Enhancing Gaze Reasoning in Vision Foundation Models for Gaze Following

Shijing Wang[1] **Yaping Huang**[1]* **Chaoqun Cui**[4] **David Wong**[5] **Yihua Cheng**[2, 3]*
Alexandros Neophytou[5] **Hyung Jin Chang**[2, 5]

[1]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence,
Beijing Jiaotong University
[2]University of Birmingham [3]Beijing Institute of Technology
[4]Institute of Automation, Chinese Academy of Sciences [5]Microsoft, UK

Abstract

Gaze following requires both scene understanding and gaze reasoning to localize the gaze target of an in-scene person. Recently, vision foundation models (VFMs) have demonstrated strong performance on this task, enabling simpler architectures while outperforming prior methods. However, we observe a key limitation of VFM-based approaches: while VFMs substantially improve scene understanding, they contribute little to gaze reasoning. As a result, existing methods often rely on semantically salient objects rather than true gaze cues, leading to degraded performance when targets are not salient. To address this, we propose a novel training mechanism to enhance gaze reasoning in VFMs for gaze following. Our method includes: (1) a head-conditioned local LoRA, which enables localized adaptation to preserve scene token learning while improving head token learning for gaze reasoning; and (2) an out-of-cone penalty, which injects gaze cues into head tokens while aligning them with scene tokens. Experiments on the GazeFollow and VAT datasets demonstrate that our method achieves state-of-the-art performance, with particularly strong improvements when gaze targets are not semantically salient. Our findings offer valuable insights for advancing future gaze following research. Code will be available at `https://github.com/shijing-wang/HCloRA`.

1 Introduction

Human gaze behavior is a fundamental component of non-verbal communication [5, 26]. The ability to follow a person's gaze is essential for understanding human behavior, and has broad applications in human-robot interaction [1, 6], assistive systems [20, 31], and behavior understanding [19, 13].

Gaze following is challenging because it requires both scene understanding and gaze reasoning. The model must understand the surrounding scene to capture contextual information, while also using gaze cues to reason where the target person is looking. Vision foundation models (VFMs) are well suited to the scene-understanding aspect of this task, as their pre-trained representations provide strong semantic and contextual priors for interpreting complex visual environments [23, 2, 30]. These properties make VFMs attractive backbones for gaze following, contributing to the remarkable progress and new state-of-the-art performance achieved in this task.

However, in this paper, we observe that the strong performance of VFM-based gaze following does not necessarily imply stronger gaze reasoning. VFMs provide rich scene-level representations and are highly effective at capturing semantic and contextual regularities in visual scenes. As shown in Fig. 1 (left), existing methods can often make plausible predictions by exploiting scene priors, such

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

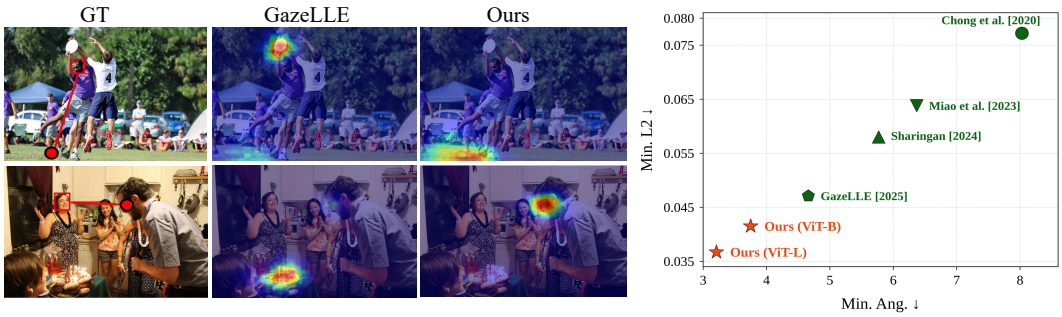


Figure 1: **Left**: Given an input image and the query head marked by the red box, existing VFM-based methods such as GazeLLE often predict gaze targets around semantically salient objects or interaction centers, even when they are not the true gaze target. In contrast, our method better localizes the gaze target by relying on true gaze cues rather than semantic saliency. **Right**: Comparison on the GazeFollow benchmark, where lower Min. L2 and Min. Ang. indicate better performance. Our method achieves state-of-the-art results, demonstrating the effectiveness of enhancing gaze reasoning in VFM-based gaze following.

as focusing on visually prominent objects, interaction centers, or semantically meaningful regions. While these priors are helpful in many common cases, they become unreliable when the true gaze target is not the most salient region in the scene. Consequently, current methods tend to rely on semantically salient objects rather than true gaze cues, assigning high confidence to visually plausible but incorrect regions and thus degrading localization performance.

To address this issue, we propose a novel training mechanism for VFM-based gaze following that enhances gaze reasoning without redesigning the main backbone-decoder pipeline. Our method consists of two complementary components. First, we introduce a head-conditioned local LoRA, which uses the queried head position to modulate low-rank residual updates inside the VFM. This enables localized adaptation of head-relevant features while preserving the backbone's scene-level representations. Second, we propose an out-of-cone penalty, which regularizes auxiliary gaze-evidence maps from adapted layers by penalizing probability mass outside the geometrically plausible gaze cone. Instead of imposing a rigid gaze-cone prior, this penalty flexibly injects gaze cues into head features. Together, these designs encourage the model to rely on true gaze cues rather than insufficient semantic saliency.

Experiments on the GazeFollow and VAT datasets demonstrate that our method achieves state-of-the-art performance, with representative results shown in Fig. 1 (right). The gains are especially significant when gaze targets are not semantically salient, showing that our method improves gaze following by strengthening gaze reasoning rather than relying on scene understanding alone. These findings highlight the importance of explicit gaze reasoning for robust VFM-based gaze following.

In summary, our main contributions are as follows:

- We identify a key limitation of current VFM-based gaze following methods: although VFMs substantially improve scene understanding, they contribute little to gaze reasoning, causing models to mainly rely on semantically salient objects rather than true gaze cues.
- We propose a novel training mechanism for VFM-based gaze following without redesigning the main backbone-decoder pipeline, including a head-conditioned local LoRA and an out-of-cone penalty to enhance gaze reasoning.
- We achieve state-of-the-art performance on the GazeFollow and VAT datasets, with strong gains in more challenging scenarios where gaze targets are not semantically salient.

2 Related Work

2.1 Gaze Following

Gaze following aims to localize the gaze point of the target person in a scene. Recasens et al. [24] first formulated this task by combining head information with scene context to predict a gaze heatmap.



Figure 2: **Left:** Failure cases of GazeLLE [25], where the predicted gaze target is attracted to semantically salient regions instead of the ground-truth gaze target. **Right:** Average L2 error on the GazeFollow consistent and inconsistent subsets across different methods. All methods show larger errors on the inconsistent subset, suggesting that current gaze-following models are still affected by semantic saliency and do not always perform robust gaze reasoning.

Following this paradigm, later methods improve gaze prediction by incorporating additional cues [12], including depth [10, 11], body pose [11], 3D head orientation [14], and temporal context [9]. More recently, transformer-based models have been introduced to better capture interactions between the queried person and the surrounding scene [28, 25]. These methods improve contextual modeling and highlight the importance of jointly modeling head information and scene context.

2.2 Vision Foundation Models for Gaze Following

Vision foundation models, such as DINOv2 [23] and MultiMAE [2], learn transferable visual representations from large-scale pretraining. Their strong semantic and contextual understanding makes them effective backbones for gaze following. Recent gaze-following methods have started to build upon such pre-trained visual representations. Sharingan [28] represents the image as scene tokens and fuses them with person-specific gaze tokens through transformer attention. GazeLLE [25] further simplifies the architecture by using a frozen VFM backbone for scene feature extraction and injecting the queried head information into a lightweight decoder, achieving stronger performance.

3 Preliminary

VFMs have brought performance breakthroughs to various vision tasks, including gaze following. In this work, we aim to further improve the performance of VFM-based gaze following models.

3.1 Observation: Semantic Shortcut on Gaze Following

We start by analyzing the failure cases of recent models. Specifically, we select GazeLLE [25], a state-of-the-art VFM-based method, as our baseline. We sort test samples by prediction error, and visualize representative failure samples. As shown in Fig. 2 (left), large-error predictions usually share a common pattern: *the prediction concentrates on semantically salient regions, such as manipulated objects or interaction centers, even when these regions lie in the opposite direction of the human gaze.* Please see Appendix A.1 for more failure cases.

This observation indicates that such models do not always capture reliable gaze cues. Instead, they tend to infer gaze targets from scene-level semantic context, which can lead to an overestimation of the model's gaze reasoning ability when the gaze target coincides with a semantically salient object. We refer to this behavior as a *semantic shortcut*, where the model relies on semantic plausibility as a proxy for gaze following rather than performing gaze reasoning. However, we argue that a robust gaze following method should possess strong and reliable gaze reasoning capability.

3.2 Quantitative Analysis: Performance Gap from Semantic Shortcut

To further validate our hypothesis, we conduct a quantitative analysis based on the results of GazeLLE. Specifically, we aim to split the test dataset into two subsets: ***consistent***, where the gaze target is consistent with a semantically salient object, and ***inconsistent***, where it is not. If we observe a performance gap between these two subsets, it supports our hypothesis.

However, it is non-trivial to precisely define and identify semantically salient objects. Therefore, we adopt a *heuristic strategy* for this analysis. Specifically, we use Qwen3-VL-32B-Instruct [3] as a semantic prober to identify contextually plausible gaze regions for one person. Importantly, the prober is not used to predict the gaze target directly. Instead, it is prompted to localize regions that are semantically likely to attract the target person's attention based on scene context. Each predicted region is represented by a bounding box. We then measure the overlap between the ground-truth gaze points and the predicted boxes. If more than half of the gaze points fall within the boxes, the sample is assigned to the *consistent* subset; otherwise, it is assigned to the *inconsistent* subset. For single annotation, the threshold is set to 1. Please see Appendices A.2 and A.3 for more details.

The results are shown in Fig. 2 (right). GazeLLE exhibits a significant performance gap between the two subsets, which validates our hypothesis. We further evaluate other gaze following methods on these subsets and observe a similar trend: all methods show a noticeable performance gap. This finding suggests that the semantic shortcut is a common issue across gaze following approaches, *highlighting the importance of enhancing gaze reasoning in current methods to alleviate this problem.*

4 Methodology

We propose a novel training mechanism to enhance gaze reasoning in VFMs without redesigning the main backbone-decoder pipeline. Our contributions include a head-conditioned local LoRA for head feature extraction and an out-of-cone penalty for injecting gaze cues into head features.

4.1 Head-Conditioned Local LoRA

Previous gaze following methods typically freeze the VFM for image tokenization [25], while using a learnable decoder for adaptation. This raises an intuitive question: *Can simply fine-tuning the VFM on gaze following datasets improve gaze reasoning?*

Naive fine-tuning remains insufficient. We conduct experiments to evaluate the effectiveness of fine-tuning. As shown in Fig. 3, we evaluate full fine-tuning (FT), LoRA [15], and visual prompt tuning (VPT) [17]. These methods fail to deliver consistent performance improvements and even degrade performance in some cases.

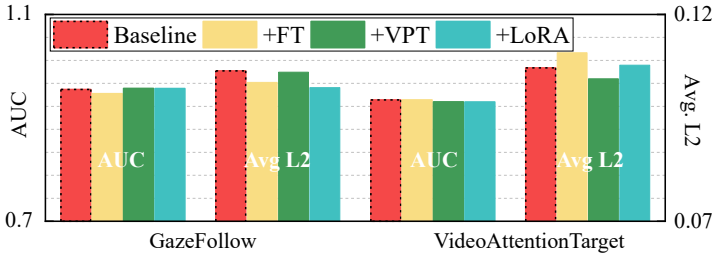


Figure 3: We fine-tune GazeLLE on two datasets using several SoTA fine-tuning methods. Naive fine-tuning fails to consistently improve performance.

We attribute this to the following insight: VFMs have strong scene understanding capabilities through large-scale pretraining on diverse image data. Fine-tuning on relatively small datasets may degrade this capability. At the same time, VFMs show limited gaze reasoning ability, meaning they struggle to precisely capture human facial and gaze cues. It suggests that an effective solution should enhance head token learning in VFMs while preserving their scene token learning capabilities.

HCLoRA: In this work, we propose Head-Conditioned Local LoRA (HCLoRA), a method that adapts the model in a spatially localized manner by focusing the trainable update on the target head region while preserving scene-level representations. To this end, we build on LoRA, which keeps the pre-trained projection frozen and introduces a lightweight trainable residual branch. Different from standard LoRA, HCLoRA conditions this residual branch on the target head position, so that the adaptation is emphasized on head tokens and has limited influence on the global scene representation.

In detail, let $\mathbf{I}$ denote the input image and let $b = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ denote the bounding box of the target head. Given an input token sequence $\mathbf{X} \in \mathbb{R}^{N \times d}$ and a frozen pre-trained projection $W(\cdot)$, the adapted projection is formulated as

$$\mathbf{Y} = W(\mathbf{X}) + \Delta(\mathbf{X}, b), \tag{1}$$

where the head-conditioned residual update is defined as

$$\Delta(\mathbf{X}, b) = \big(A(\mathbf{X}) \odot \mathbf{M}(b)\big)\mathbf{B}. \tag{2}$$

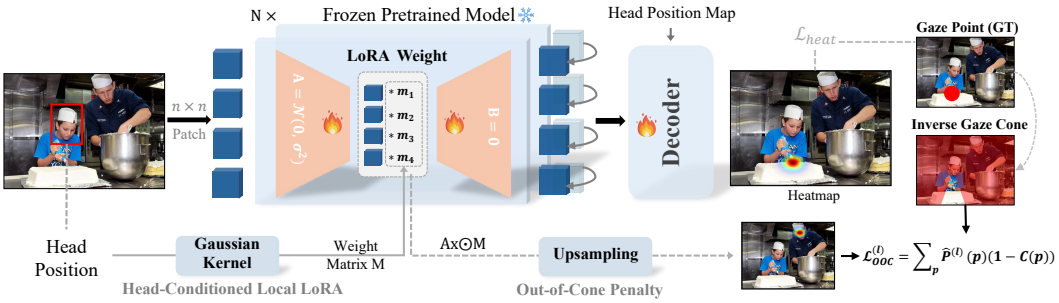


Figure 4: Overview of our proposed framework. Given an input image and the target person's head box, we construct a head-position prior to modulate the low-rank residual branch inside a frozen pretrained vision backbone. The proposed head-conditioned local LoRA selectively adapts head-relevant features while preserving scene-level representations. The adapted features are decoded into the final gaze heatmap, with an additional lightweight in/out branch omitted for clarity. During training, auxiliary gaze-evidence maps from intermediate adapted layers are regularized by the proposed out-of-cone penalty, which suppresses prediction mass outside the geometrically plausible gaze cone.

Here, $A(\mathbf{X}) = \mathbf{XA} \in \mathbb{R}^{N\times r}$ denotes the low-rank feature produced by the LoRA down-projection, where $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times d}$ are trainable low-rank matrices with $r \ll d$. The matrix $\mathbf{M}(b) \in \mathbb{R}^{N\times r}$ is generated from the target head position and controls where the residual update is applied.

Specifically, we first construct a soft spatial prior from the target head bounding box. Let (c_x, c_y) denote the box center and (s_x, s_y) denote its scale on the feature map. For each spatial location (u, v), the head-guided map is computed as

$$H(u,v) = \exp\left(-\frac{(u-c_x)^2}{2s_x^2+\epsilon} - \frac{(v-c_y)^2}{2s_y^2+\epsilon}\right). \tag{3}$$

The head-guided map is then passed through a lightweight gating network to produce $\mathbf{M}(b)$. This modulation matrix is applied to the low-rank feature $A(\mathbf{X})$ before the LoRA up-projection $\mathbf{B}$. In this way, HCLoRA localizes the trainable adaptation to head-related regions.

4.2 Out-of-Cone Penalty

HCLoRA enhances head token extraction, but lacks explicit constraints on what information the head tokens should encode. Encoding gaze cues into head tokens can improve gaze reasoning ability.

Misalignment between Head and Scene Tokens. A straightforward solution is to supervise head token learning by regressing gaze direction. However, directly regressing gaze direction does not consistently improve gaze-following performance (see Table 4 for details). We argue that this is due to the misalignment between head tokens and scene tokens. Gaze following methods typically estimate a probability heatmap to represent the gaze target, *i.e.*, each scene token encodes the likelihood of being the target location. In contrast, head tokens are supervised using gaze direction, requiring the decoder to implicitly learn a mapping from a direction vector to a spatial probability distribution. Such a transformation is non-trivial and difficult to learn.

Rigid Representation. An alternative is to use a gaze cone for supervision by generating a heatmap from head tokens, where conventional methods convert gaze direction into a cone-shaped attention prior. However, it imposes overly rigid cone-shaped supervision on representations of head tokens.

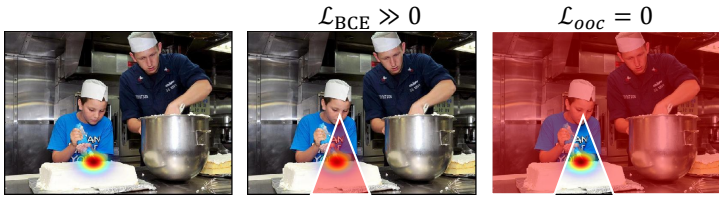


Figure 5: Comparison of gaze cone supervision and the out-of-cone penalty. Gaze cone supervision can produce a large loss even when the prediction captures the ground-truth target.

OOC Penalty: In this work, we propose an out-of-cone (OOC) penalty. The OOC penalty has two key strengths: (1) it leverages head tokens to predict a heatmap aligned with the gaze-following output; and (2) instead of enforcing rigid gaze-cone supervision, it penalizes probabilities located outside the gaze cone.

This formulation injects gaze cues into head features by discouraging predictions outside the gaze cone rather than enforcing an explicit target distribution. Figure 5 compares gaze-cone supervision with our OOC penalty. Gaze-cone supervision can produce a large loss even when the head token successfully captures the GT, whereas our OOC penalty can yield a near-zero penalty in such cases.

Specifically, we attach lightweight auxiliary heads to the last L adapted layers, where L denotes the number of adapted layers with auxiliary supervision. For each selected layer l, given hidden low-rank features $\mathbf{Z}^{(l)} \in \mathbb{R}^{N_p \times r}$, we obtain a normalized auxiliary gaze-evidence map:

$$\hat{\mathbf{P}}^{(l)} = \mathrm{Softmax}_{\mathrm{spatial}}\left(f_{\mathrm{aux}}^{(l)}(\mathbf{Z}^{(l)})\right), \tag{4}$$

where $f_{\mathrm{aux}}^{(l)}$ maps the low-rank token features to a spatial gaze-evidence map aligned with the supervision resolution.

To penalize auxiliary evidence outside the gaze-consistent region, we construct a soft cone mask centered on the ground-truth gaze direction. Let $\mathbf{h}$ and $\mathbf{g}$ denote the head center and the ground-truth gaze target, respectively, and let $\mathbf{d} = (\mathbf{g} - \mathbf{h})/(\|\mathbf{g} - \mathbf{h}\|_2 + \epsilon)$ denote the normalized gaze direction. For each spatial location (u, v) with coordinate $\mathbf{p} = \mathbf{x}_{u,v}$, we compute its signed projection on the gaze ray and perpendicular distance to the ray:

$$t(\mathbf{p}) = (\mathbf{p} - \mathbf{h})^\top \mathbf{d}, \quad d_\perp(\mathbf{p}) = \|(\mathbf{p} - \mathbf{h}) - t(\mathbf{p})\mathbf{d}\|_2 . \tag{5}$$

Given a cone angle θ, the soft cone mask is defined as

$$C(\mathbf{p}) = \sigma\left(\alpha\left(\max(t(\mathbf{p}), 0)\tan\frac{\theta}{2} - d_\perp(\mathbf{p})\right)\right) \cdot \sigma\left(\alpha t(\mathbf{p})\right), \tag{6}$$

where α controls the boundary sharpness. The first term defines the cone boundary, while the second suppresses locations opposite to the gaze direction.

Finally, we penalize the proportion of auxiliary evidence outside the gaze cone. Since $\sum_{\mathbf{p}} \hat{\mathbf{P}}^{(l)}(\mathbf{p}) = 1$ by Eq. (4), the penalty simplifies to:

$$\mathcal{L}_{\mathrm{OOC}}^{(l)} = \sum_{\mathbf{p}} \hat{\mathbf{P}}^{(l)}(\mathbf{p})\big(1 - C(\mathbf{p})\big). \tag{7}$$

4.3 Implementation details

Our method is built upon the GazeLLE framework [25]. We adopt a DINOv2 backbone with a lightweight gaze decoder, whose output is a gaze heatmap. The final prediction is supervised using a binary cross-entropy (BCE) loss $\mathcal{L}_{\mathrm{heat}}$. When required, we additionally use a binary head for in/out prediction, supervised by the BCE loss $\mathcal{L}_{\mathrm{in/out}}$. The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{heat}} + \beta\mathcal{L}_{\mathrm{in/out}} + \lambda\frac{1}{L}\sum_{l=1}^{L}\mathcal{L}_{\mathrm{OOC}}^{(l)}, \tag{8}$$

where β and λ are loss weights.

5 Experiments

5.1 Experimental Setup

We evaluate our method on GazeFollow [24] and VideoAttentionTarget (VAT) [9]. To clearly demonstrate the performance in detail, we use the heuristic strategy described in Sec. 3.2 to split the evaluation set into *consistent* and *inconsistent* subsets for both datasets. Note that this partition does not modify the evaluation set itself, but only reports results under different subsets.

Table 1: Comparison with state-of-the-art gaze-following methods on GazeFollow. Our method shows best performance, with particularly large improvements on the *inconsistent* subset.

Method	Consistent					Inconsistent				
	AUC↑	Avg. L2↓	Min. L2↓	Avg. Ang.↓	Min. Ang.↓	AUC↑	Avg. L2↓	Min. L2↓	Avg. Ang.↓	Min. Ang.↓
Chong et al. [9]	0.9402	0.1194	0.0701	15.68	7.41	0.8892	0.1731	0.0914	23.18	9.25
Miao et al. [22]	0.9503	0.1044	0.0559	13.73	5.63	0.9026	0.1601	0.0795	21.82	7.83
Sharingan [28]	0.9586	0.0958	0.0496	10.83	4.96	0.9148	0.1536	0.0745	17.39	7.37
GazeLLE [25]	0.9689	0.0834	0.0373	11.04	3.37	0.9270	0.1519	0.0664	21.09	7.22
GazeLLE + FT	0.9613	0.0835	0.0403	11.46	3.86	0.9183	0.1434	0.0655	20.67	7.02
GazeLLE + VPT	0.9700	0.0825	0.0360	11.06	3.33	0.9315	0.1527	0.0666	21.10	7.22
GazeLLE + LoRA	0.9700	0.0789	0.0352	10.76	3.22	0.9308	0.1487	0.0667	20.55	6.86
Ours (ViT-B)	0.9696	0.0777	0.0348	10.50	3.19	0.9340	0.1353	0.0547	18.00	4.85
Ours (ViT-L)	**0.9719**	**0.0709**	**0.0301**	**9.66**	**2.58**	**0.9371**	**0.1271**	**0.0498**	**17.36**	**4.44**

We use AUC and L2 distance to evaluate gaze-target localization [24, 9, 25]. For VAT, we additionally report $\mathrm{AP}_{\mathrm{in/out}}$ for in/out prediction. In this paper, we introduce angular error as an additional metric for measuring gaze reasoning. Specifically, given the head center $\mathbf{h}$, the predicted gaze point $\hat{\mathbf{g}}$, and a ground-truth gaze point $\mathbf{g}$, the angular error is computed as

$$\mathrm{Ang.} = \arccos\left(\frac{(\hat{\mathbf{g}}-\mathbf{h})^\top(\mathbf{g}-\mathbf{h})}{\|\hat{\mathbf{g}}-\mathbf{h}\|_2\|\mathbf{g}-\mathbf{h}\|_2}\right). \tag{9}$$

This metric measures whether the prediction follows the correct gaze direction from the queried head.

5.2 Comparison with State-of-the-Art Gaze-Following Methods

Tables 1 and 2 compare our method with state-of-the-art gaze-following methods on GazeFollow and VAT. On GazeFollow, our method consistently improves over GazeLLE under the same ViT-B backbone, especially on the *inconsistent* subset where semantic saliency is unreliable. Compared with GazeLLE, our method reduces Avg. L2 by 10.9%, Min. L2 by 17.6%, Avg. Ang. by 14.7%, and Min. Ang. by 32.8% on this subset. It also improves localization accuracy on the *consistent* subset, indicating that stronger gaze reasoning is achieved without sacrificing scene understanding.

The same trend holds on VideoAttentionTarget. On the *inconsistent* subset, our method reduces L2 error from 0.1670 to 0.1513 and angular error from 24.21 to 18.34. Compared with full fine-tuning, VPT, and standard LoRA, our method provides more stable gains across datasets and subsets, showing the effectiveness of head-conditioned local adaptation. Using a stronger ViT-L backbone further improves most localization metrics. For completeness, an overall comparison on the full test sets with trainable parameter counts is provided in Appendix B.2 and Table 5.

5.3 Ablation Study

We conduct component ablation on GazeFollow to examine how the proposed designs contribute to gaze reasoning. As shown in Table 3, adding HCLoRA leads to substantial gains on the *inconsistent* subset, reducing Avg. L2 from 0.1519 to 0.1383 and Min. Ang. from 7.22 to 5.72. This confirms that localizing the adaptation around head-relevant regions is crucial for improving gaze reasoning while preserving the original VFM representations. With the out-of-cone penalty, the full model further improves all metrics on the *inconsistent* subset, achieving 0.1353 Avg. L2 and 4.85 Min. Ang., while also maintaining strong performance on the *consistent* subset.

Additional ablations on hyperparameter sensitivity and architectural variants are provided in Appendix Tables 6, 7, and 8. These results further support that our method robustly improves gaze reasoning through localized adaptation.

5.4 Comparison with Alternative Gaze-Aware Supervision Strategies

We study different forms of gaze-aware supervision in Table 4. Direct gaze-vector regression, heatmap prediction, and rigid gaze-cone supervision bring limited gains. For example, although gaze-cone

Table 2: Comparison with state-of-the-art gaze-following methods on VideoAttentionTarget.

Method	All			Consistent			Inconsistent			$AP_{in/out}$ ↑
	AUC ↑	L2 ↓	Ang. ↓	AUC ↑	L2 ↓	Ang. ↓	AUC ↑	L2 ↓	Ang. ↓	
Chong et al. [9]	0.8628	0.1339	16.14	0.8808	0.1035	13.65	0.8343	0.1821	20.09	0.8510
Miao et al. [22]	0.9163	0.1099	12.35	0.9295	0.0796	9.77	0.8954	0.1580	16.45	0.9061
Sharingan [28]	0.9162	0.1038	12.09	0.9320	0.0741	9.41	0.8912	0.1509	**16.34**	0.8921
GazeLLE [25]	0.9347	0.1071	15.02	0.9528	0.0693	9.21	0.9059	0.1670	24.21	0.8979
GazeLLE + FT	0.9347	0.1107	15.32	0.9538	0.0746	9.60	0.9044	0.1680	24.38	0.8858
GazeLLE + VPT	0.9309	0.1044	13.66	0.9455	0.0709	9.31	**0.9078**	0.1576	20.56	0.8911
GazeLLE + LoRA	0.9307	0.1077	14.24	0.9477	0.0724	9.35	0.9037	0.1637	22.00	0.8775
Ours (ViT-B)	0.9352	0.1006	12.26	0.9534	0.0686	8.42	0.9064	0.1513	18.34	0.8988
Ours (ViT-L)	**0.9387**	**0.0951**	**11.75**	**0.9590**	**0.0630**	**7.29**	0.9066	**0.1459**	18.83	**0.9068**

Table 3: Ablation study on GazeFollow. Both HCLoRA and the OOC penalty improve performance, and reduced angular error indicates enhanced gaze reasoning.

Components		Consistent					Inconsistent				
HCLoRA	OOC penalty	AUC ↑	A-L2 ↓	M-L2 ↓	A-Ang ↓	M-Ang ↓	AUC ↑	A-L2 ↓	M-L2 ↓	A-Ang ↓	M-Ang ↓
×	×	0.9689	0.0834	0.0373	11.04	3.37	0.9270	0.1519	0.0664	21.09	7.22
✓	×	0.9677	0.0786	0.0366	10.66	3.37	0.9303	0.1383	0.0567	19.07	5.72
✓	✓	**0.9696**	**0.0777**	**0.0348**	**10.50**	**3.19**	**0.9340**	**0.1353**	**0.0547**	**18.00**	**4.85**

supervision improves several metrics on the *consistent* subset, it does not reduce the Avg. L2 error on the *inconsistent* subset compared with no auxiliary supervision. In contrast, the proposed out-of-cone penalty achieves the best performance on the *inconsistent* subset, improving AUC from 0.9303 to 0.9340, reducing Avg. L2 from 0.1383 to 0.1353, and reducing Min. Ang. from 5.72 to 4.85. It also remains competitive on the *consistent* subset, achieving the best AUC and tied best Avg. L2. These results indicate that softly suppressing gaze-inconsistent evidence is more effective than forcing intermediate features to match a fixed direction vector or rigid cone-shaped distribution.



Figure 6: We visualize the predictions of GazeLLE and our method for comparison. GazeLLE often assigns high responses to semantically salient but incorrect regions, while our method follows the queried person's gaze and localizes the true target. This shows improved gaze reasoning beyond semantic saliency. Zoom in for facial and gaze direction details.

Table 4: Comparison with other gaze-aware supervision strategies on GazeFollow. All variants use HCLoRA and differ only in the auxiliary supervision strategy.

Method	Consistent					Inconsistent				
	AUC↑	A-L2↓	M-L2↓	A-Ang↓	M-Ang↓	AUC↑	A-L2↓	M-L2↓	A-Ang↓	M-Ang↓
w/o supervision	0.9677	0.0786	0.0366	10.66	3.37	0.9303	0.1383	0.0567	19.07	5.72
Gaze-vector	0.9682	0.0790	0.0366	10.34	3.11	0.9318	0.1378	0.0578	18.80	5.42
Heatmap	0.9681	0.0782	0.0359	10.51	3.13	0.9293	0.1385	0.0586	18.59	5.42
Gaze-cone	0.9678	**0.0777**	**0.0347**	**10.25**	**3.00**	0.9304	0.1385	0.0578	18.59	5.26
Out-of-cone penalty	**0.9696**	**0.0777**	0.0348	10.50	3.19	**0.9340**	**0.1353**	**0.0547**	**18.00**	**4.85**

5.5 Gaze Reasoning Ability Analysis

Our goal is to improve the gaze reasoning ability. Compared with the L2 error, angular error better reflects gaze reasoning capability. Therefore, we conduct experiments to count the number of samples with large angular errors. Intuitively, methods with fewer large-error samples exhibit stronger gaze reasoning ability.

The results are shown in Fig. 7. "$> k°$" denotes the number of samples whose minimum angular error exceeds $k°$. Our method is built upon GazeLLE. The results show that both HCLoRA and OOC improve gaze reasoning ability. Under the 30° threshold, the number of large-error samples decreases from 96 to 70 with HCLoRA, and further to 52 with OOC. These results indicate that our method not only improves overall performance, but also enhances gaze reasoning ability, consistent with our original motivation.

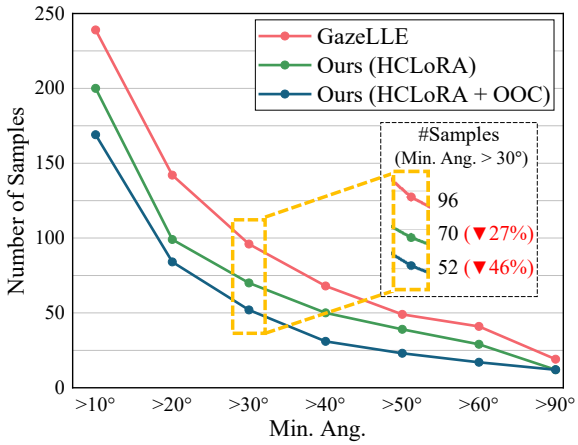


Figure 7: We count the number of samples with large angular errors. "$> k°$" denotes the number of samples whose minimum angular error exceeds $k°$. The results show that our method effectively enhances gaze reasoning ability, consistent with our motivation.

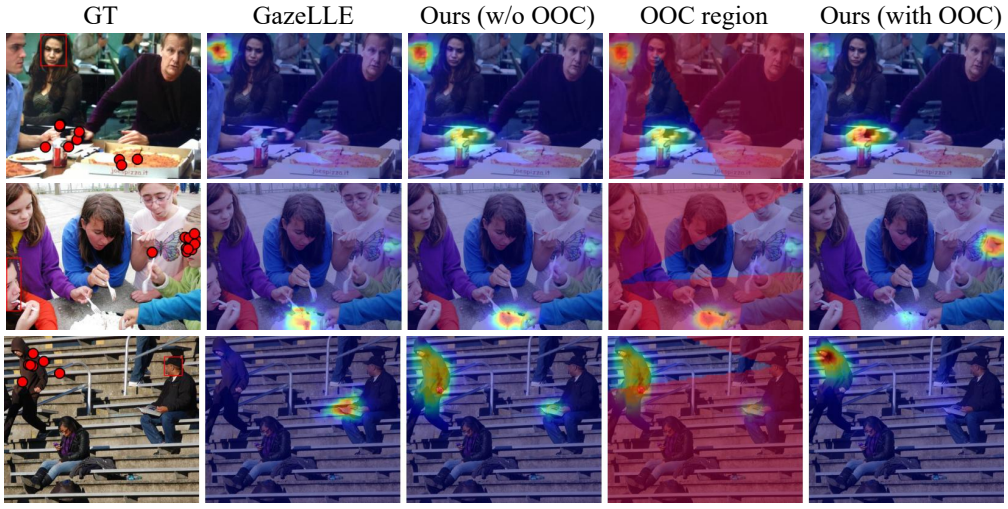


Figure 8: We visualize the impact of the OOC penalty on gaze prediction. The OOC penalty improves performance by suppressing predictions outside the gaze cone.

5.6 Visualization Results

We further provide qualitative visualizations to demonstrate the advantage of our method. As shown in Fig. 6, GazeLLE often produces high responses around semantically salient objects or interaction centers, even when these regions are inconsistent with the queried person's gaze direction. In contrast, our method more accurately follows the target person's head and gaze cues, shifting the response toward the true gaze target. This shows that our method reduces reliance on semantic saliency and improves localization of the correct gaze direction.

Fig. 8 visualizes the effect of the out-of-cone penalty. We overlay the OOC region on the prediction of HCLoRA *w/o* OOC. The OOC penalty suppresses these gaze-inconsistent responses and encourages the model to concentrate its predictions on the gaze-consistent target. It guides the model toward more accurate target localization by discouraging evidence outside the gaze-consistent region.

6 Conclusion

We study the limitations of VFM-based gaze following and show that strong benchmark performance can still be driven by semantic shortcuts rather than reliable gaze reasoning. To address this, we propose HCLoRA and OOC penalty to enhance gaze reasoning. Experiments on GazeFollow and VideoAttentionTarget demonstrate that our method achieves state-of-the-art performance while significantly improving gaze reasoning ability, consistent with our primary motivation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62672028, 62271042, 62376021), the Beijing Natural Science Foundation (L259055), and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2022-II220608). The research was conducted during Shijing Wang's visit to the University of Birmingham.

References

[1] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer Nature Switzerland, 2022.

[3] 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.

[4] 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*, pages 14106–14115, 2022.

[5] Francesca Capozzi, Cigdem Beyan, Antonio Pierro, Atesh Koul, Vittorio Murino, Stefano Livi, Andrew P. Bayliss, Jelena Ristic, and Cristina Becchio. Tracking the leader: Gaze behavior in group interactions. *iScience*, 16:242–249, 2019.

[6] Rodrigo Chacón Quesada, Fernando Estévez Casado, and Yiannis Demiris. An integrated 3D eye-gaze tracking framework for assessing trust in human–robot interaction. *ACM Transactions on Human-Robot Interaction*, 14(3):57:1–57:28, 2025.

[7] 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, 2022.

[8] 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)*, pages 397–412, 2018.

[9] 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*, pages 5395–5405, 2020.

[10] 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*, pages 11390–11399, 2021.

[11] 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 (CVPR) Workshops*, pages 5037–5046, 2022.

[12] Anshul Gupta, Pierre Vuillecard, Arya Farkhondeh, and Jean-Marc Odobez. Exploring the zero-shot capabilities of vision-language models for improving gaze following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 615–624, 2024.

[13] Roy S. Hessels, Toshiki Iwabuchi, Diederick C. Niehorster, Ren Funawatari, Jeroen S. Benjamins, Sayaka Kawakami, Marcus Nyström, Momoka Suda, Ignace T. C. Hooge, Motofumi Sumiya, et al. Gaze behavior in face-to-face interaction: A cross-cultural investigation between Japan and The Netherlands. *Cognition*, 263:106174, 2025.

[14] Nora Horanyi, Linfang Zheng, Eunji Chong, Aleš Leonardis, and Hyung Jin Chang. Where are they looking in the 3D space? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2678–2687, 2023.

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[16] Zhengxi Hu, Kunxu Zhao, Bohan Zhou, Hang Guo, Shichao Wu, Yuxue Yang, and Jingtai Liu. Gaze target estimation inspired by interactive attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8524–8536, 2022.

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, Lecture Notes in Computer Science, pages 709–727. Springer, 2022. doi: 10.1007/978-3-031-19827-4_41.

[18] Tianlei Jin, Qizhi Yu, Shiqiang Zhu, Zheyuan Lin, Jie Ren, Yuanhai Zhou, and Wei Song. Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence*, 113:104924, 2022.

[19] Nayeon Kim and Hyunsoo Lee. Assessing consumer attention and arousal using eye-tracking technology in virtual retail environment. *Frontiers in Psychology*, 12:665658, 2021.

[20] Jing Li, Zejin Chen, Yihao Zhong, Hak-Keung Lam, Junxia Han, Gaoxiang Ouyang, Xiaoli Li, and Honghai Liu. Appearance-based gaze estimation for ASD diagnosis. *IEEE Transactions on Cybernetics*, 52(7):6504–6517, 2022.

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

[22] 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*, pages 880–889, 2023.

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[24] Adrià Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, volume 28, pages 199–207, 2015.

[25] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28874–28884, 2025.

[26] Samy Tafasca, Anshul Gupta, Nada Kojovic, Mirko Gelsomini, Thomas Maillart, Michela Papandrea, Marie Schaer, and Jean-Marc Odobez. The AI4Autism project: A multimodal and interdisciplinary approach to autism diagnosis and stratification. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 414–425, 2023.

[27] 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*, pages 20878–20889, 2023.

[28] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Sharingan: A transformer architecture for multi-person gaze following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2008–2017, 2024.

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

[30] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. End-to-end human-gaze-target detection with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2192–2200, 2022.

[31] Zejia Zhang, Bo Yang, Xinxing Chen, Weizhuang Shi, Haoyuan Wang, Wei Luo, and Jian Huang. MindEye-OmniAssist: A gaze-driven LLM-enhanced assistive robot system for implicit intention recognition and task execution. In *2025 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, pages 1–6, 2025.

A Additional Details for Observation and Analysis

In this appendix, we provide additional details for the observation and analysis in Section 3. We first present more failure cases that illustrate the semantic shortcut behavior discussed in Section 3.1. We then describe the prompt used for semantic target annotation and provide qualitative examples of the semantic-consistency-based partition.

A.1 More Failure Cases of Semantic Shortcut

Figure 9 shows additional failure cases of GazeLLE [25]. These examples further support our observation that the model often predicts semantically salient regions, such as manipulated objects or interaction centers, even when these regions are inconsistent with the target person's gaze direction. This behavior indicates that the model may rely on semantic plausibility rather than true gaze cues.

A.2 Prompt for Semantic Target Annotation

We use Qwen3-VL-32B-Instruct [3] as a semantic prober to annotate contextually plausible gaze regions for the target person. The input image contains a red bounding box marking the target person's head. The model is asked to return four types of semantic regions: hand object, social target, action target, and semantic center. The full prompt is shown below.



Figure 9: Additional failure cases of semantic shortcut behavior in GazeLLE [25]. The predictions often concentrate on semantically salient regions or interaction centers, while deviating from the true gaze targets. These cases further illustrate that existing VFM-based gaze-following methods may rely on semantic saliency rather than robust gaze reasoning.

Prompt for Semantic Target Annotation

Preamble

```
<Task Description>
The red box marks the HEAD of the target person.

Your job is to identify semantic targets in the scene that are strongly
related to what this person may be attending to.
Return results for the following four slots:
```

Semantic Slots

```
<Slots>
1.  hand_object:
  ONE non-person object that this person is holding, touching with the hands,
or directly operating right now.
  Examples:  phone, cup, laptop, book, tool, remote controller, steering
wheel, toy.

2.  social_target:
  ONE other person who is directly socially interacting with the target
person.
  Examples:  the person they are talking to, hugging, facing in conversation,
handing something to, or responding to.

3.  action_target:
  ONE object that is the direct target of this person's current action.
  Examples:  the ball they just hit or throw, the skateboard they are riding,
  the obstacle they are jumping onto, the door they are opening, the screen
they are typing on.

4.  semantic_center:
  ONE event-central object/person in the scene that is most semantically
important for understanding
  what is happening, and is also plausibly something the target person may
attend to.
  Examples:  birthday cake at a birthday party, ball in a sports play,
microphone or podium in a speech,
  board or screen in a classroom, gift being opened, musical instrument being
performed.
```

Instructions

```
<Requirements>
1.  Do NOT return the target person themselves.
2.  Do NOT choose generic background regions or random furniture unless
clearly central.
3.  If a slot does not exist, return null for that slot.
```

Output Format

```
<JSON Schema>
Return ONLY valid JSON in this exact schema:
{
  "hand_object":  {"bbox_2d":  [x1, y1, x2, y2], "label":  "..." } or null,
  "social_target":  {"bbox_2d":  [x1, y1, x2, y2], "label":  "..." } or null,
  "action_target":  {"bbox_2d":  [x1, y1, x2, y2], "label":  "..." } or null,
  "semantic_center":  {"bbox_2d":  [x1, y1, x2, y2], "label":  "..." } or
null
}

Coordinates must be integers from 0 to 1000 and use [x1, y1, x2, y2].
```

A.3 Qualitative Examples of Semantic-Consistency-Based Partition

Figure 10 provides representative examples of the semantic-consistency-based partition. For each sample, the semantic prober predicts contextually plausible gaze regions, shown as yellow dashed bounding boxes. The ground-truth gaze annotations are overlaid as red circles. A sample is labeled as *consistent* if more than half of the valid ground-truth gaze annotations fall inside at least one predicted semantic region; otherwise, it is labeled as *inconsistent*.



Figure 10: Semantic-consistency-based partition examples. The red box marks the target person's head, yellow dashed boxes denote semantic regions predicted by the semantic prober, and red circles denote ground-truth gaze annotations. **(a)** Consistent samples, where most gaze annotations fall inside the predicted semantic regions. **(b)** Inconsistent samples, where gaze annotations lie outside the predicted semantic regions despite their contextual saliency.

B Additional Experimental Results

B.1 Additional Experimental Details

Dataset partition. We provide the exact subset statistics used in our semantic-consistency-based evaluation. GazeFollow contains only in-frame test samples, which are divided into 3,179 *consistent* and 1,603 *inconsistent* samples. For VideoAttentionTarget (VAT), we apply the semantic-consistency partition to in-frame samples and keep out-of-frame samples as a separate group, obtaining 12,664 *consistent*, 7,991 *inconsistent*, and 11,323 out-of-frame samples. These subset-level statistics are used for the diagnostic evaluation reported in the main experiments.

Implementation details. The input image resolution is set to 448×448. For GazeFollow, we train the model for 15 epochs with a batch size of 60. We use the Adam optimizer with an initial learning rate of 1×10^{-3} and a cosine learning-rate schedule. For VideoAttentionTarget, we initialize from the GazeFollow-trained model and jointly fine-tune for 8 epochs while keeping the original VFM

backbone frozen. We use the Adam optimizer with a learning rate of 1×10^{-2} for the in/out prediction parameters, 1×10^{-7} for the HCLoRA modules, and 1×10^{-5} for the remaining parameters.

HCLoRA-specific settings. For HCLoRA, the gaze-aware supervision loss is applied to the last $L = 4$ adapted layers with a weight of $\lambda = 0.02$. For the soft forward gaze cone, we use a cone angle of 30° and set the sharpness parameter to $\alpha = 40$.

Comparison methods. We compare HCLoRA with representative gaze-following methods, including Chong et al. [9], Miao et al. [22], Sharingan [28], and GazeLLE [25]. These methods cover both earlier CNN/transformer-based models and recent approaches built on large pre-trained vision backbones. In particular, Sharingan and GazeLLE are strong recent baselines with highly competitive performance on standard gaze-following benchmarks, making them suitable for evaluating robustness under semantic shortcuts. Since HCLoRA is built on the GazeLLE framework, we use GazeLLE as the main baseline and further compare with adaptation variants using the same backbone, including full fine-tuning, standard LoRA, and visual prompt tuning (VPT). For full fine-tuning on GazeFollow, we start from a learning rate of 1×10^{-5}, as the default learning rate is too large for updating all backbone parameters and leads to unstable training.

B.2 Overall Comparison with State-of-the-Art Methods

Table 5 summarizes the overall comparison with prior state-of-the-art methods on GazeFollow and VideoAttentionTarget. Results of prior methods are taken from GazeLLE [25], while our overall results are computed from the corresponding subset-level evaluations. Our method achieves the best overall localization performance on both benchmarks using only RGB input. Compared with GazeLLE under the same ViT-B backbone, our method reduces Avg. L2 from 0.104 to 0.097 on GazeFollow and L2 from 0.107 to 0.101 on VideoAttentionTarget. The ViT-L variant further improves performance, demonstrating that enhancing gaze reasoning benefits both diagnostic subset evaluation and standard overall evaluation.

Table 5: Overall comparison with state-of-the-art methods on GazeFollow and VideoAttentionTarget. Results of prior methods are taken from GazeLLE [25], while our overall results are computed from the corresponding subset-level evaluations. Learnable Params denotes the number of trainable parameters. Input types are denoted as I: image, D: depth, E: eye, P: pose, and O: object. With only RGB input and a small number of learnable parameters, our method achieves state-of-the-art localization performance on both benchmarks.

Method	**Learnable Params**	**Input**	**GazeFollow**			**VideoAttentionTarget**		
			AUC ↑	**Avg. L2** ↓	**Min. L2** ↓	**AUC** ↑	**L2** ↓	**AP$_{\text{in/out}}$** ↑
Recasens et al. [24]	50M	I	0.878	0.190	0.113	–	–	–
Chong et al. [8]	51M	I	0.896	0.187	0.112	0.833	0.171	0.712
Lian et al. [21]	55M	I	0.906	0.145	0.081	–	–	–
Chong et al. [9]	61M	I	0.921	0.137	0.077	0.860	0.134	0.853
Chen et al. [7]	50M	I	0.908	0.136	0.074	–	–	–
Fang et al. [10]	68M	I+D+E	0.922	0.124	0.067	0.905	0.108	0.896
Bao et al. [4]	29M	I+D+P	0.928	0.122	–	0.885	0.120	0.869
Jin et al. [18]	>52M	I+D+P	0.920	0.118	0.063	0.900	0.104	0.895
Tonini et al. [29]	92M	I+D	0.927	0.141	–	0.862	0.125	0.742
Hu et al. [16]	>61M	I+D+O	0.923	0.128	0.069	0.880	0.118	0.881
Gupta et al. [11]	35M	I+D+P	0.943	0.114	0.056	0.914	0.110	0.879
Horanyi et al. [14]	46M	I+D	0.896	0.196	0.127	0.832	0.199	0.800
Miao et al. [22]	61M	I+D	0.934	0.123	0.065	0.917	0.109	**0.908**
Tafasca et al. [27]	>25M	I+D	0.939	0.122	0.062	0.914	0.109	0.834
Tafasca et al. [28]	105M	I	0.944	0.113	0.057	–	0.107	0.891
GazeLLE (ViT-B) [25]	2.8M	I	0.956	0.104	0.045	0.933	0.107	0.897
GazeLLE (ViT-L) [25]	2.9M	I	<u>0.958</u>	0.099	<u>0.041</u>	<u>0.937</u>	0.103	0.903
Ours (ViT-B)	3.7M	I	<u>0.958</u>	<u>0.097</u>	<u>0.041</u>	0.935	<u>0.101</u>	0.899
Ours (ViT-L)	5.3M	I	**0.960**	**0.090**	**0.037**	**0.939**	**0.095**	<u>0.907</u>

B.3 Hyperparameter Sensitivity

We provide a hyperparameter sensitivity study in Table 6. We vary one hyperparameter at a time, including the out-of-cone loss weight λ, cone angle θ, and the number of last adapted layers L used for gaze-aware supervision. The results show that our method is generally robust to these choices. Our default setting, $\lambda = 0.02$, $\theta = 30^\circ$, and $L = 4$, achieves the best performance on the inconsistent subset for most metrics while maintaining competitive performance on the consistent subset.

Table 6: Hyperparameter sensitivity on GazeFollow. We vary one hyperparameter at a time, including the loss weight λ, cone angle θ, and the number of last adapted layers L used for gaze-aware supervision, with the sharpness parameter fixed as $\alpha = 40$. A-L2/M-L2 and A-Ang/M-Ang denote average/minimum L2 distance and average/minimum angular error, respectively. Bold indicates the best results within each hyperparameter group; gray shading marks the default setting.

Setting	Consistent					Inconsistent				
	AUC↑	A-L2↓	M-L2↓	A-Ang↓	M-Ang↓	AUC↑	A-L2↓	M-L2↓	A-Ang↓	M-Ang↓
$\lambda = 0.005$	0.9678	**0.0777**	0.0354	10.53	3.24	0.9311	0.1363	0.0571	19.03	5.60
$\lambda = 0.02$	0.9696	**0.0777**	**0.0348**	**10.50**	**3.19**	**0.9340**	**0.1353**	**0.0547**	**18.00**	**4.85**
$\lambda = 0.05$	**0.9705**	0.0795	0.0365	10.68	3.31	0.9325	0.1423	0.0585	19.61	6.13
$\theta = 20^\circ$	**0.9700**	**0.0756**	**0.0335**	**10.22**	**3.00**	0.9338	0.1383	0.0569	19.11	5.72
$\theta = 30^\circ$	0.9696	0.0777	0.0348	10.50	3.19	**0.9340**	**0.1353**	**0.0547**	**18.00**	**4.85**
$\theta = 40^\circ$	**0.9700**	0.0772	0.0337	10.40	3.05	0.9334	0.1388	0.0573	18.48	5.29
$L = 2$	**0.9701**	0.0802	0.0359	10.84	3.38	**0.9343**	0.1393	0.0559	19.16	5.81
$L = 4$	0.9696	0.0777	**0.0348**	10.50	3.19	0.9340	**0.1353**	**0.0547**	**18.00**	**4.85**
$L = 8$	0.9690	**0.0772**	0.0350	**10.49**	**3.15**	0.9319	0.1377	0.0565	18.61	5.30

B.4 Additional Analysis of Head Integration and Branch Design

We further analyze whether stronger architectural head integration is sufficient to improve gaze reasoning. Following the design ablation protocol of GazeLLE [25], we compare different ways of incorporating the queried head information while keeping the decoder configuration consistent with their ablation setting. Specifically, we consider early head integration before the VFM encoder, late head integration after scene feature extraction, and whether to use a separate head-and-scene branch or a single scene branch.

As shown in Table 7, early head integration substantially degrades performance on both subsets, suggesting that injecting head information before the VFM encoder can disturb the pretrained visual representation. Late integration performs much better, but adding a separate head branch brings only limited gains and still leaves a large gap on the inconsistent subset. This supports our motivation that architectural head fusion alone is insufficient for robust gaze reasoning.

Table 7: GazeLLE design variants on GazeFollow under our consistent/inconsistent evaluation. Following the ablation setting of GazeLLE [25], we compare early vs. late head integration and single scene branch (S) vs. head-and-scene branches (H+S). Early integration clearly hurts performance, while late H+S fusion still leaves a large gap on the inconsistent subset.

Variant	Head	Br.	Consistent					Inconsistent				
			AUC↑	A-L2↓	M-L2↓	A-Ang↓	M-Ang↓	AUC↑	A-L2↓	M-L2↓	A-Ang↓	M-Ang↓
GazeLLE	Early	H+S	0.9213	0.1657	0.1079	18.21	9.30	0.8847	0.1905	0.1085	<u>23.82</u>	9.97
GazeLLE	Late	H+S	**0.9695**	**0.0875**	<u>0.0400</u>	**11.54**	**3.74**	**0.9286**	**0.1599**	**0.0743**	**22.15**	**8.14**
GazeLLE	Late	S	<u>0.9682</u>	**0.0875**	**0.0399**	<u>11.99</u>	<u>3.98</u>	<u>0.9219</u>	<u>0.1670</u>	<u>0.0783</u>	24.05	<u>9.65</u>

We also examine adding a separate head-and-scene branch to our method. As shown in Table 8, this design slightly improves a few metrics on the consistent subset, but degrades the inconsistent subset. For example, Avg. L2 increases from 0.1353 to 0.1421 and Min. Ang. from 4.85 to 6.35. Therefore, our final model keeps the simpler single-branch design and uses HCLoRA to inject head-conditioned adaptation inside the VFM.

Table 8: Effect of adding a separate head-and-scene branch to our method on GazeFollow. The H+S branch slightly improves some consistent-subset metrics but degrades the inconsistent subset.

Variant	Consistent					Inconsistent				
	AUC ↑	A-L2 ↓	M-L2 ↓	A-Ang ↓	M-Ang ↓	AUC ↑	A-L2 ↓	M-L2 ↓	A-Ang ↓	M-Ang ↓
Ours	0.9696	**0.0777**	0.0348	**10.50**	3.19	**0.9340**	**0.1353**	**0.0547**	**18.00**	**4.85**
Ours + H+S	**0.9704**	0.0790	**0.0345**	10.63	**3.04**	0.9310	0.1421	0.0610	19.70	6.35

B.5 Limitations in Challenging Scenarios

By incorporating head-conditioned adaptation, our method achieves strong overall gaze localization performance, with particularly strong gains on the inconsistent subsets. Nevertheless, as illustrated in Figure 11, it can still struggle when the queried head is small, blurred, partially occluded, or viewed from behind, making gaze direction difficult to infer. These challenging scenarios leave room for further improvement in exploiting limited head cues.



Figure 11: Challenging cases for our method on GazeFollow. The top row shows the queried head boxes and ground-truth gaze annotations in red; the bottom row shows our predicted heatmaps. Weak or ambiguous head cues can limit gaze localization in these cases.

C Broader Impacts

Gaze following is important for understanding human attention and behavior, with potential benefits for human-robot interaction, assistive systems, educational technologies, and human-centered scene understanding.

At the same time, gaze following is privacy-sensitive. If deployed without safeguards, such models could be misused for surveillance, behavioral profiling, or unintended inference of personal attention patterns, especially in public spaces, workplaces, and schools where explicit consent may be absent. Gaze behavior may also vary across individuals, contexts, cultures, and visual conditions, so benchmark-trained models may not generalize equally well across populations or environments.

Our work uses existing benchmark datasets and does not collect new personal data. Real-world deployment should require clear user consent, data protection measures, and careful consideration of privacy, fairness, and potential misuse.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: Claims match the method and results.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: See Appendix B.5 for failure cases with weak or ambiguous head cues.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

   Justification: No theoretical results or formal proofs are presented.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: Sections 4–5 and Appendices A and B describe the method, evaluation protocol, and training settings.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The code link is provided in the abstract.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Training and evaluation settings are described in Section 5.1 and Appendix B.1.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No].

   Justification: We report point estimates without error bars or significance tests.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No].

   Justification: Hardware specifications are not reported.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: We respect dataset and model terms; no new participant data were collected.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: See Appendix C for potential benefits and privacy, misuse, and fairness risks.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: This work does not release high-risk datasets or pretrained foundation models.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No].

    Justification: Existing assets are cited, but their licenses are not explicitly listed.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The code includes documentation.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: No crowdsourcing or new human-subject experiments were conducted.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: No new human-subject studies or personal-data collection were conducted.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: See Section 3.2 and Appendix A.2 for Qwen3-VL usage and prompts.