Looking at Hands: A new benchmark for Hand-Centered Gaze Understanding

Nawal Haidar[1,2], Anirban Mukherjee[1,2], and Jean-Marc Odobez[1,2]

[1] Idiap Research Institute, Martigny, Switzerland
[2] École Polytechnique Fédérale de Lausanne, Switzerland
{nhaidar, amukherjee, odobez}@idiap.ch

Abstract. Gaze following is typically evaluated through spatial target localization, but an accurate gaze point does not necessarily identify the hand involved in an interaction. We introduce Looking-at-Hands (**LAHands**), a multi-label task that identifies the specific hands associated with an observer's attention, including attention directed to actively manipulated objects. We provide **LAHands** annotations for GazeFollow, ChildPlay, and VideoAttentionTarget, together with calibrated pseudo-labels for large-scale GazeFollow training. We further propose **MTGS-Hands**, which augments multi-person gaze following with candidate-hand tokens encoding hand location, ownership, pose, and optional manipulation supervision. On GazeFollow, learned hand-aware association substantially outperforms the predicted-gaze geometric baseline while preserving competitive gaze localization. Experiments on ChildPlay further show that manipulation supervision can improve hand-attention prediction under domain adaptation. These results show that gaze-point proximity alone is insufficient for hand-directed attention prediction. The **LAHands** annotations will be available at `https://www.idiap.ch/dataset/lahands`.

Keywords: Gaze following · Hand-directed attention · Multi-person gaze and manipulation understanding

1 Introduction

Visual attention is a fundamental component of human behavior, reflecting the dynamic allocation of cognitive resources to task-relevant elements in the environment. As a result, gaze direction provides rich information about a person's intentions, goals, and future actions, making it one of the most informative cues for understanding human behavior in both individual and social contexts [14,31].

Human attention is expressed through coordinated behavioral cues, including gaze, head orientation, gestures, and hand actions. Gaze and hands are particularly closely coupled during object-directed behavior: people typically fixate relevant locations before reaching or contact, maintain attention on objects during manipulation, and shift their gaze toward upcoming action subgoals [18–20].

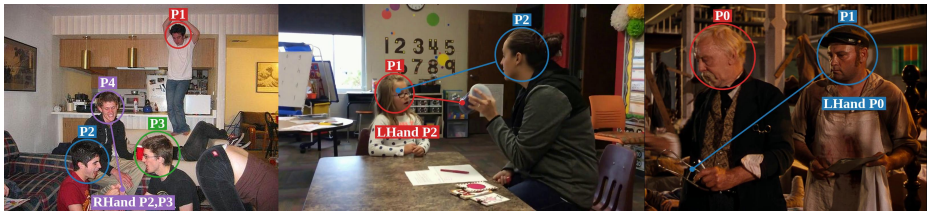


Fig. 1: Examples of our **LAHands** annotations on (left to right): GazeFollow, ChildPlay, and VideoAttentionTarget. **LHand P**k and **RHand P**k mean that the observer is looking at person Pk's left and right hand, respectively, or at an object actively held or manipulated by the corresponding hand. Multiple **LAHands** labels indicate attention involving more than one hand.

Importantly, this coordination is not only relevant for controlling one's own actions but also for understanding the actions of others. Observers use hand movements together with gaze cues to infer intentions, predict upcoming actions, and establish shared attention. Such mechanisms are also prominent for social attention, especially in child-caregiver interactions, where attention is frequently coordinated through objects being held or manipulated rather than through face-directed gaze alone [38].

In this paper, we address the gaze following task, which aims to predict the spatial target of a person's attention in an image or video [7,25,29]. Specifically, we investigate whether a person's gaze is directed toward hands involved in object manipulation or interaction.

Recent gaze-following models have steadily improved by incorporating richer contextual cues, including depth, human pose and body skeletons [13], and more recently by relying on large-scale vision foundation models [26]. However, despite these advances, most methods still formulate gaze following primarily as a *where* problem, with limited emphasis on identifying *what* behavioral entity is being attended.

Several works have begun moving in this direction. Social gaze models infer interaction-oriented relations such as eye contact and shared attention [11, 12], while semantic gaze-following methods associate gaze with object-level semantic entities [28]. Nevertheless, comparatively little work has addressed attention toward human activities and interactions.

This limitation becomes particularly evident in interaction-rich scenes involving multiple people, hands, and manipulated objects. A gaze prediction may lie near multiple hands or between a hand and the object it manipulates. Consequently, a spatially accurate prediction may still fail to determine which hand is being attended, whereas a slightly displaced prediction may nevertheless identify the correct hand-directed attention.

We therefore introduce **Looking-at-Hands (LAHands)**, a fine-grained social-attention task and a first step toward modeling attention in human interactions. For each observer and candidate hand, the task predicts whether that specific hand is being attended.

LAHands supports attention to one's own or other people's hands and multiple simultaneous hands. It also includes attention directed toward an object actively held or manipulated by a hand, which captures hand-mediated attention rather than simple gaze point-in-box overlap. Examples are shown in Figure 1.

To support this task, we provide **LAHands** annotations for GazeFollow [25], ChildPlay [29], and VideoAttentionTarget [7]. ChildPlay contains natural child–adult interactions and is complemented by ChildPlay-Hand, which provides per-hand manipulation labels [9]. The ChildPlay, VideoAttentionTarget and GazeFollow test sets have been manually annotated while the GazeFollow training set has been augmented with automatically geometrically generated **LAHands** pseudo-labels to enable large-scale training.

To benchmark the task, we propose several baselines. In particular, we introduce **MTGS-Hands**, a hand-aware extension of MTGS [12] that decodes attention towards hands in addition to conventional gaze following. This enables us to investigate whether incorporating hand cues can further improve gaze-following performance. More specifically, in addition to scene and person representations, **MTGS-Hands** introduces hand tokens encoding hand location, owner information, and, optionally, 3D hand-to-shoulder pose. A pairwise decoder predicts **LAHands** relations along the conventional gaze-following objective. Finally, we also investigate manipulation-aware hand representations through ChildPlay-Hand pretraining and an auxiliary manipulation objective during ChildPlay fine-tuning.

In summary, our contributions are:

- We introduce **LAHands**, a multi-label task for identifying directed attention toward hands or objects being actively manipulated.
- We provide **LAHands** annotations for GazeFollow, ChildPlay, and VideoAttentionTarget, and establish standardized training and evaluation protocols.
- We introduce **MTGS-Hands**, a joint gaze-following and **LAHands** model, and systematically study the impact of hand location, 3D pose, and manipulation-aware representations in both in-domain and cross-domain settings.

On GazeFollow, the **MTGS-Hands** hand-aware models substantially outperform predicted-gaze geometric association, whereas on ChildPlay, this geometric baseline remains competitive. In general, experiments show that **LAHands** is a relevant but challenging task, with room for improvements, and that 3D manipulation-aware hand representations can provide useful cues for prediction.

2 Related Work

Gaze following and social attention. Gaze following aims to localize the target of a person's attention from visual cues such as head appearance, head location, body pose, and scene context [6, 7, 25]. Subsequent works have incorporated additional scene geometry, object information, and multimodal cues to improve target localization [1, 10, 15, 35]. Recent approaches also move beyond the conventional single-person setting. Sharingan performs multi-person gaze following with a transformer architecture [30], while MTGS jointly models

multi-person temporal gaze following and social gaze relations such as looking at humans, mutual gaze, and shared attention [12]. Other work has similarly explored unified gaze-following and social-gaze prediction [11], as well as semantic characterizations of gaze targets [28]. These approaches enrich gaze prediction beyond a purely geometric formulation, but they primarily reason about spatial targets, objects, or person-level social relations. In contrast, **LAHands** introduces a finer person–hand relation that identifies the specific left or right hand involved in an observer's attention.

Gaze, hands, and action understanding. Behavioral studies have long established that gaze is closely coordinated with manual action. During object manipulation, gaze typically precedes the hand, fixates relevant contact and placement locations, and transitions toward future action subgoals [14, 18–20]. Eye–hand coordination also supports social attention. In child–caregiver interaction, participants can coordinate attention through objects held and manipulated by their hands, even without reliable face-directed gaze following [38]. This coupling has also been studied in relation to motor and developmental differences [3, 24, 32, 36].

Computational work has begun to exploit these dependencies in several related settings. Gaze has been used to improve human–object interaction prediction in videos [22] and to guide the synthesis of hand–object interactions [33]. In extended-reality environments, hand and head motion have been used as cues for gaze estimation [8, 16], while full-body pose has been investigated as a representation of eye–body coordination [17]. Related work on deictic gestures combines hand orientation, gaze, body context, and scene understanding to estimate gesture targets [4]. These studies demonstrate the predictive relationship between gaze, hands, and actions, but generally address gaze estimation, interaction recognition, or gesture grounding in settings different from third-person social gaze following. Our work instead asks which specific hand instance is attended in an unconstrained multi-person scene and learns this relation jointly with spatial gaze prediction.

Hand and manipulation representations. Large-scale hand datasets and interaction benchmarks have enabled increasingly detailed representations of hand location, pose, and activity [5, 27]. Of particular relevance, ChildPlay-Hand extends ChildPlay with per-hand manipulation annotations in natural child–adult interaction videos [9]. Its labels describe manual states and actions associated with each hand, providing supervision that is complementary to the gaze annotations of ChildPlay [29]. Broader action-recognition datasets also show the value of combining body pose, object context, and temporal structure for understanding manual activities [2, 21]. However, these datasets and methods focus on recognizing what the hands are doing rather than whether a particular hand is the target of another person's visual attention.

3 Looking-at-Hands Task and Annotations

3.1 Task Definition

We introduce **Looking-at-Hands (LAHands)**, a fine-grained social-attention task that identifies when a person's visual attention is directed toward a hand or an object actively manipulated by that hand.

Given the people $\mathcal{P} = \{p_i\}_{i=1}^{N_P}$ and hand instances $\mathcal{H} = \{h_j\}_{j=1}^{N_H}$ visible in a scene, **LAHands** assigns a label to each observer–hand pair:

$$y_{ij} = \begin{cases} 1, & \text{if person } p_i \text{ attends to hand } h_j, \\ 0, & \text{if person } p_i \text{ does not attend to hand } h_j, \end{cases} \tag{1}$$

Each hand is associated with an owner and a side; hence, the left and right hands of the same person are distinct target instances. An observer may attend to their own hand, another person's hand, or multiple hands simultaneously. A positive relation can denote attention to the hand itself or to an object actively held or manipulated by that hand.

3.2 Annotated Datasets

We construct **LAHands** annotations for GazeFollow [25], ChildPlay [29], and VideoAttentionTarget (VAT) [7]. Our annotations are built on the VSGaze [12] extensions of these datasets, using their unified multi-person annotations and person identities to construct observer–hand pairs. GazeFollow contains diverse static scenes, VAT contains video sequences from movies and TV shows, whereas ChildPlay contains natural child–adult interactions. ChildPlay is particularly suitable for studying hand-directed attention because its scenes frequently involve play, demonstrations, object exchanges, and joint manipulation. It is also accompanied by ChildPlay-Hand [9], which provides per-hand manipulation annotations in the same domain.

3.3 Human annotation protocol

LAHands annotations. We rely on the gaze targets provided by the original datasets. For each gaze annotation in a given frame, the annotator is shown the observer identity, their gaze location, and the head locations and identities of the other people in the scene. The annotator then selects, from the proposed list of hands, the hand or hands toward which the observer is directing their attention. The possible annotations include left and right self-hands, other people's hands, multiple simultaneous targets, and no-hand cases.

Selected hands are converted into positive observer–hand pairs, while non-selected hands are treated as negative pairs. The annotations were subsequently validated and corrected by the authors, showing good agreement between annotators. For instance, on the GazeFollow test set, only 2.5% of the evaluated pairs required correction. The ChildPlay, VAT, and GazeFollow test sets were annotated in this manner. Due to its size, the GazeFollow training set was pseudo-annotated using the calibrated labeling procedure described below.

Hand and body geometry. To obtain hand localization information, we applied the SAM 3D Body model [37] to all images and matched the predicted

Table 1: Source-level **LAHands** annotation statistics. A positive instance contains at least one attended hand. GazeFollow train and validation contain pseudo-labels; the remaining entries are human annotations.

Dataset	Split	Frames	Gaze inst.	Positive	Pos. frames
GazeFollow	Test	4,782	4,782	1,436	41.1%
ChildPlay	Train	100,209	211,059	104,941	67.1%
ChildPlay	Val	10,283	21,838	12,383	72.4%
ChildPlay	Test	9,986	25,030	13,111	69.3%
VAT	All	164,142	164,142	25,390	15.5%
GazeFollow	Train	103,455	108,955	26,684	25.1%
GazeFollow	Val	4,486	4,499	1,133	25.2%

head keypoints with the ground-truth head bounding boxes to associate the estimated body models with the person identities in the dataset. The resulting 3D body and hand keypoints were used both for hand bounding-box localization and to construct the hand-to-shoulder representation used by **MTGS-Hands**, comprising the hand keypoints together with the corresponding wrist, elbow, and shoulder joints.

3.4 Annotation statistics

Table 1 summarizes the scale of the **LAHands** annotations across datasets and splits, whereas Table 2 decomposes more precisely which hands are the focus of attention. The probability of attending at least one hand is computed over all gaze instances. The remaining columns describe the conditional distribution among **LAHands**-positive instances.

Hand-directed attention is considerably more frequent in the test split of ChildPlay than in that of VAT. In ChildPlay, almost half of the positive instances involve both hand sides, and attention is often directed toward another person's hands, reflecting the prevalence of play, object manipulation, and hand interactions, including hand-overs, as evidenced by the proportion of attention directed toward self-and-other hands (22%). In contrast, VAT contains fewer hand-directed gaze instances and exhibits a stronger tendency toward self-hand attention. These differences suggest that the annotations capture complementary behavioral settings.

Additional analysis in the supplementary material further relates **LAHands** to manual activity in ChildPlay-Hand. Hands are more frequently attended during interaction-heavy actions such as *operate*, *give*, and *put*, and when people hold an object, which motivates using manipulation-aware and object-aware representations for **LAHands** prediction.

3.5 Calibrated GazeFollow Pseudo-Labels

GazeFollow provides large-scale gaze annotations but no hand-level attention labels. Due to its large size, we adopt a pseudo-labeling approach using a trans-

Table 2: Breakdown of **LAHands** annotation targets for the ChildPlay and VAT test splits. "Any hand" is measured over all gaze instances; the remaining values are conditioned on **LAHands**-positive instances.

Dataset	Any hand	Right	Left	Both	Self	Other	Self+other
ChildPlay	0.523	0.281	0.237	0.482	0.322	0.456	0.222
VAT	0.113	0.301	0.262	0.430	0.641	0.322	0.035

parent geometric rule whose confidence thresholds are calibrated on human-annotated ChildPlay, as described below.

For an annotated gaze point $\mathbf{g}_i$ and candidate hand bounding box b_j, we compute the normalized distance

$$d_{ij} = \frac{\|\mathbf{g}_i - \mathbf{c}(b_j)\|_2}{\frac{1}{2}\,\mathrm{diag}(b_j)}, \tag{2}$$

where $\mathbf{c}(b_j)$ and $\mathrm{diag}(b_j)$ denote the center and diagonal length of the hand box, respectively. By normalizing with half of the box diagonal, the distance becomes relative to the apparent hand scale. Rather than relying on a single decision boundary, we introduce separate positive and negative confidence thresholds:

$$\tilde{y}_{ij} = \begin{cases} 1, & d_{ij} \le T_{\mathrm{pos}}, \\ 0, & d_{ij} \ge T_{\mathrm{neg}}, \\ -1, & T_{\mathrm{pos}} < d_{ij} < T_{\mathrm{neg}}. \end{cases} \tag{3}$$

The intermediate region is treated as uncertain and excluded from supervision.

Threshold calibration on ChildPlay. We sweep the normalized distance threshold over ChildPlay observer–hand training pairs and measure the precision of positive labels, $d_{ij} \le T$, and negative labels, $d_{ij} \ge T$, against human **LAHands** annotations. Based on Fig. 2, we set

$$T_{\mathrm{pos}} = 0.5, \qquad T_{\mathrm{neg}} = 1.25, \tag{4}$$

and mark intermediate distances as uncertain. These thresholds are transferred to GazeFollow without re-tuning.

Application to GazeFollow. For each GazeFollow observer, we enumerate candidate left and right hands, compute Eq. (2), and apply Eq. (3). Confident positive and negative pairs are retained, while intermediate and otherwise invalid pairs receive the unknown label. For the GazeFollow test analysis, where several independent gaze annotations are available for the same observer, we average the normalized distances before thresholding. GazeFollow train contains one gaze annotation per observer and therefore does not require aggregation.

Pseudo-Label Validation.

We validate the pseudo-labelling rule in two settings. First, we compare on the GazeFollow test how the rule would perform (using our manual annotations as ground-truth). Given that this test set has multiple gaze annotations, we

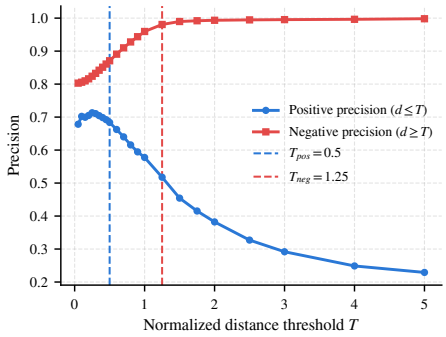


Fig. 2: Calibration of the distance-based pseudo-labeling rule on ChildPlay. Small normalized distances yield high-confidence positive labels, whereas large distances yield high-confidence negative labels. The interval between T_{pos} and T_{neg} is excluded from supervision.

Table 3: Human validation of GazeFollow pseudo-labels. Train results are computed and evaluated on the validated subset.

Evaluation set	Pos. precision	Neg. precision
GazeFollow test	0.783	0.977
GazeFollow train sample	0.730	0.989

relied on the mean to apply the pseudo-labeling rule. Second, on the training set, we manually inspected 1000 frames equally sampled from the positive and negative class. Results are given in Table 3. They confirm that positive-label quality decreases near T_{pos}, while negative labels remain highly reliable.

4 MTGS-Hands

We introduce **MTGS-Hands**, a hand-aware extension of MTGS [12] that jointly predicts gaze targets and **LAHands** relations. Given an image or short video clip containing N_P people, it constructs frame tokens, gaze tokens from head appearance and location, and separate left- and right-hand tokens combining hand-to-shoulder keypoints, hand boxes, and owner head boxes.

As shown in Fig. 3, these tokens are jointly contextualized through repeated transformer interaction blocks. Contextualized person tokens condition the gaze and in/out decoders, hand tokens support hand-center prediction, and person–hand pairs are used for **LAHands** prediction.

4.1 Hand Representation

Left and right hands are represented as separate instances. For hand h_j, let b_j^h be its bounding box, $b_{\text{own}(j)}^p$ its owner's head box, and $\mathbf{s}_j$ its hand-to-shoulder keypoints. A temporal hand processor encodes the pose sequence:

$$\mathbf{z}_j = E_{\text{hand}}(\mathbf{s}_{j,1:T}). \tag{5}$$

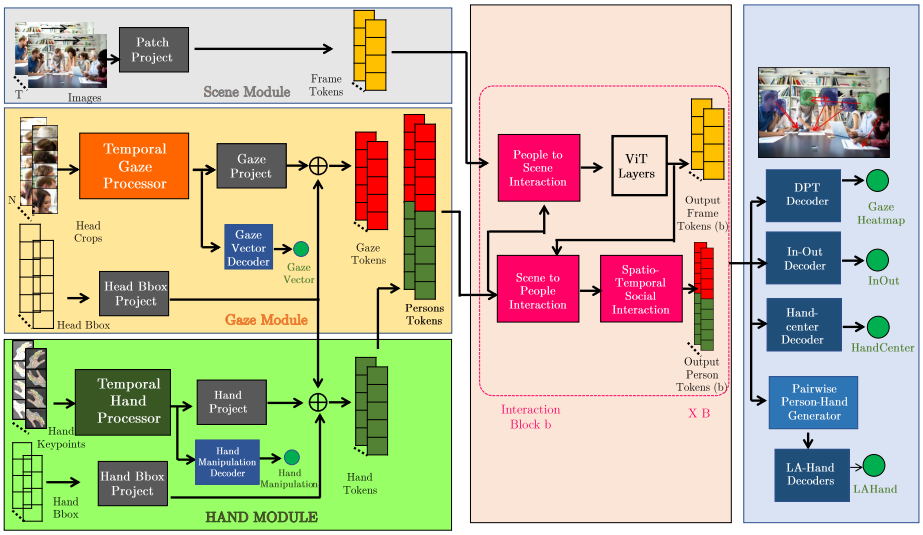


Fig. 3: Overview of **MTGS-Hands**. The scene module extracts frame tokens, while the gaze module constructs one gaze token per person from temporally processed head crops and head bounding boxes. The proposed hand module constructs one token for each left or right hand by combining encoded hand-to-shoulder keypoints, the hand bounding box, and the head bounding box of the hand owner. Frame, gaze, and hand tokens interact through repeated interaction blocks. Person tokens (gaze and hands) interact through self-attention, allowing to model social as well as manipulation aware interactions. Person tokens also are used to query scene information through cross-attention, allowing to identify salient gaze information (for the gaze token) or identify visually relevant hand information (e.g. holding objects) for the hand tokens. The contextualized gaze tokens condition the gaze heatmap and in/out decoders, contextualized hand tokens condition the hand-center decoder, and contextualized gaze–hand pairs are processed by the **LAHands** decoder. All reported experiments are static (the temporal length is $T = 1$).

The encoded pose is projected to the shared token dimension and combined with learned hand-box and owner-box projections:

$$\mathbf{h}_{j,t} = P_{\text{hand}}(\mathbf{z}_{j,t}) + P_{\text{hbox}}(b^{h}_{j,t}) + P_{\text{owner}}(b^{p}_{\text{own}(j),t}) + \mathbf{e}^{\text{side}}_{j}, \tag{6}$$

where $\mathbf{e}^{\text{side}}_{j}$ distinguishes left- and right-hand instances.

4.2 Gaze–Hand–Scene Interaction

At each time step, gaze and hand tokens form a shared social sequence:

$$\mathbf{S}_t = [\mathbf{g}_{1,t}, \ldots, \mathbf{g}_{N_P,t}, \mathbf{h}_{1,t}, \ldots, \mathbf{h}_{N_H,t}], \tag{7}$$

where $\mathbf{g}_{i,t}$ and $\mathbf{h}_{j,t}$ denote person-gaze and hand tokens. Repeated interaction blocks jointly model social-token interactions and bidirectional cross-attention with frame tokens. The final contextualized gaze tokens condition the gaze decoders, while contextualized gaze–hand pairs are used for **LAHands** prediction.

4.3 LAHands Prediction

For every valid observer–hand pair, we concatenate the contextualized gaze token of the observer with the contextualized candidate-hand token:

$$\mathbf{q}_{ij,t} = \left[\tilde{\mathbf{g}}_{i,t}; \tilde{\mathbf{h}}_{j,t}\right]. \tag{8}$$

A shared multilayer perceptron predicts the **LAHands** probability independently for each pair:

$$\hat{y}_{ij,t} = \sigma(f_{\mathbf{LAHands}}(\mathbf{q}_{ij,t})). \tag{9}$$

4.4 Manipulation-Aware Hand Encoder

To incorporate action-related information beyond hand geometry, we pretrain the 3D hand encoder on ChildPlay-Hand manipulation classification [9] using the five classes:

$$\mathcal{A} = \{\text{background/pointing}, \text{grasp}, \text{hold}, \text{operate-like}, \text{release}\}, \tag{10}$$

where *operate-like* groups the operate, give, put, drop, and throw labels of the original dataset, following the 5-class setting in [9]. The pretrained encoder initializes **MTGS-Hands**. We also apply an auxiliary manipulation classifier to the hand representation:

$$\hat{\mathbf{a}}_{j,t} = f_{\text{manip}}(\mathbf{z}_{j,t}), \tag{11}$$

that is optimized on ChildPlay using class-weighted cross-entropy, $\mathcal{L}_{\text{manip}}$.

4.5 Training Objective

For all valid observer–hand pairs, we optimize a focal binary cross-entropy loss:

$$\mathcal{L}_{\mathbf{LAHands}} = \frac{1}{|\Omega|} \sum_{(i,j,t)\in\Omega} \text{FL-BCE}\left(\hat{y}_{ij,t}, y_{ij,t}\right), \tag{12}$$

where $\Omega = \{(i,j,t) \mid y_{ij,t} \neq -1\}$ is the set of valid pairs. The focal weighting reduces the contribution of easy negative pairs and helps address the strong class imbalance in **LAHands**.

We additionally regress each valid hand's normalized center from its contextualized token:

$$\hat{\mathbf{c}}_{j,t} = f_{\text{center}}\left(\tilde{\mathbf{h}}_{j,t}\right), \tag{13}$$

using

$$\mathcal{L}_{\text{center}} = \frac{1}{|\mathcal{H}_{\text{valid}}|} \sum_{(j,t)\in\mathcal{H}_{\text{valid}}} \left\|\hat{\mathbf{c}}_{j,t} - \mathbf{c}_{j,t}\right\|_1. \tag{14}$$

where $\mathcal{H}_{\text{valid}}$ denotes the set of valid hand instances.

The complete objective combines the standard original gaze-following losses from [12] with the **LAHands** and auxiliary hand objectives:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{gaze}} + \lambda_{\text{io}}\mathcal{L}_{\text{io}} + \lambda_{\mathbf{LAHands}}\mathcal{L}_{\mathbf{LAHands}} \\ + \lambda_{\text{center}}\mathcal{L}_{\text{center}} + \lambda_{\text{manip}}\mathcal{L}_{\text{manip}}.\end{aligned} \tag{15}$$

We set $\lambda_{\mathbf{LAHands}} = 1$ and $\lambda_{\text{center}} = \lambda_{\text{manip}} = 0.1$ in the experiments.

5 Experiments

Even if the model supports temporal processing, in the following all our results rely on a static modeling approach (i.e. $T = 1$).

5.1 Experimental Setup

Datasets and evaluation settings. GazeFollow and ChildPlay are used in our experiments. VAT annotations are included in the release to support future work on video-based **LAHands**; benchmarking on VAT is left to future work. We train on GazeFollow and evaluate on its test split. For transfer evaluation, we test the GazeFollow-trained models directly on ChildPlay and after fine-tuning on the ChildPlay training split.

Model configurations. We compare *MTGS-Geom*, which applies geometric hand association to the gaze output of a hand-free MTGS model, with progressively richer **MTGS-Hands** variants:

- *Hand bbox* uses hand location and owner information;
- *+ 3D hand-to-shoulder* additionally uses upper-limb keypoints;
- *+ Pretrained hand encoder* initializes the 3D encoder from ChildPlay-Hand manipulation classification pretraining;
- *+ Manipulation auxiliary* exploits manipulation supervision during ChildPlay fine-tuning.

Implementation details. All experiments use 448×448 inputs and are trained for 20 epochs with AdamW. The DINOv2 encoder [23] is frozen. Hand-aware models use owner-associated hand boxes, $\lambda_{\textbf{LAHands}} = 1$, and $\lambda_{\text{center}} = 0.1$. ChildPlay fine-tuning uses batch size 8, four-step gradient accumulation, and learning rate 10^{-6}.

Candidate hands and detection coverage. SAM 3D Body [37] provides candidate hands for matched people, recovering 97.9% and 99.2% of annotated attended hands on GazeFollow and ChildPlay test, respectively. Undetected hands are excluded from the valid set of hands.

5.2 Evaluation Metrics and Baselines

Metrics. For GazeFollow, we report the standard heatmap AUC, minimum normalized distance to any ground-truth gaze annotation (*Min. Dist.*), and distance to their mean location (*Avg. Dist.*). For ChildPlay, we report normalized gaze distance. **LAHands** is evaluated by average precision over valid observer–hand pairs since the task is imbalanced.

Geometric baseline. The hand-free baseline predicts gaze only. **LAHands** scores are obtained by the negative normalized distance from the predicted gaze point to an extended hand center. Following [9], the center is shifted from the wrist along the forearm direction to better represent attention directed toward potential objects held or manipulated beyond the hand itself:

$$\mathbf{c}_j = \mathbf{w}_j + 0.5(\mathbf{w}_j - \mathbf{e}_j), \tag{16}$$

where $\mathbf{w}_j$ and $\mathbf{e}_j$ are the wrist and elbow locations. Distances are normalized by half the hand-box diagonal and scored independently for each observer–hand pair. We also report a ground-truth-gaze geometric reference.

Table 4: Results on GazeFollow. Published gaze-following results use their original settings, while our controlled models use 448×448 inputs. *MTGS-Geom* derives **LAHands** scores geometrically from the gaze output of our hand-free MTGS model; **MTGS-Hands** variants use a learned **LAHands** decoder. *Chance* is the mean AP of a random scorer over 200 seeds.

Method	AUC ↑	Avg. Dist. ↓	Min. Dist. ↓	**LAHands** AP ↑
Recasens et al. [25]	.878	.190	.113	–
Chong et al. [7]	.921	.137	.077	–
Tonini et al. [34]	.927	.141	–	–
Tafasca et al. [30]	.944	.113	.057	–
Ryan et al. [26]	.956	.104	.045	–
Human	.924	.096	.040	–
***LAHands** baselines*				
Chance	–	–	–	.13
MTGS-Geom	**.950**	.0944	.0410	.47
GT-gaze geometry	–	–	–	.57
MTGS-Hands				
Hand bbox	**.950**	**.0929**	**.0390**	.55
+ 3D hand-to-shoulder	.940	.0944	.0410	**.57**
+ Pretrained hand encoder	**.950**	.0936	.0399	**.57**

5.3 Main Results on GazeFollow

Table 4 shows that explicit hand representations improve person–hand attention prediction over geometric association based on predicted gaze. MTGS-Geom obtains an **LAHands** AP of .47, while the **MTGS-Hands** Hand-bbox variant reaches .55 and the 3D hand-to-shoulder and pretrained-hand-encoder variants reach .57. The latter match the GT-gaze geometric reference, despite not having access to ground-truth gaze at inference. These results show that predicted-gaze location alone is insufficient for identifying the attended hand, and learned hand-aware representations can close the gap between predicted-gaze geometry and association based on ground-truth gaze.

Adding 3D hand-to-shoulder information increases **LAHands** AP from .55 to .57, suggesting that structured upper-limb information provides useful cues beyond 2D hand location. Manipulation pretraining preserves the .57 AP while recovering most of the gaze-localization performance of the Hand-bbox variant.

The effect on conventional gaze localization is smaller and not monotonic. The Hand-bbox variant obtains the lowest average and minimum distances, while the **MTGS-Hands** variants remain competitive with the other gaze-following methods.

5.4 ChildPlay results

We evaluate three transfer settings. Zero-shot models are trained only on GazeFollow and evaluated directly on ChildPlay. The manipulation-pretrained model

Table 5: ChildPlay results. *Zero-shot* models use no ChildPlay training data. *Target-domain manipulation pretraining* uses manipulation labels from the ChildPlay-Hand training split, but no ChildPlay **LAHands** labels. *Fine-tuned* models use ChildPlay training **LAHands** annotations. MTGS-Geom derives **LAHands** scores from gaze-target/hand geometry. *Chance* is the mean AP of a random scorer over 200 seeds.

Method	Gaze Dist. ↓	**LAHands** AP ↑
Chance	–	.18
GT-gaze geometry	–	.81
Ryan et al (zero-shot) [26]	0.114	–
Ryan et al (fine-tuned) [26]	0.106	–
Human	0.048	–
Zero-shot from GazeFollow		
MTGS-Geom	**.0941**	**.62**
MTGS-Hands (Hand bbox)	.0966	.53
MTGS-Hands (+ 3D hand-to-shoulder)	.0967	**.62**
Target-domain manipulation pretraining		
MTGS-Hands (+ Pretrained hand encoder)	.0957	.63
*Fine-tuned on ChildPlay **LAHands***		
MTGS-Geom	**.0902**	.64
MTGS-Hands (Hand bbox)	.0917	.56
MTGS-Hands (+ 3D hand-to-shoulder)	.0945	.63
MTGS-Hands (+ Pretrained hand encoder)	.0918	.61
MTGS-Hands (+ Manipulation auxiliary)	.0927	**.65**

additionally uses manipulation labels from the ChildPlay-Hand training split, but no ChildPlay **LAHands** labels. Fine-tuned models are initialized from their GazeFollow checkpoints and adapted using ChildPlay training **LAHands** labels.

Table 5 shows that, among models trained only on GazeFollow, the Hand-bbox variant transfers poorly, while adding 3D hand-to-shoulder information restores **LAHands** AP to the MTGS-Geom level. Manipulation pretraining on the ChildPlay-Hand training split further increases AP to .63 without using ChildPlay **LAHands** labels. After fine-tuning, gaze distance improves for all corresponding configurations. MTGS-Geom retains the lowest gaze distance, while the Manipulation-auxiliary variant obtains the highest AP among the learned **MTGS-Hands** variants. In general, these results show that in such close interaction settings, improving gaze following could improve **LAHands** predictions (*cf*. the .81 AP using GT-gaze geometry), and that hand and manipulation aware models can lead to better results. This could be achieved using temporal models.

5.5 Qualitative Results

Figure 4 shows two examples using the best-performing model in terms of **LA-Hands** AP for each evaluation setting. For GazeFollow, we visualize the model with the pretrained hand encoder. For ChildPlay, we visualize the fine-tuned

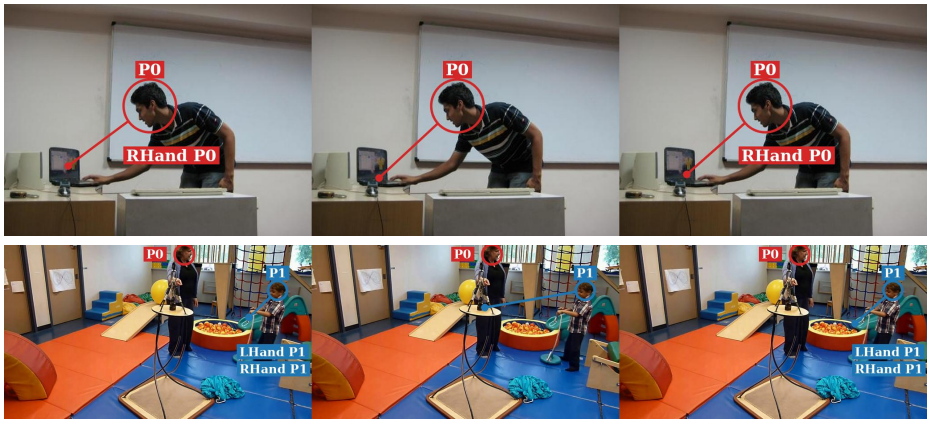


Fig. 4: Qualitative comparison on GazeFollow (top) and ChildPlay (bottom). Each row shows the ground-truth annotation, the MTGS-Geom prediction, and the **MTGS-Hands** prediction. On GazeFollow, both methods localize gaze near the target, but only **MTGS-Hands** identifies the correct attended hand. On ChildPlay, **MTGS-Hands** improves gaze localization and correctly predicts both attended hands.

model with the auxiliary manipulation objective. In the GazeFollow example, MTGS-Geom and **MTGS-Hands** localize gaze near the same target, but MTGS-Geom fails to identify the attended hand, whereas **MTGS-Hands** predicts the correct **LAHands** relation. In the ChildPlay example, **MTGS-Hands** improves the gaze prediction and also recovers the correct attended hands. These examples show that **LAHands** errors can arise either from gaze-localization error or from the limitations of a fixed geometric association rule. Additional examples are provided in the supplementary material.

6 Conclusion

We introduced Looking-at-Hands (**LAHands**), a fine-grained task for identifying the specific hands associated with an observer's attention, together with annotations for GazeFollow, ChildPlay, and VideoAttentionTarget and calibrated GazeFollow training pseudo-labels. We also proposed **MTGS-Hands**, which jointly predicts spatial gaze targets and observer–hand relations.

On GazeFollow, explicit hand representations substantially improve **LAHands** prediction over predicted-gaze geometric association, with the strongest variants matching the ground-truth-gaze geometric reference, while preserving competitive gaze localization. On ChildPlay, geometric association remains competitive, while manipulation supervision provides the strongest learned **LAHands** result. These findings show that explicit hand-aware reasoning complements spatial gaze localization and motivate future work on temporal, object-aware, and manipulation-aware modeling.

Acknowledgement. This research has been supported by the TUNASBE project (Towards a Unified Understanding of Attention and Social Behaviors in the Wild, grant no. 10002186) of the Swiss National Science Foundation (SNSF).

References

1. Bao, J., Liu, B., Yu, J.: ESCNet: Gaze target detection with the understanding of 3D scenes. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14106–14115 (2022). https://doi.org/10.1109/CVPR52688.2022.01373
2. Ben-Shabat, Y., Yu, X., Saleh, F., Campbell, D., Rodriguez-Opazo, C., Li, H., Gould, S.: The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 846–858 (2021). https://doi.org/10.1109/WACV48630.2021.00089
3. Bloch, C., Tepest, R., Jording, M., Vogeley, K., Falter-Wagner, C.M.: Intrapersonal synchrony analysis reveals a weaker temporal coherence between gaze and gestures in adults with autism spectrum disorder. Scientific Reports **12**, 20417 (2022). https://doi.org/10.1038/s41598-022-24605-8
4. Cao, X., Virupaksha, P., Lee, S., Lai, B., Jia, W., Chen, J., Rehg, J.M.: Toward human deictic gesture target estimation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
5. Cheng, T., Shan, D., Hassen, A., Higgins, R., Fouhey, D.: Towards a richer 2d understanding of hands at scale. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 30453–30465. Curran Associates, Inc. (2023). https://doi.org/10.52202/075280-1327
6. Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., Rehg, J.M.: 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) (September 2018)
7. Chong, E., Wang, Y., Ruiz, N., Rehg, J.M.: Detecting attended visual targets in video. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5395–5405 (2020). https://doi.org/10.1109/CVPR42600.2020.00544
8. Emery, K.J., Zannoli, M., Xiao, L., Warren, J., Talathi, S.S.: Estimating gaze from head and hand pose and scene images for open-ended exploration in vr environments. In: 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). pp. 554–555 (2021). https://doi.org/10.1109/VRW52623.2021.00159
9. Farkhondeh*, A., Tafasca*, S., Odobez, J.M.: Childplay-hand: A dataset of hand manipulations in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2024), * Equal contribution
10. Guan, J., Yin, L., Sun, J., Qi, S., Wang, X., Liao, Q.: Enhanced gaze following via object detection and human pose estimation. In: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II. p. 502–513. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-37734-2_41
11. Gupta, A., Tafasca, S., Chutisilp, N., Odobez, J.M.: A unified model for gaze following and social gaze prediction. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–9 (2024). https://doi.org/10.1109/FG59268.2024.10581955
12. Gupta, A., Tafasca, S., Farkhondeh, A., Vuillecard, P., Odobez, J.M.: Mtgs: A novel framework for multi-person temporal gaze following and social gaze prediction. Advances in Neural Information Processing Systems **37**, 15646–15673 (2024)

13. Gupta, A., Tafasca, S., Odobez, J.M.: 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 Workshops (CVPRW). pp. 5041–5050 (2022)
14. Hayhoe, M., Ballard, D.: Eye movements in natural behavior. Trends in Cognitive Sciences **9**(4), 188–194 (2005). https://doi.org/10.1016/j.tics.2005.02.009
15. Hou, Y., Zhang, Z., Horanyi, N., Moon, J., Cheng, Y., Chang, H.J.: Multi-modal gaze following in conversational scenarios. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1175–1184 (2024). https://doi.org/10.1109/WACV57701.2024.00122
16. Hu, Z., Haeufle, D., Schmitt, S., Bulling, A.: Hoigaze: Gaze estimation during hand-object interactions in extended reality exploiting eye-hand-head coordination. SIGGRAPH Conference Papers '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3721238.3730692
17. Hu, Z., Xu, J., Schmitt, S., Bulling, A.: Pose2gaze: Eye-body coordination during daily activities for gaze prediction from full-body poses. IEEE Transactions on Visualization and Computer Graphics **31**(9), 4655–4666 (2025). https://doi.org/10.1109/TVCG.2024.3412190
18. Johansson, R., Westling, G., Bäckström, A., Flanagan, J.: Eye–hand coordination in object manipulation. The Journal of neuroscience : the official journal of the Society for Neuroscience **21**, 6917–32 (10 2001). https://doi.org/10.1523/JNEUROSCI.21-17-06917.2001
19. Land, M., Mennie, N., Rusted, J.: The roles of vision and eye movements in the control of activities of daily living. Perception **28**, 1311–28 (02 1999). https://doi.org/10.1068/p2935
20. Land, M.F., Hayhoe, M.: In what ways do eye movements contribute to everyday activities? Vision Research **41**(25), 3559–3565 (2001). https://doi.org/10.1016/S0042-6989(01)00102-X
21. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(10), 2684–2701 (2020). https://doi.org/10.1109/TPAMI.2019.2916873
22. Ni, Z., Valls Mascaró, E., Ahn, H., Lee, D.: Human–object interaction prediction in videos through gaze following. Computer Vision and Image Understanding **233**, 103741 (2023). https://doi.org/10.1016/j.cviu.2023.103741
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
24. Rane, D., Singh, M., Lahiri, U.: Automated assesment of eye-hand coordination skill using a vertical tracing task on a gaze-sensitive human computer interaction platform for children with autism. Proceedings of the ACM on Human-Computer Interaction **8**, 1–18 (05 2024). https://doi.org/10.1145/3655598
25. Recasens*, A., Khosla*, A., Vondrick, C., Torralba, A.: Where are they looking? In: Advances in Neural Information Processing Systems (NIPS) (2015), * indicates equal contribution
26. Ryan, F., Bati, A., Lee, S., Bolya, D., Hoffman, J., Rehg, J.M.: Gaze-lle: Gaze target estimation via large-scale learned encoders. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28874–28884 (2025). https://doi.org/10.1109/CVPR52734.2025.02689

27. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9866–9875 (2020). https://doi.org/10.1109/CVPR42600.2020.00989
28. Tafasca, S., Gupta, A., Bros, V., Odobez, J.M.: Toward semantic gaze target detection. Advances in Neural Information Processing Systems **37**, 121422–121448 (2024)
29. Tafasca*, S., Gupta*, A., Odobez, J.M.: Childplay: A new benchmark for understanding children's gaze behaviour. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20935–20946 (October 2023), * Equal contribution
30. Tafasca, S., Gupta, A., Odobez, J.M.: Sharingan: A transformer architecture for multi-person gaze following. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2008–2017 (June 2024)
31. Tatler, B.W., Hayhoe, M.M., Land, M.F., Ballard, D.H.: Eye guidance in natural vision: Reinterpreting salience. Journal of Vision **11**(5), 5:1–5:23 (2011). https://doi.org/10.1167/11.5.5
32. Thorup, E., Nyström, P., Gredebäck, G., Bölte, S., Falck-Ytter, T.: Reduced alternating gaze during social interaction in infancy is associated with elevated symptoms of autism in toddlerhood. Journal of Abnormal Child Psychology **46**, 1547–1561 (03 2018). https://doi.org/10.1007/s10802-017-0388-0
33. Tian, J., Ji, R., Yang, L., Ni, S., Ma, B., Xu, L., Yu, J., Shi, Y., Wang, J.: Gaze-guided hand-object interaction synthesis: Dataset and method. IEEE Transactions on Multimedia pp. 1–11 (01 2026). https://doi.org/10.1109/TMM.2026.3710081
34. Tonini, F., Beyan, C., Ricci, E.: Multimodal across domains gaze target detection. In: Proceedings of the 2022 International Conference on Multimodal Interaction. p. 420–431. ICMI '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3536221.3556624
35. Tonini, F., Dall'Asen, N., Beyan, C., Ricci, E.: Object-aware gaze target detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21860–21869 (2023)
36. Wilmut, K., Wann, J., Brown, J.: Problems in the coupling of eye and hand in the sequential movements of children with development coordination disorder. Child: care, health and development **32**, 665–78 (12 2006). https://doi.org/10.1111/j.1365-2214.2006.00678.x
37. Yang, X., Kukreja, D., Pinkus, D., Sagar, A., Fan, T., Park, J., Shin, S., Cao, J., Liu, J., Ugrinovic, N., Feiszli, M., Malik, J., Dollar, P., Kitani, K.: Sam 3d body: Robust full-body human mesh recovery. arXiv preprint arXiv:2602.15989 (2026)
38. Yu, C., Smith, L.B.: Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. PLOS ONE **8**, 1–10 (11 2013). https://doi.org/10.1371/journal.pone.0079659

Supplementary Material for Looking at Hands: A new benchmark for Hand-Centered Gaze Understanding

1 Relationship Between LAHands and Hand Activity

We further analyze how hand-centered attention relates to manual activity in ChildPlay-Hand. Figure 1 reports, under two complementary views, the probability that a given hand is marked as attended by at least one person in the **LAHands** annotations.

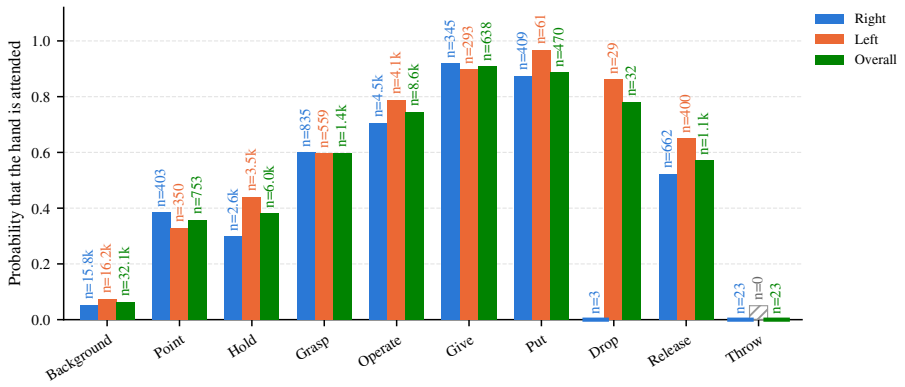
The per-class analysis in Fig. 1a shows that a hand is more likely to be attended during interaction-heavy manipulation classes such as *give*, *put*, and *operate*, whereas this probability is lower for *hold*, *point*, and background instances. This indicates that hand-directed attention is closely associated with active manual interaction rather than merely with the presence of a visible hand.

The object-based analysis in Fig. 1b shows a higher probability of attention for hands holding an object than for hands without an object. The same pattern is observed for left and right hands. These results motivate richer manipulation-aware and object-aware representations for **LAHands** prediction.

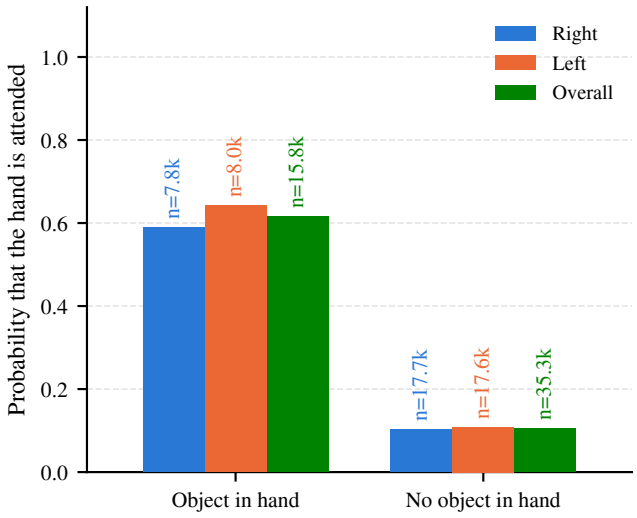
2 Additional Qualitative Results

Figure 2 shows additional successful cases in scenes containing several nearby hands. MTGS-Geom produces false positives for hands that are close to the predicted gaze target in the image plane but are not involved in the observer's attention. It can also miss the correct hand when the attended hand lies slightly away from the gaze prediction. In contrast, **MTGS-Hands** more accurately recovers the complete set of attended hands by reasoning jointly about the observer, candidate hands, and surrounding scene. In these examples, **MTGS-Hands** also improves gaze localization over the hand-free MTGS model underlying MTGS-Geom.

Figure 3 presents some failure cases. In some examples, **MTGS-Hands** predicts an incorrect or incomplete set of attended hands, even when MTGS-Geom succeeds. Other cases reveal inconsistencies between the predicted gaze location and the **LAHands** output, suggesting that the two prediction heads do not always produce aligned decisions. We also observe examples in which **MTGS-Hands** degrades gaze localization relative to the hand-free MTGS model. These cases motivate richer object- and manipulation-aware representations and stronger consistency between the gaze and **LAHands** predictions.



(a) Probability that a hand is attended, according to the **LAHands** annotations, conditioned on the manipulation class of that hand.



(b) Probability that a hand is attended, according to the **LAHands** annotations, conditioned on whether that hand holds an object.

Fig. 1: Relationship between **LAHands** and manual activity in ChildPlay-Hand. For each hand instance, we measure the probability that it is marked as attended by at least one person in the **LAHands** annotations, either by its owner or by another observer. Right and Left condition on the manipulation label or object-in-hand status of the corresponding hand side, while Overall pools individual hand instances from both sides. Sample counts are reported above each bar; hatched placeholders marked with $n = 0$ indicate unavailable estimates.

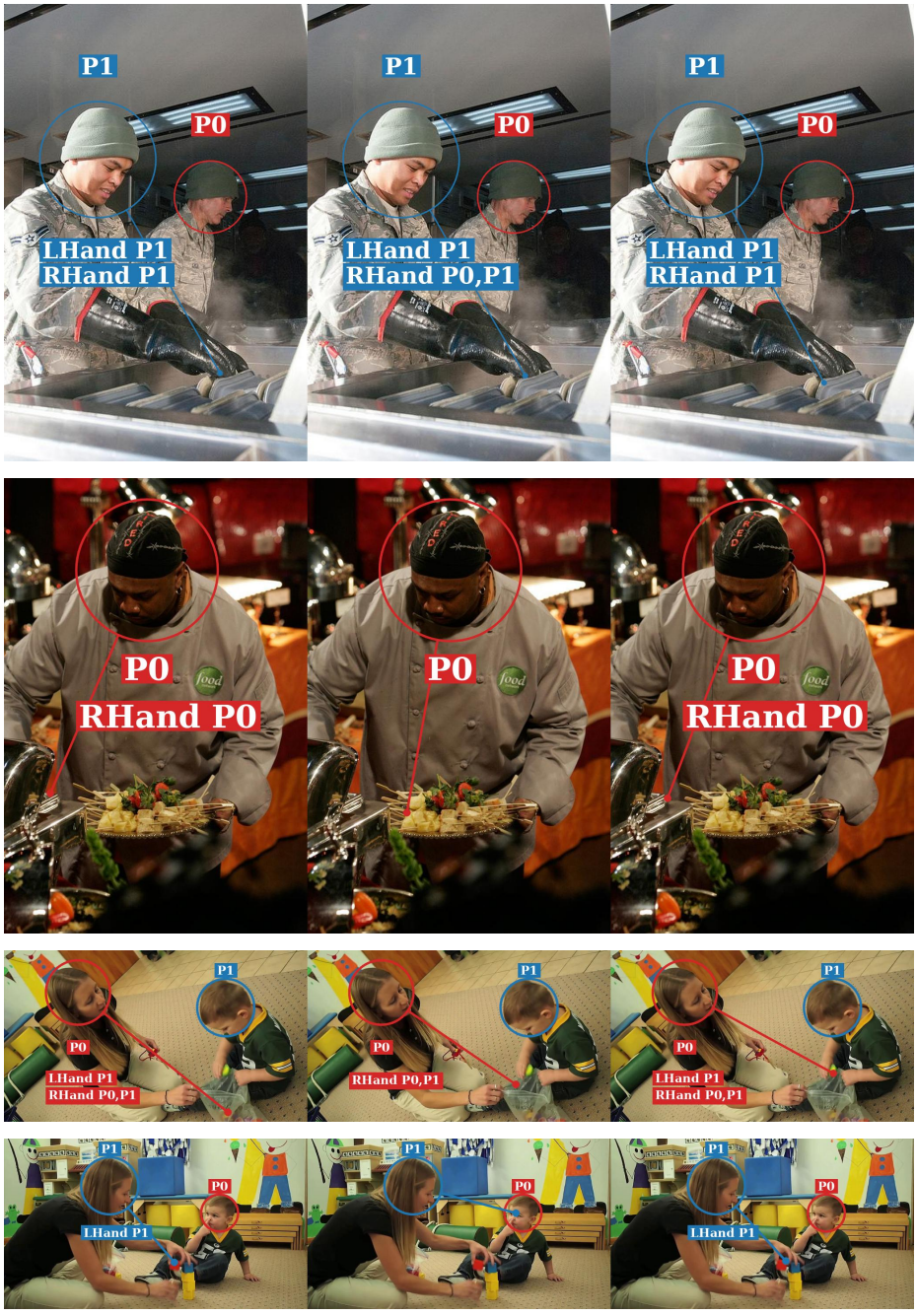


Fig. 2: Additional successful qualitative results on GazeFollow (rows 1–2) and Child-Play (rows 3–4). Each row shows, from left to right, the ground-truth annotation, the MTGS-Geom prediction, and the **MTGS-Hands** prediction. **MTGS-Hands** more accurately identifies the attended hands in scenes with multiple nearby candidates.



Fig. 3: Example failure cases of **MTGS-Hands** on GazeFollow (rows 1–2) and ChildPlay (rows 3–4). These examples include incorrect observer–hand associations, cases in which MTGS-Geom performs better, and inconsistencies between gaze localization and **LAHands** prediction. Each row shows, from left to right, the ground-truth annotation, the MTGS-Geom prediction, and the **MTGS-Hands** prediction.