

Learning Ego-Exo Visual Representations for Conversational Gaze Estimation

Supplementary Material

1. Head Bounding Box Identification

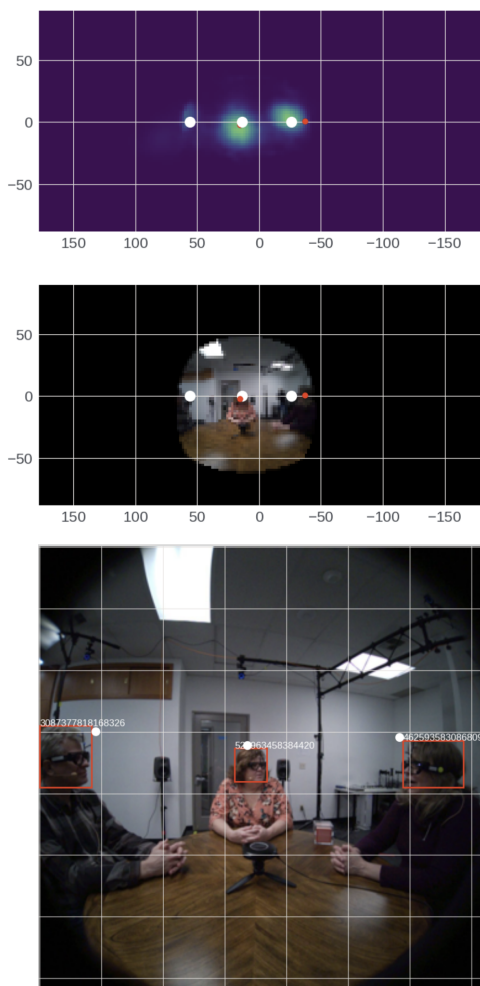


Figure 1. Overall pipeline for identifying head bounding boxes. We first obtain the average locations of other participants (white dots) in the conversation in 360 degrees world-locked coordinates. These are then mapped to the head-locked FoV coordinates and matched to the nearest head bounding box within a threshold.

To reliably identify the head bounding boxes of individuals visible in a participant’s field of view (FoV), we leverage a pre-trained MAV-ASL model [4] to obtain active speaker

heatmaps for each egocentric image frame. The MAV-ASL model produces two types of heatmaps: one that indicates the direction of the active speaker over a full 360-degree span, and another that provides the 2D location of the active speaker when present in the FoV. Both heatmaps are initially computed in a head-locked coordinate system.

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

Subsequently, we match the detected head bounding boxes to these averaged locations by selecting the nearest match within a threshold of 200 pixels. Although this thresholding process means that not all head bounding boxes are assigned an identity, the matches that are made have been verified to be of high quality. The overall pipeline is illustrated in Figure 1.

2. Discussion

Implicit Matching vs. Explicit Matching. In Figure 2, we examine how well the Implicit Matching method learns to align egocentric and exocentric gaze features. The model successfully matches features in several cases (top row), demonstrating its ability to capture meaningful ego-exo correspondences. However, it also exhibits failure cases (bottom row), where mismatches occur. In some cases, such as the bottom right example, failure is expected because the corresponding exocentric person is outside the field of view. However, the model also fails in other scenarios (e.g., bottom left), indicating that implicit matching alone may not always be sufficient for robust alignment.

Impact of number of people. Table 1 provides a detailed breakdown of egocentric gaze estimation performance on different splits of the RLR-CHAT Golden Subset, based on the number of participants in the included sessions. As expected, performance generally declines in sessions with a higher number of people due to the increased number of potential gaze targets and the resulting complexity of the task. Notably, there is an apparent spike in performance for sessions with 5 participants; however, since this

Subset	Initialization	Distance		Precision	LAH	
		Mean	Median		Recall	F1
Full	Standard Training	0.102	0.057	0.538	0.819	0.650
	Synchronization	0.100	0.055	0.536	0.843	0.656
	Implicit Matching	0.101	0.056	0.533	0.833	0.650
	Explicit Matching	0.101	0.055	0.545	0.836	0.660
≥ 3 people	Standard Training	0.111	0.067	0.524	0.790	0.630
	Synchronization	0.110	0.064	0.519	0.815	0.634
	Implicit Matching	0.111	0.065	0.512	0.805	0.626
	Explicit Matching	0.110	0.065	0.532	0.803	0.640
≥ 4 people	Standard Training	0.110	0.074	0.466	0.754	0.576
	Synchronization	0.107	0.069	0.461	0.771	0.578
	Implicit Matching	0.111	0.074	0.438	0.750	0.553
	Explicit Matching	0.106	0.069	0.473	0.773	0.587
≥ 5 people	Standard Training	0.096	0.054	0.542	0.820	0.653
	Synchronization	0.090	0.049	0.546	0.849	0.664
	Implicit Matching	0.101	0.058	0.524	0.796	0.632
	Explicit Matching	0.092	0.052	0.554	0.827	0.664

Table 1. Evaluation results on different splits of the RLR-CHAT Golden Subset based on the number of people in the session. Best results for each split are given in bold.

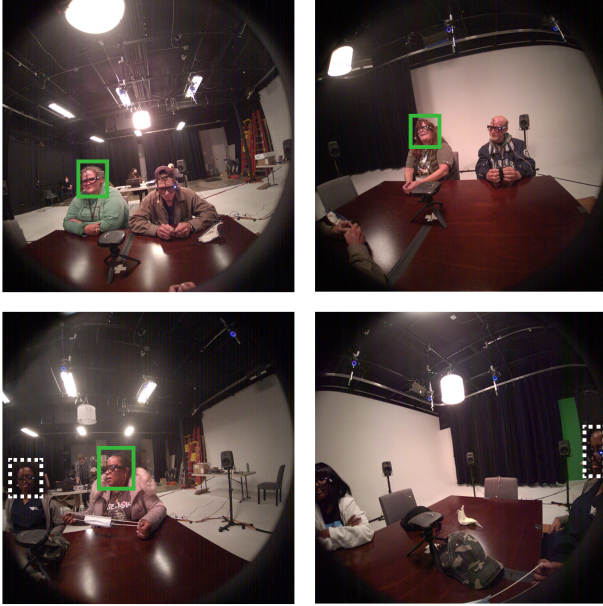


Figure 2. Egocentric and exocentric feature alignment results for the Implicit Matching approach. The correct exocentric person is highlighted with a green box. In the top row, these are correctly selected by the model. Incorrect selections made by the model are indicated with dotted white boxes in the bottom row.

split comprises only 2 sessions, the result is likely subject to high variance and may not be representative.

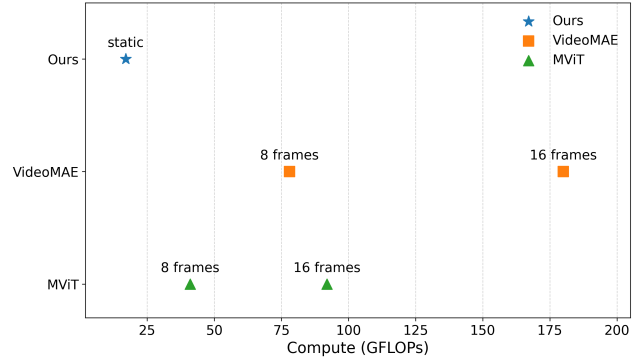


Figure 3. FLOPs comparison of our model against video-based models. Our model requires significantly fewer FLOPs while still achieving competitive performance.

Impact of using exocentric views at inference. Since there is no feature sharing between the two branches of our siamese architecture, using exocentric views does not improve performance during inference. During training, the architecture enables self-supervised learning of exocentric representations through *ego-exo alignment*, but at inference each branch operates independently. The architecture was deliberately designed this way to avoid the need for additional views during inference.

Memory and Compute of Static vs. Temporal Models. As motivated in the paper introduction, static models provide benefits for practical applications as they require

lower compute and memory compared to temporal models. First, temporal models require maintaining a frame buffer. For a model processing even just 8 frames, this multiplies storage requirements by $8\times$. They also incur substantially higher computational costs. In Figure 3, we compare the FLOPs consumed by our model, VideoMAE-based models [7], and MViT-based models [2].

We observe that all video-based models consume significantly more compute than our static model, which is reasonable as they process a larger number of tokens. MViT, used in [6], employs a more efficient attention mechanism and thus requires less compute than VideoMAE. Increasing the number of frames from 8 to 16 further raises compute requirements. Actual FPS would then depend on the hardware and platform on which these models are deployed.

3. Qualitative Comparisons

We provide qualitative comparisons of our models for egocentric gaze prediction in Figures 4 and 5. In Figure 4, all models perform similarly, as these cases involve either a single salient target or a target positioned near the image center, reducing task complexity.

In contrast, Figure 5 highlights scenarios where our ego-exo alignment approaches improve egocentric gaze prediction. These cases are more ambiguous, featuring multiple salient targets near the center, making gaze estimation more challenging.

- *Row 1:* The Standard Training and Implicit Matching approaches incorrectly identify the target person.
- *Row 2:* The Synchronization and Implicit Matching approaches struggle to differentiate between two people. The Standard Training approach confidently selects the wrong target, whereas the Explicit Matching approach correctly identifies the target with high confidence.
- *Row 3:* The Standard Training and Synchronization approaches misidentify the target, while the Implicit Matching and Explicit Matching approaches show uncertainty between two possible targets.
- *Row 4:* The Standard Training approach produces a diffused heatmap due to confusion, whereas the ego-exo alignment approaches correctly select the target. The Explicit Matching approach still exhibits some uncertainty.

These results suggest that ego-exo alignment helps disambiguate complex scenarios by leveraging exocentric gaze cues, leading to more precise egocentric gaze predictions.

4. Training and Evaluation

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

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

5. Implementation Details

The model uses a ViT-B encoder initialized with masked autoencoder (MAE) pretraining [3]. The Prediction Module consists of four transformer layers, each with a token dimension of 384—half the dimension of the ViT tokens. Input images are processed at a resolution of 224×224 , and the predicted gaze heatmap is generated at the same resolution, following the MAE architecture. The ground truth gaze heatmap \mathbf{H}_{gt} is constructed by placing a Gaussian centered at the gaze point, with a standard deviation of 9.35 pixels. It is converted to two channels ($\mathbf{H}_{gt}, 1 - \mathbf{H}_{gt}$) to facilitate training using the cross-entropy loss detailed in Section 4.4.

6. Glossary

We provide definitions for key terms used in the paper.

- **Ego view:** For a person A wearing augmented reality glasses, this refers to the scene as observed from their own perspective.
- **Exo view:** This refers to the third-person observation of person A from another person’s (e.g., person B ’s) perspective.
- **Looking at heads:** Defined for a pair of people. A person’s gaze point falls within another person’s head bounding box.
- **Eye contact:** Defined for a pair of people. A symmetric version of looking at heads, where the gaze points of both people fall within each other’s head box.
- **Shared attention:** Defined for two or more people. Occurs when multiple people look at the same object or person. In RLR-CHAT, this typically refers to multiple people looking at the same other person.

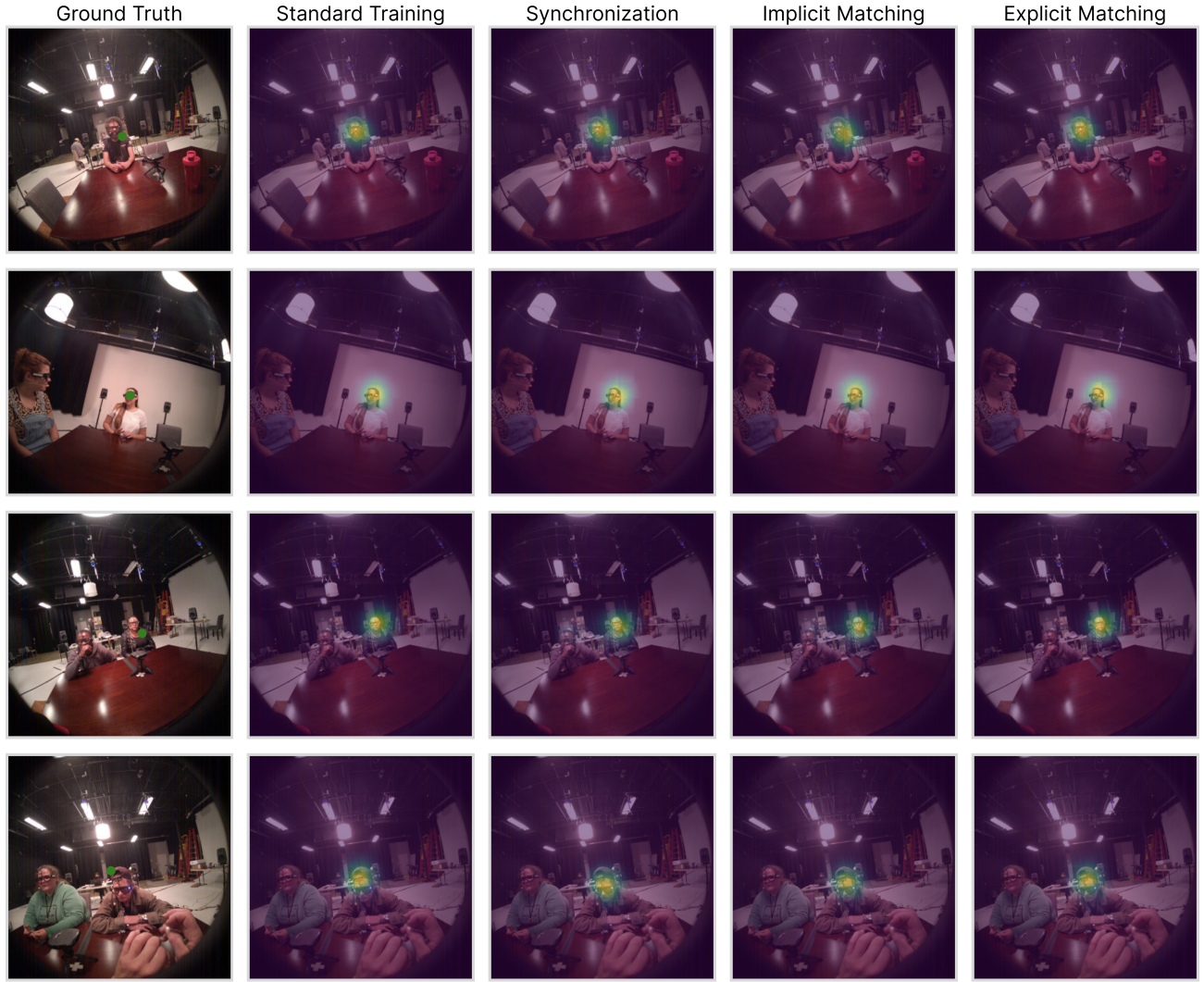


Figure 4. Qualitative results on images from RLR-CHAT where all approaches yield similar predictions. In these cases, either a single salient target dominates the scene, making gaze estimation straightforward, or the gaze target is near the image center, reducing ambiguity across models.

- **CLS token:** Standard transformer architectures, particularly the Vision Transformer (ViT) [1] used in our work, include an additional learnable token called the *CLS* token. This token is appended to the extracted image tokens to capture global information and is often used for tasks, such as classification, that require holistic image understanding.

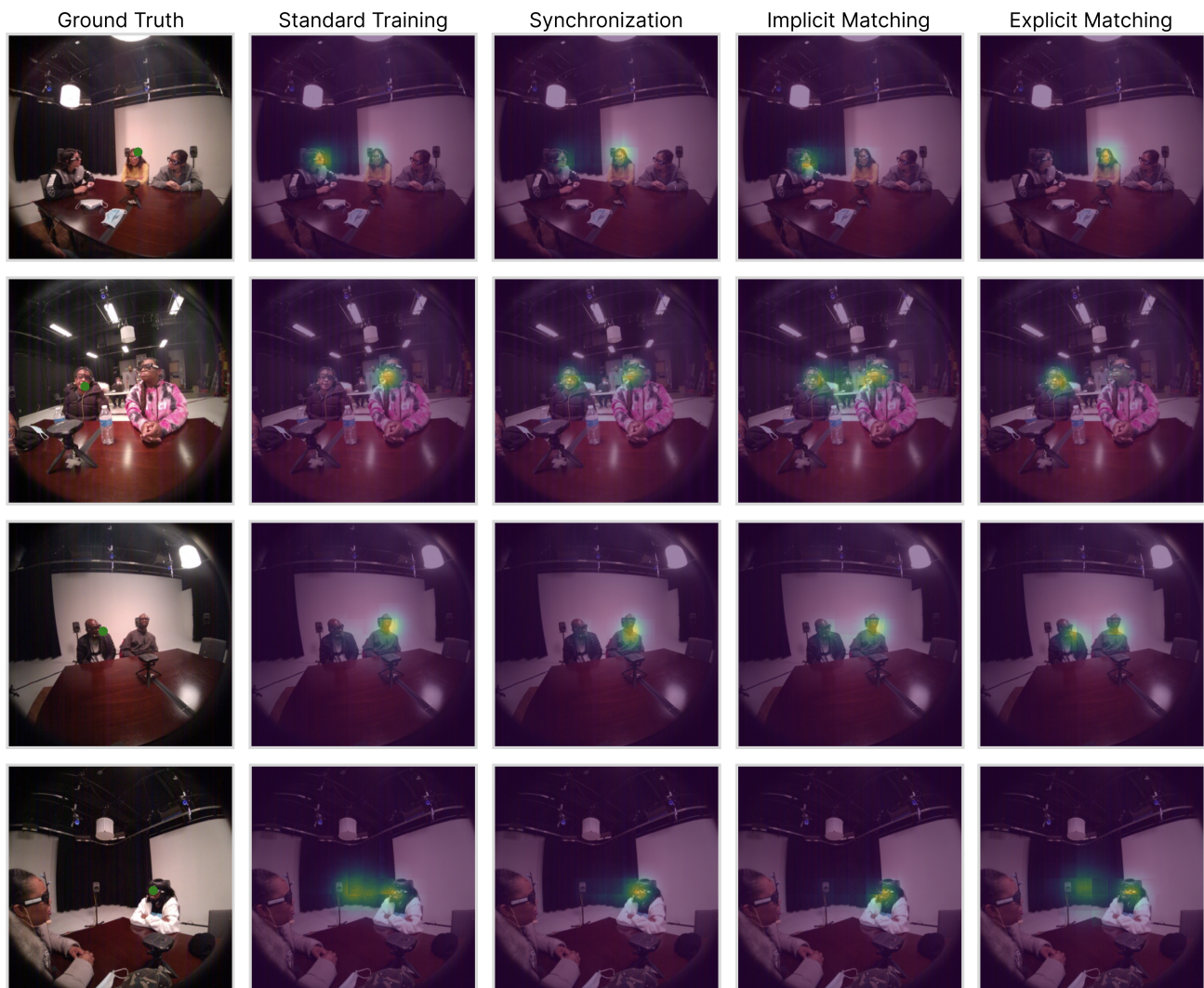


Figure 5. Qualitative results on images from RLR-CHAT where our proposed ego-exo alignment approaches improve egocentric gaze prediction compared to standard training. These cases involve greater ambiguity, with multiple salient targets positioned near the center, making gaze estimation more challenging.

References

- [1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021. [4](#)
- [2] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021. [3](#)
- [3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. [3](#)
- [4] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Ego-centric 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. [1](#)
- [5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [3](#)
- [6] 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. [3](#)
- [7] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Video-mae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. [3](#)