LEARNING WORLD-SPACE REPRESENTATIONS OF HUMAN ATTENTION FROM UNCALIBRATED VIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human gaze target estimation (GTE) remains challenging due to the limited camera field of view. While multi-view setups can mitigate this issue, prior attempts either operate in fixed camera setups or heavily rely on explicit bounding boxes. In this work, we present GazeWorld, a unified framework dedicated to learning world-space representations of human attention directly from uncalibrated multi-view images in the wild. By constructing a unified 3D-aware representation that jointly encodes scene geometry and semantics, GazeWorld leverages pre-trained geometry priors, establishes prompt-based subject association, and predicts per-view gaze targets using features expressed in a shared, estimated 3D coordinate frame. This world-space modeling allows our approach to process a variable number of uncalibrated views, helping resolve ambiguities across views. To support this paradigm, we curated a large-scale dataset from Ego-Exo4D containing over 260K verified multi-view GTE groups, alongside a self-collected in-the-wild benchmark featuring challenging head-invisible scenarios. Experiments show strong GTE performance on multiple benchmarks under both single-view and multi-view scenarios, and demonstrate consistent gains with increasing input views. Code, models, and benchmarks will be released upon acceptance.

1 INTRODUCTION

Gaze target estimation (GTE) aims to predict the specific location that a human subject is looking at within a scene. Beyond its established roles in human-robot interaction (Saran et al., 2018), cognitive psychology (Henderson, 2003; Guterstam et al., 2020) and autism research (Chong et al., 2017; Jones et al., 2023), GTE is gaining renewed importance in embodied AI: where a person looks provides a cue to *what* they may intend to interact with and *where* it is, providing a rich supervisory signal for learning from human demonstrations (Chuang et al., 2025; Pani & Yang, 2026a;b). Exploiting this signal, however, requires attention to be grounded in the shared 3D world space rather than in a single image plane.

Despite decades of research, accurate GTE remains challenging, with severe occlusion and limited camera field of view (FOV) serving as primary bottlenecks. When a subject's face is occluded in the camera view, directional cues become less informative; when the target lies outside the camera view, that image alone does not directly show the attended location. To address these issues, a recent study introduced a multi-view GTE dataset and an algorithm designed to enhance prediction robustness (Miao et al., 2025). However, their solution is limited to two-view estimation and relies on reasonably accurate and pre-associated head bounding boxes across views. More importantly, their pipeline requires precisely calibrated camera parameters, which restricts its applicability in unconstrained in-the-wild environments, such as monocular video captured from mobile devices.

In this paper, we introduce GazeWorld, a novel multi-view framework that learns human attention representation in the world-space directly from uncalibrated views. GazeWorld accommodates a variable number of observations from a static scene, using a prompt-driven architecture that enables subject specification through both the traditional head bounding boxes and natural-language descriptions, which specify the subject where boxes are unavailable, such as egocentric views and crowded multi-person scenes. Leveraging the geometric and semantic priors from pre-trained foundation models, GazeWorld grounds the prompted subject, uses spatial structure in 3D latent to locate the gaze target, and can be easily adapted to both egocentric and exocentric perspectives. Extensive

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

experiments have been conducted to compare GazeWorld with specialized GTE models and vision-language models (VLMs) under single-view GTE and multi-view GTE, where GazeWorld achieves state-of-the-art multi-view performance and competitive single-view accuracy. We also introduce datasets for out-of-domain (OOD) evaluation, and GazeWorld exhibits remarkable generalization capability. In summary, our main contributions are threefold:

- We formulate and study unconstrained GTE with uncalibrated multi-view cameras, which remains unaddressed.
- We present *GazeWorld*, a multi-view GTE framework with jointly trained fusion and prediction modules that constructs world-space attention representations using flexible prompt control without supplied camera calibration.
- We introduce two complementary datasets: *EgoExo4D-Gaze*, featuring over 260K verified multi-view GTE groups with paired egocentric and exocentric views, and *UMGaze*, a challenging real-world dataset for out-of-domain evaluation.

2 Related Work

Gaze Target Estimation in 2D and 3D. Gaze following, established by foundational benchmarks (Recasens et al., 2015; Chong et al., 2020; Tomas et al., 2021; Tafasca et al., 2023), has been dominated by multi-branch fusion architectures that consume auxiliary priors such as head crops, depth, pose or text before regressing a 2D target (Chong et al., 2018; Fang et al., 2021; Tonini et al., 2022; Yang et al., 2024; Ryan et al., 2025; Ye et al., 2026; She et al., 2026). Recent end-to-end models relax these dependencies with open-vocabulary prediction (Tafasca et al., 2024) and promptable subject localization (Cao et al., 2026a), yet all remain confined to a single image plane. A parallel line of 3D gaze estimation regresses a direction vector from the face (Kellnhofer et al., 2019; Zhang et al., 2020b), and extensions to panoramic or 360° settings (Dai et al., 2025; Qu et al., 2025) still operate within one coordinate frame; a direction without depth cannot pinpoint a target in 3D (Lian et al., 2018a; Kawana et al., 2025). The MVGT benchmark (Miao et al., 2025) marks a shift toward multi-view GTE that triangulates targets in a global world frame, but its baseline processes pairs of calibrated views with pre-aligned head boxes. We tackle multi-view GTE without supplied camera calibration or pre-aligned head detections, and defer a broader survey of gaze estimation to Appendix A.

Learning 3D Representation. Learning scene geometry supports reasoning about spatial gaze targets (Chen et al., 2025). Classical multi-view stereo depends on calibration from Structure-from-Motion, which can be difficult to obtain for moving or wearable cameras. Feed-forward models such as DUSt3R (Wang et al., 2024), MASt3R (Leroy et al., 2024) and VGGT (Wang et al., 2025a) instead predict aligned point maps and camera parameters directly from uncalibrated images (Wang et al., 2025b; Yang et al., 2025); VGGT-Ω (Wang et al., 2026a) scales this paradigm with an efficient architecture and self-supervised training, and Vista4D (Lin et al., 2026) extends it to 4D point clouds of dynamic videos for downstream reshooting. These models provide geometric representations, and VGGT features have been adopted for 3D detection (Cao et al., 2026b), manipulation (Ni et al., 2025), and co-visibility (Ziliotto et al., 2026), but none of these works predicts the gaze target of a prompted subject. Related work such as LERF (Kerr et al., 2023) embeds language features in a neural radiance field to support open-ended language queries in 3D. GazeWorld uses geometric features together with prompt-conditioned semantic features for per-view gaze prediction, adding subject grounding, 3D distance-biased fusion, and geometric consistency objectives.

3 World-Space GTE from Uncalibrated Views

We formulate *Unconstrained Multi-view Gaze Target Estimation* (UM-GTE) as follows. Let $\mathcal{V} = \{I^{(k)}\}_{k=1}^{N}$ be $N \geq 1$ RGB images of a shared scene, $I^{(k)} \in \mathbb{R}^{3\times H_k \times W_k}$, captured at the same time instant from arbitrary and unknown viewpoints, possibly by moving or wearable cameras. The target subject is specified by a single prompt $\mathcal{P}$, either a head bounding box in *one* view or a short text description (e.g., "the man in a plaid shirt and red pants"). The subject must be visible in at least one view, but may be partially or entirely absent from others. The N supplied views of the scene and the subject prompt are processed jointly in one forward pass to estimate the subject's single scene-level gaze target. As such, we do not differentiate views as either primary or reference, but jointly predict

the 2D projected locations if the target is inside the view for all the input images, otherwise indicate the target as out of the image. The goal is to learn a mapping:

$$\{(\hat{\mathbf{g}}^{(k)}, \hat{o}^{(k)})\}_{k=1}^{N} = \mathcal{F}(\mathcal{V}, \mathcal{P}; \theta), \tag{1}$$

where $\hat{\mathbf{g}}^{(k)} = \pi_k(\mathbf{g}) \in [0,1]^2$ is the 2D projection, represented as a heatmap, of the 3D gaze target $\mathbf{g}$ in camera π_k, and $\hat{o}^{(k)} \in [0,1]$ is the probability of target being out of the frame. The in/out probability indicates whether the decoded location is valid. With $\hat{\mathbf{g}}^{(k)}$ from all the views, the 3D target gaze $\mathbf{g}$ can be recovered through triangulation. For evaluation we use ground-truth calibration (Appendix D). In the ideal geometric formulation, the in-frame targets are *geometrically consistent*: they are the projections $\mathbf{g}^{(k)} = \pi_k(\mathbf{g})$ of a single, unobserved 3D target $\mathbf{g} \in \mathbb{R}^3$ under the unknown cameras π_k. This latent $\mathbf{g}$ is what we mean by a world-space attention target.

Compared with the existing multi-view setting (Miao et al., 2025), UM-GTE removes two critical constraints for optimal practicality. **(1) Calibration-free:** at inference, $\mathcal{F}$ receives no intrinsic or extrinsic parameters; cross-view geometry must be inferred from pixels alone. **(2) Unified subject prompting:** Similar to the prior-art, the model can take as input pre-aligned head boxes across views to identify the subject. When there are multiple images with multiple subjects in the scene, and the bounding box detection and association become less reliable, the model remains effective out-of-the-box given subject description in natural language.

4 METHOD: GAZEWORLD

Figure 1 illustrates the overall architecture of GazeWorld. Given N uncalibrated views, GazeWorld integrates representations from three frozen foundation encoders and a prompt embedding layer into a Transformer-based Geometry-Semantic Fusion backbone, ultimately decoding 2D gaze targets and per-view in/out probabilities. We detail each component below and defer extended configuration specifications to Appendix E.

4.1 MODEL ARCHITECTURE

Frozen Encoders. GazeWorld builds upon two complementary visual foundation streams to ground attention in uncalibrated physical environments. The geometric stream ϕ_G uses VGGT-Ω (Wang et al., 2026a), a feed-forward 3D reconstruction model that processes all N views jointly via interleaved frame-local and global attention. From ϕ_G, we extract patch representations from four intermediate aggregator layers. Its frozen prediction heads also provide per-view camera pose encodings and dense depth maps with pixel-wise confidence. The semantic stream comprises the image tower ϕ_S and text tower ϕ_P of a language-aligned DINOv3 (Siméoni et al., 2025), where ϕ_S outputs patch tokens aligned with text via DINO Meets Text (dino.txt) (Jose et al., 2025) and ϕ_P encodes natural language subject descriptions. All foundation encoders remain frozen during training, and predicted geometric primitives receive no gradients.

Unified Promptable Subject Specification. GazeWorld supports either text description prompts or bounding box prompts without retraining.

The text description input is processed by the frozen text encoder, and the semantic patch tokens from visual encoder query the text prompt embeddings through a 3-layer cross-attention block in the semantic fusion module. This operation injects text-driven subject constraints directly into the image representation while suppressing background distractors. We denote the resulting text-aligned semantic features for view v as $Z_S'^{(v)}$.

The box prompt is rasterized onto the patch grid as a binary head mask, combined with a per-view presence indicator, and injected into the geometry branch through an MLP whose final layer is zero-initialized. Thus, the box pathway initially leaves the pretrained features unchanged. In addition, a box prompt gives access to a high-resolution view of the subject: following MVGT and PAGE, a 256×256 crop around the prompted head is encoded by the same frozen semantic encoder, its patch tokens are placed at the head's location in the scene through grid-sampled positional embeddings, and a light bidirectional module (one self-attention block, three scene←head and two head←scene cross-attention blocks with LayerScale initialized near zero) lets the scene tokens read the head's

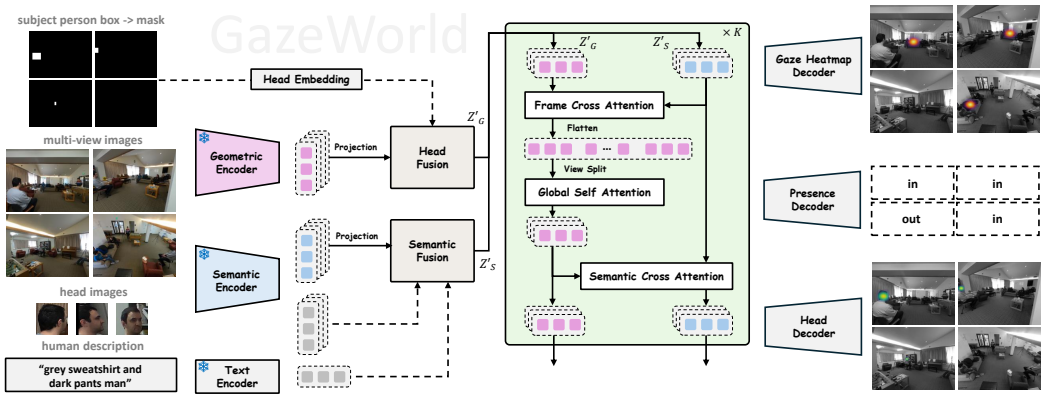


Figure 1: GazeWorld architecture overview. Frozen encoders supply cross-view geometric features (VGGT-Ω) and text-aligned semantic features (DINOv3 with dino.txt). Semantic Fusion grounds the subject prompt, and Geometry-Semantic Fusion exchanges information across views. Three lightweight heads decode the per-view gaze heatmap, head heatmap, and in/out probability. The foundation encoders are frozen; feature projections, prompt and positional embeddings, fusion modules, and task heads are trained. See Appendix E for a detailed description.

orientation before Geometry-Semantic Fusion. Views without a box prompt, including all views under text-only prompting, skip this head branch. Prompt dropout during training randomly omits either prompt type, yielding a unified model capable of handling text-only, box-only, or multi-modal prompts (Appendix E.2).

Geometry-Semantic Fusion. The fusion module operates primarily on geometry tokens $Z_G'^{(v)}$, which are enriched with two dense 3D positional cues prior to fusion: a per-patch ray embedding derived from estimated camera poses, and a downsampled 3D world-point map with its confidence channel. To reduce variation in translation and scale across scenes, world points are centered by their scene centroid and divided by the mean distance to that centroid. The resulting coordinates remain rotation-dependent, while their pairwise distances are invariant to rigid transformations. Each view sequence is prepended with a learnable in/out token and register tokens.

The core fusion backbone consists of $L_A = 6$ stacked blocks, where each block executes a sequential three-step interaction: (1) **Frame-level cross-attention**, where geometry tokens $Z_G'^{(v)}$ query corresponding text-aligned semantic features $Z_S'^{(v)}$ to bind identity to spatial structure locally; (2) **Global spatially-biased self-attention**, where tokens across all N views exchange information globally, with attention logits between tokens i and j modulated by a learned 3D distance prior:

$$A_{ij} = \frac{Q_i K_j^\top}{\sqrt{d}} - \mathrm{softplus}(\gamma_l) \cdot d_{ij}, \tag{2}$$

where Q_i and K_j are query and key vectors for one attention head, d is their dimension, A_{ij} is a pre-softmax logit, and γ_l is a learnable scalar for layer l. The distance d_{ij} is measured in 3D rather than in the image: each patch token takes the average predicted world point of its pixels, so patches from different views that image the same surface are biased together. As the predicted world frame differs in origin and scale across scenes, these points are scene-normalized by the group's centroid and mean radius. Special-token distances are defined in Appendix E.1. The final step is (3) **Semantic cross-attention**, where semantic tokens query globally updated geometry tokens $\tilde{Z}_G'^{(v)}$. Executing global geometric fusion prior to semantic updates ensures that the semantic branch receives cross-view context essential for reasoning under severe occlusions.

Task Decoders. Three lightweight parallel heads decode the fused representations for every view. The gaze heatmap head upsamples geometry tokens into a 2D gaze heatmap of per-pixel sigmoid scores, supervised with BCE. The head heatmap head upsamples semantic tokens into a 2D heatmap identifying the subject's head center, serving as both an auxiliary localization target and a grounding output. This dense head-center representation provides patch-level subject grounding. Finally, a two-layer MLP decodes each view's in/out token into the in-frame probability $1 - \hat{o}^{(v)}$. Here $\hat{o}^{(v)}$

denotes the complementary out-of-frame probability. Because this token attends across both global geometry and text-aligned semantics, in/out prediction uses multi-view contextual evidence.

4.2 Learning Objectives

GazeWorld is trained across three progressive stages (Section 6.1): single-view pretraining, multi-view pretraining on EgoExo4D-Gaze, and multi-view fine-tuning on MVGT. In all stages, 2D heatmaps are supervised using pixel-wise Binary Cross-Entropy (BCE) against Gaussian targets, and in/out predictions are optimized via Focal Loss (Lin et al., 2017). Two forms of geometric regularization use the estimated 3D geometry to constrain gaze predictions, including cross-view agreement in the multi-view stages.

Geometric Alignment Loss. For the training losses, ground truth gaze target coordinates $\mathbf{g}^{(v)} = (x, y)$ are normalized to the preprocessed input canvas: $\mathbf{u}^{(v)} = T_v(\mathbf{g}^{(v)})$, where T_v is the image-to-canvas transform. We obtain the predicted canvas coordinate $\hat{\mathbf{u}}^{(v)}$ from each predicted heatmap by differentiable soft-argmax and bilinearly sample the frozen world-point map $\mathcal{M}^{(v)} : [0,1]^2 \to \mathbb{R}^3$ there. These sampled points and both consensus targets below are expressed in the encoder's estimated reconstruction frame. Averaging these points over eligible source views $\mathcal{S}_{\text{pred}}$ gives the consensus $\hat{P}_{\text{world}}$. For a nonempty source set, the alignment loss penalizes its reprojection error:

$$\hat{P}_{\text{world}} = \frac{1}{|\mathcal{S}_{\text{pred}}|} \sum_{v \in \mathcal{S}_{\text{pred}}} \mathcal{M}^{(v)}(\hat{\mathbf{u}}^{(v)}), \qquad \mathcal{L}_{\text{align}} = \frac{\sum_{v=1}^{N} a_v \|\bar{\Pi}^{(v)}(\hat{P}_{\text{world}}) - \hat{\mathbf{u}}^{(v)}\|_1}{\max(\sum_{v=1}^{N} a_v, \varepsilon)}. \quad (3)$$

Here a_v selects in-frame views with sufficiently confident sampled geometry, $\varepsilon = 10^{-6}$ prevents division by zero, and $\bar{\Pi}^{(v)}$ is the numerically stabilized, image-normalized projection under the estimated camera in the same canvas coordinates. If no source view is eligible, we define the alignment loss to be zero without evaluating the consensus expression. Source-view selection and confidence gating are detailed in Appendix E.3. Estimated cameras are used both for feature construction and for training losses; supplied calibration is not required at inference.

3D Gaze Consistency Loss. The alignment loss encourages agreement among predictions, while the consistency loss anchors sampled 3D predictions to the 2D annotations. A detached pseudo target P^*_{world} is the mean world point sampled at ground-truth locations over eligible source views $\mathcal{S}_{\text{gt}}$. With no eligible source, we set this target and all residual weights to zero. We minimize

$$\mathcal{L}^{\text{gaze}}_{\text{cons}} = \frac{\sum_{v=1}^{N} w_v \, \text{SmoothL}_{1,\beta}\left(\frac{\mathcal{M}^{(v)}(\hat{\mathbf{u}}^{(v)}) - P^*_{\text{world}}}{s}\right)}{\max(\sum_{v=1}^{N} w_v, \varepsilon)}, \quad \beta = 0.05, \quad (4)$$

where s is the per-scene mean radius and w_v combines annotation validity, predicted-point confidence, and pseudo-target reliability. Smooth-L_1 is summed over the three coordinates. The analogous head loss $\mathcal{L}^{\text{head}}_{\text{cons}}$ uses head centers and head-specific validity masks. Frozen geometry receives no gradients, but bilinear sampling propagates gradients to the predicted 2D locations. Appendix E.3 specifies the masks, reliability weights, and stage-specific activation of these losses.

5 Datasets

EgoExo4D-Gaze. Supervised multi-view GTE data is notoriously expensive to collect because the same subject and target must be annotated consistently across views. Consequently, the previous MVGT dataset contains only 2,281 samples. We reduce manual annotation effort by deriving gaze labels from wearable eye tracking, followed by MLLM verification and targeted manual audits. Our data source is Ego-Exo4D (Grauman et al., 2024), which captures skilled activities using an egocentric Aria device alongside multiple time-synchronized, co-registered exocentric cameras. Although the wearer's 3D gaze target recorded by the device eye tracker can be projected into all exocentric views simultaneously, physical occlusions and calibration drift sometimes induce projection misalignments. Gemini 3.8 Flash is used to address these issues. After cleaning all data, we construct **EgoExo4D-Gaze**, containing 264,523 groups (1,286,762 views) from 3,840 takes and 80 activity categories. Groups contain two to six views (averaging 4.86 views), with 99.97% including an

Table 1: Comparison on MVGT with leave-one-scene-out evaluation. Box-prompt GazeWorld receives boxes in all views containing the subject's head. Text-prompt GazeWorld receives the subject text description. N/A denotes an evaluation setting not applicable to the model.

Method	Calib.-free	Prompt	AUC ↑	L2 Dist ↓		Inout AP ↑	3D Metric [m]↓
				w/ Subject Head	w/o Subject Head		
Random	N/A	N/A	0.500	0.456	0.456	0.594	N/A
Qwen3.8-27B (zero-shot)	✓	bbox	N/A	0.256	0.264	0.611	2.183
Qwen3.8-27B (zero-shot)	✓	text	N/A	0.253	0.261	0.626	2.119
Gemini 3.8 Flash (zero-shot)	✓	bbox	N/A	0.289	0.301	0.622	2.193
Gemini 3.8 Flash (zero-shot)	✓	text	N/A	0.276	0.289	0.670	2.049
Qwen3.5-9B (SFT)	✓	bbox	N/A	0.232	0.235	0.604	1.936
Qwen3.5-9B (SFT)	✓	text	N/A	0.250	0.254	0.599	1.999
GazeAnywhere	N/A	text	0.908	0.139	N/A	0.895	1.304
Gaze-LLE	N/A	bbox	0.916	0.136	N/A	0.907	1.016
Gaze-LLE (RexSeek + SAM3)	N/A	text	0.912	0.140	N/A	0.903	1.051
PAGE	N/A	bbox	0.946	0.122	N/A	0.910	0.993
PAGE (RexSeek + SAM3)	N/A	text	0.939	0.129	N/A	0.892	1.028
MVGT Baseline	✗	bbox	0.850	0.118	0.188	0.918	0.750
MVGT Baseline (RexSeek + SAM3)	✗	text	0.827	0.133	0.201	0.905	0.887
GazeWorld (ours)	✓	bbox	**0.966**	**0.082**	**0.152**	**0.935**	**0.546**
GazeWorld (ours)	✓	text	0.954	0.098	0.169	0.932	0.673

egocentric view. Annotations provide gaze out-of-frame labels, 2D gaze targets when in-frame, subject descriptions, and exocentric head boxes with reliability states. Overall, 98.5% of views have in-frame targets and 79.4% have a nonempty head-box annotation. The official take-level split ensures zero video-level overlap, yielding $205{,}030$ training and $59{,}493$ held-out groups. Appendix F.2 provides detailed statistics.

UMGaze. UMGaze contains 200 six-view samples (1,200 images at 4080×3072) collected with five volunteers for text-prompted OOD evaluation. We first photograph the scene from multiple viewpoints using a handheld camera, then photograph a selected participant in the scene, and select six matching images. Half of the cases have more than one person. This sequential capture makes UMGaze a stringent test: scene-only views must supply context for a target observed by the participant in other views. Participants look at the center of a designated target. No calibration or head boxes are supplied. UMGaze will be released under CC BY-NC 4.0.

6 Experiments

We evaluate GazeWorld on MVGT under leave-one-scene-out cross-validation and on the EgoExo4D-Gaze, comparing against specialized single-view and multi-view GTE methods and VLMs. UMGaze provides an OOD evaluation setting. Ablation studies examine the encoder choices and the effect of the number of input views. Qualitative results are shown in Figure 2.

6.1 Implementation Details

Training Schedule. Our training pipeline has three stages. First, we pretrain on the single-view Gaze-Co corpus (Cao et al., 2026a) to learn prompt-conditioned subject grounding and gaze prediction. Second, we pretrain on EgoExo4D-Gaze to learn cross-view geometry and gaze semantics. Third, we fine-tune on MVGT with leave-one-scene-out cross-validation, mixing in EgoExo4D-Gaze and single-view data to limit forgetting (Appendix E.3). Training uses 4 NVIDIA H100 GPUs. All baselines follow the same multi-stage training except the zero-shot VLMs.

Feature Extractor. We implement the GazeWorld framework using DINOv3 (Siméoni et al., 2025) as the semantic backbone and VGGT-Ω as the geometry encoder. For text-conditioned alignment, we use dino.txt (Jose et al., 2025) features with a language-aligned DINOv3 image tower. Single-view Stage-1 training uses square 512×512 inputs with a 256×256 crop of the prompted subject's head for the head branch. The multi-view stages letterbox every view to 448×592.

Baselines. We compare GazeWorld against SOTA single-view and multi-view GTE methods. For single-view baselines, we evaluate Gaze-LLE and PAGE, which take bounding-box prompts,

Table 2: Evaluation on EgoExo4D-Gaze (after stage 2 finetuning) and UMGaze. EgoExo4D-Gaze reports both egocentric and exocentric L2 distance. Every method is prompted by text only in these datasets. L2 is averaged over the views with an in-frame target and in/out AP over all views.

Method	EgoExo4D-Gaze test set			UMGaze	
	Ego L2 Dist (text) ↓	Exo L2 Dist (text) ↓	Inout AP ↑	L2 Dist (text) ↓	Inout AP ↑
Gemini 3.8 Flash (zero-shot)	0.179	0.096	0.982	0.366	0.772
Qwen3.8-27B (zero-shot)	0.165	0.123	0.986	0.284	0.863
Qwen3.5-9B (SFT)	0.097	0.057	0.986	0.297	0.801
GazeWorld (ours)	**0.087**	**0.052**	**0.999**	**0.095**	**0.973**

GazeAnywhere, which takes text prompts. Single-view baselines are not evaluated on views that do not contain the subject (N/A in Table 1). For the multi-view baseline, we use the official MVGT framework, which requires explicit camera calibration; it pairs each primary view with each of the other five views in separate two-view forward passes and averages the resulting predictions. Additionally, we evaluate leading VLMs, including Gemini 3.8 Flash, Qwen3.8-27B, and Qwen3.5-9B.

Evaluation Metrics. All heatmap predictions are decoded with DARK, whereas VLMs directly output pixel coordinates. Coordinates are mapped back to the original image and normalized for L2 evaluation. We report heatmap area under the ROC curve (AUC), L2 distance in normalized image coordinates, and average precision (AP) with in-frame targets as the positive class. AUC is unavailable for VLMs, which output only a target point (N/A in Table 1). On MVGT, localization is evaluated on unoccluded, in-frame targets and split into views where the subject's head is visible (w/ Subject Head) or invisible (w/o Subject Head). For MVGT, we follow its leave-one-scene-out protocol: each fold holds out an entire indoor environment (commons room, convenience store, kitchen, or research lab), including its layout, background, subjects, and camera setups, so every fold evaluates generalization to an unseen environment. Metrics are averaged equally over the four held-out environments. EgoExo4D-Gaze reports separate L2 errors for egocentric and exocentric views. Note that the wearer's head is never visible in the egocentric image, which corresponds to the w/o Subject Head case in MVGT. The 3D metric triangulates per-view predictions using ground-truth calibration solely for evaluation (Appendix D).

6.2 Comparison to State-of-the-art Methods

Experiments on Multi-view GTE Benchmarks. Table 1 reports the results of all methods on MVGT. Among them, only the MVGT baseline requires camera calibration. We evaluate two prompt settings: the ground-truth head bounding box (bbox) of the subject, given in every view where the head is visible, and a text description of the subject; Appendix B.2 additionally evaluates the minimal setting with a box in a single view. For GTE methods that accept only box prompts, we build a strong multi-stage pipeline for the text setting: SAM 3 (Carion et al., 2026) detects person boxes, RexSeek (Jiang et al., 2025) selects the described person, and SAM 3 then extracts the head region. We also compare with VLMs, which have been explored for single-view GTE (Pani & Yang, 2026a; Mathew et al., 2025). These models can accept all six views and subject prompts without calibration. We evaluate zero-shot Qwen3.8-27B (Bai et al., 2025) and Gemini 3.8 Flash (Google, 2025), and fine-tune Qwen3.5-9B on our training data. GazeWorld outperforms all baselines by a large margin, achieving the lowest L2 on both the w/ Subject Head and w/o Subject Head subsets as well as the lowest 3D error. Even with the harder text prompt, GazeWorld outperforms every baseline, including those given box prompts. Appendix G further compares GazeWorld with MVGT and Self-MVGTE (Chen et al., 2026) under their respective evaluation protocols. Though calibration-free, GazeWorld achieves much lower localization error in all four reference-view visibility categories of the original two-view MVGT protocol (Table 9), and outperforms the calibration-dependent Self-MVGTE on all reported metrics under its protocol (Table 10).

Experiments on EgoExo4D-Gaze. The left part of Table 2 reports results on the EgoExo4D-Gaze test set. Because egocentric views never contain the wearer's head and some exocentric head boxes are missing or unreliable (Section 5), we evaluate all methods with text prompts on this test set. GazeWorld outperforms all VLMs and achieves the lowest L2 on both egocentric and exocentric views, even though the subject's head is never visible in the egocentric view.

Table 3: Single-view GTE on Gaze-Co benchmarks (Cao et al., 2026a) with ViT-L visual encoders, except PAGE, which uses DINOv3 ViT-H+; the VGGT-Ω encoder in GazeWorld is also a DINOv3 ViT-L. bbox rows use the head box of the evaluated view, text rows the appearance description only. TextGaze needs both text and bbox input.

Gaze Model	Backbone	Prompt	GazeFollow-Concept			VAT-Concept		ChildPlay-Concept	
			AUC ↑	Avg L2 ↓	Min L2 ↓	L2 ↓	AP ↑	L2 ↓	AP ↑
ViTGaze	ViT-L	bbox	0.955	0.097	0.045	0.098	0.879	0.113	0.905
Sharingan	ViT-L	bbox	0.944	0.114	0.059	0.109	0.852	0.118	0.854
Gaze-LLE	ViT-L	bbox	0.961	0.099	0.045	0.101	0.875	0.113	0.912
PAGE	ViT-H+	bbox	**0.970**	0.074	0.028	**0.076**	**0.930**	0.082	0.918
GazeWorld	ViT-L	bbox	0.968	**0.072**	**0.027**	0.084	0.916	**0.079**	**0.935**
GazeAnywhere	ViT-L	text	**0.958**	**0.099**	**0.050**	0.123	0.879	0.098	0.906
TextGaze	ViT-L	text+bbox	0.949	0.123	0.061	0.105	0.843	0.127	0.904
GazeWorld	ViT-L	text	0.955	0.107	0.055	**0.101**	**0.901**	**0.096**	**0.918**

Table 4: Geometry and semantic encoder module ablation on MVGT. All variants use the same three-stage training recipe.

Geometry	Semantic	w/ Head L2↓	w/o Head L2↓
VGGT	DINOv2 (dino.txt)	0.101	0.211
VGGT-Ω	-	0.095	0.178
-	DINOv3 (dino.txt)	0.114	0.277
VGGT-Ω	shared (adapter)	0.100	0.171
VGGT-Ω	DINOv3 (dino.txt)	**0.082**	**0.152**

Table 5: Effect of the number of input views on MVGT and GazeWorld (L2 w/ Subject Head).

#Views	MVGT Baseline	GazeWorld
1	0.150	0.109
2	0.130	0.097
3	0.123	0.090
4	0.121	0.085
5	0.119	0.084
6	**0.118**	**0.082**

Experiments on Single-view GTE Benchmarks (Gaze-Co). Unlike the MVGT baseline, GazeWorld is a unified model: a single set of weights accepts any number of input views, including a single view. Table 3 evaluates GazeWorld on the concept-annotated GazeFollow, VAT, and ChildPlay benchmarks, showing strong performance in the conventional single-view setting. With a box prompt, GazeWorld surpasses Gaze-LLE on every metric of all three benchmarks. GazeWorld further outperforms PAGE, the current single-view SOTA method in the bbox condition, on four of the seven metrics (Avg/Min L2 on GazeFollow and L2/AP on ChildPlay), although PAGE uses a larger ViT-H+ backbone. With a text prompt, GazeWorld improves on GazeAnywhere in both L2 and AP on VAT and ChildPlay, while GazeAnywhere is slightly better on GazeFollow. These results show that GazeWorld handles a variable number of input views while remaining strong in the single-view setting.

6.3 Ablation Study

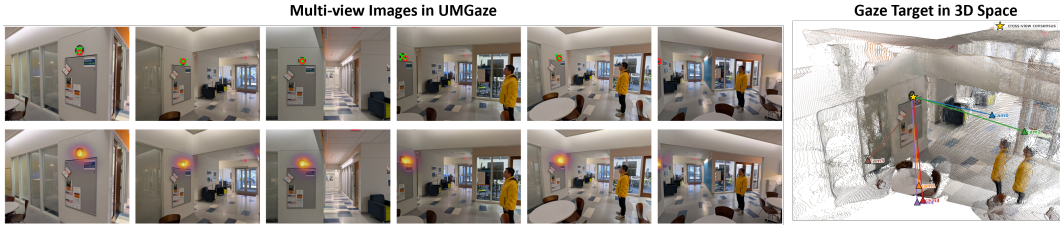
We first quantify how much each additional camera contributes. As shown in Table 5, GazeWorld improves when more views are added, from 0.109 with a single view to 0.082 with all six, showing that the model benefits from additional views. Table 4 compares the backbone choices along two axes: the encoder generation and the tower layout. Among the dual-tower variants, VGGT-Ω/DINOv3 outperforms VGGT/DINOv2. Removing the geometry tower and keeping only DINOv3 (zero world points, identity poses, no 3D losses) raises the w/o Head L2 from 0.152 to 0.277: without the geometry tower and its associated cues and losses, the model has higher error in views that do not show the subject's head, while the absolute increase is smaller in head-visible views (0.082 to 0.114). Removing the language-aligned semantic tower and feeding the geometry tower's internal DINOv3 tokens as the semantic stream increases L2 on all subsets, but this variant can only take box prompts, since the internal features are not aligned with dino.txt text embeddings. We additionally test collapsing the two towers into one while keeping text prompting: a lightweight distillation adapter maps the DINOv3 features inside VGGT-Ω back to the dino.txt-aligned space, removing an entire ViT-L forward pass. This single-tower variant has higher L2 than the dual-tower model (0.100 vs. 0.082 w/ Head and 0.171 vs. 0.152 w/o Head), trading accuracy for one fewer visual-encoder



(a) Multi-view GTE in EgoExo4D-Gaze test set, input with text prompt



(b) Multi-view GTE on MVGT with multi-person scenario, input with text prompt



(c) Out-of-domain UMGaze, input with text prompt

Figure 2: Qualitative results of GazeWorld. For each group we show the multi-view inputs with the ground-truth target (green cross) and our prediction (red circle), the predicted gaze heatmaps, and the VGGT-Ω reconstruction with camera positions. Per-view predicted viewing rays pass near the cross-view consensus (yellow star). All predictions are made without camera calibration.

forward pass. Appendix B further examines the effects of explicit geometry, the geometric losses, prompt selection, and multi-person scenes.

6.4 Real-world Out-of-Domain Experiment

We evaluate on all 200 UMGaze samples (6×200 images) using text prompts and no adaptation to the capture site. UMGaze is strictly under the UM-GTE setting and we choose to only use text prompt as condition. GazeWorld results average the four MVGT leave-one-scene-out models. Table 2 reports L2/AP of 0.095/0.973, compared with 0.284/0.863 for zero-shot Qwen3.8-27B and 0.297/0.801 for fine-tuned Qwen3.5-9B. The results demonstrate transfer to separately captured scene and subject views; Figure 2(c) illustrates selected predictions.

7 Conclusion

In this paper, we introduced *GazeWorld*, a unified promptable framework for multi-view gaze target estimation from uncalibrated views in the wild. GazeWorld integrates frozen geometric and semantic foundation models through a 3D distance-biased Geometry-Semantic Fusion and geometric consistency objectives, grounding the subject from either a head box or a text description and predicting per-view gaze targets without camera calibration. To support this setting, we curated *EgoExo4D-Gaze* for large-scale multi-view training and *UMGaze* for out-of-domain evaluation. Experiments show that GazeWorld achieves state-of-the-art multi-view performance, surpassing even calibration-dependent methods in 3D accuracy, while remaining strong in the single-view setting. By bridging uncalibrated multi-view perception and world-space attention modeling, GazeWorld establishes a practical foundation for understanding human attention and social interactions in unconstrained physical environments.

AI USE STATEMENT

We used Gemini to assist with dataset annotation verification, including screening projected gaze annotations and assessing candidate head-box reliability in EgoExo4D-Gaze. Gaze targets originate from recorded eye-tracking measurements and geometric projection; Gemini provides verification judgments rather than direct measurements of gaze. Appendix F describes the automated checks and targeted manual audits, including their limitations. We also used generative AI assistants (Claude Code, OpenAI Codex) to edit manuscript text and software code, check mathematical definitions against the implementation, search for related work, and provide feedback on experimental methodology, including the controlled geometry ablation. The authors manually reviewed the AI-assisted work and take responsibility for the final text, mathematical statements, experimental claims, and artifacts, including those developed with AI assistance.

ETHICS STATEMENT

EgoExo4D-Gaze is derived from Ego-Exo4D (Grauman et al., 2024). Access to and use of the source data remain governed by the Ego-Exo4D license. We obtained permission to process the data with Gemini for annotation verification. The EgoExo4D-Gaze annotation pipeline combines recorded gaze measurements, geometric projection, automated verification, and targeted manual audits (Appendix F); the resulting annotations can retain errors and biases from the source data and automated tools. The UMGaze data collection was approved by the institutional review board (IRB) at the collecting institution. UMGaze contains data from five volunteers (Section 5) who consented to public release of their identifiable imagery. We will make UMGaze publicly available under CC BY-NC 4.0.

Multi-view gaze estimation can support research on human attention and interaction, but it can also enable unwanted monitoring across cameras, including when a person's head is absent from a target view. Gaze estimates are uncertain and do not establish a person's intentions or mental state. Applications should respect informed consent, privacy, and the relevant data-use restrictions; the reported results do not establish suitability for covert surveillance or consequential decisions about individuals.

REPRODUCIBILITY STATEMENT

Section 4 and Appendix E describe the architecture, prompt encoding, geometric objectives, and three-stage training procedure. Section 5 and Appendix F document the datasets, annotation pipeline, filtering, and manual quality checks. Section 6 and Appendices D and G specify the evaluation protocols, including the MVGT leave-one-scene-out splits, coordinate conventions, metric aggregation, and calibrated triangulation used for evaluation. These descriptions distinguish predicted geometry used by the model from calibration used to construct annotations or evaluate 3D predictions. We will release the code and make the complete EgoExo4D-Gaze and UMGaze datasets publicly available, subject to the applicable source-data licenses and access requirements.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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*, pp. 14126–14135, 2022.

Xu Cao, Houze Yang, Vipin Gunda, Zhongyi Zhou, Tianyu Xu, Adarsh Kowdle, Inki Kim, and James M Rehg. Gaze target estimation anywhere with concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Yang Cao, Feize Wu, Dave Zhenyu Chen, Yingji Zhong, Lanqing Hong, and Dan Xu. VGGT-Det: Mining VGGT internal priors for sensor-geometry-free multi-view indoor 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Andy Catruna, Adrian Cosma, and Emilian Radoi. Crossgaze: A strong method for 3d gaze estimation in the wild. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5. IEEE, 2024.

Keqi Chen, Vinkle Srivastav, and Nicolas Padoy. Self-supervised multi-view 3d gaze target estimation via probabilistic ray marching. *arXiv preprint arXiv:2609.07415*, 2026.

Wenhe Chen, Hui Xu, Chao Zhu, Xiaoli Liu, Yinghua Lu, Caixia Zheng, and Jun Kong. Gaze estimation via the joint modeling of multiple cues. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1390–1402, 2021.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 3341–3347. IEEE, 2022.

Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M Jones, Agata Rozga, and James M Rehg. Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–20, 2017.

Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. 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)*, pp. 383–398, 2018.

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*, pp. 5396–5406, 2020.

Ian Chuang, Jinyu Zou, Andrew Lee, Dechen Gao, and Iman Soltani. Look, focus, act: Efficient and robust robot learning via human gaze and foveated vision transformers. *arXiv preprint arXiv:2507.15833*, 2025.

Zhuangzhuang Dai, Vincent Gbouna Zakka, Luis J Manso, and Chen Li. Gazetarget360: Towards gaze target estimation in 360-degree for robot perception. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 18291–18296. IEEE, 2025.

Bardia Doosti, Ching-Hui Chen, Raviteja Vemulapalli, Xuhui Jia, Yukun Zhu, and Bradley Green. Boosting image-based mutual gaze detection using pseudo 3d gaze. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1273–1281, 2021.

Nuno Ferreira Duarte, Jovica Tasevski, Moreno Coco, Mirko Raković, Aude Billard, and José Santos-Victor. Action anticipation: Reading the intentions of humans and robots. *IEEE Robotics and Automation Letters*, 2018. doi: 10.1109/LRA.2018.2861569.

Nathan J Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604, 2000.

Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6460–6468, 2018.

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*, pp. 11390–11399, 2021.

Alexey Gavryushin, Dingxi Zhang, Zhao Huang, Alexandros Delitzas, Jiaqi Chen, Ben Ellis, Cedric Zöllner, Manthan Patel, Manuel Kaufmann, Marc Pollefeys, et al. CoMind: Understanding collaborative human activity from multiple minds and views. In *European Conference on Computer Vision (ECCV)*, 2026.

Google. Gemini 3 pro: the frontier of vision ai. Online, 2025. URL `https://blog.google/innovation-and-ai/technology/developers-tools/gemini-3-pro-vision/`.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

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*, pp. 5041–5050, 2022.

Arvid Guterstam, Andrew I Wilterson, Davis Wachtell, and Michael SA Graziano. Other people's gaze encoded as implied motion in the human brain. *Proceedings of the National Academy of Sciences*, 117(23):13162–13167, 2020.

John M Henderson. Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11):498–504, 2003.

Yoichiro Hisadome, Tianyi Wu, Jiawei Qin, and Yusuke Sugano. Rotation-constrained cross-view feature fusion for multi-view appearance-based gaze estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5985–5994, 2024.

Haoyu Huang and Yoichi Sato. Egocentric gaze estimation via neck-mounted camera. *arXiv preprint arXiv:2602.11669*, 2026.

Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *European conference on computer vision*, pp. 789–804. Springer, 2018.

Qing Jiang, Lin Wu, Zhaoyang Zeng, Tianhe Ren, Yuda Xiong, Yihao Chen, Liu Qin, and Lei Zhang. Referring to any person. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21667–21678. IEEE, 2025.

Warren Jones, Cheryl Klaiman, Shana Richardson, Christa Aoki, Christopher Smith, Mendy Minjarez, Raphael Bernier, Ernest Pedapati, Somer Bishop, Whitney Ence, Allison Wainer, Jennifer Moriuchi, Sew-Wah Tay, and Ami Klin. Eye-tracking-based measurement of social visual engagement compared with expert clinical diagnosis of autism. *JAMA*, 330(9):854–865, 2023. doi: 10.1001/jama.2023.13295.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24905–24916, 2025.

Yuki Kawana, Shintaro Shiba, Quan Kong, and Norimasa Kobori. Ga3ce: Unconstrained 3d gaze estimation with gaze-aware 3d context encoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3081–3090, 2025.

Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6912–6921, 2019.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19729–19739, 2023.

Yantao Lai, Rong Quan, Dong Liang, and Jie Qin. Clipgaze: Zero-shot goal-directed scanpath prediction using clip. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Ji Woo Lee, Chul Woo Cho, Kwang Yong Shin, Eui Chul Lee, and Kang Ryoung Park. 3d gaze tracking method using purkinje images on eye optical model and pupil. *Optics and Lasers in Engineering*, 50(5):736–751, 2012.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Yunhao Li, Wei Shen, Zhongpai Gao, Yucheng Zhu, Guangtao Zhai, and Guodong Guo. Looking here or there? gaze following in 360-degree images. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 3742–3751, 2021.

Dongze Lian, Lina Hu, Weixin Luo, Yanyu Xu, Lixin Duan, Jingyi Yu, and Shenghua Gao. Multi-view multitask gaze estimation with deep convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 30(10):3010–3023, 2018a.

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

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, et al. Vista4d: Video reshooting with 4d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32671–32682, 2026.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Zhi-Yi Lin, Jouh Yeong Chew, Jan van Gemert, and Xucong Zhang. Gazehta: End-to-end gaze target detection with head-target association. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9447–9454. IEEE, 2025.

Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3477–3485, 2019.

Athul M Mathew, Haithem Hermassi, Thariq Khalid, Arshad Ali Khan, and Riad Souissi. Gazevlm: A vision-language model for multi-task gaze understanding. *arXiv preprint arXiv:2511.06348*, 2025.

Qiaomu Miao, Minh Hoai, and Dimitris Samaras. Patch-level gaze distribution prediction for gaze following. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 880–889, 2023.

Qiaomu Miao, Vivek Raju Golani, Jingyi Xu, Progga Paromita Dutta, Minh Hoai, and Dimitris Samaras. Multi-view gaze target estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5371–5381, 2025.

Vikrant Nagpure and Kenji Okuma. Searching efficient neural architecture with multi-resolution fusion transformer for appearance-based gaze estimation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 890–899, 2023.

Zehao Ni, Yonghao He, Lingfeng Qian, Jilei Mao, Fa Fu, Wei Sui, Hu Su, Junran Peng, Zhipeng Wang, and Bin He. VO-DP: Semantic-geometric adaptive diffusion policy for vision-only robotic manipulation. *arXiv preprint arXiv:2510.15530*, 2025.

Soma Nonaka, Shohei Nobuhara, and Ko Nishino. Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2192–2201, 2022.

Anupam Pani and Yanchao Yang. Gaze-vlm: Bridging gaze and vlms through attention regularization for egocentric understanding. *Advances in Neural Information Processing Systems*, 38: 163544–163577, 2026a.

Anupam Pani and Yanchao Yang. Gaze-regularized vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2603.23202*, 2026b.

Hongyu Qu, Jianan Wei, Xiangbo Shu, Yazhou Yao, Wenguan Wang, and Jinhui Tang. Omnigaze: Reward-inspired generalizable gaze estimation in the wild. *arXiv preprint arXiv:2510.13660*, 2025.

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

Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1435–1443, 2017.

Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M Rehg. Gaze-lle: Gaze target estimation via large-scale learned encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28874–28884, 2025.

Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Human gaze following for human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8615–8621. IEEE, 2018.

Junhui She, Fei Wang, Kun Li, Yiqi Nie, Yuxin Liu, Zhangling Duan, and Xun Yang. Textgaze: Prompting gaze target estimation with textual scene cues. *arXiv preprint arXiv:2607.10130*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Yuehao Song, Xinggang Wang, Jingfeng Yao, Wenyu Liu, Jinglin Zhang, and Xiangmin Xu. Vitgaze: gaze following with interaction features in vision transformers. *Visual Intelligence*, 2 (1):1–15, 2024.

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*, pp. 20935–20946, 2023.

Samy Tafasca, Anshul Gupta, Victor Bros, and Jean-Marc Odobez. Toward semantic gaze target detection. *Advances in Neural Information Processing Systems*, 37:121422–121448, 2024.

Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto. Goo: A dataset for gaze object prediction in retail environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3125–3133, 2021.

Francesco Tonini, Cigdem Beyan, and Elisa Ricci. Multimodal across domains gaze target detection. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 420–431, 2022.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-ω. *arXiv preprint arXiv:2605.15195*, 2026a.

Jun Wang, Hao Ruan, Liangjian Wen, Yong Dai, and Mingjie Wang. Gazeclip: Enhancing gaze estimation through text-guided multimodal learning. *Neurocomputing*, pp. 133584, 2026b.

Kang Wang and Qiang Ji. 3d gaze estimation without explicit personal calibration. *Pattern Recognition*, 79:216–227, 2018.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21924–21935, 2025.

Yaokun Yang and Feng Lu. Gaze target detection based on head-local-global coordination. In *European Conference on Computer Vision*, pp. 305–322. Springer, 2024.

Yaokun Yang, Yihan Yin, and Feng Lu. Gaze target detection by merging human attention and activity cues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6585–6593, 2024.

Zhoutong Ye, Chengwen Zhang, Zhaibin Cui, Mingze Sun, Jiaqi Liu, Xiangwu Li, Qingyang Wan, Chang Liu, Xutong Wang, Huan-ang Gao, et al. Page: Towards practical human-level gaze target estimation. *arXiv preprint arXiv:2607.04860*, 2026.

Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. Human gaze assisted artificial intelligence: A review. In *IJCAI: Proceedings of the Conference*, volume 2020, pp. 4951, 2020a.

Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European conference on computer vision*, pp. 365–381. Springer, 2020b.

Hao Zhao, Ming Lu, Anbang Yao, Yurong Chen, and Li Zhang. Learning to draw sight lines. *International Journal of Computer Vision*, 128(5):1076–1100, 2020.

Xinyuan Zhao, Xianrui Chen, and Ahmad Chaddad. Gazeformer-moe: Context-aware gaze estimation via clip and moe transformer. *arXiv preprint arXiv:2601.12316*, 2026.

Filippo Ziliotto, Luciano Serafini, Lamberto Ballan, and Tommaso Campari. What VGGT knows about overlap: Probing geometric foundation models for co-visibility. *arXiv preprint arXiv:2607.09503*, 2026.

APPENDIX

A Extended Related Work: Gaze Estimation Paradigms **17**
A.1 Single-view Gaze Target Estimation 17
A.2 3D Gaze Direction Estimation 17

B Ablation **17**
B.1 Ablation on Explicit Geometry and Loss Functions 18
B.2 Ablation on Prompt Selection 19
B.3 Grounding and Cross-view Association with Multiple People 20

C Limitations **21**

D Formal Definition and Evaluation Protocol of UM-GTE **21**

E Model Details **23**
E.1 Model Architecture 23
E.2 Promptable Input Implementation 24
E.3 Model Training 25

F EgoExo4D-Gaze Dataset Details **26**
F.1 EgoExo4D-Gaze Data Annotation 26
F.2 EgoExo4D-Gaze Data Statistics 27

G Additional Experiments and Details **28**
G.1 Comparison under Existing Multi-view Protocols 28
G.2 Qualitative Analysis 30
G.3 Prompt Controllability and Subject Grounding in Single-View Scenes 31

H Broader Impact **31**

A Extended Related Work: Gaze Estimation Paradigms

A.1 Single-view Gaze Target Estimation

Interpreting gaze is a cornerstone of human behavior understanding (Zhang et al., 2020a; Emery, 2000), which motivated the gaze-following task and its benchmarks (Recasens et al., 2015; Chong et al., 2020; Tomas et al., 2021; Tafasca et al., 2023). The dominant paradigm relies on multi-branch, fusion-based architectures that explicitly extract intermediate priors, including head position (Yang & Lu, 2024; Song et al., 2024; Yang et al., 2024; Ryan et al., 2025), pose (Bao et al., 2022), text (Wang et al., 2026b), facial expressions (Lai et al., 2025) and depth maps (Tonini et al., 2022), before fusing them to regress gaze coordinates (Cheng & Lu, 2022; Lin et al., 2025; Chong et al., 2018; Fan et al., 2018; Fang et al., 2021; Gupta et al., 2022; Lian et al., 2018b; Marin-Jimenez et al., 2019; Chen et al., 2021; Doosti et al., 2021; Zhao et al., 2020). These pipelines use auxiliary inputs to supply subject and scene cues; their input requirements differ in the amount of detection, pose estimation, or other preprocessing required. The field has recently moved toward end-to-end paradigms: GazeHOI (Tafasca et al., 2024) introduced open-vocabulary prediction, and GazeAnywhere (Cao et al., 2026a) pioneered promptable subject localization. These methods predict from a single view and therefore do not directly use observations of the same subject or target from other cameras to resolve occlusion and limited field of view.

A.2 3D Gaze Direction Estimation

Traditional 3D gaze estimation regresses a gaze direction, often parameterized by pitch and yaw from the subject's eyes or face (Lee et al., 2012; Wang & Ji, 2018). Early geometric approaches relied on explicit 3D eye models, whereas modern learning-based methods regress these vectors directly from appearance to handle in-the-wild variability (Zhang et al., 2020b; Li et al., 2021; Catruna et al., 2024), typically on benchmarks such as Gaze360 (Kellnhofer et al., 2019). A directional vector from a single view, however, lacks depth and cannot pinpoint a target object in 3D space (Lian et al., 2018a; Kawana et al., 2025). Works exploring panoramic settings or head-pose-only inputs still largely operate within a single coordinate frame (Nonaka et al., 2022; Nagpure & Okuma, 2023; Hisadome et al., 2024; Dai et al., 2025; Qu et al., 2025; Zhao et al., 2026). Beyond a single image, VideoGaze (Recasens et al., 2017) follows a movie character's gaze into neighboring frames of the same video, relating a source and a target frame through a learned planar affine transformation without calibration. Its frames are edited shots taken at different times, in which the camera, the scene, and the people change between shots, rather than multiple views of a shared scene. VideoGaze is therefore a pseudo-multi-view setting that differs from the UM-GTE definition. Moreover, film editing often cuts to what a character is looking at, so part of its signal can come from target-frame content rather than from the person's gaze. The MVGT benchmark (Miao et al., 2025) is the first to triangulate gaze targets in a global 3D world coordinate system from multiple views, which our work extends to prediction without supplied camera calibration or pre-aligned head boxes across views. The concurrent CoMind dataset (Gavryushin et al., 2026) also pairs egocentric and exocentric views with eye tracking, but targets joint attention in collaborative cooking rather than per-subject gaze target estimation.

B Ablation

This appendix complements the architectural ablations of Section 6 with three studies: the contribution of explicit geometry and its training objectives (Appendix B.1), the sensitivity of GazeWorld to how the subject is specified (Appendix B.2), and subject grounding in scenes with several people (Appendix B.3). The first two follow the protocol of Table 1: MVGT with leave-one-scene-out evaluation, all six views as input, and four columns, namely the L2 distance on evaluated views with the subject head visible (w/ Head) and without it (w/o Head), the in/out AP, and the triangulated 3D error. AUC, L2, and AP are first computed per held-out scene and then averaged with equal weight over the four scenes; the 3D error is the per-scene median averaged in the same way. Unless stated otherwise the model is prompted with the head box, and all variants share the three-stage recipe of Appendix E.3.

B.1 Ablation on Explicit Geometry and Loss Functions

Controlled geometry baseline. Table 6(A) compares the full model with a baseline without explicit geometry, while retaining the same frozen VGGT-Ω and DINOv3 (dino.txt) encoders. The baseline disables the world-point/confidence embeddings, camera/ray embeddings, 3D distance attention bias, and 3D head-center conditioning, together with all geometric losses throughout training. It retains the encoder patch features, ordinary cross-view attention, text and 2D box prompting, 2D positional information, and the 2D heatmap and in/out objectives, with the same fusion width and depth. VGGT-Ω features may still encode geometry implicitly; this comparison isolates the joint contribution of our explicit geometric mechanisms and supervision beyond those pretrained features. Both variants are trained separately through all three stages from the same pretrained encoders and matched initialization of shared trainable parameters, using the same data, schedule, iteration counts, and seeds. The baseline never inherits a checkpoint trained with explicit geometry; the full model follows the stage- and dataset-specific loss activation of Appendix E.3.

Stage-3 loss ablation. Table 6(B) isolates the contribution of each geometric objective during final fine-tuning. The objectives are the geometric alignment loss $\mathcal{L}_{\text{align}}$ (Eq. 3) and the gaze and head consistency losses $\mathcal{L}_{\text{cons}}^{\text{gaze}}$ and $\mathcal{L}_{\text{cons}}^{\text{head}}$ (Eq. 4). All explicit geometry modules remain enabled. Every variant starts from the same full-model Stage-2 checkpoint and is fine-tuned on each of the four leave-one-scene-out folds with one objective removed, or with all three removed; data, schedule, iteration counts, and seeds are matched. A check mark means that an objective follows its usual Stage-3 dataset gate, while a cross disables it throughout Stage 3. The all-cross row therefore measures the effect of removing geometric supervision during fine-tuning, whereas the baseline in (A) removes explicit geometry throughout training.

Evaluation. Both panels use the same six input views, box-prompt protocol, and metric aggregation as Table 1. Triangulated 3D errors are evaluated on a common set of eligible frames and views across variants using the same triangulation procedure. Evaluation calibration is used only to compute the 3D metric.

Results of the geometry ablation. Table 6(A) shows that explicit geometry improves every metric over the controlled baseline, which shares the same frozen encoders: the w/ Head L2 decreases from 0.102 to 0.082 (-20%), the w/o Head L2 from 0.198 to 0.152 (-23%), and the 3D error from 0.908 to 0.546 m (-36.2 cm, -39.9%), while the AP rises from 0.923 to 0.935. The largest relative gain occurs in 3D, followed by views without the subject's head, where subject and target information must be transferred across views; this matches Table 4, where removing the geometry tower mainly increases the w/o Head error. Because the baseline retains the VGGT-Ω patch features, these gains are attributable to the explicit geometric mechanisms and supervision rather than to the geometry encoder alone. Without explicit geometry, the model has lower 2D error than the calibrated MVGT baseline in Table 1 (w/ Head L2 0.102 vs. 0.118) but a higher w/o Head L2 (0.198 vs. 0.188) and 3D error (0.908 vs. 0.750 m); explicit geometry reverses both, which indicates that it is what allows the calibration-free model to surpass the calibrated baseline in 3D.

Results of the loss ablation. Table 6(B) shows that each geometric objective contributes and that the three are complementary. Removing any single one in Stage 3 degrades every metric, and the 3D error rises in all three cases compared with 0.546 m for the full model: to 0.610 m ($+6.4$ cm) without $\mathcal{L}_{\text{align}}$, to 0.595 m ($+4.9$ cm) without $\mathcal{L}_{\text{cons}}^{\text{gaze}}$, and to 0.578 m ($+3.2$ cm) without $\mathcal{L}_{\text{cons}}^{\text{head}}$. The alignment loss has the largest effect on localization and 3D accuracy (w/ Head L2 0.091 vs. 0.082, w/o Head L2 0.172 vs. 0.152), the gaze consistency loss causes a similar drop in AP (0.924 vs. 0.935), and the head consistency loss still improves every metric, including 3.2 cm in 3D. Removing all three at once degrades every metric most, even though all explicit geometry modules are retained: the w/ Head L2 rises from 0.082 to 0.095 ($+16\%$), the w/o Head L2 from 0.152 to 0.178 ($+17\%$), the AP drops from 0.935 to 0.912, and the 3D error grows by 14.2 cm to 0.688 m ($+26\%$). The drop in AP is the largest among all ablations in Table 6, which indicates that geometric supervision also helps the model decide in which views the target is in frame. Taken together with panel (A), explicit geometry and geometric supervision contribute complementarily to 3D accuracy: relative to the baseline without either (0.908 m), retaining the explicit geometry modules and the geometric supervision of Stages 1–2 reduces the 3D error to 0.688 m, and the Stage-3 geometric objectives reduce it further to 0.546 m.

Table 6: Explicit geometry and geometric loss ablations on MVGT (leave-one-scene-out, box prompt). (A) Matched dual-encoder models trained separately through all three stages. (B) Geometric loss ablations during Stage-3 fine-tuning only, starting from the same full-model Stage-2 checkpoint, with explicit geometry modules retained.

Configuration				w/ Head L2↓	w/o Head L2↓	Inout AP↑	3D Metric [m]↓
(A) Controlled geometry baseline: all three stages							
Same encoders, w/o explicit geometry or geometric losses				0.102	0.198	0.923	0.908
Full model				0.082	0.152	0.935	0.546
(B) Geometric loss ablation							
Explicit geometry	$\mathcal{L}_{\text{align}}$	$\mathcal{L}_{\text{cons}}^{\text{gaze}}$	$\mathcal{L}_{\text{cons}}^{\text{head}}$				
✓	✗	✗	✗	0.095	0.178	0.912	0.688
✓	✗	✓	✓	0.091	0.172	0.925	0.610
✓	✓	✗	✓	0.089	0.168	0.924	0.595
✓	✓	✓	✗	0.085	0.159	0.933	0.578
✓	✓	✓	✓	0.082	0.152	0.935	0.546

Table 7: Inference-time prompt ablation on MVGT (leave-one-scene-out, same weights for all rows). "Box views" is the number of views that receive the head-box prompt; "all" prompts every view in which the head is visible, as in Table 1.

Box views	Text	w/ Head L2↓	w/o Head L2↓	Inout AP↑	3D Metric [m]↓
0	✓	0.098	0.169	0.932	0.673
1	✗	0.095	0.170	0.930	0.640
2	✗	0.091	0.163	0.932	0.610
3	✗	0.087	0.160	0.932	0.603
all	✗	0.082	0.152	0.935	0.546
1	✓	0.095	0.165	0.933	0.638
all	✓	0.083	0.150	0.936	0.555

Of the total 36.2 cm reduction, 22.0 cm is thus obtained with explicit geometry and the Stage 1–2 geometric supervision, and a further 14.2 cm with the Stage-3 objectives.

B.2 Ablation on Prompt Selection

Design. UM-GTE designates the subject either by a head box in one prompted view or by a free-form description (Appendix D), whereas Table 1 reports the box prompt given in every view where the head is visible. We therefore evaluate, with the same weights and no retraining, how the result changes with the amount of prompt information: a text prompt alone; a head box in $k \in \{1, 2, 3\}$ views; a box in every head-visible view; and the combinations of text with a box in one view and in every view. For k-view prompts the prompted views are sampled among the views in which the subject's head is visible, with a seed derived from the frame's camera set, so the same cameras are prompted regardless of which view's prediction is being scored and all configurations see the same selection. A view that is not prompted receives the "withheld" presence value -1 (Appendix E.2), including views in which the head is not visible. Only the all-views setting, which is the protocol of Table 1, additionally tells the model in which views the subject is absent. When k is smaller than the number of head-visible views, a head-visible view whose prediction is scored may itself be unprompted, and the model must associate the subject across views on its own. In the MVGT test split the subject's head is visible in 4.9 of the six views of a w/ Head frame on average (in all six for 52% of frames, but only 2.8 on average in Shop), so even k=3 withholds the box from about two head-visible views per frame.

Table 8: Grounding and identity association on MVGT frames with several annotated people (1,435 frames; same weights as Table 1), measured only on views that received no box. *Wrong person*: the predicted head lies in a distractor's box and outside the target's. *Target absent*: only a distractor is visible and the prediction lands on it. Single-person frames (733) are given for reference.

Frames	Prompt	In target box ↑	Wrong person ↓	Target absent ↓	w/ Head L2 ↓
multi-person	text	98.7%	0.1%	4.5%	0.097
multi-person	one box	97.6%	0.1%	4.7%	0.094
single-person	text	99.7%	N/A	N/A	0.098
single-person	one box	98.6%	N/A	N/A	0.094

Results. Table 7 shows that, for the final model, accuracy grows with the number of views that receive a head box: the w/ Head L2 decreases monotonically from 0.095 with one box to 0.091, 0.087, and 0.082 with two, three, and all head-visible boxes, the w/o Head L2 from 0.170 to 0.152, and the 3D error from 0.640 to 0.546 m, while the AP stays within 0.930–0.936. This dependence is expected from the architecture: a box in a view also enables the subject-head branch for that view (Section 4), so each additional prompted view contributes a high-resolution head crop, and views without a box fall back to the scene tokens alone. The text description alone (0.098 / 0.169) performs on par with a single box (0.095 / 0.170), i.e., the model grounds the subject from text about as well as from one box. The single-box setting is also the one in which the model must propagate the subject's identity to the five unprompted views, and it does so without a loss relative to text prompting although prompt dropout during training removes the box from all views of a sample at once (Appendix E.2). Adding text on top of boxes changes the 2D metrics by at most 0.005 (w/o Head L2 0.150 vs. 0.152 with all boxes), so the two prompt types carry largely redundant subject information once a box is given. The spread across prompt configurations (0.082–0.098 w/ Head L2) remains smaller than the spread across held-out scenes under the all-box setting, which ranges from 0.069 in Kitchen to 0.097 in Commons.

B.3 Grounding and Cross-view Association with Multiple People

Design. The aggregate gaze metrics of Table 7 do not show whether the model actually locates the prompted person and carries that identity into the views that received no prompt. MVGT makes this measurable: several recording sessions contain two or three annotated participants who share the same physical frames (e.g., Commons 03_27 has three), and each participant's annotation file gives that person's head box in every camera. The other participants of a frame therefore provide ground-truth *distractor* heads. We split the leave-one-scene-out test frames of each scene into *multi-person* frames, in which the target's head and at least one distractor head are visible in the same view (1,435 frames), and *single-person* frames from sessions with one participant (733 frames); frames from multi-participant sessions in which the two never share a view are excluded from both groups (124 frames). Only annotated participants count as distractors, so the multi-person group is a lower bound on the number of people present.

For each frame we run the model once on all six views in a fixed camera order with the same weights as Table 1 under two prompts that require the model to find the person on its own in some views: the appearance text alone, and a head box in a single view chosen with the deterministic rule of Appendix B.2. Boxes in all head-visible views serve as the fully prompted reference and are not used to assess association. The model's estimate of the subject's head in a view is the peak of its head heatmap (Section 4), mapped back to image coordinates. Statistics are computed *only on views that received no box*. *Grounding* is measured on unprompted views in which the target's head is visible: the distance between the predicted and annotated head centers and the rate at which the prediction falls inside the annotated head box. *Identity* is measured on the subset of those views that also contain a distractor head: the prediction is classified as inside the target box only, inside both the target and a distractor box, inside a distractor box only, or inside neither. We also report, for unprompted views in which the target's head is *not* visible but a distractor is, how often the prediction falls on the distractor, i.e., how often another person is mistaken for the absent subject. Confidence intervals are 95% percentile bootstrap intervals that resample physical frames, since the views of one frame are not independent.

Results. Table 8 shows that, on multi-person frames, the model places the subject's head inside the annotated target box in 98.7% of the unprompted views with the text prompt (95% bootstrap interval over frames $[98.3, 99.0]$) and in 97.6% $[97.1, 98.2]$ with a single box, against 99.7% and 98.6% on single-person frames. In the $6{,}252$ unprompted views where a distractor is visible alongside the target, the prediction lies in a distractor box *and outside* the target box in 0.1% $[0.0, 0.2]$ of cases for both prompts (per scene 0.0–0.4%), which is the direct measure of identity confusion; a further 8–9% of predictions lie inside both boxes because the annotated boxes overlap in 21–26% of these views (e.g., two people seated side by side), and cannot be attributed to either person. The single-box setting requires the identity given in one view to be carried into the five others, and it reaches the same confusion rate as text prompting, in which every view is grounded from the description. The failure mode is different: when the target is out of view and only a distractor is visible (771 views), the prediction falls on the distractor in 4.5% $[3.1, 6.1]$ of cases with text and 4.7% with one box (13% in Kitchen, $\leq 5\%$ elsewhere), a case that the in/out head, rather than the head heatmap, addresses. Gaze accuracy on the same frames follows the prompt trend of Table 7 and is no worse on multi-person than on single-person frames. The model thus grounds and associates the subject on its own, without any external detector or association module.

C Limitations

GazeWorld supports multi-view GTE without supplied calibration or pre-aligned head detections, building on the geometric and semantic priors of foundation encoders.

Temporal modeling and data. GazeWorld reasons over individual multi-view frame groups and does not yet model temporal dynamics. Extending it calls for temporally dense multi-view gaze supervision, whereas EgoExo4D-Gaze deliberately keeps only frames whose cross-view targets pass verification. A practical path forward is an iterative active-learning process, in which the current model proposes gaze labels and human effort is spent only on verifying the uncertain ones.

Unexplored applications. GazeWorld may also be useful beyond perception, in directions this paper does not explore. Maintaining consistent, physically plausible character gaze across shots and temporal frames remains difficult. A calibration-free, promptable gaze estimator such as GazeWorld could serve as a supervision signal or an automatic critic for gaze consistency in generated videos with complex multi-person social scenarios. We leave this direction to future work.

D Formal Definition and Evaluation Protocol of UM-GTE

This appendix expands the formulation of Section 3 with the camera conventions, ground-truth semantics, and evaluation protocol that the main text omits for brevity. A sample is a scene-level group of N supplied views, processed jointly with one subject prompt. These are the views selected for the input group, not necessarily every image or camera available for that scene; the two-view comparison in Table 9 uses $N = 2$. The model estimates a single gaze target for that subject and expresses it through jointly predicted per-view 2D locations and in/out probabilities. No view has a primary or reference role in UM-GTE; the view containing a box prompt only specifies the subject. The geometry encoder introduces an estimated reconstruction frame, distinct from the calibrated world frame used for evaluation.

Input. A sample consists of $N \geq 1$ frames $\mathcal{V} = \{I^{(k)}\}_{k=1}^{N}$, $I^{(k)} \in \mathbb{R}^{3\times H_k \times W_k}$, captured at the same time instant by cameras with arbitrary, unknown, and possibly time-varying poses, together with a prompt $\mathcal{P}$ designating one subject. Resolutions and aspect ratios may differ across views, as is typical when egocentric and exocentric cameras are mixed. UMGaze extends this definition to sequential capture: scene photographs precede the subject photographs, and the relevant scene structure and target location are assumed to remain fixed, so all views can refer to the same 3D target.

Cameras. Each view has unknown intrinsics $\mathbf{K}_k$, distortion parameters $\mathbf{d}_k$, and world-to-camera extrinsics $\mathbf{P}_k = [\mathbf{R}_k \,|\, \mathbf{t}_k] \in \mathbb{R}^{3\times 4}$. For a world point $\mathbf{x}$, let $\mathbf{x}_{c,k} = \mathbf{R}_k\mathbf{x} + \mathbf{t}_k$ and $z_k(\mathbf{x}) = (\mathbf{x}_{c,k})_3$.

Write $\mathrm{dehom}([u, v, w]^\top) = [u/w, v/w]^\top$. For $z_k(\mathbf{x}) > 0$, define

$$\pi_k(\mathbf{x}) = \mathrm{diag}(W_k^{-1}, H_k^{-1})\ \mathrm{dehom}\left(\mathbf{K}_k \begin{bmatrix} \mathrm{dist}_{\mathbf{d}_k}(\mathrm{dehom}(\mathbf{x}_{c,k})) \\ 1 \end{bmatrix}\right). \tag{5}$$

Here $\mathrm{dist}_{\mathbf{d}_k} : \mathbb{R}^2 \to \mathbb{R}^2$ acts on normalized camera-plane coordinates and is the identity for a rectified pinhole view. Thus π_k returns image-normalized coordinates, which lie in $[0, 1]^2$ only for in-frame points. The calibration $\mathcal{C} = \{(\mathbf{K}_k, \mathbf{d}_k, \mathbf{P}_k)\}_{k=1}^N$ is never supplied to $\mathcal{F}$ at inference. It is used for dataset construction and the calibrated 3D evaluation. GazeWorld's geometric losses instead use frozen, predicted geometry; its task training does not supervise camera parameters with ground-truth calibration.

Subject prompt. $\mathcal{P}$ takes one of two forms: a box prompt $\mathcal{P} = (k_0, \mathbf{b})$, where $\mathbf{b} = (c_x, c_y, w, h) \in [0, 1]^4$ gives the normalized center, width, and height of the subject's head box in a single prompted view k_0, or a text prompt $\mathcal{P} = \tau$, a free-form description of the subject's appearance. Both designate exactly one physical person $\mathcal{S}$; text descriptions are verified to identify the subject uniquely in the scene. $\mathcal{S}$ must be visible in at least one view, but may be partially visible (e.g., head occluded) or entirely absent in any subset of the remaining views, including $\mathcal{S}$'s own egocentric view when $\mathcal{S}$ wears a camera. In this minimal-prompt setting, no cross-view correspondence of $\mathcal{S}$ is provided. Experiments that supply boxes in additional views use a richer prompt setting and state it explicitly: the bbox rows of Table 1 supply a box in every head-visible view, and Appendix B.2 varies the number of prompted views down to this minimal single-view setting.

Ground truth. For a sample with a finite gaze target, $\mathbf{g} \in \mathbb{R}^3$ denotes the attended scene point in the calibration's world frame. The ideal geometric in/out label is

$$o^{(k)} = \begin{cases} 0, & z_k(\mathbf{g}) > 0 \text{ and } \pi_k(\mathbf{g}) \in [0, 1]^2, \\ 1, & \text{otherwise}, \end{cases} \qquad \mathbf{g}^{(k)} = \pi_k(\mathbf{g}) \quad \text{when } o^{(k)} = 0. \tag{6}$$

In-frame status is distinct from visibility: let $v^{(k)} = 1$ indicate that the target is also unoccluded in camera k, so $v^{(k)} \le 1 - o^{(k)}$. Projected annotations can therefore have $o^{(k)} = 0$ and $v^{(k)} = 0$. Actual supervision follows each dataset's annotation protocol: EgoExo4D-Gaze includes projected exocentric targets behind occluders; the main MVGT localization and triangulation metrics use unoccluded, human-annotated targets. On MVGT, $\mathbf{g}$ is the official triangulated 3D annotation. When a finite target or its projection is unannotated, the corresponding localization metric is undefined; missing annotation alone does not imply an out-of-frame target.

Output. A model returns, for every view, a 2D target $\hat{\mathbf{g}}^{(k)} \in [0, 1]^2$ (decoded from the predicted heatmap) and an out-of-frame probability $\hat{o}^{(k)} \in [0, 1]$. Predictions are required for all N supplied views, including views in which the subject is not visible. The decoded coordinate localizes the target when it is in frame, and the out-of-frame probability indicates when it is not. The latent 3D target $\hat{\mathbf{g}}$ is not a required output; it is recovered from the per-view predictions at evaluation time, as described next.

2D metrics. For views with valid localization annotations, we report $\|\hat{\mathbf{g}}^{(k)} - \mathbf{g}^{(k)}\|_2$ in normalized image coordinates and heatmap AUC under the standard GTE protocol (Recasens et al., 2015; Chong et al., 2020). In the main MVGT table, AUC and L2 are computed from eligible per-view predictions with unoccluded, in-frame targets; L2 is split by whether the subject's head is visible in the evaluated image. AUC and AP use the head-visible view subset; the head-invisible subset supplies the second L2 column. These subsets select outputs for metric computation and do not assign input roles: GazeWorld jointly predicts all views of each scene group. AP uses the in-frame score $1 - \hat{o}^{(k)}$ against the label $1 - o^{(k)}$, including out-of-frame examples. Predictions and annotations are mapped from the preprocessed canvas back to the original image coordinates before computing L2 or performing calibrated triangulation. AUC, L2, and AP are computed per held-out scene and averaged equally over the four scenes. On EgoExo4D-Gaze, L2 is reported separately for egocentric and exocentric views because image-normalized distances have different meanings across their fields of view. Protocol-specific exceptions are described in Appendix G.

3D metric via triangulation. The 3D metric is reported on MVGT, the benchmark on which we evaluate against official triangulated 3D annotations; EgoExo4D-Gaze and UMGaze are evaluated with 2D metrics only. Given per-view predictions, each $\hat{\mathbf{g}}^{(k)}$ is mapped back to pixel coordinates, undistorted with $(\mathbf{K}_k, \mathbf{d}_k)$ to normalized camera coordinates $\hat{\mathbf{p}}_k = [x_k, y_k]^\top$, and a single 3D point is recovered by the direct linear transform (DLT): view k contributes the two rows $x_k\mathbf{P}_k^{(3)} - \mathbf{P}_k^{(1)}$ and $y_k\mathbf{P}_k^{(3)} - \mathbf{P}_k^{(2)}$, where $\mathbf{P}_k^{(i)}$ is the i-th row of $\mathbf{P}_k$, to a matrix $\mathbf{A} \in \mathbb{R}^{2M\times 4}$; $\hat{\mathbf{g}}$ is the first three components of the right singular vector of $\mathbf{A}$ with the smallest singular value, divided by its fourth component. The 3D error is $\|\hat{\mathbf{g}} - \mathbf{g}\|_2$ in the calibration's units, converted to meters for the main tables, with one value per physical frame. To keep the protocol symmetric with the ground truth, GazeWorld triangulation uses the M eligible unoccluded, in-frame views in the converted annotations, i.e., views with an occluded or out-of-frame target are removed; samples with $M < 2$, without an official 3D annotation, or with a numerically zero homogeneous denominator are skipped. We compute the median over eligible frames within each scene and average the four scene medians with equal weight. The separate Self-MVGTE comparison instead reports mean error in centimeters (Appendix G). This metric measures the *accuracy* of the triangulated 3D target under the true cameras; it does not depend on any depth or camera estimated by the model, and is therefore applicable to single-view and calibrated baselines alike.

Relation to existing settings. With $N = 1$ and a head-box prompt, the output task reduces to single-view GTE (Recasens et al., 2015; Chong et al., 2020), including its in/out decision. Supplying calibration and head boxes in both views recovers the inputs of the two-view MVGT baseline (Miao et al., 2025). Evaluating a wearer's egocentric view relates to egocentric gaze estimation (Huang et al., 2018; Huang & Sato, 2026), with exocentric views providing additional subject context. GazeWorld supports these view configurations through one promptable architecture; single-view baselines can still be evaluated independently on views where their required subject inputs are available.

E Model Details

This appendix expands the component descriptions of Section 4 with the tokenization, dimensions, and training configuration omitted from the main text. The configuration below describes the dual-tower model.

E.1 Model Architecture

Tokenization and dimensions. Input images are letterboxed to 512×512 or 448×592 and patchified at stride 16, giving a 32×32 grid (N_P=1024) for 512×512 inputs or a 28×37 grid (N_P=1036) for 448×592 inputs. For a given input size, both encoders use the same grid, so the two branches align without resampling. The semantic tower outputs 1024-dimensional text-aligned patch tokens, projected to D_S=256; the geometry features are 2048-dimensional (the aggregator concatenates frame-local and global tokens), projected to D_G=512. The two branches keep separate widths throughout fusion and interact only through cross-attention.

Geometry features. Rather than the last aggregator layer alone, we take features from four intermediate layers $\{4, 11, 17, 23\}$ of VGGT-Ω and combine them with a learned softmax-weighted sum: shallow layers retain boundaries and details of the small objects that often serve as gaze targets, while deep layers carry the aggregated cross-view context. VGGT-Ω exposes no point-map head; we obtain per-view world-point maps by unprojecting its predicted depth through its predicted camera pose, with the associated confidence retained as an extra channel. For the positional embeddings, the box-center conditioning, and the distance bias, world points are normalized by the per-scene centroid and mean radius; the geometric losses instead operate in the unnormalized reconstruction frame, in which the estimated cameras are defined, and divide residuals by the scene scale s (Appendix E.3). Both the pose and the world points are computed without gradients.

Geometric positional encoding. Each geometry token receives two additive embeddings: (1) the normalized world-point map average-pooled to the patch grid, concatenated with its relative confidence and passed through a two-layer MLP followed by layer normalization; (2) a ray

embedding produced by lifting the 9-dimensional pose encoding with Fourier features, broadcasting it over a normalized pixel grid, and applying an MLP. The fused sequence per view is [in/out token | 4 register tokens | N_P patch tokens], with rotary position embeddings on the patch part.

Fusion block. Writing $\text{Attn}(Q, K, V \mid B) = \text{Softmax}(QK^\top/\sqrt{d} + B)V$, with row-wise softmax and $B = 0$ when omitted, each of the L_A=6 blocks applies, in order,

$$Z_G'^{(v)} \leftarrow Z_G'^{(v)} + \text{Attn}(Z_G'^{(v)}W_Q,\ Z_S'^{(v)}W_K,\ Z_S'^{(v)}W_V), \tag{7}$$

$$Z_G' \leftarrow Z_G' + \text{Attn}(Z_G'W_Q,\ Z_G'W_K,\ Z_G'W_V \mid B_l), \tag{8}$$

$$Z_S'^{(v)} \leftarrow Z_S'^{(v)} + \text{Attn}(Z_S'^{(v)}W_Q,\ \tilde{Z}_G'^{(v)}W_K,\ \tilde{Z}_G'^{(v)}W_V), \tag{9}$$

where the first and third steps run per view, the second runs over the concatenation $Z_G' = [Z_G'^{(1)}, \ldots, Z_G'^{(N)}]$ of all views, and $\tilde{Z}_G'^{(v)}$ denotes the patch tokens of $Z_G'^{(v)}$ (the in/out and register tokens are excluded from the third step's key/value). $(B_l)_{ij} = -\text{softplus}(\gamma_l)\, d_{ij}$ is the 3D distance bias (Eq. 2): d_{ij} is the pairwise Euclidean distance between token positions in scene-normalized world coordinates, and γ_l is a learnable per-layer scalar. Entries d_{ij} in which i or j is a special token are filled with the per-sample mean patch-to-patch distance rather than zero—a zero entry would be the row-wise maximum bias and would systematically attract patch queries to the special tokens as γ_l grows, the opposite of the intended locality prior. These equations summarize the attention updates; layer normalization, output projections, and feed-forward residual updates are omitted for readability. The matrices W_Q, W_K, W_V are specific to each step and block. In this shorthand, Q, K have the same per-head width and the value/output projections return the query branch's width, so each residual addition preserves dimensions. All attention uses 8 heads.

Decoders. The gaze heatmap head reads the fused geometry patch tokens and the head heatmap head reads the semantic patch tokens; both reshape their tokens onto the input patch grid (32×32 in Stage 1 and 28×37 in the multi-view stages), apply convolutional upsampling, and bilinearly resize to the 64×64 output. The in/out head is a two-layer MLP on each view's in/out token. All heads run per view; no view is privileged.

E.2 Promptable Input Implementation

Text prompt. The description is tokenized and padded to a context length of 77; the frozen text tower yields token-level features whose `[EOS]` embedding summarizes the prompt. Text enters the model only through the Semantic Fusion module.

Box prompt. The head box is rasterized onto the input patch grid (32×32 or 28×37, depending on the stage) as a binary mask (a patch is inside if its center falls in the box), stacked with a per-view presence channel, and mapped by a two-layer MLP ($2 \to 128 \to 512$) whose final layer is zero-initialized, so at the start of training the box pathway is an exact no-op on the pretrained features. The presence channel is three-valued: 1 for a view with a box, 0 for a view with no valid supplied head box, and -1 when the box prompt is withheld for that view. The -1 sentinel distinguishes a withheld box prompt from a supplied annotation indicating no valid head box.

The box-conditioned configuration also samples scene-normalized world points at the supplied box centers and averages them across prompted views. A presence flag is concatenated with this 3D center and passed through an MLP whose last layer is zero-initialized; the resulting embedding is added to all geometry patch tokens. Both box-conditioning paths use the same prompt-dropout mask.

Prompt dropout. During training each sample keeps both prompts, only the text, or only the box with probabilities $(0.35, 0.45, 0.20)$ in the single-view stage and $(0.50, 0.25, 0.25)$ in the multi-view stages; the case with neither prompt is excluded, as the target subject would be undefined. Evaluation fixes the prompt configuration per protocol so that numbers are comparable across tables.

E.3 Model Training

Heatmap targets. The gaze target map places a Gaussian of fixed σ=3 (on the 64×64 grid) at the ground-truth location. The head target uses a size-adaptive $\sigma = \mathrm{clamp}(0.7 \cdot \frac{1}{2}\sqrt{wh} \cdot 64,\ 1,\ 8)$, where w, h are the box dimensions normalized to the input canvas (using the unclipped annotation when available), and the map is masked to a box-centered rectangle whose width and height are each at least 3 grid cells, rounded outward to pixel boundaries and clipped to the grid. A fixed σ is inadequate here because head sizes differ systematically across our training corpora (median half-widths of roughly 4.2, 2.5, and 1.7 grid cells in the single-view corpus, EgoExo4D-Gaze, and MVGT respectively), so a constant width would encode inconsistent semantics across stages.

Per-stage objectives. Stage 1 establishes single-view gaze prediction and subject grounding. In the current dual-tower configuration, it uses the 2D losses together with the gaze alignment and gaze consistency terms, with head consistency disabled. With one view, these geometric terms operate within that view rather than imposing cross-view agreement. Stage 2 (EgoExo4D-Gaze) uses $\mathcal{L}_{\text{align}}$, whose residual compares model predictions without using ground-truth gaze coordinates, and the head consistency loss; the gaze consistency loss is disabled, because EgoExo4D-Gaze's exocentric labels are projections that remain defined when the target is occluded, and sampling the world-point map at an occluded target's pixel returns the occluder's surface—supervision that is consistent but wrong. Stage 3 trains on a mixture of MVGT (0.5), EgoExo4D-Gaze (0.35), and the single-view corpus (0.15) to prevent forgetting, with all loss terms enabled on MVGT, while EgoExo4D-Gaze batches keep the Stage-2 configuration ($\mathcal{L}_{\text{align}}$ and head consistency enabled, gaze consistency disabled). In the multi-view stages, head consistency uses valid head-center annotations on all source corpora.

Confidence gating and scale. We use the following masks for Eqs. 3–4. Let m_v indicate an eligible annotation: an unoccluded target when visibility labels are used, and otherwise an in-frame target. For head consistency, m_v instead indicates a valid head annotation with an in-frame head center. At normalized canvas location $\mathbf{u}$, define

$$c_v(\mathbf{u}) = \mathbb{1}\left[C^{(v)}(\mathbf{u}) \geq 0.75\,\overline{C}^{(v)}\right]. \tag{10}$$

Here $C^{(v)}$ is the bilinearly sampled depth-confidence map and $\overline{C}^{(v)}$ its image mean; the confidence gates are computed without gradients. For a collection of per-view sampling locations $\{\mathbf{q}^{(v)}\}_{v=1}^{N}$, the source set $\mathcal{S}(\{\mathbf{q}^{(v)}\}_{v=1}^{N})$ contains views with $m_v c_v(\mathbf{q}^{(v)}) = 1$; if this set is empty, it falls back to all views with $m_v = 1$. We use $\mathcal{S}_{\text{pred}} = \mathcal{S}(\{\hat{\mathbf{u}}^{(v)}\}_{v=1}^{N})$ and $\mathcal{S}_{\text{gt}} = \mathcal{S}(\{\mathbf{u}^{(v)}\}_{v=1}^{N})$. Alignment residuals use $a_v = (1 - o^{(v)})c_v(\hat{\mathbf{u}}^{(v)})$ when a source exists, and zero otherwise. Thus reprojection residuals can include occluded in-frame targets even when those views are excluded from the source consensus. The stabilized projection clamps camera-space depth below at 0.01 and clips output coordinates to $[0, 1]^2$; this numerical operation is distinct from the true projection π_k in Eq. 5.

For consistency, P^{*}_{world} is the mean of $\mathcal{M}^{(v)}(\mathbf{u}^{(v)})$ over $\mathcal{S}_{\text{gt}}$, computed without gradients. With at least two source views, its reliability r is 1 only if their mean distance to this mean is below $0.1s$, and 0 otherwise. With one source view, $r = 0.5$; with none, $r = 0$. Residual weights are $w_v = m_v r c_v(\hat{\mathbf{u}}^{(v)})$. A zero total weight gives zero loss. For residual vector $\mathbf{e}$,

$$\mathrm{SmoothL}_{1,\beta}(\mathbf{e}) = \sum_{j=1}^{3} \begin{cases} e_j^2/(2\beta), & |e_j| < \beta, \\ |e_j| - \beta/2, & \text{otherwise}, \end{cases} \qquad \beta = 0.05. \tag{11}$$

The scale s is the mean distance of subsampled scene points to their centroid, clamped below at 10^{-3}. Eqs. 3–4 show a single frame group for readability. In a batch, the weighted residual sums and weight sums are accumulated across groups before division, and each group uses its own consensus, reliability, and scale. During training, differentiable coordinate extraction for the convolutional heatmap heads described above uses a spatial softmax of 50 times the predicted sigmoid heatmap scores and takes the expectation of pixel-center coordinates. Sampling converts $[0, 1]^2$ canvas coordinates to $[-1, 1]^2$ for bilinear grid sampling with `align_corners=False`. For quantitative evaluation, heatmap predictions are decoded with DARK; VLMs instead directly output gaze-target pixel coordinates, as described in Section 6.

Weights and schedule. The total loss is $100\,(\mathcal{L}_{\text{gaze}}+\mathcal{L}_{\text{head}})+\mathcal{L}_{\text{io}}+\mathcal{L}_{\text{align}}+\mathcal{L}_{\text{cons}}^{\text{gaze}}+\mathcal{L}_{\text{cons}}^{\text{head}}$; the factor 100 compensates for the small magnitude of pixel-averaged BCE on nearly empty maps. The in/out loss is a focal loss with balancing factor 0.25 and focusing parameter 2. All stages use AdamW on 4 H100 GPUs. The stated geometric weights apply only where the stage and dataset gates above enable them; disabled terms have zero weight.

F EgoExo4D-Gaze Dataset Details

F.1 EgoExo4D-Gaze Data Annotation

Rectification. Every view is undistorted into a pinhole frame before downstream processing because the projection described below assumes a pinhole camera model. Placing labels directly on raw fisheye frames would introduce significant distortion at the periphery, where human heads and gaze targets frequently appear. The egocentric Aria RGB stream uses the `FISHEYE624` model. We construct its rectification map by projecting a target pinhole pixel grid directly through the device camera model, ensuring the mapping is an exact numerical inverse rather than an approximate fitted polynomial. This stream is also rotated upright. Exocentric GoPro views are undistorted using the Kannala-Brandt fisheye model to target pinhole intrinsics and exported at 1920×1080. These target intrinsics are stored alongside the annotations for each camera, ensuring consistent geometry across labels and images. Because rectification crops peripheral regions, original fisheye frames are retained for border-case validation.

Targets by projection. For each sampled frame, we take the wearer's 3D gaze target $\mathbf{g} \in \mathbb{R}^3$, recovered by the Aria eye tracker and provided by Ego-Exo4D, and project it into every exocentric view using the calibrated intrinsics and extrinsics, $\mathbf{p}_k = \pi_k(\mathbf{g})$ (Eq. 5). A view is labeled as in-frame when the target has positive camera-space depth and $\mathbf{p}_k \in [0,1]^2$. Data provenance is tracked per view: *measured* for egocentric views and *projected* for exocentric views. This allows training losses to down-weight projected supervision when needed. Frames are sampled at 2 FPS and grouped by timestamp. A group is retained only if at least two views are valid.

Cross-view verification of projected targets. Target projections rely entirely on camera poses, temporal synchronization, and wearable eye tracking. Even when a target is geometrically valid, it may still be semantically incorrect due to mid-saccade motions, vergence collapse onto hands, or rapid head rotations. We therefore cross-verify each candidate group against visual evidence in *the egocentric view itself*, where raw eye-tracker measurements reveal where the wearer was looking.

Each multi-view group is compiled into a single composite image containing individual tiles for each exocentric camera (showing the head bounding box, projected target, and a zoomed crop) alongside the egocentric frame with its raw gaze point. Gemini 3.8 Flash then inspects this collage and assigns a per-view verdict of *correct*, *wrong*, or *cannot tell*. The vision-language model evaluates individual views, while final group-level acceptance is determined locally by a rule-based aggregator so the filtering logic remains fully transparent. In parallel, four geometric sanity checks reject a group if any condition is met: a 3D target without surface support, a near-body target with zero support (indicating vergence collapse near the hands), head angular velocity exceeding $120^\circ/\text{s}$, or a target landing inside the wearer's own head box.

This quality control step represents the largest share of filtering. Roughly 81% of the $1{,}495{,}217$ candidate groups were rejected. This high rejection rate is intentional: while the eye tracker provides raw measurements continuously, only a fraction of frames offer targets that are well-projected, physically grounded, and actively attended.

Head boxes by calibrated projection. Standard detectors cannot reliably label the wearer because multiple bystanders often appear in the scene. We instead derive wearer head boxes geometrically. A naive estimate places the box center at the projected Aria camera center $\mathbf{C}$, with the box size determined by a fixed physical head dimension scaled by camera distance. However, this produces a systematic offset because the headset sits in front of the head, causing the projected camera center to deviate from the true head center by a median of 0.43 box widths.

We model this displacement as a constant offset in the Aria coordinate frame and fit it, together with a box-size scale, on a subset of groups whose head centers are triangulated from SAM 3 (Carion et al., 2026) head detections across synchronized views. This single fitted correction reduces the center residual from 0.43 to 0.12 box widths and is applied to all 1.0M exocentric boxes.

We record both image-clipped and unclipped boxes, ensuring that heads along image boundaries retain supervision peaks at their true locations. For views with a candidate box, the reliability labels are: *reliable* ($954{,}703$ views, 93.4%) supervised by the head loss, and *masked* ($67{,}592$ views, 6.6%) whose images remain as visual context but whose boxes are withheld from prompting and head supervision. A third state denotes confirmed head absence: these views have no box and use an all-zero head heatmap target. Egocentric views contain no head box. These annotation states are distinct from the prompt-presence values described in Appendix E.2.

MLLM verification of borderline head boxes. Calibration removes the systematic bias but leaves per-view residual errors. We score each box by its center distance d (normalized by box width) to the segmenter's proposal. Distance thresholds are reliable only at the extremes; for the ambiguous intermediate range, roughly 5% of exocentric views, Gemini 3.8 Flash is shown the original annotation, the calibrated candidate, and the segmented mask, and judges whether the calibrated box bounds the wearer's head. On a human-annotated validation split, it agrees with the ground truth in 96.3% of these cases, compared with 14.8% for fixed thresholding. The verdict sets the reliability flag without modifying the box coordinates.

Because pinhole rectification crops image boundaries, heads near the border may be cut off despite remaining visible in the original fisheye footage; these borderline views are therefore verified directly on the original fisheye images with the same prompt. Rejected boxes are masked rather than deleted, preserving the corresponding gaze supervision.

Residual audit. The steps above filter the majority of candidate samples. After all preceding filters, the remaining $264{,}705$ groups were reviewed through three targeted auditing steps to reach the final release count. First, groups with no valid head box in any view were removed (107 groups), as subject association is ill-defined when no camera observes the wearer. Second, an active-learning ranker flagged high-error exocentric views for manual inspection. The top 1% of groups by error were reviewed manually; 75 were deleted and the rest were corrected. Third, we aggregated errors per camera take and compared each camera against its peers, which identified and corrected view-specific anomalies. Through these steps, the $264{,}705$ candidate groups were refined to the final $264{,}523$ dataset groups ($107 + 75 = 182$ removed).

F.2 EgoExo4D-Gaze Data Statistics

Figure 3 illustrates the overall composition of the $264{,}523$ released multi-view groups ($1{,}286{,}762$ annotated views).

Activities. The dataset spans 80 distinct activity categories from the Ego-Exo4D task taxonomy. The category distribution is heavy-tailed: the single largest category (*Cooking Noodles*, $34{,}439$ groups) accounts for 13% of the total data, while the top 20 categories collectively cover 81.7%, leaving 60 categories in a sparse tail. The head of this distribution consists primarily of procedural tabletop activities, such as cooking, sample preparation, and bike repair. The tail adds sports and performance activities such as basketball, bouldering, and dance, in which targets are distant and head-target geometry differs markedly, broadening the coverage beyond tabletop scenes.

Views per group. The number of cameras per group ranges from 2 to 6, with an average of 4.86 views. Five-view configurations represent the most common layout (57.5%), followed by four-view (18.9%) and six-view (17.5%) groups. This view variability arises naturally from the physical camera arrangements in the original takes, where the visibility of the wearer fluctuates across exocentric viewpoints. This variable-view setup reflects realistic in-the-wild deployments and motivates a framework that accepts variable numbers of views.

Target provenance. Among the 1.29M view-level annotations, 20.6% are egocentric targets measured directly by the wearable eye tracker, 78.0% are exocentric targets obtained by reprojecting

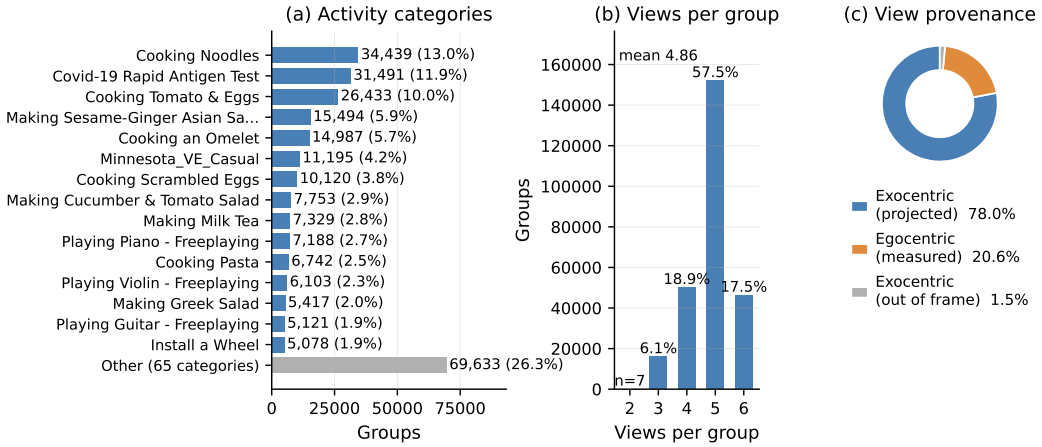


Figure 3: Composition of EgoExo4D-Gaze computed from the released annotations. (a) Activity categories: the distribution over the 80 categories has a long tail, so the 15 largest are displayed individually, annotated with their group count and share, and the remaining 65 are grouped together. (b) Views per group: unlike benchmarks that assume fixed camera rigs, the group size ranges from 2 to 6, with a mean of 4.86. (c) Provenance of the 1.29M view-level gaze targets, with the share of each source in the legend.

the corresponding 3D gaze location into calibrated camera views, and the remaining 1.5% represent exocentric viewpoints where the 3D gaze target lies outside the field of view. Of the 264,523 groups, 264,451 retain one egocentric view and 72 contain only exocentric views. The targets within each group derive from the same eye-tracker measurement and its 3D location, so all views of a group share one physically consistent target. We maintain explicit provenance metadata for every view, enabling training objectives to flexibly balance or down-weight projected supervision relative to direct sensor measurements.

Annotation format. Figure 4 illustrates representative multi-view groups across diverse domains, including basketball, bouldering, cooking, and dance. Each instance supplies per-view 2D gaze targets, in-or-out-of-frame indicators, and wearer head bounding boxes for exocentric perspectives, accompanied by a global natural-language prompt describing the target individual. This prompt specifies the intended wearer in scenes that may contain other people; detecting a person and selecting the person whose gaze is queried are distinct operations. Moreover, visualizing the camera viewing rays illustrates the cross-view geometry of 2D target annotations: rays through corresponding target pixels meet at a common 3D location under the calibrated projections, while the target occupies different image locations across views. These examples motivate using shared spatial structure for multi-view gaze estimation.

G Additional Experiments and Details

G.1 Comparison under Existing Multi-view Protocols

To situate GazeWorld within the evaluation conventions established by prior multi-view work, we additionally report results under two externally defined protocols: the original two-view protocol of MVGT (Miao et al., 2025) (Table 9) and the protocol of the concurrent, self-supervised Self-MVGTE (Chen et al., 2026) (Table 10). In both cases we follow the respective protocol.

Two-view MVGT protocol (Table 9). The terms *primary* and *reference* in this paragraph refer only to the original MVGT evaluation protocol, not to roles in the GazeWorld task or architecture. For this comparison, GazeWorld receives each two-view group jointly and predicts both views; the evaluator selects the output required by MVGT. MVGT evaluates pairs of views: the error is measured in the *primary* view, whose subject head is visible; localization error additionally requires

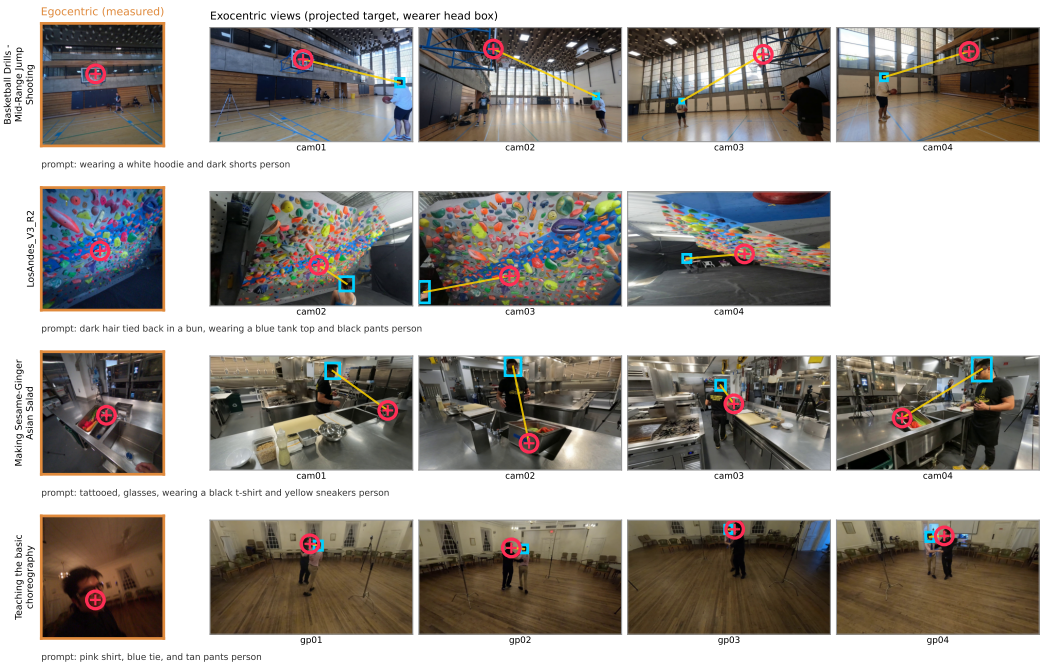


Figure 4: Visualization of annotation samples, arranged one group per row. Left (orange outline): egocentric view with direct eye-tracker target. Right: exocentric views showing the wearer's head box (cyan), projected target (red), and the 3D-to-2D projected gaze line (yellow). The subject-identifying text prompt is shown beneath each group. Rows are sampled across four representative categories (basketball, bouldering, cooking, and dance) under deterministic criteria: unique category, valid text prompt, at least three exocentric views with reliable head boxes, and in-frame targets across all shown cameras.

a valid in-frame target, while the four column groups of the table correspond to the joint (head, target) visibility categories of the *reference* view. Following MVGT, scores over the five possible pairings of each primary view are first averaged, then averaged over primary views, and finally leave-one-scene-out averaged over the four scenes. GazeWorld is trained with four to six views per group yet evaluated here with only two, whereas the MVGT baseline is trained natively for the two-view setting. Under this protocol GazeWorld attains the best distance and the best AP in all four reference-view categories; the AP margins are largest when the reference view's head is not visible ($+0.061$ and $+0.046$ over MVGT), whereas the distance margins are similar across categories (0.023–0.036).

Self-MVGTE protocol (Table 10). Self-MVGTE predicts gaze targets directly in 3D and evaluates with four metric families: 2D distance and in/out AP over all views (unmarked), the same metrics restricted to views where the subject head is visible (marked *), the mean 3D distance to the official target over physical frames (in cm), and the angular error measured at the annotated 3D eye position. Three protocol differences distinguish the two settings. **(1) Sample removal:** Self-MVGTE removes every frame whose head is visible in fewer than two views from *both training and testing*, which discards 25.2% of the Shop scene. For comparability, the GazeWorld row applies the same removal. **(2) Calibration:** Self-MVGTE requires full camera intrinsics and extrinsics at inference for eye triangulation, ray casting, and per-view projection, whereas GazeWorld is calibration-free. **(3) Prompting:** Self-MVGTE does not provide text-based or single-view-box subject prompting, while GazeWorld is promptable by text or by a single-view head box; the table reports the bbox prompt. Under these differing input and supervision assumptions, GazeWorld surpasses both the fully supervised MVGT baseline and the self-supervised Self-MVGTE on every reported metric, although it receives no camera calibration at inference. Relative to Self-MVGTE, it reduces the 2D distance by 12.9% on head-visible views ($0.093 \rightarrow 0.081$) and by 9.3% over all views ($0.108 \rightarrow 0.098$), improves AP*/AP by 0.009/0.013, and lowers the mean 3D distance by 6.9 cm ($74.09 \rightarrow 67.23$ cm, 9.3%) and the angular error from $13.08°$ to $12.78°$. Relative to the

Table 9: Comparison under the original two-view MVGT evaluation protocol; rows above the line, including the asterisk marker, are taken from MVGT (Miao et al., 2025), Table 1. N/A: the Center baseline makes no in/out prediction.

Method	Head Visible				Head Not Visible			
	Target Visible		Target Not Visible		Target Visible		Target Not Visible	
	Dist.↓	AP↑	Dist.↓	AP↑	Dist.↓	AP↑	Dist.↓	AP↑
Random	0.451	0.555	0.456	0.546	0.464	0.502	0.457	0.564
Center	0.259	N/A	0.261	N/A	0.272	N/A	0.253	N/A
Chong (Chong et al., 2020)	0.159	0.855	0.157	0.862	0.191	0.792	0.174	0.866
Miao (Miao et al., 2023)	0.141	0.886	0.140	0.892	0.164	0.831	0.163	0.886
Tafasca* (Tafasca et al., 2023)	0.149	0.893	0.148	0.895	0.166	0.831	0.154	0.884
MVGT-Single (Miao et al., 2025)	0.151	0.877	0.148	0.878	0.179	0.758	0.154	0.855
MVGT (Miao et al., 2025)	0.129	0.909	0.122	0.912	0.161	0.836	0.152	0.868
GazeWorld (ours)	**0.093**	**0.929**	**0.099**	**0.920**	**0.132**	**0.897**	**0.124**	**0.914**

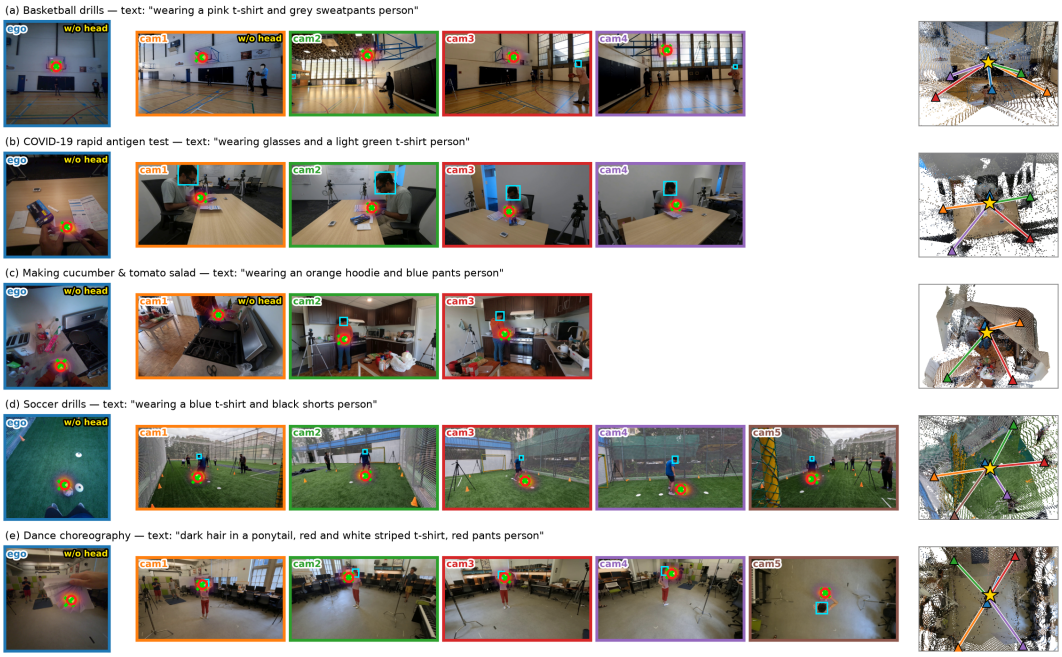


Figure 5: Qualitative results on the EgoExo4D-Gaze test set, one group per row from five different activities. Left: all views with the predicted heatmap, ground truth (green cross), prediction (red circle), and annotated subject head box (cyan); "w/o head" marks views where the subject's head is not visible but the target is in frame, which includes every egocentric view. Right: VGGT-Ω reconstruction with the predicted cameras (triangles, colored as the view frames), the viewing ray from each camera to its sampled 3D target, and the mean of the sampled points (yellow star). Rows are selected as described in Appendix G.2.

fully supervised MVGT baseline, which also requires calibration, the 3D distance drops by 30.6 cm (31%) and the angular error by 5.9°. The 3D values in this table are means over frames, whereas Table 1 reports the per-scene median; the two are therefore not directly comparable.

G.2 Qualitative Analysis

Figures 5 and 6 extend Figure 2 to more scenes. Each row is one multi-view group of synchronized views. The left part shows every input view with the predicted gaze heatmap, the ground-truth target (green cross), the predicted target (red circle), and the annotated head box of the subject (cyan); the text description of the subject is shown above each row; a view in which the subject's head is not

Table 10: Comparison with the concurrent self-supervised Self-MVGTE under its evaluation protocol on MVGT; rows above the line are taken from Self-MVGTE (Chen et al., 2026), Table 1.

Method	Calib.-free	2D Dist.* ↓	AP* ↑	2D Dist. ↓	AP ↑	3D Dist. [cm] ↓	Ang. [°] ↓
Gaze-LLE (Ryan et al., 2025)	-	0.200	0.831	-	-	-	-
MVGT baseline (Miao et al., 2025)	✗	0.119	0.907	-	-	97.86	18.69
Self-MVGTE (Chen et al., 2026)	✗	0.093	0.929	0.108	0.910	74.09	13.08
GazeWorld (ours)	✓	**0.081**	**0.938**	**0.098**	**0.923**	**67.23**	**12.78**

visible but the target is in frame is tagged "w/o head". The right part is the VGGT-Ω reconstruction of the same group, rendered from a virtual viewpoint: each camera is drawn as a triangle in the color of its view's frame, a ray connects it to the 3D point sampled from its own predicted world-point map at its predicted target, and the yellow star is the mean of these points. Cameras, depths, and world points are all predicted by the model; no calibration is used.

Selection protocol. The models are the ones evaluated in Tables 1 and 2: the Stage-2 model for EgoExo4D-Gaze, with examples drawn from a subset of the test split (3,722 groups covering all 835 test takes), and, for each MVGT scene, the leave-one-scene-out model that never saw that scene. For qualitative example selection only, every group is scored by the mean L2 distance between the argmax prediction and the ground truth over its annotated views; the quantitative tables use DARK decoding for heatmap outputs. On EgoExo4D-Gaze we take the best group of each coarse activity category among groups with at least five views, all of them annotated, and show five of these categories; on MVGT we show one frame per scene, chosen among frames with at least five annotated views (three in Shop, where fewer cameras see the subject). To include views without the subject's head, row (c) of Figure 5 and rows (b) and (d) of Figure 6 are instead chosen among groups that contain such an exocentric view, without the five-view requirement, scored by the error on those views together with the group mean; row (a) of Figure 5 contains one as well.

In the selected examples from the three datasets (UMGaze in Figure 2(c)), the per-view predictions agree approximately in 3D: the viewing rays from the cameras pass near a common point in the reconstruction, although the model never receives ground-truth camera poses and the 3D point of each view is read from that view's own predicted depth. This holds for views in which the subject is absent. The egocentric view of every EgoExo4D-Gaze group contains no head by construction, yet its prediction lands on the same object that the exocentric views point to, e.g., the rim in (a) and the test kit in (b). In row (c), exocentric camera 1 sees only the subject's arms above the stove, and in MVGT the Kitchen camera in (b) and the Shop camera in (d) observe the target but not the subject; in these views the other views provide subject cues for target estimation through the shared representation. The same holds in crowded scenes such as the basketball court in (a) and the dance studio in (e), where several people are visible in each view and the prompt must single out the subject in every camera. On UMGaze, a site never seen in training, the text prompt alone locates the subject and the predictions stay consistent across the six cameras (Figure 2(c)). The in/out head also behaves as intended: in row (d) of Figure 6, the target lies outside cameras 3 and 5, and the model predicts no in-frame target in exactly these views (no red circle is drawn).

G.3 Prompt Controllability and Subject Grounding in Single-View Scenes

Figures 7–10 show selected visualizations of GazeWorld on the single-view benchmarks VideoAttentionTarget, GazeFollow, and ChildPlay under text-only prompting. Figures 7 and 8 prompt different people in the same image, and Figures 9 and 10 show crowded scenes. In these examples, GazeWorld grounds the described subject and localizes their gaze target accurately, showing strong performance on text-prompted GTE.

H Broader Impact

Supporting research on early autism screening and assessment. Quantifying where children look may help researchers characterize social attention during interactions with caregivers and their surroundings. Clinical research has shown that eye-tracking measures of social visual engagement

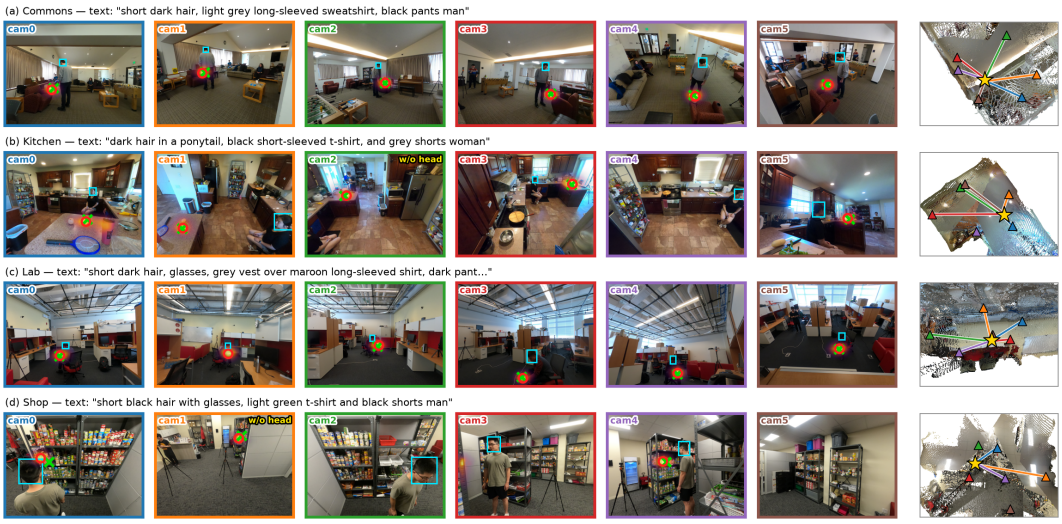


Figure 6: Qualitative results on MVGT under the leave-one-scene-out protocol, one frame per scene; each row is predicted by the model that was not trained on that scene. Rows (b) and (d) contain a view in which the subject's head is not visible but the target is ("w/o head"). Notation as in Figure 5.

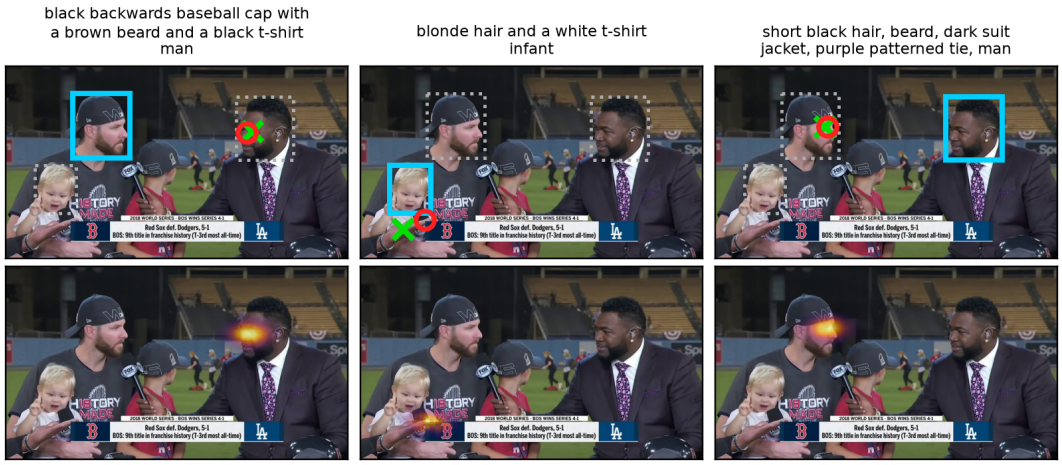


Figure 7: Prompt controllability on a single image with several annotated people (VideoAttentionTarget, `MLB Interview`). The same model and the same image are used in all three columns; only the text prompt changes, shown above each column. Top: the prompted person's head box (cyan) is drawn for reference and the other annotated heads are dotted grey; neither is given to the model under text-only prompting. Green cross is the ground-truth target, red circle the prediction. Bottom: the predicted heatmap for the same prompt.

can support assessment of autism in young children referred to specialty clinics (Jones et al., 2023). GazeWorld could contribute a complementary measurement tool for studying such behavior in natural settings: multiple views can capture a child and the attended target even when they do not appear together in one image, without requiring camera calibration as model input. With additional temporal modeling and validation, these estimates could help quantify attention to people and objects or responses to another person's attentional cues, supporting research toward earlier screening and clinician-led assessment. Our experiments evaluate gaze localization, however, and provide no evidence of autism screening or diagnostic accuracy. Translation to clinical use would require prospective studies in the intended population, validation against appropriate behavioral and clinical reference measures, and assessment of errors across ages and demographic groups. Gaze patterns alone should not be used to diagnose autism or characterize an individual's social abilities.

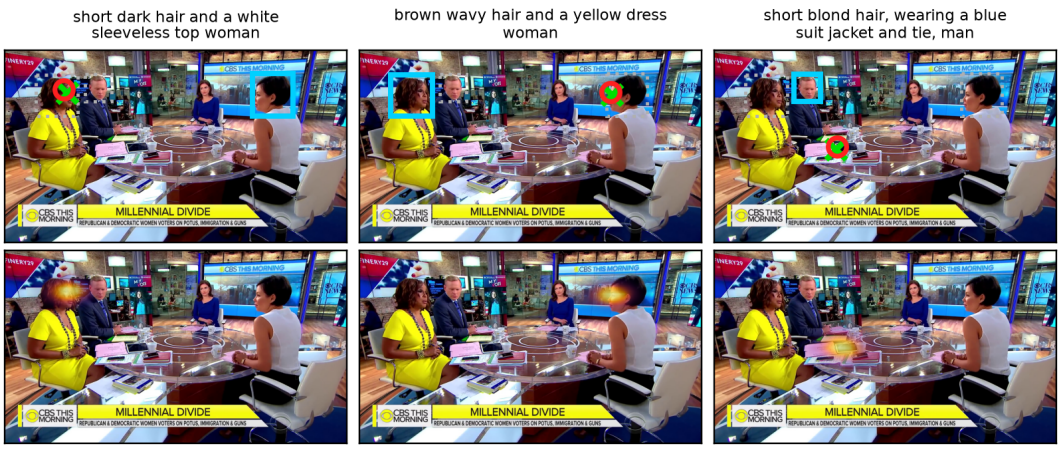


Figure 8: A second prompt-switching example (VideoAttentionTarget, `CBS This Morning`), with four people seated at the table and three of them annotated. Notation as in Figure 7.



Figure 9: Text grounding in crowded GazeFollow images, where only the prompted person is annotated. Grey boxes are all people found by SAM 3 (Carion et al., 2026) with the prompt "person"; they measure how crowded the scene is and are *not* given to the model, which receives only the description above each column. Cyan is the annotated head of the prompted person, the green cross is the target (mean over the ten test annotators), the red circle the prediction. Bottom: the predicted heatmap.

Human intention understanding and social embodied AI. Gaze targets can provide context for interpreting actions and communication. Prior work has combined gaze and body-motion cues to model action anticipation in human–robot interaction (Duarte et al., 2018). GazeWorld could supply a spatial attention cue to future socially interactive embodied agents, linking a prompted person to a potential target across camera views. Combined with language, gesture, and task context, this could help a robot resolve an ambiguous object reference, establish joint attention, or identify a candidate object during collaborative assistance. These are potential downstream uses rather than capabilities demonstrated by our gaze-estimation benchmarks. Looking at an object does not necessarily imply an intention to act on it; an interactive system should allow users to correct an inferred target and request clarification before consequential actions. Temporal reasoning, evaluation in closed-loop interactions, and alignment of the estimated scene with the robot's coordinate frame would also be needed for practical deployment.

Responsible development. The ability to associate a person with targets across views also creates privacy risks, particularly in clinical environments and spaces shared with bystanders. Future appli-



Figure 10: Text grounding in crowded ChildPlay classrooms. Notation as in Figure 9; ChildPlay provides a single target annotation per subject.

cations should use consent-based data collection, limit retention and access to identifiable imagery, and evaluate performance with the communities they are intended to serve. Clinical and assistive systems should support individual needs and human judgment without treating differences in gaze behavior as evidence of incapacity or lack of engagement.