GAZEGGT: GEOMETRY-GROUNDED TRANSFORMER FOR UNCALIBRATED MULTI-VIEW GAZE ESTIMATION

Anonymous authors
Paper under double-blind review

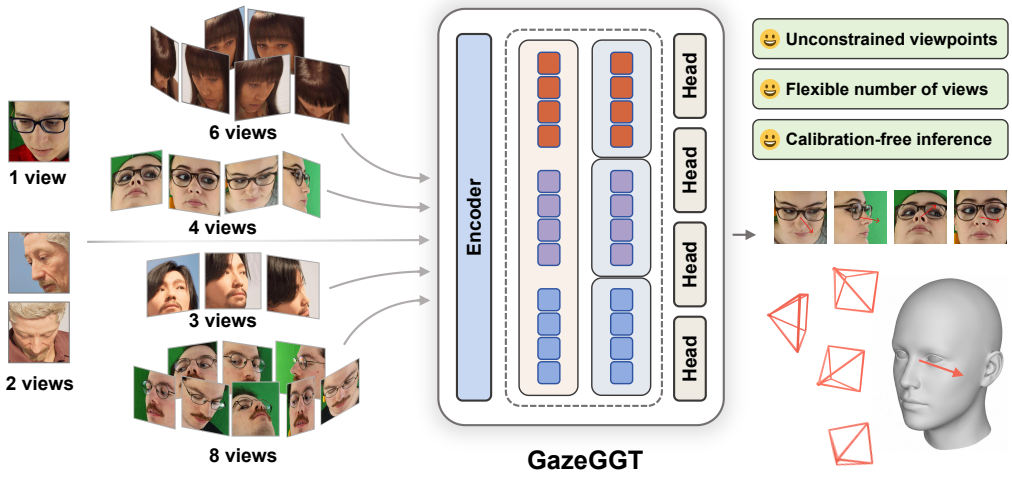

Figure 1: **Feed-forward multi-view gaze estimation via GazeGGT.** Without camera calibration or extrinsics, our multi-view gaze transformer supports unconstrained camera layouts and arbitrary view counts, where adding views naturally expands coverage while boosting gaze accuracy.

ABSTRACT

Robust 3D gaze estimation remains challenging because monocular appearance cues are inherently under-constrained: eye visibility, head pose, and self-occlusion can make the same gaze direction appear substantially different across views. Multi-view observations offer a natural remedy, but existing multi-view gaze estimators typically depend on fixed camera layouts, calibrated cameras, or test-time adaptation, limiting their applicability in flexible real-world environments. We study the underexplored problem of uncalibrated multi-view gaze estimation, where multiple synchronized face observations are available but camera intrinsics, extrinsics, and view configurations are unknown at inference. We propose *GazeGGT*, a geometry-grounded transformer for unconstrained, calibration-free multi-view gaze estimation. GazeGGT encodes an arbitrary view number of face crops into per-view features and employs alternating multi-view attention to fuse information across views, followed by separate decoders for per-view 3D gaze directions and camera poses. Unlike prior multi-view methods, GazeGGT requires no camera parameters and naturally supports a variable number of input views, enabling flexible camera configurations. We further introduce a *Geometry-Grounded Supervision Paradigm* that jointly supervises gaze and camera pose predictions, providing the geometric grounding necessary for accurate calibration-free gaze estimation. Experiments show that GazeGGT achieves competitive performance in both single-view and dual-view settings and consistently benefits from increasing numbers of input views. Ablation studies further demonstrate the importance of the Geometry-Grounded Supervision Paradigm for calibration-free estimation. Our results establish a practical and extensible approach to gaze estimation in unconstrained multi-camera environments.

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.

Table 1: **Comparison with existing methods.** *Calibration-free:* camera extrinsics are not required at inference. *Unseen camera layouts:* generalization of a single model to arbitrary new setups, demonstrated by transfer between ETH-XGaze and GazeGene rigs. *Flexible:* supporting an arbitrary number of views with the same single model (trained on 1 to 9 views, evaluated up to 18 views).

Method	Number of views	Calibration-free	Unseen camera layouts
MVMT Lian et al. (2018)	3-view	×	×
DVGaze Cheng & Lu (2023)	2-view	×	×
UVAGaze Liu & Lu (2024)	2-view	✓	×
Rot-MVGaze Hisadome et al. (2024)	2-view	×	✓
GazeGGT (ours)	Flexible	✓	✓

1 INTRODUCTION

Estimating where a person is looking in 3D is a fundamental capability with broad applications in human-computer interaction Majaranta & Bulling (2014); Cai et al. (2025), behavioral analysis Morales et al. (2017); Sharma & Chakraborty (2024), robotics Wang et al. (2015); Aronson et al. (2018); Terzioğlu et al. (2020), and XR Kytö et al. (2018); Wang et al. (2020). Despite significant progress in appearance-based gaze estimation Kellnhofer et al. (2019); Cheng & Lu (2022); Qin et al. (2026), single-view methods remain fundamentally limited. In particular, eye-region cues become unreliable under extreme head poses, and monocular observations provide insufficient geometric constraints. As a result, robust 3D gaze estimation from a single view remains challenging in real-world scenarios. Multi-view setups offer a natural solution. By observing the same subject from multiple viewpoints, they not only expand pose coverage but also introduce complementary geometric constraints that can improve estimation accuracy. However, effectively exploiting such multi-view information in practice remains an open challenge.

Existing dual-view and multi-view gaze estimation methods Lian et al. (2018); Cheng & Lu (2023); Liu & Lu (2024) typically rely on retraining or adaptation to handle new camera configurations, making them difficult to generalize to flexible multi-camera setups. Other approach Hisadome et al. (2024) explicitly takes calibrated camera poses as input to enable flexible dual-view reasoning. While effective, this approach is limited to two-view settings and depends on accurate camera calibration, which is often unavailable or costly to maintain in real-world deployments. Consequently, multi-view gaze estimation without camera calibration remains largely underexplored. Table 1 summarizes this gap from a deployment-oriented perspective.

In this work, we revisit multi-view gaze estimation from a different perspective: *Can a single feed-forward model accept an arbitrary number of views and achieve increasingly accurate gaze estimation as more views become available?* Furthermore, *can it jointly infer multi-view gaze directions and relative camera poses, particularly rotations, thereby enabling uncalibrated estimation?*

To answer these questions, we propose *GazeGGT*, a geometry-grounded transformer for unconstrained, calibration-free multi-view gaze estimation. GazeGGT takes an arbitrary number of face crops as input and represents each view with a view-specific feature. These features are then fused through alternating multi-view attention, allowing information to flow across views without requiring a predefined number or configuration of cameras. Separate decoders predict the 3D gaze direction for each view and the relative camera pose between views.

To enable calibration-free gaze estimation, we further introduce a *Geometry-Grounded Supervision Paradigm* that jointly supervises gaze and camera pose predictions. This geometric supervision substantially improves the accuracy of relative rotation estimation without compromising gaze accuracy, allowing gaze estimates to be transformed into a reference camera coordinates without calibration.

Extensive experiments show that GazeGGT achieves competitive performance in both single-view and calibration-free multi-view settings and consistently benefits from additional views. These findings are validated in both in-domain and cross-domain settings. Notably, our cross-domain experiments between ETH-XGaze and GazeGene involve substantial differences not only in image appearance but also in multi-camera layouts, providing a challenging test of the model's generalization to different camera configurations. Ablation studies further demonstrate the importance of the Geometry-Grounded Supervision Paradigm for calibration-free estimation.

2 Related Work

Single-view Gaze Estimation. Appearance-based gaze estimation has been studied primarily in the single-view setting, where 3D gaze direction is predicted from eye or face images from one camera. Early datasets and deep learning models such as MPIIGaze Zhang et al. (2015) and iTracker Krafka et al. (2016) established the modern in-the-wild formulation. Later large-scale benchmarks, including Gaze360 Kellnhofer et al. (2019) and ETH-XGaze Zhang et al. (2020), substantially expanded variation in head pose, scene conditions, and gaze direction. Recent work has explored transformer-based architectures Cheng & Lu (2022), gaze-specific generalization objectives Xu et al. (2023), and large-scale pre-training Qin et al. (2026); Zheng et al. (2022) to improve robustness across datasets. Nevertheless, these methods remain limited to monocular inference at test time and cannot leverage complementary observations when multiple views are available.

Geometry Priors for Gaze Estimation. Early eye-tracking systems heavily rely on geometric modeling, where gaze direction is explicitly inferred from 3D eye geometry under calibrated setups Morimoto & Mimica (2005); Hansen & Ji (2009), corneal reflection models Cornsweet & Crane (1973); Guestrin & Eizenman (2006), or personalized eyeball fitting Tsukada & Kanade (2012). These close-range systems highlight the importance of geometric constraints for accurate gaze estimation. With the rise of appearance-based methods, various works have introduced geometric priors into learning-based gaze estimation, including explicit eye-region models Sugano et al. (2014); Wood et al. (2016); Kuang et al. (2022), calibration-aware formulations Liu et al. (2019), and intermediate representations encoding eye geometry Park et al. (2018). However, these methods are primarily designed for single-view input or depend on strong calibration assumptions, which limits their applicability in unconstrained scenarios. In contrast, GazeGGT leverages a multi-view gaze transformer with a geometry-grounded supervision paradigm to enable multi-view gaze estimation.

Learning-based Multi-view Gaze Estimation. Multi-view gaze estimation leverages synchronized observations from multiple cameras to reduce ambiguity caused by self-occlusion, limited eye visibility, and weak geometric constraints. Recent learning-based methods can be broadly categorized into three lines. The first assumes fixed multi-view setups and trains deep models on 2- or 3-view inputs Lian et al. (2018); Cheng & Lu (2023). The second relaxes calibration requirements, such as UVAGaze Liu & Lu (2024), which performs uncalibrated dual-view gaze estimation via unsupervised adaptation. The third incorporates camera geometry for flexible dual-view deployment under calibrated settings Hisadome et al. (2024). However, existing methods are still either restricted to fixed multi-view configurations or require accurate calibration, and they do not generalize to arbitrary numbers of input views within a single unified model.

Geometry-Grounded Foundation Models. Recent progress in 3D vision shows that transformers trained on large-scale image and video data can learn geometry-grounded representations capable of estimating **relative camera poses** and pointmaps from multiple images. DUSt3R Wang et al. (2024) first studies pointmap estimation from image pairs within a unified coordinate system. Subsequent works such as Fast3R Yang et al. (2025) and VGGT Wang et al. (2025) extend this paradigm to longer image sequences and enable joint estimation of camera poses from multiple views. Follow-up work Wang et al. (2026) further improves model architectures and training objectives. This line of work demonstrates that feed-forward models can directly infer relative camera poses from an arbitrary number of views, opening up the possibility of calibration-free multi-view gaze estimation.

3 Method

3.1 Problem Formulation

Given a synchronized set of face crops $\mathcal{I} = \{I_i\}_{i=1}^{N}$ from N cameras observing the same subject at the same time, our goal is to jointly estimate a 3D gaze direction $\hat{\mathbf{g}}_i$ for each view without using camera parameters. The gaze direction $\mathbf{g}_i$ is defined in the camera coordinate system and is parameterized by pitch and yaw. When $N = 1$, the task reduces to monocular gaze estimation. When $N > 1$, the model is expected to leverage complementary observations across views while remaining calibration-free and flexible to the number of available inputs.

In addition, we estimate the extrinsic parameters $\hat{\mathbf{T}}_i$ of each camera to facilitate transforming gaze directions from individual camera coordinate systems to a reference camera with known pose.

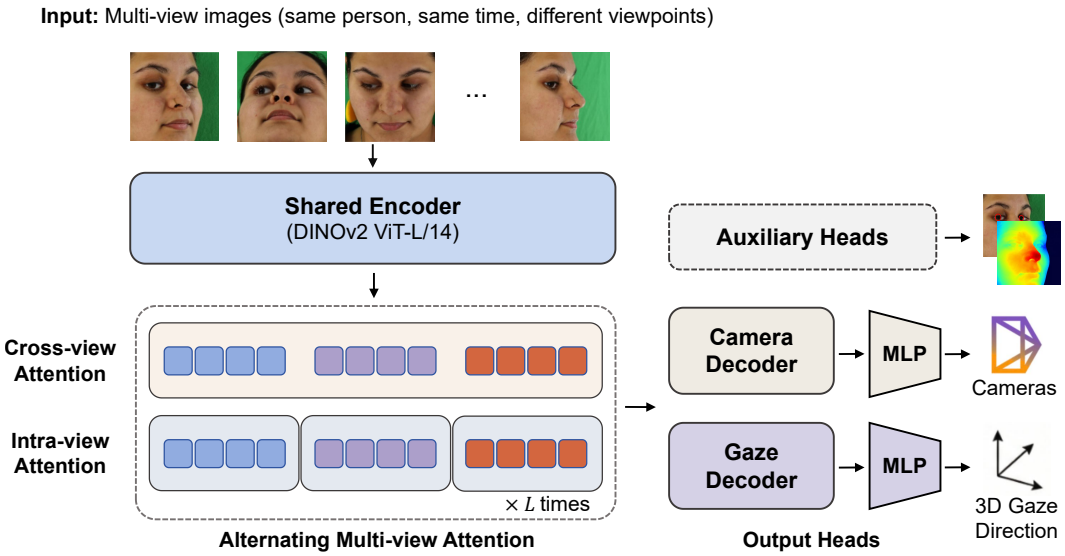

Figure 2: Overview of GazeGGT. Given N synchronized face crops, an image encoder extracts per-view patch tokens, which are processed by alternating multi-view attention layers to enable cross-view interaction. The resulting shared representations are used by separate heads to predict per-view gaze directions, camera extrinsics, and auxiliary supervision signals.

Overall, we aim to train an end-to-end model parameterized by θ:

$$f_\theta : \mathcal{I} \rightarrow \left\{ \{\hat{\mathbf{g}}_i\}_{i=1}^N, \{\hat{\mathbf{T}}_i\}_{i=1}^N \right\}. \tag{1}$$

3.2 Model Architecture: Multi-View Gaze Transformer

Our multi-view gaze transformer is a geometry-grounded transformer designed to process an arbitrary number of views. A shared encoder produces per-view tokens, and a decoder with alternating intra-view and inter-view attention enables information exchange across all views. Task-specific heads then predict gaze directions and camera poses. Since the architecture is independent of the number of input views, it requires no view padding and can naturally handle different numbers of views. Although the model is trained with 1–9 views, the same model weights can be applied to a larger number of views at inference time.

Per-view encoding. A shared ViT independently encodes each face crop into patch tokens. We prepend register tokens but do not introduce a gaze token or view-index embedding. Spatial patch tokens use 2D rotary positional embeddings, while register tokens are assigned a null position.

Alternating multi-view attention. The decoder alternates between intra-view and inter-view attention to capture both view-specific facial information and cross-view geometric relationships. Intra-view blocks reshape the tokens into (BN, T, C) and perform self-attention independently within each view, preserving local facial information. Inter-view blocks reshape them into (B, NT, C) and perform full self-attention across all views, allowing information to flow between arbitrary pairs of views. Here, B denotes the batch size, N the number of views, T the number of tokens per view, and C the feature dimension. Both attention patterns use the same weights, and no cross-view mask or explicit view pairing is introduced. Since the model does not use view-index embeddings, the same architecture and parameters can be applied to different numbers and configurations of views.

Multi-task heads. We use separate task-specific heads for gaze estimation, camera pose estimation, and auxiliary tasks. The gaze and camera heads each average-pool patch tokens from a dedicated decoder: an MLP regresses the 3D gaze direction represented by pitch and yaw, and the camera head predicts the camera poses used after inference to transform gaze predictions into a reference camera coordinate system. The auxiliary heads predict confidence, pointmap, and facial landmarks to provide additional supervision.

3.3 GEOMETRY-GROUNDED SUPERVISION PARADIGM

Per-view gaze labels do not identify the rotation that relates the cameras. A model trained only with $\mathcal{L}_{\text{gaze}}$ can fit each crop and drop the relative geometry. We therefore add geometric terms in the order of the constraint they impose: relative rotation on the camera head, agreement of the gaze vectors in one world frame, and distillation that keeps the geometric representation in place. Landmarks are an auxiliary cue when annotations exist. Every camera quantity below is a training signal.

$$\mathcal{L} = \mathcal{L}_{\text{gaze}} + \lambda_{\text{rot}}\mathcal{L}_{\text{rot}} + \lambda_{\text{wc}}\mathcal{L}_{\text{wc}} + \lambda_{\text{distill}}\mathcal{L}_{\text{distill}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}. \tag{2}$$

Gaze supervision. On each labeled view we compare unit vectors with an angular term, a cosine term, and an ℓ_2 term. The angular term matches the reported metric; the other two smooth the direction.

$$\mathcal{L}_{\text{gaze}} = \lambda_{\text{ang}} \arccos(\langle \hat{\mathbf{g}}_i, \mathbf{g}_i \rangle) + \lambda_{\cos}(1 - \langle \hat{\mathbf{g}}_i, \mathbf{g}_i \rangle) + \lambda_2 \|\hat{\mathbf{g}}_i - \mathbf{g}_i\|_2^2. \tag{3}$$

Relative-rotation supervision. This is the direct geometric constraint. For a pair (i, j) the target is $\mathbf{R}_{ij} = \mathbf{R}_i^\top \mathbf{R}_j$, and $\mathcal{L}_{\text{rot}}$ is the mean geodesic angle between the predicted and ground-truth relative rotations. Translation is not supervised: a gaze direction is invariant to translation, so the transform that aligns views is the rotation. The term is warmed up during training and is absent at inference.

World-frame consistency. Given the same training rotations, each prediction is lifted by $\hat{\mathbf{g}}_i^w = \mathbf{R}_i \hat{\mathbf{g}}_i$, and $\mathcal{L}_{\text{wc}}$ penalizes the mean pairwise angle over valid pairs $\mathcal{P}$. The term does not use gaze labels. It requires the per-view predictions to describe one direction once those rotations are removed. $\mathbf{R}_i$ is not an inference input.

$$\mathcal{L}_{\text{wc}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \arccos\left(\langle \hat{\mathbf{g}}_i^w, \hat{\mathbf{g}}_j^w \rangle\right). \tag{4}$$

Geometry distillation. Gaze labels are a thin constraint on a dense geometric backbone, so we distill a frozen geometry-grounded transformer teacher on the same images into log-depth, confidence, pointmap, metric scale, camera translation, and ℓ_2-normalized patch tokens. Camera rotation is not distilled: $\mathcal{L}_{\text{rot}}$ already supervises it, and a teacher rotation would compete with that signal.

Landmark supervision. When eyeball annotations exist, $\mathcal{L}_{\text{aux}}$ is a Smooth-L_1 loss on eyeball centers, pupil centers, and iris points. Datasets without those labels do not contribute this term.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate our method on several commonly used gaze estimation datasets, including two multi-view datasets, ETH-XGaze Zhang et al. (2020) and GazeGene Bao et al. (2025), as well as three single-view datasets: MPIIFaceGaze Zhang et al. (2017), EyeDiap Funes Mora et al. (2014), and Gaze360 Kellnhofer et al. (2019). **ETH-XGaze**: we use 3DGazeNet's 64/16 subject split of the public 80 subjects Ververas et al. (2024) and the updated gaze labels used by UniGaze Qin et al. (2026). **GazeGene**: we follow the last-10-subject split for testing. **MPIIFaceGaze**: the first 10 subjects are used for training. **EyeDiap**: the first 8 subjects are used for training. **Gaze360**: we follow the original split provided in the paper.

Implementation Details. We initialize the model with pretrained weights from Pi3X, the latest version of π^3 Wang et al. (2026), and unfreeze both the encoder and the camera head, with only the point head kept frozen. We use the AdamW optimizer with a weight decay of 0.05, a gaze learning rate of 10^{-4}, an encoder learning rate of $2{\times}10^{-5}$, and a decoder learning rate of 10^{-5}, following a cosine annealing schedule. All datasets use 224×224 face crops. For multi-view datasets, we randomly sample between 1 and 9 views per sample. Training is conducted on 4 GPUs for 2 epochs. More details are provided in the appendix.

Table 2: Single-view quantitative results of GazeGGT under different training configurations.

Dataset	UniGaze-H Qin et al. (2026)	ETH-XGaze-trained	GazeGene-trained	SV+MV-joint-trained
ETH-XGaze Zhang et al. (2020)	4.16	**3.66**	9.99	3.93
GazeGene Bao et al. (2025)	–	11.38	3.86	**3.72**
MPIIFaceGaze Zhang et al. (2017)	5.12	6.21	5.78	**4.98**
EyeDiap Funes Mora et al. (2014)	5.82	5.67	7.57	**4.90**
Gaze360 Kellnhofer et al. (2019)	10.03	13.46	13.45	**9.78**

Table 3: Multi-view results under different training configurations, reporting per-view averaged errors. ETH and GG denote evaluations on ETH-XGaze and GazeGene, respectively. Since ETH-XGaze contains 18 views and GazeGene contains 9 views, results without * use all available views, while results with * are computed using the first 16 (ETH-XGaze) or 8 (GazeGene) views for consistency. Percentages indicate relative error reduction compared to the corresponding single-view baseline. Grouping is fixed, non-overlapping, and consecutive in dataset order. ETH-XGaze $N \in \{2, 3, 6, 9\}$ partitions all 18 cameras; $N \in \{4, 8\}$ partitions the first 16. GazeGene $N \in \{3, 9\}$ uses all 9 cameras; $N \in \{2, 4, 8\}$ uses the first 8. Test order is not shuffled.

Views N	ETH-XGaze-trained		GazeGene-trained		SV+MV-joint-trained	
	ETH	GG	ETH	GG	ETH	GG
1	3.90	11.38	9.19	3.86	4.03	3.72
1*	3.69	11.04	8.84	3.63	3.85	3.47
2	3.45 (-11.54%)	10.48 (-5.07%)*	8.70 (-5.33%)	3.03 (-16.53%)*	3.59 (-10.92%)	2.87 (-17.29%)*
3	3.07 (-21.28%)	10.71 (-5.89%)	8.35 (-9.14%)	3.01 (-22.02%)	3.20 (-20.60%)	2.84 (-23.66%)
4	2.82 (-23.58%)*	9.67 (-12.41%)*	7.67 (-13.24%)*	2.30 (-36.64%)*	2.89 (-24.94%)*	2.15 (-38.04%)*
6	2.66 (-31.79%)	–	7.46 (-18.82%)	–	2.80 (-30.52%)	–
8	2.52 (-31.71%)*	9.42 (-14.67%)*	7.19 (-18.67%)*	2.08 (-42.70%)*	2.60 (-32.47%)*	1.92 (-44.67%)*
9	2.53 (-35.13%)	9.53 (-16.26%)	7.32 (-20.35%)	2.07 (-46.37%)	2.62 (-34.99%)	1.92 (-48.39%)

Configurations. We evaluate three training configurations, each trained for 2 epochs: (1) training on ETH-XGaze, (2) training on GazeGene, and (3) joint training on multiple datasets, including two multi-view and three single-view datasets as described above. We denote this setting as *SV+MV-joint-trained* in the tables. All results are reported in terms of angular gaze error in degrees.

4.2 Single-view Evaluation

Table 2 reports single-view ($N = 1$) performance under different training configurations. All results are evaluated on unseen subjects in the test sets. Joint training achieves the lowest error on GazeGene, MPIIFaceGaze, EyeDiap, and Gaze360. Across all datasets, GazeGGT consistently achieves slightly better performance than UniGaze-H. Several details are worth noting when interpreting these results. GazeGGT has approximately 0.89B parameters, compared with approximately 0.63B for UniGaze-H. More importantly, our primary focus is multi-view gaze estimation rather than single-view performance. We therefore provide these single-view results as a reference for the multi-view comparisons in the following tables, rather than as a claim of state-of-the-art single-view performance. We use UniGaze-H as a reference because it is also a large-scale model trained jointly across multiple gaze datasets, making it a relevant comparison for our joint-training setting. GazeGene is outside the UniGaze training set, so the corresponding cell is omitted. For ETH-XGaze, we use only the five subjects held out by both our 64/16 split and UniGaze's 60/20 split. For the ETH-XGaze-trained and GazeGene-trained columns, entries evaluated on a different dataset from the training set are **cross-dataset** results, while all joint-training results are in-domain.

4.3 Multi-view Evaluation

Alignment-Free Evaluation. Table 3 reports **per-view averaged gaze errors** without cross-view alignment, where predictions from each view are evaluated in their own coordinate systems and then averaged. We compare multi-view inference against the single-view setting under the same training configurations. Across all configurations, multi-view inference consistently outperforms single-view inference, and the relative gain generally increases as the number of views grows.

Table 4: Multi-view results on ETH-XGaze using predicted camera poses. **Ref-Avg** denotes averaging gaze predictions from all views after transforming them to the reference view using predicted camera poses. **Ref-Sel** denotes selecting the view with the most frontal head pose, followed by transformation to the reference view. For each training configuration, the better result (lower error) between Ref-Avg and Ref-Sel is highlighted in green, while the worse is highlighted in orange.

Views N	ETH-XGaze-trained		GazeGene-trained		SV+MV-joint-trained	
	Ref-Avg	Ref-Sel	Ref-Avg	Ref-Sel	Ref-Avg	Ref-Sel
1	3.90	3.90	9.19	9.19	4.03	4.03
1*	3.69	3.69	8.84	8.84	3.85	3.85
2	3.49 (-10.51%)	3.50 (-10.26%)	8.63 (-6.09%)	8.59 (-6.53%)	3.62 (-10.17%)	3.63 (-9.93%)
3	3.08 (-21.03%)	3.09 (-20.77%)	7.90 (-14.04%)	7.87 (-14.36%)	3.22 (-20.10%)	3.24 (-19.60%)
4*	2.88 (-21.95%)	2.88 (-21.95%)	7.28 (-17.65%)	7.32 (-17.19%)	2.92 (-24.16%)	2.93 (-23.90%)
6	2.64 (-32.31%)	2.67 (-31.54%)	6.34 (-31.01%)	6.50 (-29.27%)	2.80 (-30.52%)	2.81 (-30.27%)
8*	2.55 (-30.89%)	2.56 (-30.62%)	7.11 (-19.57%)	7.12 (-19.46%)	2.64 (-31.43%)	2.65 (-31.17%)
9	2.51 (-35.64%)	2.52 (-35.38%)	7.15 (-22.20%)	7.13 (-22.42%)	2.59 (-35.73%)	2.61 (-35.24%)

Table 5: Two-view comparison on ETH-XGaze (mean angular error, degrees). Consecutive pairs are the nine adjacent pairs and large-baseline pairs are $(i, i{+}9)$. All entries use the subjects held out by both our 64/16 split and the last-10 split used to train Rot-MVGaze and to adapt UVAGaze.

Pairing	Rot-MVGaze	UVAGaze Dual-A	UVAGaze Dual-S	GazeGGT
Consecutive $(i, i{+}1)$	4.10	3.01	**2.94**	3.03
Large-baseline $(i, i{+}9)$	3.84	3.12	**2.51**	2.79
Test-time pose	yes	no	no	no
Models for 9 pairs	1	9	9	1
Extends past $N{=}2$	no	no	no	yes

Notably, in the first two configurations, we evaluate cross-dataset generalization, i.e., training on ETH-XGaze and testing on GazeGene, and vice versa. Even under this cross-domain setting, multi-view inference still provides clear improvements, although the gains are smaller than those observed in the in-domain setting. ETH-XGaze uses a fixed rig and GazeGene uses randomly placed cameras. Transfer between these two rigs is evidence that the same weights generalize to an unseen camera. The relative reductions at $N{=}9$ versus $N{=}1$ are $11.38° \to 9.53°$ (-16.26%) from ETH-XGaze to GazeGene and $9.19° \to 7.32°$ (-20.35%) in the other direction. Table 3 uses fixed consecutive groups, not random subsets. Starred rows use a prefix when N does not divide the rig: the first 16 ETH-XGaze views for $N \in \{4, 8\}$, and the first 8 GazeGene views for $N \in \{2, 4, 8\}$. Unstarred ETH-XGaze rows with $N \in \{2, 3, 6, 9\}$ partition all 18 views. GazeGene $N \in \{3, 9\}$ uses all 9. Test-time view order is not shuffled.

This metric is practically meaningful: although camera parameters are not required as input, when they are available, gaze directions can be transformed into a shared coordinate system and averaged to achieve the maximum multi-view gain reported in Table 3.

Reference-frame Evaluation with Predicted Camera Poses. Table 4 reports results when using predicted camera poses to transform gaze directions into a reference coordinate system. Every column is measured on ETH-XGaze. The ETH-XGaze-trained and joint columns are in-domain, and the GazeGene-trained column is cross-dataset. Appendix Table **??** reports Ref-Avg on both test sets. We assume the first view serves as the reference camera in practical deployment. We evaluate two aggregation strategies: (1) averaging gaze directions from all views after transforming them into the reference coordinate system using predicted poses (denoted as **Ref-Avg**), and (2) selecting the view with the most frontal head pose and transforming only its gaze direction (denoted as **Ref-Sel**).

Ref-Avg and Ref-Sel are both below the matching single-view baseline on every measured row. Ref-Avg is lower for the ETH-XGaze-trained and jointly trained models at every N. For the GazeGene-trained model, Ref-Sel is lower at $N{=}2$, $N{=}3$, and $N{=}9$, while Ref-Avg is lower at $N{=}4$, $N{=}6$, and $N{=}8$. The largest gap between them is $0.16°$ at $N{=}6$.

Table 6: Ablations trained on ETH-XGaze. GGSP denotes the geometry-grounded supervision.

	ETH-XGaze (Within-Domain)			GazeGene (Cross-Domain)		
	N=1	N=9 (Avg)	N=9 (Ref-Avg)	N=1	N=9 (Avg)	N=9 (Ref-Avg)
GazeGGT (full)	3.90	2.53	2.51	11.38	9.53	10.43
DINOv2 init	3.79	2.85	2.80	14.21	14.28	16.33
From scratch	10.47	7.86	6.88	53.94	53.95	49.26
w/o GGSP	3.91	2.60	4.36	11.13	9.52	13.44



Figure 3: Qualitative results on 4-, 6-, and 12-view samples from ETH-XGaze. Row 1–2: 4-view samples; Row 3: 6-view samples; Row 4: 12-view samples. Red marks predictions and yellow marks ground truth. The angular error is printed in the top-left corner. Blue boxes highlight examples where single-view inference suffers from large gaze errors due to extreme head poses and limited eye-region cues, while multi-view joint inference significantly improves the estimation results.

With predicted poses, error still falls as N grows on these rigs. The extrinsics used for this aggregation are predicted, not supplied at test time.

Comparison at N=2. Table 5 is the external comparison missing from the original submission. On ETH-XGaze, consecutive pairs are $(i, i{+}1)$ and large-baseline pairs are $(i, i{+}9)$. The numbers are restricted to subject0104 and subject0106, the subjects held out by both our 64/16 split and the last-10 split of Rot-MVGaze and UVAGaze. GazeGGT uses one model and no test-time pose. It is lower than Rot-MVGaze on both pairings ($3.03°$ vs. $4.10°$ consecutive, $2.79°$ vs. $3.84°$ large-baseline) even though Rot-MVGaze receives relative rotation at test time. UVAGaze Dual-S uses nine pair-specific adapted models and is lower still ($2.94°$ consecutive, $2.51°$ large-baseline). Dual-A is $3.01°$ on consecutive pairs and $3.12°$ on large-baseline pairs. No baseline here extends past N=2, so larger N is supported only by the internal scaling and the cross-rig transfer above.

Relative pose after gaze training. Table 7 is relative rotation at N=2 on the fixed adjacent pairs. The frozen π^3 teacher has no gaze training on either dataset ($4.62°$ on ETH-XGaze, $12.14°$ on GazeGene). ETH-trained GazeGGT is lower on both ($0.99°$ and $4.52°$). Joint training is $1.04°$ and

Table 7: Relative camera rotation error at $N{=}2$, in degrees.

Model	ETH-XGaze	GazeGene
π^3 teacher	4.62	12.14
w/o GGSP	9.00	11.98
ETH-trained	**0.99**	4.52
GazeGene-trained	2.27	1.44
Joint	1.04	**1.38**

Table 8: Inference latency of GazeGGT under different numbers of input views.

	1	2	4	8	9
A100 latency (ms)	53.1	53.5	53.8	63.9	69.2
RTX 5090 latency (ms)	59.0	60.3	61.1	82.3	83.2
Peak memory (GB)	6.33	6.34	6.37	6.53	6.59

$1.38°$. Keeping the π^3 initialization and dropping the geometry-grounded supervision raises the ETH-XGaze error to $9.00°$. This is relative pose between views, not absolute camera pose from a single face.

Qualitative Analysis. Figure 3 presents qualitative results across different numbers of views. These examples provide additional insights into model behavior. In cases with large head poses and severe self-occlusion, single-view predictions are often ambiguous, while multi-view inference effectively resolves this ambiguity by leveraging complementary observations across views, leading to more accurate gaze estimates.

4.4 Ablation Studies

Table 6 ablates initialization and the geometry-grounded supervision. All rows train on ETH-XGaze. GazeGGT (full) is the π^3 initialization with the full loss. From scratch falls from $10.47°$ at $N{=}1$ to $7.86°$ at $N{=}9$ in domain and stays near $53.9°$ on GazeGene. DINOv2 init is slightly lower than GazeGGT (full) at ETH-XGaze $N{=}1$ ($3.79°$ vs. $3.90°$), while GazeGGT (full) is lower at $N{=}9$ ($2.53°$ vs. $2.85°$). On GazeGene, GazeGGT (full) is $11.38°$ at $N{=}1$ and $9.53°$ at $N{=}9$; DINOv2 init is $14.21°$ and $14.28°$. w/o GGSP keeps the π^3 initialization and matches GazeGGT (full) on per-view error ($3.91°/2.60°$ on ETH-XGaze, $11.13°/9.52°$ on GazeGene). The two separate once views are fused with the predicted pose: at $N{=}9$, Ref-Avg is $4.36°$ on ETH-XGaze and $13.44°$ on GazeGene, against $2.51°$ and $10.43°$ for GazeGGT (full).

4.5 Runtime Analysis

Table 8 reports inference latency and peak memory under different numbers of input views. On an A100, latency moves from 53.1 ms at $N{=}1$ to 69.2 ms at $N{=}9$. On an RTX 5090 the same sweep is 59.0 ms to 83.2 ms. Peak memory stays between 6.33 GB and 6.59 GB.

5 Conclusion

Limitations and Future Work. While GazeGGT demonstrates multi-view gaze estimation without camera parameters at inference, several limitations remain. First, the current model has a relatively large parameter size, and further work could focus on improving efficiency and real-time performance. Second, more accurate estimation of relative camera poses could further improve multi-view aggregation when predicted poses are used to change reference frames. Third, inputs must be synchronized crops of the same person. Appendix experiments show that a minority of mismatched or temporally mixed views is absorbed by the majority, but those views themselves fail. We therefore recommend an association check. The method is aimed at controlled rigs, vehicle cabins, and telepresence, not general in-the-wild deployment. Appearance noise on a single view changes error by at most about $0.8°$ in our proxy tests. Dropped views behave like inference at a smaller N.

GazeGGT is a geometry-grounded transformer for calibration-free multi-view gaze estimation. Alternating attention fuses an arbitrary number of synchronized face crops, and separate decoders predict per-view gaze and camera pose without camera parameters at inference. The geometry-grounded supervision paradigm supplies the cross-view geometry that gaze labels alone do not retain. GazeGGT is competitive in the single-view and two-view settings and improves as more views are added.

AI USE STATEMENT

We used generative AI tools to edit the manuscript, format references, and assist with experimental scripts. We did not use them to generate data, design the method or experiments, create figures, or produce the reported measurements. We reviewed the assisted text and code against our own results, and we take responsibility for the final content.

REFERENCES

Reuben M Aronson, Thiago Santini, Thomas C Kübler, Enkelejda Kasneci, Siddhartha Srinivasa, and Henny Admoni. Eye-hand behavior in human-robot shared manipulation. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, pp. 4–13, 2018.

Yiwei Bao, Zhiming Wang, and Feng Lu. Gazegene: Large-scale synthetic gaze dataset with 3d eyeball annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18749–18759, 2025.

Zhuojiang Cai, Jingkai Hong, Zhimin Wang, and Feng Lu. Gazeswipe: Enhancing mobile touchscreen reachability through seamless gaze and finger-swipe integration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2025.

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

Yihua Cheng and Feng Lu. Dvgaze: Dual-view gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20632–20641, 2023.

Tom N Cornsweet and Hewitt D Crane. Accurate two-dimensional eye tracker using first and fourth purkinje images. *Journal of the Optical Society of America*, 63(8):921–928, 1973.

Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*, pp. 255–258, 2014.

Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6):1124–1133, 2006.

Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2009.

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.

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.

Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2176–2184, 2016.

Chenyi Kuang, Jeffery O Kephart, and Qiang Ji. Towards an accurate 3d deformable eye model for gaze estimation. In *International conference on pattern recognition*, pp. 109–123. Springer, 2022.

Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A Lee, and Mark Billinghurst. Pinpointing: Precise head-and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–14, 2018.

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

Gang Liu, Yu Yu, Kenneth A Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1092–1099, 2019.

Ruicong Liu and Feng Lu. Uvagaze: Unsupervised 1-to-2 views adaptation for gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3693–3701, 2024.

Päivi Majaranta and Andreas Bulling. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*, pp. 39–65. Springer, 2014.

José M Morales, Carolina Díaz-Piedra, Héctor Rieiro, Joaquín Roca-González, Samuel Romero, Andrés Catena, Luis J Fuentes, and Leandro L Di Stasi. Monitoring driver fatigue using a single-channel electroencephalographic device: A validation study by gaze-based, driving performance, and subjective data. *Accident Analysis & Prevention*, 109:62–69, 2017.

Carlos H Morimoto and Marcio RM Mimica. Eye gaze tracking techniques for interactive applications. *Computer vision and image understanding*, 98(1):4–24, 2005.

Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 721–738, 2018.

Jiawei Qin, Xucong Zhang, and Yusuke Sugano. Unigaze: Towards universal gaze estimation via large-scale pre-training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5809–5820, 2026.

Pavan Kumar Sharma and Pranamesh Chakraborty. A review of driver gaze estimation and application in gaze behavior understanding. *Engineering Applications of Artificial Intelligence*, 133: 108117, 2024.

Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1821–1828, 2014.

Yunus Terzioğlu, Bilge Mutlu, and Erol Şahin. Designing social cues for collaborative robots: the role of gaze and breathing in human-robot collaboration. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pp. 343–357, 2020.

Akihiro Tsukada and Takeo Kanade. Automatic acquisition of a 3d eye model for a wearable first-person vision device. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 213–216, 2012.

Evangelos Ververas, Polydefkis Gkagkos, Jiankang Deng, Michail Christos Doukas, Jia Guo, and Stefanos Zafeiriou. 3dgazenet: Generalizing 3d gaze estimation with weak-supervision from synthetic views. In *European Conference on Computer Vision*, pp. 387–404, 2024.

Haofei Wang, Xujiong Dong, Zhaokang Chen, and Bertram E Shi. Hybrid gaze/eeg brain computer interface for robot arm control on a pick and place task. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1476–1479. IEEE, 2015.

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, 2025.

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.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, 2026.

Zhimin Wang, Huangyue Yu, Haofei Wang, Zongji Wang, and Feng Lu. Comparing single-modal and multimodal interaction in an augmented reality system. In *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 165–166. IEEE, 2020.

Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. A 3d morphable eye region model for gaze estimation. In *European conference on computer vision*, pp. 297–313. Springer, 2016.

Mingjie Xu, Haofei Wang, and Feng Lu. Learning a generalized gaze estimator from gaze-consistent feature. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3027–3035, 2023.

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.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4511–4520, 2015.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 51–60, 2017.

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, 2020.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18697–18709, 2022.