GAZE @ CVPR 2026

Learning Ego-Exo Visual Representations for Conversational Gaze Estimation

Anshul Gupta Yijun Qian Ruohan Gao Ishwarya Ananthabhotla
Jean-Marc Odobez Vamsi Krishna Ithapu Calvin Murdock

Meta

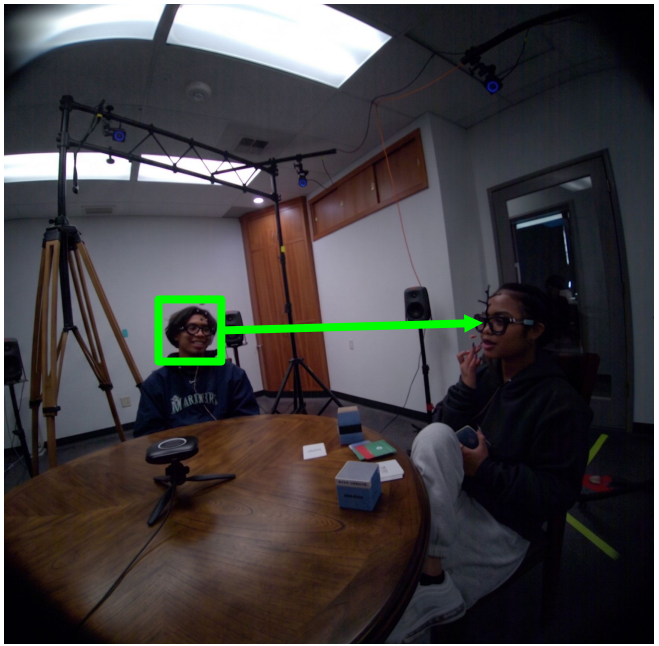
idiap RESEARCH INSTITUTE

EPFL

UNIVERSITY OF MARYLAND 18 56

Other people's gaze can tell you where the wearer is looking

Where is the person looking?



Ambiguity

In a scene with several people in view, predicting which person the wearer looks at is often ambiguous.

Insight

Other people's gaze — eye contact, shared attention — is a strong cue for the wearer's own gaze target.

Can we improve egocentric gaze estimation by learning others' exo gaze representations through self-supervision?

Contributions

01 Ego-exo alignment

Three SSL approaches — Time Sync, Implicit Match, Explicit Match — jointly learn ego and exo gaze from paired views to improve single-view egocentric gaze estimation.

02 Probing for exo gaze

Probes confirm the encoder genuinely captures exo gaze representations, significantly improving over standard training.

03 Single-frame is competitive; spatial audio can help

Modern ViT/CNN encoders match temporal baselines without sequence inputs. Spatial audio further helps by identifying the active speaker.

04 New evaluation metrics

LAH-based precision / recall / F1 alongside distance for richer analysis.

Prior work shows auxiliary signals help

Method	Input	Extra signal
Huang et al. [1]	video	hand actions
Thakur et al. [2]	video	IMU
Lai et al. [3] (GLC)	video	—
Lai et al. [4]	video	audio (single-channel)
Ours	***single frame***	**exo gaze (SSL), multi-channel audio**

[1] Huang et al., TIP 2020. [2] Thakur et al., ICMI 2021. [3] Lai et al., IJCV 2023. [4] Lai et al., ECCV 2024.

Auxiliary signals help.

Hand actions, IMU, and audio have all been shown to improve performance — room for new signals.

Our direction: new signals.

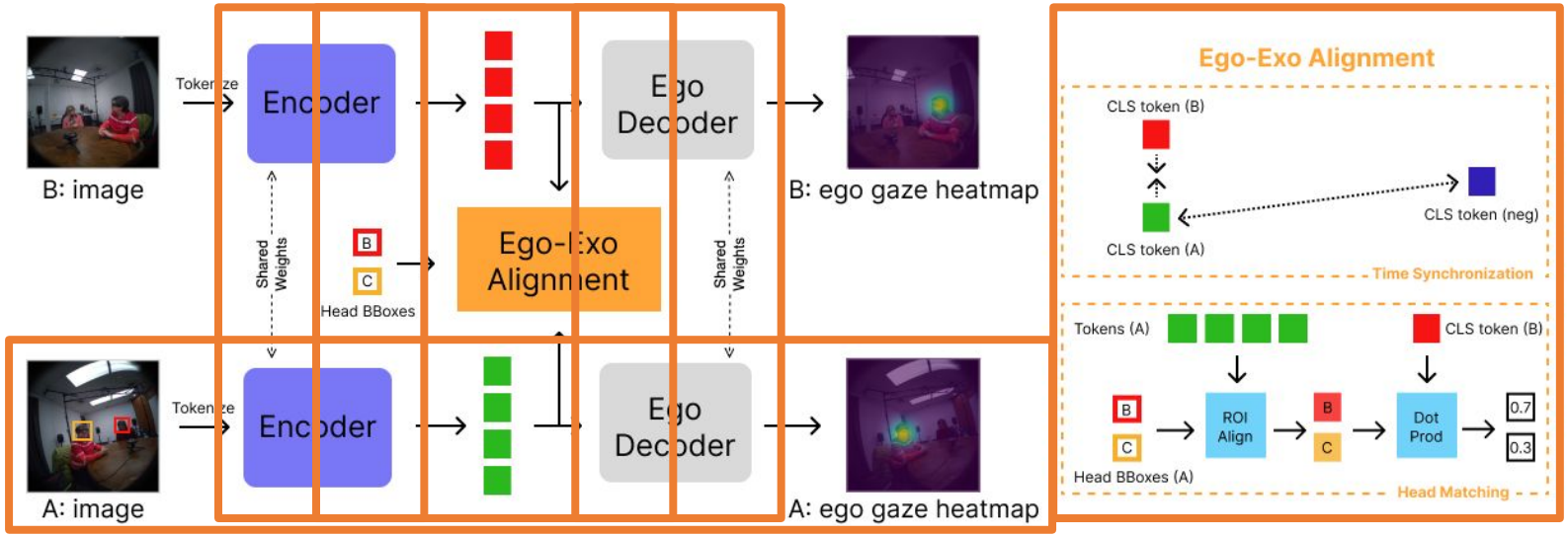
Exo gaze via self-supervision; preliminary results with multi-channel audio.

Sequences dominate.

Every major prior method relies on video, paying an order-of-magnitude memory/compute cost.

NOTE: Ego-exo alignment is not specific to single-frame. We study it here because the setting is practical and, as we'll show, competitive — but the idea extends naturally to temporal models.

Siamese ego-exo alignment



1 **Self-supervised exo**

Person B's ego features supervise B's exo features (from A's view) — no exo gaze labels needed.

2 **Symmetric, shared weights**

Same encoder captures both ego and other people's exo gaze.

3 **Single view at inference**

Using a single branch.

Three ego-exo alignment variants

Time Synchronization

Align: A's and B's CLS tokens at the same timestamp.

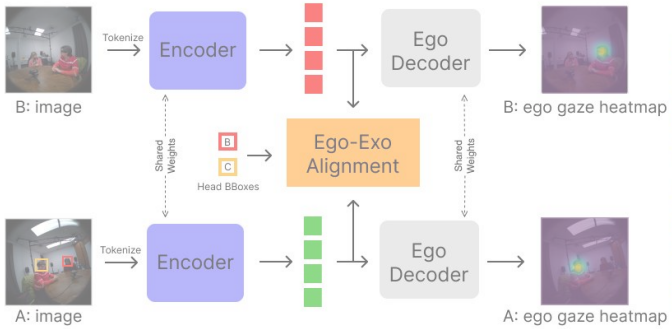
Loss: Triplet loss; negatives sampled from the batch.

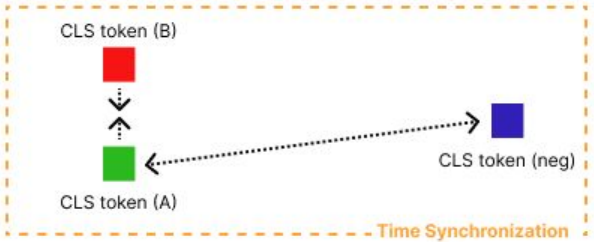
Head Matching

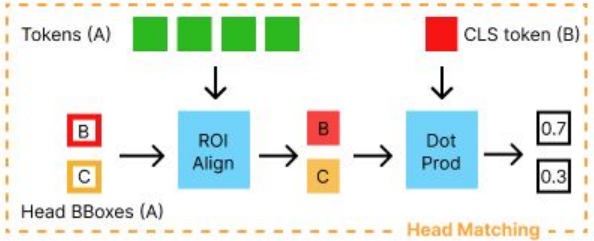
Align: A's CLS token vs ROI-aligned head-box features in B's view.

Loss: Maximize similarity on matched head; minimize otherwise.

Variants: **Implicit** (no head IDs) and **Explicit** (GT head IDs).



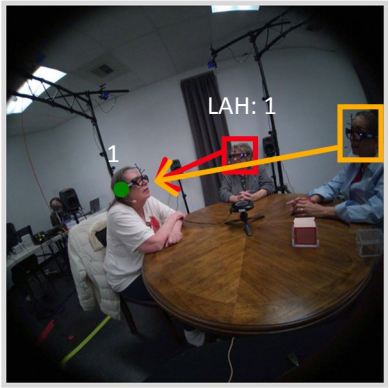




RLR-CHAT dataset

RLR-CHAT num frames and exo LAH stats

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



170

conversation sessions

~1 hour each, 2-5 people (mostly 2)

~40%

of gaze targets are faces

Participants recruited to ensure no individual appears in more than one session.

What enables ego-exo alignment

Simultaneous egocentric views from every participant — RGB at 5Hz, eye-tracking at 30Hz, 7-ch spatial audio.

Train/val: Head identities assigned via spatial audio

Test: OptiTrack head tracking + manually corrected head bounding boxes.

Ego gaze + head ID mapped to other views to obtain exo gaze labels.

https://ai.meta.com/datasets/rlr-chat/

Gaze-following inspired metrics

Gaze point = *argmax of predicted heatmap.*

LAH (Looking At Heads)

Semantic metric: precision, recall, and F1 for detecting gaze directed at other people's heads.

FN case

Distance

L2 distance between predicted and ground-truth gaze points, on a normalized unit square.

Recasens et al. (2015). Where are they looking?
Gupta et al. (2024). MTGS: A novel framework for multi-person temporal gaze following and social gaze prediction.

Ego-exo alignment helps egocentric gaze

Within-dataset (RLR-CHAT)

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**

Cross-dataset (Ego4D)

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

All three SSL variants improve or match standard training on RLR-CHAT, with Explicit Matching showing consistent gains across scene crowding.

Synchronization shows the strongest cross-dataset generalization on Ego4D.

Implicit Matching underperforms cross-dataset — likely overfits to scene geometry on RLR-CHAT.

The encoder genuinely learns exocentric gaze

LAH AP — frozen-encoder probe

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

↑ ***~2× over standard training.*** *Time Synchronization yields the strongest exo representations.*

The SSL objective works.

Strong exocentric gaze representations emerge through ego-exo alignment.

Future work: scaling gaze-following labels.

Paired egocentric data could generate exo gaze annotations automatically — labels that are otherwise expensive.

Where the bottleneck isn't.

Modest ego gains alongside strong exo gains suggest headroom lies elsewhere — data, new methods.

Single-frame methods are strong

Within-dataset results (Ego4D)

Initialization/Model	Distance↓		Heatmap↑		
	Mean	Median	Prec	Recall	F1
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

EgoGazeViT trained on Ego4D (Explicit Matching init). Improves over several temporal baselines.

Not far behind temporal SoTA, despite using a single video frame.

Spatial audio can help

Subset of RLR-CHAT

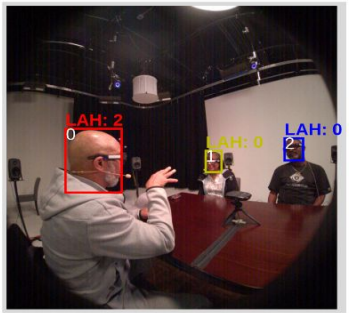
Model	Distance↓		LAH↑		
	Mean	Median	Prec	Recall	F1
Heuristic Baselines					
Predict center	0.107	0.093	0.633	0.146	0.237
Predict avg of train data	0.105	0.092	**0.638**	0.130	0.216
Predict closest head to center	0.131	0.073	0.396	**0.863**	0.543
CNN Baselines					
U-Net	0.105	0.072	0.520	0.610	0.561
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

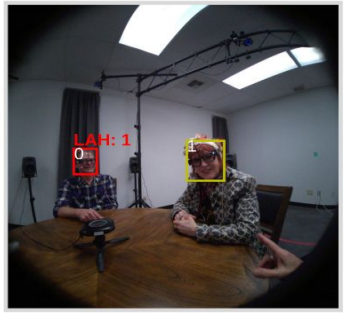
MAV-Gaze (image + 7-channel audio) significantly improves over image-only U-Net.

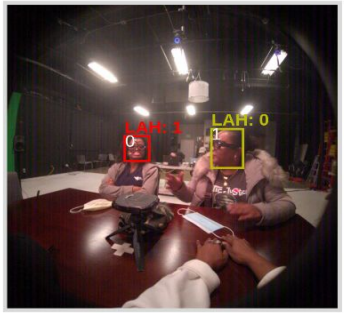
Achieves best LAH F1 — audio helps identify the active speaker.

Qualitative Results

Ground Truth
Prediction







Takeaways

01

Ego-exo SSL improves single-frame egocentric gaze.

02

The encoder learns real exocentric representations — a path to scaling gaze-following labels.

03

Temporal modeling and spatial audio are promising next directions.

Thank you!

Poster #306

3:25-4:15 pm

Exhibit Hall A

Paper

Anshul Gupta · anshul.gupta@tri.global

Examples where ego-exo alignment methods improve over standard training