automatically assigning identities to these boxes using spatial audio cues (details in supplementary). The test set includes manually corrected head bounding boxes and high-quality, OptiTrack-based head identity matching. 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 (denoted as $\text{LAH}_{A\to B}$). This annotation process is applied to all pairs of individuals present in the scene. The resulting LAH statistics are summarized in Table 1.

4. Method

Our training architecture, illustrated in Figure 3, 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 V. $\mathbf{F}^A$ captures A's egocentric features, as well as others' exocentric features (including B) and vice-versa for $\mathbf{F}^B$. We then align the ego-exo gaze features of the same individual across views. We explore two approaches: (1) *Time Synchronization*, which encodes others' exocentric features within a single global representation, and (2) *Head Matching*, which encodes others' 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.

4.1. Feature Extraction

The feature extraction module is responsible for obtaining gaze-relevant features from an input egocentric image frame, denoted as $\mathbf{I}$. We employ a Vision Transformer (ViT)

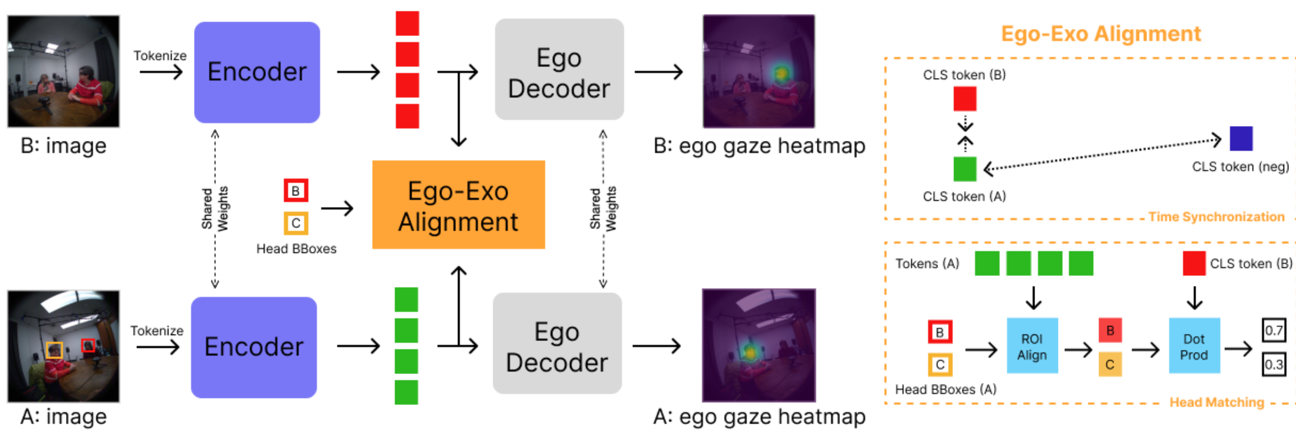

Figure 3. Our proposed architecture for ego-exo gaze representation learning. The Encoder first extracts features for each person's view, which are then aligned using one of two Ego-Exo Alignment techniques: (1) time synchronization or (2) head matching. Finally, the Ego Decoder processes the aligned features to predict each person's egocentric gaze heatmap. During inference, we leverage a single branch of the architecture, termed **EgoGazeViT**, for egocentric gaze estimation.

encoder, denoted as V, to extract these features. Specifically, we utilize the output from the last layer of the ViT.

$$\mathbf{F} = V(\mathbf{I}) \tag{1}$$

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 $\mathbf{F}^B$.

4.2. 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.

In practice, persons A and B are randomly sampled from a given timestamp. We explore two alignment approaches:

Time Synchronization. Inspired by prior work [48], we align egocentric features of two participants from the same timestamp in a session. Egocentric features for each person (e.g., person A) are obtained from the CLS token of the ViT output:

$$\mathbf{G}^A_{ego} = \text{CLS}(\mathbf{F}^A) \tag{2}$$

Here, the exocentric features for person A are directly captured in the egocentric features from person B:

$$\mathbf{G}^A_{exo} = \mathbf{G}^B_{ego} \tag{3}$$

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

$$S = \|\mathbf{G}^A_{ego} - \mathbf{G}^A_{exo}\|_2 \tag{4}$$

Our triplet loss (Section 4.4) encourages high similarity between egocentric features from the same timestamp. Simultaneously, it minimizes similarity against negative samples drawn from the 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 others' exo gaze information.

Head Matching. We explore a novel approach that aligns an individual's egocentric features with their exocentric features encoded locally in the region corresponding to their head box in another person's view. Specifically, given a participant B, we first extract exocentric features for all people visible in B's FoV (including A) using ROI-Align [13]:

$$\mathbf{G}^B_{exo} = \text{ROI Align}(\mathbf{F}^B, \mathbf{B}^B) \tag{5}$$

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

$$\mathbf{G}^A_{ego} = \text{CLS}(\mathbf{F}^A) \tag{6}$$

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

$$\mathbf{S}^A = \mathbf{G}^B_{exo} \cdot \mathbf{G}^A_{ego} \tag{7}$$

Our loss function (Section 4.4) 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 others' exocentric (in tokens corresponding to head box regions) gaze information.

4.3. Prediction

The Prediction Module processes the features from the Feature Extraction module, integrating both egocentric and others' 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 $\mathbf{F}$ from the Feature Extraction module, the egocentric gaze heatmap $\mathbf{H}$ is predicted as follows:

$$\mathbf{H} = D_{\text{ego}}(\mathbf{F}) \tag{8}$$

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 $\mathbf{H}^A, \mathbf{H}^B$.

4.4. Losses

Our training objective combines an egocentric gaze estimation loss ($\mathcal{L}_{\text{gaze}}$) and an ego-exo alignment loss ($\mathcal{L}_{\text{ego-exo}}$). $\mathcal{L}_{\text{gaze}}$ 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:

$$\mathcal{L} = \mathcal{L}^A_{\text{gaze}} + \mathcal{L}^B_{\text{gaze}} + \mathcal{L}_{\text{ego-exo}} \tag{9}$$

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

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

$$\mathcal{L}_{\text{ego-exo}} = \frac{e^{S^+}}{e^{S^+} + e^{S^-}} \tag{10}$$

where S^+ is the distance between matched ego features from simultaneous views, and S^- is the distance from $\mathbf{G}^A_{ego}$ 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. Specifically, 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 $\mathcal{L}_{\text{ego-exo}}$ is the sum of the losses computed independently for persons A and B.

5. Experiments

5.1. Datasets

We perform experiments on two datasets:

RLR-CHAT. The RLR-CHAT dataset, as described in Section 3, 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 [8]. It 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. [24], which consists of 27 approximately hour-long videos featuring 80 participants engaged in social interactions such as playing board games.

We selected Ego4D over alternative datasets such as EGTEA Gaze [26] and Aria [30], 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.

5.2. 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 heuristic baselines, as well as CNN and transformer-based approaches:

- **Heuristic Baselines:** Predicting the image center, using the average gaze point from training, 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 [21], 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 4, 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:** Explicitly aligns ego-exo features using ground truth head-box identities.

5.3. Metrics

Previous works on egocentric gaze estimation evaluate the quality of the predicted gaze heatmap by performing a pixel-level comparison against the generated ground truth heatmap after binarizing both of them [24] (referred to as the **Heatmap** metric). However, this approach is highly sensitive to the choice of the heatmap's standard deviation and the threshold used for binarization. Similar concerns have led to the avoidance of such metrics in recent gaze following research [9, 42]. Moreover, these metrics assess only localization accuracy, and may not always reflect semantic performance, which is often more valuable for downstream applications.

To address these limitations, we propose a new set of evaluation metrics inspired from gaze following [35, 41]:

- **Distance:** The predicted gaze point is obtained by taking the argmax of the predicted gaze heatmap. We then compute the L2 distance between the predicted and ground truth gaze points, normalized to a unit square (1×1). Both mean and median distances are reported.
- **Looking at Heads (LAH):** This semantic metric evaluates how well the model detects gaze directed at other people's heads. A prediction is classified as follows:
  - **True Positive:** Both the predicted and ground truth gaze points fall in the same head box.
  - **False Positive:** The predicted gaze point falls in a head box, but the ground truth does not.
  - **False Negative:** The ground truth gaze point falls in a head box, but the predicted gaze point falls on a different head or object.
  - **True Negative:** Neither the predicted nor the ground truth gaze points fall in a head box.

  We compute precision, recall, and F1-scores.

Model	Distance↓		LAH↑		
	Mean	Median	Prec	Recall	F1
Heuristic 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 (Standard Training)	**0.096**	**0.057**	0.507	0.798	0.620

Table 2. 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 3. Results for egocentric gaze estimation on the full RLR-CHAT golden subset. We leverage EgoGazeViT with different initializations. Best results are in bold.

6. Results

6.1. Egocentric Gaze Estimation Baselines

We trained and evaluated the baselines listed in Section 5.2 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 2.

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

6.2. Learning Exo Gaze Representations

We leverage the entire RLR-CHAT dataset by training our models on the designated train split and evaluating them on

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 4. 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/Model	Distance↓		Heatmap↑		
	Mean	Median	Prec	Recall	F1
Cross-Dataset Evaluation					
Standard training	0.174	0.151	0.260	0.520	0.347
Synchronization	**0.169**	**0.144**	**0.286**	0.506	**0.365**
Implicit Matching	0.183	0.154	0.259	**0.537**	0.349
Explicit Matching	0.170	0.147	0.274	0.493	0.352
Within-Dataset Evaluation					
GBVS [12]	-	-	0.111	0.472	0.180
Attention Transition [16]	-	-	0.295	0.476	0.364
I3D-R50 [7]	-	-	0.292	0.525	0.375
MViT [24]	-	-	0.317	**0.574**	0.409
GLC [24]	**0.156**	**0.123**	**0.347**	0.570	**0.431**
EgoGazeViT (Explicit Matching init)	0.163	0.131	0.315	0.562	0.404

Table 6. Results for egocentric gaze estimation on the Ego4D dataset. Best results are in bold.

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 (Table 3) indicate that performance on the Distance metric for all methods is comparable. However, Synchronization and Explicit Matching yield some improvements over Standard Training for the LAH metric, with Explicit Matching having the best performance.

As seen in Table 4, Explicit Matching consistently improves over Standard 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 eye contact, 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 4. Specifically, we freeze the trained encoder and train a new exocentric decoder D_{exo} to predict LAH labels for RLR-CHAT. D_{exo} 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 [9, 11]. For instance, to predict whether person B is looking at person C within person A's FoV, the model proceeds as follows:

$$\text{LAH}_{B\rightarrow C} = D_{\text{exo}}(\mathbf{G}^{B}_{exo}, \mathbf{G}^{C}_{exo}) \tag{11}$$

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

We present the results in Table 5. 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.

6.3. Evaluation on Ego4D

We provide cross-dataset evaluation results for our RLR-CHAT trained models on Ego4D in Table 6. 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 it may be overfitting to scene geometry in order to learn head-identity correspondences.

For comparison with state-of-the-art methods, we additionally train one of our models—EgoGazeViT initialized 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

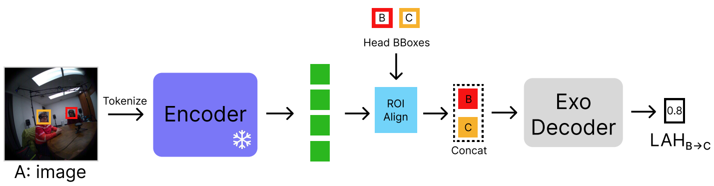

Figure 4. Architecture for probing learned exocentric gaze representations. We freeze the encoder, which was initially trained for egocentric gaze estimation, and train a 2-layer MLP probe to predict Looking at Heads (LAH).

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

Table 5. Results for exocentric gaze probing on the full RLR-CHAT golden subset. Best results are in bold.

Figure 5. 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.

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.

6.4. Qualitative Results

We present qualitative results for egocentric and exocentric gaze prediction in Figure 5, using the encoder initialized with Explicit Matching. 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 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 in the supplementary material.

7. Conclusion

In this work, we introduced novel self-supervised learning approaches for egocentric gaze estimation that leverage ego-exo alignment to learn exocentric gaze representations. Our methods improve egocentric gaze prediction in challenging single-frame setting across RLR-CHAT and Ego4D by 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.

References

[1] S.O. Ba and J.-M Odobez. Recognizing human visual focus of attention from head pose in natural meetings. *IEEE Trans. on System, Man and Cybernetics: part B, Cybernetics*, 39(1): 16–34, 2009. 1

[2] Jun Bao, Buyu Liu, and Jun Yu. 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, 2022. 2

[3] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5396–5406, 2020. 2

[4] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 3

[5] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5125–5133, 2017. 3

[6] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11390–11399, 2021. 2

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 7

[8] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 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, 2022. 5

[9] 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. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2, 6, 7

[10] Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. 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, 2022. 2

[11] Anshul Gupta, Samy Tafasca, Naravich Chutisilp, and Jean-Marc Odobez. 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, 2024. 7

[12] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19, 2006. 7

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[14] Seongsil Heo, Calvin Murdock, Michael Proulx, and Christi Miller. Gaze-enhanced multimodal turn-taking prediction in triadic conversations. In *Proceedings of Interspeech*, 2025. 3

[15] Yuqi Hou, Zhongqun Zhang, Nora Horanyi, Jaewon Moon, Yihua Cheng, and Hyung Jin Chang. Multi-modal gaze following in conversational scenarios. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1186–1195, 2024. 2

[16] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 754–769, 2018. 1, 2, 7

[17] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020. 2

[18] Apple Inc. Apple vision pro. `https://www.apple.com/apple-vision-pro/`, . Accessed: 2025-01-13. 1

[19] Snap Inc. Spectacles '24. `https://www.spectacles.com/spectacles-24`, . Accessed: 2025-01-13. 1

[20] Wenqi Jia, Miao Liu, Hao Jiang, Ishwarya Ananthabhotla, James M Rehg, Vamsi Krishna Ithapu, and Ruohan Gao. The audio-visual conversational graph: From an egocentric-exocentric perspective. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[21] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. 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, 2022. 6

[22] Tianlei Jin, Zheyuan Lin, Shiqiang Zhu, Wen Wang, and Shunda Hu. 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, 2021. 2

[23] Tianlei Jin, Qizhi Yu, Shiqiang Zhu, Zheyuan Lin, Jie Ren, Yuanhai Zhou, and Wei Song. Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence*, 113:104924, 2022. 2

[24] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision*, pages 1–18, 2023. 1, 2, 5, 6, 7

[25] Bolin Lai, Fiona Ryan, Wenqi Jia, Miao Liu, and James M Rehg. Listen to look into the future: Audio-visual egocentric gaze anticipation. In *European Conference on Computer Vision*, pages 192–210. Springer, 2025. 2

[26] Yin Li, Miao Liu, and James M Rehg. 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, 2018. 5

[27] Yin Li, Miao Liu, and James M Rehg. 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, 2021. 2

[28] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 3

[29] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018. 2

[30] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, Kiran Somasundaram, Luis Pesqueira, Mark Schwesinger, Omkar Parkhi, Qiao Gu, Renzo De Nardi, Shangyi Cheng, Steve Saarinen, Vijay Baiyya, Yuyang Zou, Richard Newcombe, Jakob Julian Engel, Xiaqing Pan, and Carl Ren. Aria everyday activities dataset, 2024. 3, 5

[31] Qiaomu Miao, Alexandros Graikos, Jingwei Zhang, Sounak Mondal, Minh Hoai, and Dimitris Samaras. Diffusion-refined vqa annotations for semi-supervised gaze following. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 2

[32] Calvin Murdock, Ishwarya Ananthabhotla, Hao Lu, and Vamsi Krishna Ithapu. Self-motion as supervision for egocentric audiovisual localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7835–7839. IEEE, 2024. 3

[33] Cheng Peng and Oya Celiktutan. Visual saliency guided gaze target estimation with limited labels. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9. IEEE, 2024. 2

[34] Ray-Ban. Ray-ban meta smart glasses. `https://www.ray-ban.com/usa/ray-ban-meta-smart-glasses`. Accessed: 2025-01-13. 1

[35] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *Advances in neural information processing systems*, 28, 2015. 2, 6

[36] Federico Rossano. Gaze in conversation. *The handbook of conversation analysis*, pages 308–329, 2012. 1

[37] Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M Rehg. Gaze-lle: Gaze target estimation via large-scale learned encoders. *arXiv preprint arXiv:2412.09586*, 2024. 2

[38] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. 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, 2018. 3

[39] Yuehao Song, Xinggang Wang, Jingfeng Yao, Wenyu Liu, Jinglin Zhang, and Xiangmin Xu. Vitgaze: Gaze following with interaction features in vision transformers. *arXiv preprint arXiv:2403.12778*, 2024. 2

[40] Hyun Soo Park and Jianbo Shi. Social saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785, 2015. 2

[41] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20935–20946, 2023. 2, 6

[42] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. 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, 2024. 2, 6

[43] Hamed Rezazadegan Tavakoli, Esa Rahtu, Juho Kannala, and Ali Borji. Digging deeper into egocentric gaze prediction. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 273–282. IEEE, 2019. 1, 2

[44] Sanket Kumar Thakur, Cigdem Beyan, Pietro Morerio, and Alessio Del Bue. Predicting gaze from egocentric social interaction videos and imu data. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 717–722, 2021. 2

[45] Yangming Wen, Krishna Kumar Singh, Markham Anderson, Wei-Pang Jan, and Yong Jae Lee. 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, 2021. 3

[46] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. 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, 2018. 3

[47] Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36:53688–53710, 2023. 3

[48] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. 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, 2020. 3, 4

[49] Heeseung Yun, Ruohan Gao, Ishwarya Ananthabhotla, Anurag Kumar, Jacob Donley, Chao Li, Gunhee Kim, Vamsi Krishna Ithapu, and Calvin Murdock. Spherical world-locking for audio-visual localization in egocentric videos. In *European Conference on Computer Vision*, pages 256–274. Springer, 2024. 3

[50] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. 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, 2017. 2