Multi-scale Object-Aware Gaze Estimation via Geometric Reasoning

Jiajie Mi*, Xinyu Liu*, Mengke Song, and Chenglizhao Chen†

Qingdao Institute of Software & School of Computer Science and Technology, China
University of Petroleum (East China), China
Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, China
cclz123@163.com

Abstract. Gaze target estimation aims to predict the semantic object an observer fixates upon within an image, a task deeply rooted in the object-oriented nature of human gaze. Observers tend to select a specific semantic entity as the attentional target, rather than responding randomly across arbitrary regions of the image. However, existing methods typically model this task as a direct mapping from global features to gaze heatmaps, essentially treating it as a pixel-level regression problem. This approach fails to explicitly represent the gazed object as a distinct entity, making it difficult to produce stable and semantically consistent predictions in complex scenes. To address this, we propose a two-stage gaze estimation framework guided by object semantics, reformulating gaze target estimation as a hierarchical reasoning process. Our method incorporates object-level representations during feature encoding to align image features with discrete semantic entities, then introduces multi-scale feature fusion and geometric constraints from head pose and gaze direction for fine-grained localization and object-level discrimination. Extensive experiments on GazeFollow, VideoAttentionTarget, ChildPlay, and GOO-Real demonstrate that our method achieves AUC of 0.961, 0.948, 0.987, and 0.977 respectively, delivering strong performance across all benchmarks while maintaining a compact parameter size of 7.1M.

Keywords: Gaze Target Estimation · Object-Aware Fusion · Multi-Scale Feature Fusion · Geometric Constraints

1 Introduction

Gaze target estimation aims to predict the location of the object that an observer is looking at in a scene. It is a fundamental task for understanding human visual attention and social interaction [13] [1] [15] [14]. In autonomous driving [11] [62] [30], a driver's gaze target reflects attention state and awareness of potential risks. In human-computer interaction [32] [1] [31] [4] and social robotics [61] [2] [23] [41], accurate understanding of gaze targets is essential for joint attention

* Equal contribution. † Corresponding author.

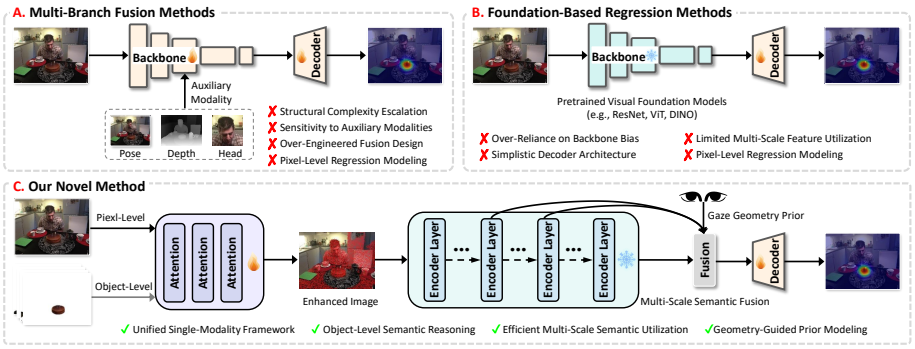

Fig. 1: Comparison of existing gaze estimation paradigms and the proposed method. (A) Multi-branch fusion methods. (B) Foundation-based regression methods. (C) Our method reformulates gaze target estimation as hierarchical reasoning, enabling object-level modeling and multi-scale localization in a unified single-modality framework.

[18] [38] [50] and natural interaction [17] [5]. Collectively, these scenarios establish gaze target estimation as a fundamental research problem in computer vision.

Although significant progress has been made in recent years, existing methods commonly formulate gaze target estimation as a direct regression from image features to spatial heatmaps. Current approaches can be broadly divided into two categories. As shown in Fig. 1(A), one line of work introduces auxiliary modalities such as depth [3] [16] [19] [26] [37], human pose [3] [60], or three-dimensional head information [3] [35] and builds multi-branch networks [56] [3] [10] [16] [21] [24] [26] to enhance spatial understanding. However, as more modalities are added, model complexity increases continuously. Prediction performance becomes highly dependent on the accuracy of upstream auxiliary models, which often leads to limited generalization across different scenes or real world environments. As shown in Fig. 1(B), another line of work adopts pretrained visual foundation models as unified feature extractors. This design improves parameter efficiency and simplifies network architecture, but prediction is usually performed using only the final layer features through shallow decoding. As a result, important spatial structure information contained in hierarchical representations is largely ignored. More importantly, both categories share the same underlying modeling assumption, where gaze prediction is treated as a pixel-level regression problem.

This modeling paradigm overlooks a fundamental property of human visual attention that human gaze is naturally object oriented. In real visual behavior, an observer first selects a specific semantic object as the focus of attention rather than responding randomly to arbitrary locations in the image. In other words, the gaze process consists of two consecutive stages. The observer first determines which object is being attended, and then forms a precise gaze location within that object region. When models lack explicit object-level decision modeling, predictions often become spatially scattered or semantically unstable in complex scenes where multiple candidate targets compete for attention. This

observation suggests that the main limitation of existing methods does not arise from insufficient feature representation, but from the absence of a structured modeling mechanism for object selection in task formulation.

Based on the above observations, this work reformulates gaze target estimation as a hierarchical reasoning process. The model first constructs a semantic hypothesis space of potential gaze objects within the scene and then incorporates geometric constraints of the observer to determine which regions are visually reachable. Precise localization is subsequently performed within candidate regions that satisfy both semantic and geometric consistency. This perspective transforms gaze prediction from single-step pixel regression into a reasoning problem composed of object-level discrimination and region-level localization, making the modeling process more consistent with human visual behavior.

Following this formulation, we propose a multi-scale object-aware framework for gaze target estimation. As illustrated in Fig. 1(C), the framework adopts a pretrained visual foundation model as a unified feature extractor to maintain parameter efficiency while enabling structured reasoning through three key mechanisms. First, object-level semantic representations are introduced during feature encoding to establish explicit associations between scene features and discrete semantic entities. Second, multi-scale feature fusion is applied to fully utilize spatial and semantic information across different hierarchical layers of the foundation model. Third, a field-of-view geometric constraint derived from gaze direction is introduced to provide a physiologically plausible spatial prior. These components together form a complete reasoning pipeline from candidate object identification to precise gaze localization in complex real-world scenes.

We evaluate our method on GazeFollow [43], VideoAttentionTarget [10], ChildPlay [48], and GOO-Real [51]. Our main contributions are as follows:

- We revisit gaze target estimation from a cognitive modeling perspective and reformulate it from pixel-level regression into a hierarchical reasoning problem that jointly performs object selection and spatial localization.
- We develop a unified single-modality framework that realizes the proposed hierarchical reasoning formulation through object-level semantic modeling, multi-scale feature representation, and gaze direction guidance.
- We demonstrate the proposed framework achieves strong and consistent performance on four benchmark datasets, with AUC scores of 0.961, 0.948, 0.987, and 0.977 on GazeFollow, VideoAttentionTarget, ChildPlay, and GOO-Real respectively, while maintaining a compact 7.1M-parameter model.

2 Related Work

Pixel-Level Gaze Estimation. Early gaze target estimation methods formulate gaze prediction as a direct mapping from visual features to spatial heatmaps. Representative approaches adopt dual-branch architectures [8] [9] [10] [29] [43] [45] to jointly encode head appearance and scene context in order to infer gaze locations. Subsequent works enhance this paradigm by incorporating attention mechanisms [47] [55] [49], global context modeling [58] [49] [55], or temporal

aggregation [20] [36] [37] to improve feature interaction between observers and surrounding regions. Despite architectural improvements, these methods fundamentally treat gaze estimation as a pixel-level regression problem, where predictions are generated directly from global feature responses without explicitly modeling the attended object. Such formulations often lead to spatially diffused predictions when multiple candidate objects coexist within complex scenes.

Object-Level Gaze Representation. Recognizing the object-oriented nature of human attention, several studies attempt to introduce object information into gaze estimation. Object-aware approaches [39] [27] leverage detected regions or semantic cues to guide attention modeling and improve interpretability. These methods demonstrate that incorporating object-level cues can reduce ambiguity caused by cluttered backgrounds or competing targets. However, object information is typically used as auxiliary guidance rather than being integrated into the core decision process. As a result, gaze prediction remains largely dependent on feature aggregation or regression-based decoding, lacking an explicit mechanism that models gaze estimation as object selection followed by spatial localization.

Foundation Model Based Gaze Estimation. Recent advances in visual foundation models [47] [44] [28] [36] [59] enable gaze estimation frameworks to benefit from strong pretrained representations. Methods built upon large-scale encoders improve parameter efficiency and generalization capability by leveraging frozen backbones. Nevertheless, most approaches rely primarily on final-layer features for prediction, underutilizing hierarchical representations available within deep networks. More importantly, these methods continue to adopt regression-based formulations, where gaze prediction is inferred from feature responses rather than structured reasoning over candidate objects. Consequently, the object-oriented nature of gaze behavior remains insufficiently modeled.

3 Method

3.1 Hierarchical Reasoning Overview

Gaze target estimation is reformulated as a hierarchical reasoning process that progressively narrows the search space from global scenes to precise fixation locations. Rather than directly regressing gaze heatmaps from global image features, the proposed framework models gaze prediction as structured inference over semantically meaningful and geometrically feasible regions. The model first establishes object-level semantic hypotheses by identifying candidate entities that may attract visual attention, while observer-dependent geometric cues derived from head position and gaze direction constrain attention to spatially plausible regions consistent with human visual behavior and natural attention patterns.

Building upon these semantic and geometric priors, multi-scale representations are jointly exploited to refine gaze localization within valid candidate regions. As illustrated in Fig. 2, the entire reasoning process is implemented within a unified single-modality framework based on a frozen DINOv3 backbone, where lightweight learnable modules progressively transform object-aware representations into precise gaze predictions with interpretable steps.

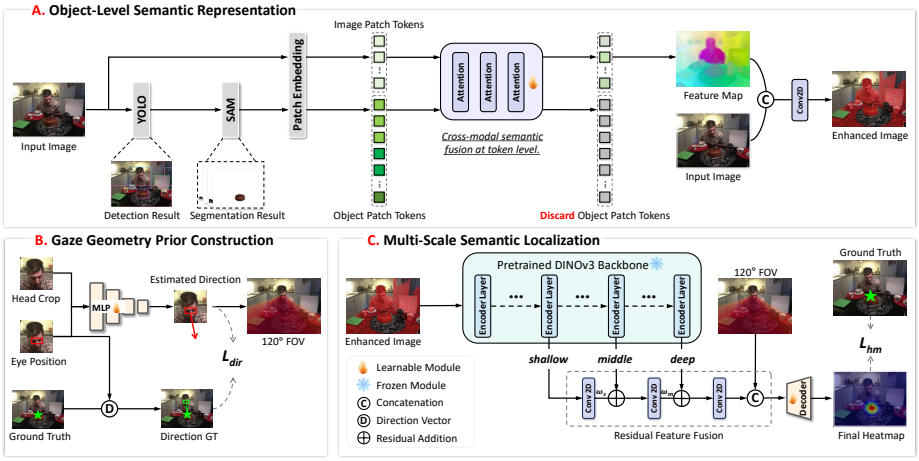

Fig. 2: Overview of the proposed framework. (A) Object-level semantic representation introduces object-aware tokens to enhance features through cross-modal fusion. (B) Gaze geometry prior construction estimates gaze direction from head appearance and eye position to generate a FOV prior. (C) Multi-scale semantic localization integrates hierarchical backbone features with the FOV prior to produce the gaze heatmap.

3.2 Object-Level Semantic Representation

Existing gaze target estimation methods typically predict gaze locations directly from global scene features, lacking explicit modeling of potential gaze objects. As a result, predictions often become unstable in complex scenes containing multiple competing targets or strong background distractions. Considering that human visual attention is inherently object-oriented, we introduce an object-level semantic representation to explicitly construct candidate gaze hypotheses during feature encoding, allowing subsequent reasoning to operate over discrete semantic entities rather than unconstrained spatial responses.

As illustrated in Fig. 2(A), given an input image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, we first obtain N candidate object masks $\{\mathbf{M}_i\}_{i=1}^{N}$, where $\mathbf{M}_i \in \{0,1\}^{H\times W}$, using pre-trained object detection and segmentation models. The image is divided into non-overlapping patches of size $p \times p$, which are projected into image tokens:

$$\mathbf{T}_{\text{img}} \in \mathbb{R}^{N_p\times d}, \quad N_p = \frac{H}{p}\cdot\frac{W}{p}, \tag{1}$$

where d denotes the token embedding dimension.

For each object region, masked image features are encoded into an object token sequence:

$$\mathbf{T}_{\text{obj}}^{(i)} \in \mathbb{R}^{N_o\times d}, \tag{2}$$

where N_o denotes the number of tokens used to represent each object, obtained through fixed-grid pooling with $N_o = n^2$. All object tokens are concatenated as:

$$\mathbf{T}_{\text{obj}} = [\mathbf{T}_{\text{obj}}^{(1)}; \ldots; \mathbf{T}_{\text{obj}}^{(N)}] \in \mathbb{R}^{(N \cdot N_o) \times d}. \tag{3}$$

The image tokens and object tokens are combined into a unified sequence:

$$\mathbf{T}_{\text{fuse}} = [\mathbf{T}_{\text{img}}; \mathbf{T}_{\text{obj}}] \in \mathbb{R}^{(N_p + N \cdot N_o) \times d}, \tag{4}$$

which is processed by a Transformer encoder to enable cross-token interaction. Through this interaction, semantic information associated with individual objects is propagated into the global visual representation, transforming pixel-based scene features into object-aware semantic representations and explicitly establishing candidate gaze hypotheses.

Finally, an object-aware spatial response $\mathbf{O} \in \mathbb{R}^{1 \times H \times W}$ is reconstructed from the fused representation and integrated with the original image through a lightweight fusion mapping:

$$\mathbf{I}_{\text{enh}} = \mathcal{F}(\mathbf{I}, \mathbf{O}), \tag{5}$$

where $\mathcal{F}(\cdot)$ denotes a feature fusion operation realized through channel concatenation followed by a 1×1 convolution. The enhanced representation $\mathbf{I}_{\text{enh}}$ preserves original appearance information while incorporating structured object-level semantics, providing a semantic foundation for subsequent geometry-guided reasoning and multi-scale gaze localization.

3.3 Gaze Geometry Prior Construction

After obtaining object-level semantic representations, scene semantics alone remain insufficient to uniquely determine gaze locations, since human gaze behavior is inherently constrained by the geometric state of the observer. In practice, visual attention is limited by head orientation and gaze direction, restricting attention to a finite forward-facing visual field. To incorporate such constraints, we explicitly estimate the observer's gaze direction and construct a geometry-aware spatial prior that limits the gaze inference space.

Given the head region image and the corresponding eye position coordinates (x_e, y_e), we first extract head appearance features $\mathbf{f}_{\text{head}}$. The head feature and eye position are jointly fed into a multilayer perceptron to predict a normalized two-dimensional gaze direction vector:

$$\hat{\mathbf{g}} = \frac{\text{MLP}_{dir}([\mathbf{f}_{\text{head}}, x_e, y_e])}{\|\text{MLP}_{dir}([\mathbf{f}_{\text{head}}, x_e, y_e])\|}. \tag{6}$$

The predicted vector represents the estimated gaze direction of the observer, as illustrated in Fig. 2(B).

To stabilize direction learning, a supervision signal is constructed from the ground-truth target. Let $\mathbf{p}_e$ denote the eye position and $\mathbf{p}_t$ denote the ground-truth gaze target position. The ground-truth gaze direction is defined as:

$$\mathbf{g}^{gt} = \frac{\mathbf{p}_t - \mathbf{p}_e}{\|\mathbf{p}_t - \mathbf{p}_e\|}. \tag{7}$$

The predicted direction is constrained using a directional loss:

$$\mathcal{L}_{\text{dir}} = 1 - \hat{\mathbf{g}} \cdot \mathbf{g}^{gt}. \tag{8}$$

Based on the estimated gaze direction, a field-of-view (FOV) geometric prior is constructed. For any pixel location (x, y), let $\hat{\mathbf{v}}_{(x,y)}$ denote the unit vector from the eye center. The directional response is computed via cosine similarity:

$$c_{(x,y)} = \hat{\mathbf{v}}_{(x,y)} \cdot \hat{\mathbf{g}}. \tag{9}$$

Considering the bounded nature of human vision, responses are thresholded by the FOV range θ, where $\theta = 120^\circ$ (Table 6), to obtain the cone map:

$$C_{(x,y)} = \begin{cases} c_{(x,y)}, & c_{(x,y)} > \cos\theta, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

To mitigate discontinuities near the cone boundary, the cone response is smoothed using Gaussian filtering, yielding FOV representation C_{smooth}. A learnable FOV embedding vector $\mathbf{e}_{fov} \in \mathbb{R}^{d'}$ is then broadcast-multiplied with the smoothed cone map to produce a geometry-aware representation:

$$\mathbf{E}_{fov} = C_{\text{smooth}} \odot \mathbf{e}_{fov}. \tag{11}$$

The resulting feature map $\mathbf{E}_{fov}$ explicitly encodes spatial regions that are visually reachable by the observer and serves as the gaze geometry prior injected into the subsequent multi-scale localization stage, guiding gaze prediction within geometrically feasible regions.

3.4 Multi-Scale Semantic Localization

After establishing object-level semantic hypotheses and constraining visually reachable regions through gaze geometry priors, precise gaze prediction requires fine-grained localization within semantically consistent and geometrically feasible regions. As illustrated in Fig. 2(C), we jointly integrate semantic representations and observer-dependent geometric constraints within a multi-scale feature space, enabling geometry-guided gaze localization and completing the hierarchical reasoning process from object selection to spatial refinement.

The enhanced representation $\mathbf{I}_{\text{enh}}$ is fed into a frozen DINOv3 backbone to extract hierarchical visual features from layers. Features at different depths encode complementary information, where shallow layers preserve spatial structures while deep layers capture high-level semantics. All feature maps are projected into a shared embedding space to facilitate cross-scale interaction.

Instead of directly aggregating multi-scale features, we treat deep semantic features as the dominant representation and introduce intermediate and shallow features as bounded residual complements:

$$\mathbf{F}_{\text{fused}} = \mathbf{F}_{\text{deep}} + \alpha_{\text{mid}}\mathbf{F}_{\text{mid}} + \alpha_{\text{shallow}}\mathbf{F}_{\text{shallow}}, \tag{12}$$

where α_{mid} and α_{shallow} denote multi-scale fusion coefficients, with hyperparameter analysis reported in Table 7.

The gaze geometry priors constructed in Sec. 3.3 are subsequently injected into the fused representation. Specifically, the field-of-view feature $\mathbf{E}_{fov}$, encoding geometrically reachable regions along the estimated gaze direction, is fused with semantic representations, allowing gaze prediction to be simultaneously guided by object semantics and gaze-reachable spatial constraints:

$$\mathbf{F}_{\text{guided}} = \text{Fusion}(\mathbf{F}_{\text{fused}}, \mathbf{E}_{fov}). \tag{13}$$

Finally, the geometry-guided representation is processed by a lightweight decoding network to progressively recover spatial resolution and produce the final gaze probability heatmap $\mathbf{H}$. The resulting prediction no longer relies on global pixel-wise regression, but instead emerges from the joint interaction between object-level semantic hypotheses and observer-dependent geometric constraints, enabling structured gaze localization.

3.5 Training Objective

The proposed framework is optimized under joint supervision for gaze localization and gaze existence prediction, enabling consistent learning of spatial reasoning under semantic and geometric constraints.

For gaze localization, the ground-truth fixation point is represented as a two-dimensional Gaussian heatmap $\mathbf{H}^{gt}$ centered at the annotated gaze target. The predicted heatmap $\mathbf{H}$ is optimized using pixel-wise binary cross-entropy loss:

$$\mathcal{L}_{\text{hm}} = -\sum_{x,y} \Big(\mathbf{H}^{gt}_{x,y} \log \mathbf{H}_{x,y} + (1 - \mathbf{H}^{gt}_{x,y}) \log(1 - \mathbf{H}_{x,y}) \Big). \tag{14}$$

To handle cases where the gaze target lies outside the image, an additional in/out prediction branch is introduced. Let p denote the predicted probability indicating whether the gaze target is inside the image and $y \in \{0, 1\}$ the corresponding ground-truth label. The in/out classification loss is defined as:

$$\mathcal{L}_{\text{inout}} = -y \log p - (1 - y) \log(1 - p). \tag{15}$$

The overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{hm}} + \lambda_{\text{dir}} \mathcal{L}_{\text{dir}} + \lambda_{\text{inout}} \mathcal{L}_{\text{inout}}, \tag{16}$$

where $\mathcal{L}_{\text{dir}}$ denotes the direction supervision defined in Sec. 3.3, and λ_{dir} is detailed in Table 8. $\lambda_{\text{inout}} \in \{0, 1\}$ is a dataset-dependent binary hyperparameter, set to 1 when in/out-of-frame annotations are available and 0 otherwise. The weighting coefficients balance supervision signals during optimization.

4 Experiments

4.1 Experimental Setup, Datasets, and Metrics

Implementation Details. All experiments are conducted on a single NVIDIA RTX 5090 GPU. A pretrained DINOv3 ViT-L/16 backbone is employed and kept entirely frozen during training. Object masks are generated offline using YOLO11x and SAM2-hiera-large [42]. Input images are resized to 512×512, and gaze heatmaps are predicted at 64×64 resolution. Training is performed using Adam [12] with an initial learning rate of 1×10^{-3}. The overall training objective jointly optimizes heatmap regression, direction supervision, and in/out classification losses. More details are provided in the *supplementary material*.

Datasets. Experiments are conducted on four public gaze estimation benchmarks: GazeFollow [43], VideoAttentionTarget [10], ChildPlay [48], and GOO-Real [51]. GazeFollow provides image-based gaze annotations with multiple observers per sample, enabling robust evaluation under gaze ambiguity. VideoAttentionTarget extends the task to video sequences and includes gaze in/out-of-frame annotations. ChildPlay evaluates gaze prediction in child-interaction scenarios characterized by larger behavioral variability. GOO-Real focuses on real-world scenarios with notable domain shifts, serving as a benchmark for cross-domain generalization and distribution shift robustness.

Evaluation Metrics. Performance is evaluated using standard metrics including AUC and L2 distance under official evaluation protocols for fair and consistent comparison. For GazeFollow, both Avg L2 and Min L2 are reported due to multiple gaze annotations per image. Average Precision (AP) is additionally adopted for datasets involving gaze in/out-of-frame prediction tasks.

4.2 Comparison with Existing Methods

We comprehensively compare our method with a broad range of existing gaze target estimation approaches across all evaluated benchmarks. On GazeFollow and VideoAttentionTarget, comparisons include Fang et al. [29], Jin et al. [26], Tonini et al. [52], Hu et al. [25], HGTTR [54], GTR [55], Horanyi et al. [24], Miao et al. [37], Sharingan [49], Gaze-LLE [44], Chen et al. [7], GazeVLM [33], and FGI-Gaze [28]. On ChildPlay, we further compare with Gupta et al. [21], Tafasca et al. [48], Sharingan [49], MTGS [20], and Gaze-LLE [44]. For the cross-domain benchmark GOO-Real, comparisons involve Tonini et al. [52], Miao et al. [37], Gaze-LLE [44], Mathew et al. [34], and Wang et al. [57], covering representative methods under diverse evaluation settings and experimental conditions.

Quantitative Comparisons. As shown in Table 1, the proposed method achieves consistently strong performance across all evaluated benchmarks. On GazeFollow, our framework obtains the highest AUC together with lower Avg L2 and Min L2 errors, indicating improved spatial consistency and more precise gaze localization. On VideoAttentionTarget, the proposed method maintains competitive performance in both localization accuracy and in/out-of-frame prediction,

Table 1: Quantitative comparisons on (A) GazeFollow, (B) VideoAttentionTarget, (C) ChildPlay, and (D) GOO-Real. Higher AUC/AP and lower L2 indicate better performance. Bold and underline denote best and second-best.* estimated parameters.

Model	Params	A. GazeFollow			B. VideoAttentionTarget		
		AUC ↑	Avg L2 ↓	Min L2 ↓	AUC ↑	L2 ↓	AP ↑
Fang et al. [2021]	68M	0.922	0.124	0.067	0.905	0.108	0.896
Jin et al. [2022]	>52M*	0.920	0.118	0.063	0.900	0.104	0.895
Tonini et al. [2022]	92M	0.927	0.141	-	0.862	0.125	0.742
Hu et al. [2022]	>61M*	0.923	0.128	0.069	0.880	0.118	0.881
HGTTR [2022]	35M*	0.917	0.133	0.069	0.893	0.137	0.821
GTR [2023]	48.2M*	0.928	0.114	0.057	0.925	0.093	0.923
Horanyi et al. [2023]	46M	0.896	0.196	0.127	0.832	0.199	0.800
Miao et al. [2023]	61M	0.934	0.123	0.065	0.917	0.109	0.908
Sharingan [2024]	>135M*	0.944	0.113	0.057	-	0.107	0.891
Gaze-LLE [2025]	**2.8M**	0.956	0.104	0.045	0.933	0.107	0.897
Chen et al. [2025]	49.4M*	0.934	0.108	0.070	-	-	-
GazeVLM [2025]	2G	0.929	0.131	0.076	0.926	0.112	0.898
FGI-Gaze [2026]	35M*	0.950	0.099	0.044	**0.958**	**0.091**	0.911
Ours	7.1M	**0.961**	**0.094**	**0.038**	0.948	0.095	**0.923**

Model	C. ChildPlay		
	AUC ↑	L2 ↓	AP ↑
Gupta et al. [2022]	0.919	0.113	0.983
Tafasca et al. [2023]	0.935	0.107	0.986
Sharingan [2024]	-	0.106	0.990
MTGS [2024]	-	0.113	0.993
Gaze-LLE [2025]	0.951	0.101	**0.994**
Ours	**0.987**	**0.084**	0.990

Model	D. GOO-Real	
	AUC ↑	L2 ↓
Tonini et al. [2022]	0.840	0.238
Miao et al. [2023]	0.869	0.202
Gaze-LLE [2024]	0.901	0.174
Mathew et al. [2025]	0.954	0.130
Wang et al. [2025]	0.957	**0.081**
Ours	**0.977**	0.092

demonstrating robustness under dynamic scene variations without explicit temporal modeling. For ChildPlay, which involves complex interaction behaviors and increased appearance diversity, our approach achieves clear improvements in both AUC and L2 metrics, suggesting enhanced capability in resolving competing attention targets. On the cross-domain benchmark GOO-Real, our method achieves the best AUC with only lightweight fine-tuning, highlighting strong generalization ability to real-world scenarios. Notably, these results are achieved with only 7.1M parameters, providing a favorable balance between prediction accuracy and model efficiency compared with existing approaches.

Qualitative Analysis. The qualitative results shown in Fig. 3 further validate the effectiveness of the proposed method in complex real-world scenarios. Compared with existing approaches such as Tonini et al. [53] and Gaze-LLE [44], which often produce scattered or shifted responses under multiple candidate

Fig. 3: Qualitative comparison of gaze prediction results on representative examples.

targets or background distractions, our method generates more compact predictions that closely align with the true gaze targets. This advantage arises from reformulating gaze target estimation as a hierarchical reasoning process, where object-level semantic modeling identifies potential attention entities while geometrically reachable regions constrain the spatial search space. Such a design effectively suppresses irrelevant distractions and alleviates attention ambiguity. As a result, the predicted heatmaps exhibit clearer spatial concentration and stronger object consistency, indicating that the model performs stable structured gaze decision making rather than relying on pixel-level regression. These visualization results are consistent with quantitative findings, further demonstrating the robustness and reliability of the proposed framework in complex interaction and real-world environments.

4.3 Ablation Study

Component Ablation. Table 2 presents ablation results by progressively enabling object-level fusion, multi-scale representation, and gaze geometry prior modules (FOV). Starting from the baseline model without additional components (No. 0), introducing object-level semantic fusion improves performance on both datasets, indicating that explicit object-aware representations reduce attention ambiguity. Incorporating multi-scale feature modeling further improves localization accuracy, demonstrating the importance of leveraging hierarchical visual representations beyond single-layer features. When the gaze FOV prior is introduced, performance improves by constraining predictions within geometrically feasible regions. Combining semantic modeling with geometric guidance produces more stable spatial responses than using either component alone.

Table 2: Ablation study on key components across GazeFollow and VideoAttentionTarget, with bold entries denoting the best results.

No.	Object Fusion	Multi-Scale	FOV	GazeFollow			VideoAttentionTarget		
				AUC ↑	Avg L2 ↓	Min L2 ↓	AUC ↑	L2 ↓	AP ↑
0				0.920	0.188	0.118	0.886	0.162	0.865
1	✓			0.932	0.166	0.101	0.903	0.140	0.887
2		✓		0.937	0.148	0.087	0.906	0.134	0.891
3	✓	✓		0.954	0.111	0.049	0.937	0.108	0.912
4	✓		✓	0.945	0.132	0.069	0.922	0.122	0.903
5		✓	✓	0.946	0.125	0.063	0.919	0.124	0.904
6	✓	✓	✓	**0.961**	**0.094**	**0.038**	**0.948**	**0.095**	**0.923**

Table 3: Ablation study on the impact of multi-scale feature levels.

No.	S	M	D	AUC ↑	Avg L2 ↓	Min L2 ↓
1	✓			0.931	0.139	0.068
2		✓		0.936	0.132	0.063
3			✓	0.943	0.122	0.056
4	✓	✓		0.940	0.126	0.059
5	✓		✓	0.952	0.108	0.047
6	✓	✓	✓	**0.961**	**0.094**	**0.038**

Table 4: Comparison of different backbone architectures.

Backbone	AUC ↑	Avg L2 ↓	Min L2 ↓
Resnet-50	0.932	0.166	0.101
Dinov1	0.937	0.148	0.087
Dinov2	0.946	0.125	0.063
Dinov3	**0.961**	**0.094**	**0.038**

The complete model integrating all three components achieves the best performance across all evaluation metrics, confirming that object-level reasoning, multi-scale representation, and geometry-aware priors contribute complementarily to accurate gaze target estimation on tested benchmarks.

Multi-scale Ablation. Table 3 studies the impact of selecting backbone features from different depth levels for multi-scale semantic localization, where S, M, and D denote shallow, middle, and deep features, respectively. Using a single feature level provides reasonable performance, while deeper representations yield stronger semantics and lower localization error. Performance improves progressively as features from multiple depth levels are incorporated, indicating that spatial details from shallow layers and semantic cues from deeper layers are complementary. The best results are achieved when all feature levels are utilized, validating the effectiveness of the proposed multi-scale fusion strategy.

Backbone Analysis. Table 4 compares different backbone architectures for feature extraction. Performance improves consistently from ResNet-50 [22] to DINO-based models [6] [40] [46], highlighting the advantage of pretrained vision foundation models in capturing semantic representations relevant to gaze understanding. The DINOv3 backbone achieves the best results across all metrics, suggesting that stronger hierarchical representations further significantly enhance the effectiveness of the proposed reasoning framework.

Table 5: Ablation study on attention layer numbers in cross-modal fusion.

Layers	AUC ↑	Avg L2 ↓	Min L2 ↓
1	0.939	0.126	0.059
2	0.950	0.111	0.049
3	**0.961**	**0.094**	**0.038**
4	0.959	0.097	0.040

Table 6: Impact of different FOV ranges on gaze estimation performance.

Range	AUC ↑	Avg L2 ↓	Min L2 ↓
60°	0.940	0.125	0.059
90°	0.951	0.109	0.048
120°	**0.961**	**0.094**	**0.038**
150°	0.953	0.107	0.046

Table 7: Ablation study on the residual fusion weights defined in Eq. 12.

$\alpha_{shallow}$ \ α_{mid}	**0**	**0.05**	**0.10**	**0.15**	**0.20**
0	0.943	0.944	0.947	0.943	0.940
0.05	0.947	0.955	**0.961**	0.956	0.949
0.10	0.945	0.952	0.958	0.953	0.947
0.15	0.942	0.948	0.953	0.949	0.943
0.20	0.939	0.944	0.949	0.945	0.940

Table 8: Ablation study on the direction supervision coefficient λ_{dir} in Eq. 16.

λ_{dir}	AUC ↑	Avg L2 ↓	Min L2 ↓
0.2	0.948	0.104	0.044
0.3	0.953	0.099	0.041
0.4	0.958	0.096	0.039
0.5	**0.961**	**0.094**	**0.038**
0.6	0.957	0.097	0.040

4.4 Parameter Sensitivity Analysis

To further analyze the robustness of the proposed framework, we conduct parameter sensitivity experiments on key components in semantic fusion, geometric prior modeling, multi-scale feature integration, and training optimization.

Effect of Attention Layer Depth. Table 5 evaluates the influence of the number of attention layers in the cross-modal semantic fusion module. Increasing attention depth enhances object-level semantic interaction and improves gaze localization accuracy. Performance peaks when three attention layers are adopted, indicating that sufficient semantic reasoning is achieved at moderate depth. Further increasing the layer number slightly degrades performance, which may be attributed to feature over-smoothing and redundant token interaction.

Effect of FOV Range. Table 6 studies the impact of different field-of-view (FOV) ranges in the geometry prior construction stage. A narrow FOV overly restricts the search space, while an excessively large FOV weakens geometric guidance. The best performance is achieved with a 120° FOV, suggesting that a balanced geometric constraint effectively limits attention to visually reachable regions without suppressing valid gaze targets. Visualizations of the FOV cone map are provided in the *supplementary material*.

Effect of Multi-scale Residual Fusion Weights. Table 7 analyzes the residual fusion coefficients defined in Eq. 12, where deep semantic features serve as the primary representation while shallow and middle features are introduced as residual refinements controlled by α_{shallow} and α_{mid}. The results show that moderate residual contributions improve localization accuracy, whereas excessive weighting introduces feature interference and slightly degrades performance.

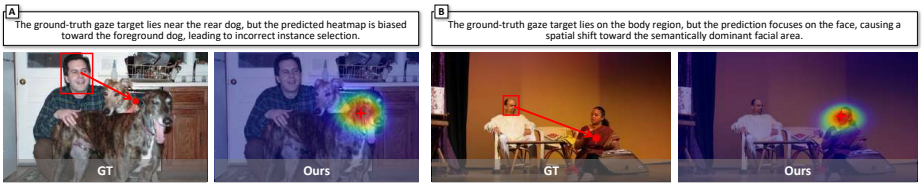

Fig. 4: Failure cases of the proposed method.

Effect of Direction Supervision Weight. Table 8 investigates the influence of the direction supervision coefficient λ_{dir} in the overall training objective (Eq. 16). Increasing the weight initially improves performance by strengthening geometric consistency during optimization. However, excessive weighting slightly harms localization accuracy due to over-constraining the learning process. The optimal performance is obtained at $\lambda_{\text{dir}} = 0.5$, demonstrating an effective balance between heatmap supervision and geometric direction learning.

Overall, the proposed framework exhibits stable performance across diverse parameter settings, indicating robustness and balanced interactions among semantic reasoning, geometric constraints, and optimization objectives. Further ablation results are provided in the *supplementary material*.

5 Limitation and Future Work

Despite the strong performance of the proposed framework, several limitations remain, mainly arising under two challenging scenarios.

As illustrated in Fig. 4(A), when multiple semantically similar objects appear in close spatial proximity, visually dominant instances may generate stronger responses despite consistent geometric cues, leading to incorrect target selection. This reflects the difficulty of resolving semantic competition among candidate objects in complex scenes. As shown in Fig. 4(B), ambiguity commonly occurs in human-centered scenes where gaze targets correspond to extended regions rather than precise points. In such cases, the model tends to focus on semantically salient areas, such as faces, instead of annotated locations, revealing a mismatch between discrete supervision and continuous human attention distribution.

Future work explores adaptive object reasoning mechanisms and uncertainty-aware gaze modeling to better address semantic ambiguity and multi-object competition. Extending the framework to multi-person interaction scenarios and temporal gaze reasoning in videos also represents a promising direction.

6 Conclusion

In this work, we reformulate gaze target estimation as a hierarchical reasoning problem instead of direct pixel-level regression. The proposed framework performs object-level semantic selection together with geometry-guided spatial

localization within a unified single-modality architecture. By integrating multi-scale feature fusion and gaze geometry priors, the method achieves stable and interpretable gaze prediction in complex scenes. Extensive experiments demonstrate consistent improvements over existing approaches, highlighting the effectiveness of structured semantic and geometric reasoning for gaze estimation.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 62172246, in part by Excellent Young Scientists Fund of Natural Science Foundation of Shandong Province under Grant ZR2024YQ071, in part by the Key Laboratory of Forensic Examination for Sichuan Provincial Universities under Grant 2024YB01, and in part by the Fundamental Research Funds for the Central Universities under Grant 22CX06037A.

References

1. Admoni, H., Scassellati, B.: Social eye gaze in human-robot interaction: a review. Journal of Human-Robot Interaction **6**(1), 25–63 (2017)
2. Arreghini, S., Abbate, G., Giusti, A., Paolillo, A.: Predicting the intention to interact with a service robot: the role of gaze cues. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 993–999. IEEE (2024)
3. Bao, J., Liu, B., Yu, J.: Escnet: Gaze target detection with the understanding of 3d scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14126–14135 (2022)
4. Beyan, C., Vinciarelli, A., Bue, A.D.: Co-located human–human interaction analysis using nonverbal cues: A survey. ACM Computing Surveys **56**(5), 1–41 (2023)
5. Cañigueral, R., Hamilton, A.F.d.C.: The role of eye gaze during natural social interactions in typical and autistic people. Frontiers in psychology **10**, 560 (2019)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Chen, W., Chai, Y., Wu, X.J., Zhu, H., Yu, Q., Du, Z.M., Han, F., Gao, W., Zheng, C., Fan, H.: Privileged information-guided multitask mutualistic transformer for gaze prediction. IEEE Transactions on Multimedia (2025)
8. Chen, W., Xu, H., Zhu, C., Liu, X., Lu, Y., Zheng, C., Kong, J.: Gaze estimation via the joint modeling of multiple cues. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1390–1402 (2021)
9. Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., Rehg, J.M.: Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In: Proceedings of the European conference on computer vision (ECCV). pp. 383–398 (2018)
10. Chong, E., Wang, Y., Ruiz, N., Rehg, J.M.: Detecting attended visual targets in video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5396–5406 (2020)
11. Deng, T., Yang, K., Li, Y., Yan, H.: Where does the driver look? top-down-based saliency detection in a traffic driving environment. IEEE Transactions on Intelligent Transportation Systems **17**(7), 2051–2062 (2016)
12. Diederik, K.: Adam: A method for stochastic optimization. (No Title) (2014)

13. Emery, N.J.: The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience & biobehavioral reviews **24**(6), 581–604 (2000)
14. Fan, L., Chen, Y., Wei, P., Wang, W., Zhu, S.C.: Inferring shared attention in social scene videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6460–6468 (2018)
15. Fan, L., Wang, W., Huang, S., Tang, X., Zhu, S.C.: Understanding human gaze communication by spatio-temporal graph reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5724–5733 (2019)
16. Fang, Y., Tang, J., Shen, W., Shen, W., Gu, X., Song, L., Zhai, G.: Dual attention guided gaze target detection in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11390–11399 (2021)
17. Fischer, T., Chang, H.J., Demiris, Y.: Rt-gene: Real-time eye gaze estimation in natural environments. In: Proceedings of the European conference on computer vision (ECCV). pp. 334–352 (2018)
18. Frischen, A., Bayliss, A.P., Tipper, S.P.: Gaze cueing of attention: visual attention, social cognition, and individual differences. Psychological bulletin **133**(4), 694 (2007)
19. Guan, J., Yin, L., Sun, J., Qi, S., Wang, X., Liao, Q.: Enhanced gaze following via object detection and human pose estimation. In: International Conference on Multimedia Modeling. pp. 502–513. Springer (2019)
20. Gupta, A., Tafasca, S., Farkhondeh, A., Vuillecard, P., Odobez, J.M.: Mtgs: A novel framework for multi-person temporal gaze following and social gaze prediction. Advances in Neural Information Processing Systems **37**, 15646–15673 (2024)
21. Gupta, A., Tafasca, S., Odobez, J.M.: A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5041–5050 (2022)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
23. Holman, B., Anwar, A., Singh, A., Tec, M., Hart, J., Stone, P.: Watch where you're going! gaze and head orientation as predictors for social robot navigation. In: 2021 IEEE international conference on robotics and automation (ICRA). pp. 3553–3559. IEEE (2021)
24. Horanyi, N., Zheng, L., Chong, E., Leonardis, A., Chang, H.J.: Where are they looking in the 3d space? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2678–2687 (2023)
25. Hu, Z., Zhao, K., Zhou, B., Guo, H., Wu, S., Yang, Y., Liu, J.: Gaze target estimation inspired by interactive attention. IEEE Transactions on Circuits and Systems for Video Technology **32**(12), 8524–8536 (2022)
26. Jin, T., Yu, Q., Zhu, S., Lin, Z., Ren, J., Zhou, Y., Song, W.: Depth-aware gaze-following via auxiliary networks for robotics. Engineering Applications of Artificial Intelligence **113**, 104924 (2022)
27. Jin, Y., Guo, G., Wang, B.: Gatector+: A unified head-free framework for gaze object and gaze following prediction. arXiv preprint arXiv:2510.25301 (2025)
28. Lan, E., Yang, Y., Zhao, C., Liu, D.: Fgi-gaze: Gaze target detection. In: Pattern Recognition and Computer Vision: 8th Chinese Conference, PRCV 2025, Shanghai, China, October 15–18, 2025, Proceedings, Part VII. p. 471. Springer Nature (2026)
29. Lian, D., Yu, Z., Gao, S.: Believe it or not, we know what you are looking at! In: Asian Conference on Computer Vision. pp. 35–50. Springer (2018)

30. Liu, C., Chen, Y., Tai, L., Ye, H., Liu, M., Shi, B.E.: A gaze model improves autonomous driving. In: Proceedings of the 11th ACM symposium on eye tracking research & applications. pp. 1–5 (2019)
31. Madhusanka, B., Ramadass, S., Rajagopal, P., Herath, H.: Biofeedback method for human–computer interaction to improve elder caring: Eye-gaze tracking. In: Predictive Modeling in Biomedical Data Mining and Analysis, pp. 137–156. Elsevier (2022)
32. Majaranta, P., Räihä, K.J., Hyrskykari, A., Špakov, O.: Eye movements and human-computer interaction. In: Eye movement research: An introduction to its scientific foundations and applications, pp. 971–1015. Springer (2019)
33. Mathew, A.M., Hermassi, H., Khalid, T., Khan, A.A., Souissi, R.: Gazevlm: A vision-language model for multi-task gaze understanding. arXiv preprint arXiv:2511.06348 (2025)
34. Mathew, A.M., Khan, A.A., Khalid, T., Al-Tam, F., Souissi, R.: Leveraging multi-modal saliency and fusion for gaze target detection. arXiv preprint arXiv:2504.19271 (2025)
35. Miao, Q., Golani, V.R., Xu, J., Dutta, P.P., Hoai, M., Samaras, D.: Multi-view gaze target estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5371–5381 (2025)
36. Miao, Q., Graikos, A., Zhang, J., Mondal, S., Hoai, M., Samaras, D.: Diffusion-refined vqa annotations for semi-supervised gaze following. In: European Conference on Computer Vision. pp. 439–457. Springer (2024)
37. Miao, Q., Hoai, M., Samaras, D.: Patch-level gaze distribution prediction for gaze following. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 880–889 (2023)
38. Moore, C., Dunham, P.J., Dunham, P.: Joint attention: Its origins and role in development. Psychology Press (2014)
39. Nieva-Suárez, Á., Marron-Romera, M., Losada-Gutiérrez, C., Guardiola-Luna, I.: Towards fusing gaze estimation and object prediction: What are you looking at? Engineering Applications of Artificial Intelligence **157**, 111113 (2025)
40. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
41. Prada, J.D.P., Lee, M.H., Song, C.: A gaze-speech system in mixed reality for human-robot interaction. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 7547–7553. IEEE (2023)
42. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
43. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? Advances in neural information processing systems **28** (2015)
44. Ryan, F., Bati, A., Lee, S., Bolya, D., Hoffman, J., Rehg, J.M.: Gaze-lle: Gaze target estimation via large-scale learned encoders. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28874–28884 (2025)
45. Saran, A., Majumdar, S., Short, E.S., Thomaz, A., Niekum, S.: Human gaze following for human-robot interaction. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8615–8621. IEEE (2018)
46. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)

47. Song, Y., Wang, X., Yao, J., Liu, W., Zhang, J., Xu, X.: Vitgaze: gaze following with interaction features in vision transformers. Visual Intelligence **2**(1), 31 (2024)
48. Tafasca, S., Gupta, A., Odobez, J.M.: Childplay: A new benchmark for understanding children's gaze behaviour. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 20935–20946 (2023)
49. Tafasca, S., Gupta, A., Odobez, J.M.: Sharingan: A transformer architecture for multi-person gaze following. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2008–2017 (2024)
50. Thoermer, C., Sodian, B.: Preverbal infants' understanding of referential gestures. First Language **21**(63), 245–264 (2001)
51. Tomas, H., Reyes, M., Dionido, R., Ty, M., Mirando, J., Casimiro, J., Atienza, R., Guinto, R.: Goo: A dataset for gaze object prediction in retail environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3125–3133 (2021)
52. Tonini, F., Beyan, C., Ricci, E.: Multimodal across domains gaze target detection. In: Proceedings of the 2022 International Conference on Multimodal Interaction. pp. 420–431 (2022)
53. Tonini, F., Dall'Asen, N., Beyan, C., Ricci, E.: Object-aware gaze target detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21860–21869 (2023)
54. Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., Shen, W.: End-to-end human-gaze-target detection with transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2192–2200. IEEE (2022)
55. Tu, D., Shen, W., Sun, W., Min, X., Zhai, G., Chen, C.: Un-gaze: A unified transformer for joint gaze-location and gaze-object detection. IEEE Transactions on Circuits and Systems for Video Technology **34**(5), 3271–3285 (2023)
56. Wang, B., Guo, C., Cui, J., Xia, H., Guo, G., Li, Z.: Vl-htr: Learning human–target representation from vision–language model. IEEE Transactions on Cybernetics (2026)
57. Wang, B., Guo, C., Jin, Y., Xia, H., Liu, N.: Transgop: Transformer-based gaze object prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 10180–10188. No. 9 (2024)
58. Wang, B., Hu, T., Li, B., Chen, X., Zhang, Z.: Gatector: A unified framework for gaze object prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19588–19597 (2022)
59. Yang, Y., Lu, F.: Gazellm: a plug-and-play zero-shot llm reasoning framework for boosting gaze target detection. Visual Intelligence **3**(1), 26 (2025)
60. Yang, Y., Yin, Y., Lu, F.: Gaze target detection by merging human attention and activity cues. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6585–6593 (2024)
61. Zhang, Q., Hu, Z., Song, Y., Pei, J., Liu, J.: The human gaze helps robots run bravely and efficiently in crowds. In: 2023 IEEE international conference on robotics and automation (ICRA). pp. 7540–7546. IEEE (2023)
62. Zhao, Y., Lei, C., Shen, Y., Du, Y., Chen, Q.: Improving autonomous vehicle visual perception by fusing human gaze and machine vision. IEEE Transactions on Intelligent Transportation Systems **24**(11), 12716–12725 (2023)