From Gaze to Meaning: A Training-Free AI Agent for Unified Grounding and Explanation

Shayan Nasiriboukani[1], Sara Atito[1,2⋆], Mohammad Nezamipour[1], and Muhammad Awais[1,2⋆]

[1] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[2] Surrey Institute for People-Centred AI, University of Surrey, UK
{s.nasiriboukani,sara.atito,muhammad.awais}@surrey.ac.uk
monezamipoor@gmail.com
https://shayan137.github.io/gta-project-page/

Abstract. Understanding human attention is fundamental for scene interpretation, yet existing approaches often rely on heavily trained models that lack interpretability. Prior methods struggle to jointly reason about gaze targets, attended objects, and visual grounding without extensive supervision. To the best of our knowledge, this work introduces the first training-free **G**aze **T**arget **A**gent (GTA) for gaze-guided reasoning across tasks such as gaze target prediction, attention localization, and object identification. This is achieved by leveraging pretrained vision–language models, augmenting them with visually guided prompts, and employing a memory-based retrieval strategy for high-uncertainty samples to improve performance without additional training. We evaluate our approach using both quantitative metrics and qualitative results. Quantitatively, our method achieves state of the art performance on the GazeFollow and GazeHOI benchmarks. Qualitatively, our agent provides detailed semantic predictions, predicts the correct targets even when ground truth labels are wrong, and remains flexible without vocabulary constraints.

Keywords: Gaze Estimation · Gaze-Guided AI Agent · vision language models · Training free · Visual Prompting · Grounding · RAG

1 Introduction

Understanding where people direct their attention is central to visual scene interpretation [11,44,45]. Humans naturally organize their perception around gaze cues [5]: we first identify who is present in the scene, then infer where each person is looking, and finally determine the object or region that is being attended to. This sequence of reasoning allows observers to interpret the focus of attention within a scene and to identify elements that are likely to be important or behaviorally relevant.

Beyond simply locating visual targets, gaze cues provide rich social information that supports higher-level reasoning about human behavior. By analyzing where individuals direct their gaze, observers can infer intentions, anticipate

⋆ Corresponding authors.

actions, and understand interactions between people and objects. Furthermore, shared or coordinated gaze among multiple individuals can reveal joint attention, which is a key signal of communication, collaboration, and social engagement [12, 51]. As a result, modeling gaze and attention has become an important component in computer vision systems that aim to interpret complex human-centered scenes.

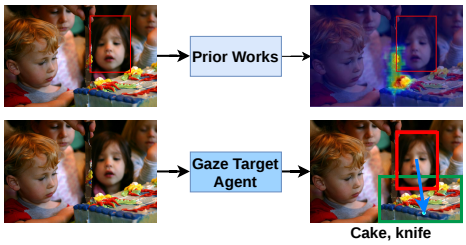

Fig. 1: Comparison with prior approaches: prior methods predict gaze heatmaps indicating where a person is looking, whereas our approach not only estimates gaze but also grounds the attended object and identifies it.

Despite the rapid progress of vision language models [2,3,13,75], most systems still reason about scenes without explicitly modeling human attention. Existing approaches typically treat gaze estimation, object grounding, and explanation as separate problems, often relying on task-specific supervision and training pipelines that do not generalize well across domains. As a result, models may produce semantically fluent descriptions that are only weakly grounded in the visual evidence, and object hallucination remains a persistent challenge [8, 28].

In this work, we investigate whether an AI agent can leverage gaze as a guiding signal for visual reasoning within a fully training-free framework, and formulate this setting as a gaze target agent. We build a simple gaze-in-the-loop pipeline(Fig. 1) in which a head detector [67] identifies people and a zero-shot gaze estimator [48] predicts where they are looking. The predicted gaze is rendered as an arrow and used as a visual prompt for a vision language model [1,3], which selects an object label from a predefined vocabulary [50] and estimates prediction uncertainty, enabling reasoning about attended objects and shared attention. When uncertainty is high, the agent retrieves visually similar examples from memory and conditions the model on them to refine its prediction, improving robustness without updating any parameters. Importantly, gaze also serves as a grounding signal by aligning predicted attention with candidate object regions, thereby reducing object hallucination and producing more robust and interpretable predictions.

We evaluate our agent on GazeFollow [44] and GazeHOI [50], achieving state of the art performance on gaze target prediction across both datasets. In addition, we compare several grounding models on GazeHOI, which provides object-level bounding boxes, to identify the most reliable grounding component for gaze-guided reasoning. Beyond quantitative results, we conduct qualitative analyses to examine the semantic capabilities of our gaze target agent. Our model produces richer semantic predictions of attended objects, correctly identifies gaze targets in cases where ground truth annotations are inaccurate, and demonstrates flexible predictions when vocabulary constraints are removed. These results highlight the ability of a training-free gaze target agent to integrate gaze estimation, ground-

ing, uncertainty handling, and reasoning within a single framework, improving both accuracy and interpretability without requiring additional supervision.

2 Related Works

Gaze Target Prediction. Early learning-based gaze methods relied on scene or activity priors that restricted where a target could be, which worked in constrained settings but struggled in open-world scenes. A shift to direct coordinate or heatmap prediction removed such assumptions and encouraged designs that combine person-specific and scene-level cues. The seminal two-branch framework of Recasens et al. on GazeFollow [44] multiplied a viewer-independent scene saliency with a head-conditioned gaze mask to yield person-specific maps, and motivated many extensions on data and fusion [11, 12, 23, 24, 45, 51, 52, 71, 72]. Building on this, multi-branch models introduced additional modalities to reduce ambiguity and improve generalization. Depth was added either via monocular prediction or sensors so that near objects in the image but far in 3D could be separated; representative methods reconstruct point clouds or derive geometric cues such as front-most surfaces and angular offsets that guide head-conditioned decoding [4,17,21,32,36,54]. Pose-aware designs complemented head appearance with body keypoints and depth, often with a human branch that predicts a 2D gaze vector and a differentiable cone prior, and a scene branch that encodes RGB, depth, and pose maps with attention-based fusion and modality dropout for robustness [20,43,69]. A related direction estimated 3D head orientation and combined a gaze cone with depth rebasing so that only geometrically consistent regions remain, which also supports in or out of frame decisions [17, 21]. To better handle extreme head poses or partial occlusion, face plus left and right eye streams preserved fine ocular detail before regressing pitch and yaw; attention and transformer-based fusion adaptively weighted facial versus ocular evidence [6,9–11,18,46,58,73]. Object-aware variants first detected heads and objects, then restricted reasoning to items that fall inside a fixed-angle field of view and biased transformer attention using cone to object alignment scores, which improved heatmaps and out-of-frame classification in clutter [23, 55, 61, 74].

End-to-End and 3D Gaze Estimation. In parallel, alternative formulations simplified or unified the pipeline. DETR-style set prediction removed upstream head detectors by jointly predicting a fixed-size set of human and gaze instances including head box, in or out decision, and heatmap, trained with Hungarian matching across boxes, classification, and regression [55,56]. A distinct track regressed 3D gaze direction without identifying a target, which is attractive for AR or driver monitoring but provides less semantics; these methods align 3D context in an egocentric frame and encode direction and distance relations among pose and objects with a transformer to refine the vector [16, 22, 26, 27, 31, 39, 41, 60]. To improve cross-domain robustness, contrastive approaches shaped feature geometry so that samples with similar gaze semantics align while identity or quality factors are suppressed; recent methods combine appearance-aware regres-

sion with language-driven differential contrast or distill toward vision–language spaces [15, 25, 64, 65, 68, 70].

Social and Multi-Person Gaze. Researchers then moved beyond single-person localization to semantic and social gaze. Mutual gaze or looking at each other was recognized by fusing temporal head pose with spatial context [14, 34]. Joint attention estimated a shared focus using interaction-aware transformers over per-person attributes together with scene-based attention maps [12,37]. Multi-person temporal frameworks went further by producing per-person heatmaps, in or out predictions, and pairwise social labels such as looking at head, looking at each other, and shared attention through people–scene and spatio–temporal interaction modules [19]. Unified token-based encoders achieved compact inference by fusing gaze tokens derived from head crops and bounding boxes with image tokens and decoding heatmaps and in or out labels in a single pass [52].

Vision Language Models for Gaze. With the rise of vision language models [1–3, 13, 29, 30, 33, 42, 53, 63], gaze modeling has begun to benefit from open-vocabulary and zero-shot capabilities. These models enable richer multimodal reasoning by aligning visual representations with natural language, allowing systems to interpret attention in a more flexible and semantically meaningful way. As a result, gaze prediction can move beyond predefined object categories and instead relate gaze targets to descriptive concepts expressed in language.

Semantic Gaze Understanding. Several works have attempted to integrate gaze estimation with broader visual understanding pipelines. Some approaches combine gaze transformers with image–text matching or object detection frameworks to associate predicted gaze with candidate objects in the scene. Other methods adopt unified architectures that jointly predict both the gaze location and the identity of the attended object within a single model [50, 57]. More recently, GazeLLM [66] introduces a zero-shot reasoning framework that leverages large language models to infer gaze targets through spatial–semantic reasoning over structured scene representations. Despite these advances, existing approaches remain constrained to predefined object categories or depend on intermediate object detection pipelines, limiting their ability to leverage open-vocabulary reasoning. As a result, several challenges in multimodal gaze-based scene understanding remain unresolved.

3 Architecture

Our goal is to develop a fully training-free agent that leverages human gaze as a guiding signal for attention-aware and open-vocabulary reasoning. As illustrated in Figure 2, the proposed framework integrates pretrained components into a unified pipeline that jointly performs gaze estimation, object reasoning, and grounding. Given an input image, the agent first detects individuals

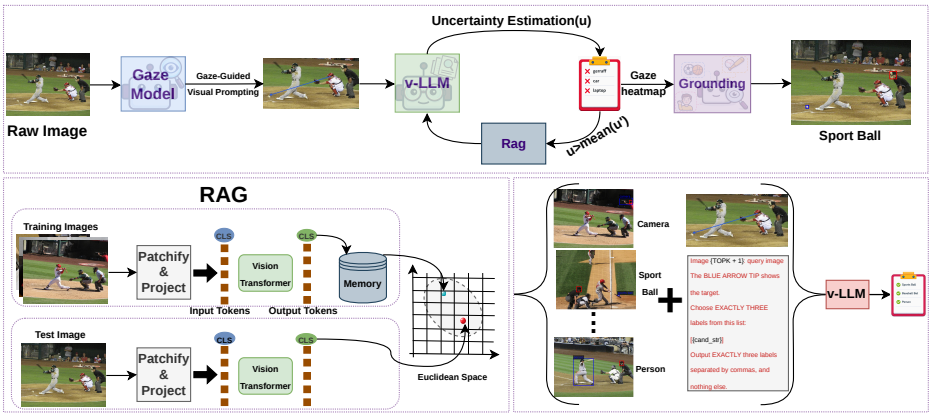

Fig. 2: Overview of the proposed gaze-target reasoning agent. Given a raw image, the agent first detects head bounding boxes and estimates the gaze direction of each person to infer where they are looking. The predicted gaze is converted into a visual prompt that highlights the attended region and is provided to a vision language model (VLM). The VLM predicts the object being observed and outputs an uncertainty score. When the prediction uncertainty is high, a retrieval augmented generation (RAG) module retrieves visually similar images and their labels from a memory bank and feeds them, together with the query image, back to the VLM to improve reasoning. Finally, a grounding module localizes the predicted object in the image.

and estimates their gaze, which serves as an explicit attentional cue for subsequent reasoning. These gaze signals guide a vision–language model to identify attended objects and interpret scene interaction, while a grounding module localizes relevant regions in the image. To further support reasoning without additional training, the framework incorporates a lightweight memory mechanism that stores visual representations from the training data. During inference, the agent evaluates the uncertainty of its predictions, and when the uncertainty exceeds a predefined threshold, the system retrieves the most relevant examples from this memory to provide contextual guidance for interpreting the scene. This retrieval-based refinement helps the model resolve ambiguous gaze predictions by leveraging similar visual contexts.

3.1 Gaze-Guided Visual Encoding

The first stage of our gaze target agent(GTA) extracts person-specific attention signals from the input image. Given an RGB image, we begin by detecting all visible heads using a pretrained YOLOv8 detector [67] without additional fine-tuning. Accurate head localization is essential, as gaze estimation is conditioned on the detected head regions. We adopt a head detector rather than a face detector because it is more robust to profile views, back views, and partial occlusions, enabling more consistent subject identification in unconstrained

multi-person scenes. For each detected head, we estimate gaze using the zero-shot Gaze-LLE model [48]. A frozen DINOv2 encoder [40] processes the image to produce a shared scene representation that is reused for all individuals, improving efficiency in crowded scenes. The head bounding box is converted into a binary mask and passed to a lightweight decoder composed of ViT blocks [59], which attends to the shared features and predicts both a fixation heatmap and an in-frame or out-of-frame decision. When the predicted fixation lies within the image, the heatmap is reduced to a single coordinate representing the estimated point of gaze.

To integrate gaze into subsequent reasoning without modifying the vision language model, we convert each predicted fixation into an explicit visual cue. For every detected individual, we draw a blue arrow from the center of the head box toward the predicted fixation location. If the gaze is predicted to lie outside the frame, the arrow extends toward the nearest image boundary to indicate outward attention. Given an image containing N individuals, we generate N prompted images $\{I'_i\}_{i=1}^{N}$, where I'_i contains only the gaze arrow corresponding to person i. This allows the model to reason about each individual's gaze target independently. Unlike traditional gaze-following approaches that represent attention as heatmaps or coordinates alone [12, 44], our framework explicitly encodes gaze as structured geometric cues that can be directly interpreted by modern vision language models [1, 3, 29, 33, 53]. This design preserves interpretability while enabling gaze-aware reasoning without additional training or architectural modification.

3.2 Vision Language Reasoning and Retrieval

The final stage translates gaze-conditioned visual inputs into structured language outputs. For each prompted image, the vision language model (VLM) generates three candidate textual predictions together with their corresponding logits. These logits allow us to estimate the confidence of each prediction rather than relying solely on the decoded text. At each decoding step t, the model produces logits $\mathbf{z}_t \in \mathbb{R}^{|V_{\mathrm{VLM}}|}$ over the vocabulary V_{VLM}. These logits are converted into a probability distribution using softmax:

$$P(w_t = w \mid \mathbf{x},\, w_{1:t-1}) = \frac{\exp\left(z_t^{(w)}\right)}{\sum_{w' \in V_{\mathrm{VLM}}} \exp\left(z_t^{(w')}\right)}. \tag{1}$$

We compute the probability of the first generated token $\hat{w}_1$:

$$p_{\mathrm{first}} \;=\; P(\hat{w}_1 \mid \mathbf{x}), \tag{2}$$

The uncertainty score is then defined as $u = 1 - p_{\mathrm{first}}$, where a lower value of p_{first} indicates reduced model confidence in its initial prediction. Retrieval refinement is activated when the uncertainty exceeds the mean uncertainty $\mathbb{E}[u]$ computed over the model's previous predictions. For the training images, we

compute global representations using a pretrained Vision Transformer (ViT) encoder [42], extracting the final-layer `[CLS]` token and storing these embeddings in a memory bank. When a test sample is flagged as uncertain, the GTA activates a retrieval-based reasoning module. The gaze-conditioned test image is encoded using the same ViT encoder to obtain a query embedding. The GTA then compares this query embedding with feature embeddings stored in an external memory, which contains embeddings of gaze-conditioned training images together with their corresponding object labels. Similarity is measured by computing the Euclidean distance between the query embedding and the stored training embeddings, allowing the system to retrieve visually similar examples from the memory bank.

Based on this similarity measure, the agent retrieves the six nearest neighbors from the memory bank. Each retrieved item consists of a gaze-conditioned image and its associated object label. These examples are appended to the vision–language model prompt as in-context demonstrations, allowing the model to leverage similar previously observed attention patterns when resolving ambiguous cases. This retrieval-based conditioning enables the system to refine uncertain predictions without updating any model parameters, preserving the training-free nature of the gaze target agent. Through this uncertainty-aware and retrieval-augmented reasoning process, the agent combines gaze estimation, confidence estimation, and external memory to refine uncertain gaze-target predictions without updating any model parameters.

3.3 Gaze-Guided Object Grounding

While gaze estimation provides a fixation point indicating the direction of attention, it does not precisely localize the full spatial extent of the attended object. In cluttered scenes where multiple similar objects appear close together, a gaze prediction may indicate the correct region but not the exact instance. This is particularly important for shared attention analysis, as predicting the object category alone does not guarantee that multiple individuals are attending to the same physical object. Accurate instance level grounding is therefore necessary to ensure spatial consistency in social gaze reasoning. To localize the attended object, we incorporate a dedicated object grounding stage that generates candidate object regions in the scene. We evaluate several pretrained detection and segmentation-based grounding models, including RF-DETR [47], YOLO26 [49], and SAM3 [7]. These models produce candidate object regions in the image, either as bounding boxes or segmentation masks. Given the predicted gaze heatmap, we measure its spatial overlap with each candidate region and select the region with the highest interaction with the gaze distribution. This gaze-heatmap alignment enables the system to identify the object instance most consistent with the subject's visual attention, particularly in scenes containing multiple nearby objects. Table 1b reports the grounding performance of RF-DETR within our gaze-guided framework. Additional comparisons with other grounding models are provided in Appendix A.1 of the supplementary material.

Table 1: Performance comparison on *GazeFollow* and GazeHOI. We compare GGVL and the proposed GTA against the previous state-of-the-art (SOTA). For *GazeFollow*, we report Acc@1, Acc@3, and MultiAcc@1, while for GazeHOI we report Acc@1, Acc@3, AP@50, and mIoU. Grey marks the previous SOTA, while blue highlights GTA. Here, I and V denote image and vocabulary, respectively.

(a) *GazeFollow*. Params denotes the number of learnable parameters.

Method	Params	Input	Acc@1 ↑	Acc@3 ↑	MultiAcc@1 ↑
Tafasca et al. [50]	116M	I+V	0.447	0.642	0.516
$\text{Qwen3}_{\text{[2B (GGVL)]}}$	0	I	0.450	0.684	0.506
$\text{Qwen3}_{\text{[2B (GTA)]}}$	0	I	**0.493**	**0.728**	**0.558**

(b) GazeHOI. The † symbol indicates the previous state-of-the-art training-free result.

Method	GazeAcc ↑	Acc@1 ↑	Acc@3 ↑	AP@50 ↑	mIoU ↑
Tafasca et al.† [50]	0.652	0.306	0.463	–	–
Tafasca et al. [50]	0.723	0.583	0.706	–	–
$\text{Qwen3}_{\text{[2B (GGVL)]}}$	**0.731**	0.517	0.721	**0.192**	**0.408**
$\text{Qwen3}_{\text{[2B (GTA)]}}$	**0.731**	**0.545**	**0.745**	**0.192**	**0.408**

4 Experiments

In this section, we evaluate the proposed **G**aze **T**arget **A**gent **(GTA)**, which combines gaze-guided visual prompting with uncertainty aware reasoning and memory based retrieval. To analyze the contribution of the retrieval and uncertainty components, we ablate on a simplified variant referred to as **G**aze **G**uided **V**ision **L**anguage **(GGVL)** [38]. This variant removes the uncertainty estimation and memory-based retrieval modules and relies solely on gaze conditioned visual prompts to guide the vision language model in predicting the attended object. Since localization performance is determined by the gaze estimator, we focus our analysis on the recognition capability of the proposed agent. Additionally, we compare GTA with state-of-the-art gaze target detection method.

4.1 Quantitative Evaluation

We evaluate recognition performance using three metrics: Acc@1, Acc@3, and MultiAcc@1. Acc@1 and Acc@3 measure whether the correct object label appears within the top 1 or top 3 predicted categories, respectively. MultiAcc@1 accounts for ambiguous cases where multiple object categories may plausibly correspond to the same gaze target, in such cases, a prediction is considered correct if the top 1 prediction matches any of the plausible object labels associated with the gaze region.

GazeFollow For fair comparison with prior work, we adopt the same target vocabulary as Tafasca et al. [50], ensuring that performance differences are not influenced by discrepancies in the label space. Under this controlled setting, as shown in Table 1a, our baseline GGVL [38] already achieves a clear improvement over previous methods, increasing Top 1 accuracy by 0.3% and Top 3 accuracy by 4.2%. These gains highlight the effectiveness of directional prompting for semantic gaze recognition, which enables the vision language model to better align gaze cues with meaningful object categories. Building upon this baseline, the gaze target agent further improves recognition performance across all reported

metrics. Specifically, it achieves an additional gain of 4.3% in Top 1 accuracy and 4.4% in Top 3 accuracy compared to GGVL, while also improving multi-label accuracy by 5.2%. This consistent improvement across metrics indicates that the proposed agent framework not only enhances single instance predictions but also performs reliably in scenes containing multiple interacting subjects.

GazeHOI We further evaluate our method on *GazeHOI* (Table 1b). In addition to recognition accuracy, we also report grounding performance on this dataset, measuring how well the predicted semantic labels align with the annotated human object interaction targets. For object localization, we use the RF-DETR [47] detector to generate candidate object bounding boxes and select the bounding box whose region overlaps most with the predicted gaze heatmap, treating it as the predicted gaze target object.

Even when compared against fine-tuned methods, our approach remains competitive. Notably, GTA achieves a 3.9% higher Top 3 accuracy than the baseline, further demonstrating the effectiveness of our design. These results highlight the effectiveness of the GTA framework as a unified zero-shot system. The gains mainly come from stronger semantic reasoning over the gaze prior, rather than task-specific training or changes to the localization mechanism.

4.2 Ablation Study

We conduct ablation studies to analyze the main components of the proposed gaze target agent. We first evaluate whether GTA generalizes across different VLM backbones and model scales. For the initial ablation setting, we use a blue arrow visual prompt with a width of 4 px, retrieve the top 6 nearest examples, and apply dynamic uncertainty thresholding. We then study uncertainty threshold, visual prompt design, retrieval size, open-vocabulary recognition, and computational cost. Except for the backbone and scale analysis, all ablations are conducted using Qwen3-VL 2B on the GazeFollow dataset. The best-performing configuration identified from these ablations is used for the final GTA result reported in Table 1a.

Generalization Across VLM Backbones and Model Scales. To assess whether the proposed agent-based reasoning is robust across different model families and capacities, we compare the baseline GGVL framework with GTA under identical settings, as reported in Table 2. In this experiment, the gaze localization module remains fixed, allowing us to isolate the effect of the recognition and reasoning components. Our evaluation includes both public and private VLMs. We evaluate both public VLMs, including LLaVA and Qwen3 at different scales, and private VLMs, including Gemini 2.5 Flash and Pro.

Across all public VLMs and parameter sizes, GTA consistently improves over the GGVL baseline on both GazeFollow and GazeHOI. The gains are stable from smaller to larger models, indicating that the benefits of structured reasoning are not dependent on model scale. Notably, LLaVA shows a particularly strong boost when moving from GGVL to GTA in both the 4B and 8B configurations,

Table 2: Ablation study on *GazeFollow* and GazeHOI comparing GGVL and the proposed GTA across different public and private VLM backbones. Results are reported under identical settings to show the effect of reasoning across model scales and architectures.

Method	GazeHOI		GazeFollow		
	Acc@1 ↑	Acc@3 ↑	Acc@1 ↑	Acc@3 ↑	MultiAcc@1 ↑
Public					
Qwen3$_{[2b\ (GGVL)]}$	0.517	0.721	0.450	0.684	0.506
Qwen3$_{[2b\ (GTA)]}$	**0.545**	**0.745**	**0.488**	**0.717**	**0.550**
Llava 1.5$_{[4b\ (GGVL)]}$	0.528	0.691	0.380	0.630	0.437
Llava 1.5$_{[4b\ (GTA)]}$	**0.572**	0.750	0.466	0.700	0.535
Qwen3$_{[4b\ (GGVL)]}$	0.516	0.742	0.543	0.722	0.613
Qwen3$_{[4b\ (GTA)]}$	0.564	**0.758**	**0.559**	**0.769**	**0.636**
Llava$_{[8b\ (GGVL)]}$	0.548	0.718	0.409	0.676	0.468
Llava$_{[8b\ (GTA)]}$	0.557	0.771	0.484	0.726	0.553
Qwen3$_{[8b\ (GGVL)]}$	0.533	0.742	0.584	0.755	0.650
Qwen3$_{[8b\ (GTA)]}$	**0.602**	**0.772**	**0.602**	**0.778**	**0.668**
Private					
Gemini 2.5 Flash$_{[GGVL]}$	0.617	0.741	0.594	0.773	0.655
Gemini 2.5 Flash$_{[GTA]}$	**0.621**	**0.753**	**0.622**	**0.789**	**0.679**
Gemini 2.5 pro$_{[GGVL]}$	0.625	0.728	0.584	0.767	0.649
Gemini 2.5 pro$_{[GTA]}$	0.631	0.739	0.608	0.787	0.665

with clear improvements across Top 1, Top 3, and multi-label accuracy metrics. Gemini results further show that the gains are not limited to open-source models. Overall, the ablation study shows that the observed performance gains are consistent across architectures, parameter scales, and both public and private VLMs, validating the generality of the proposed gaze target agent.

Uncertainty Threshold The results in Table 3 show that, excluding the retrieve-all setting at threshold 0, performance remains stable across different uncertainty thresholds. The weaker performance of the retrieve-all setting suggests that applying retrieval indiscriminately can introduce noisy or confusable candidates for samples that the model already predicts correctly and confidently, leading the model away from its correct prediction. The dynamic threshold achieves the best top 1 and multi-label accuracy, while the fixed threshold of 25 gives the best top 3 and closely matches the dynamic setting across all metrics. This is because the dynamic threshold converges to an optimal value close to 25 after around 20–30 steps, which explains the similar performance of the fixed 25 threshold.

Table 3: Uncertainty threshold ablation. We compare fixed thresholds with the dynamic uncertainty threshold used in GTA.

Threshold	Acc@1 ↑	Acc@3 ↑	MultiAcc@1 ↑
0	42.64	71.25	49.23
15	47.85	<u>71.77</u>	54.33
25	<u>48.49</u>	**71.83**	<u>54.77</u>
35	48.47	71.54	54.75
45	47.76	70.85	53.97
55	47.09	70.03	53.09
Dynamic	**48.80**	71.70	**55.50**

Different Visual Prompt Designs. Different visual prompt designs have a clear impact on gaze-target recognition, as shown in Table 4. The head bounding box performs poorly with only 28.69 top-1 accuracy, indicating that head location alone is insufficient without gaze direction. Non-arrow prompts such as the blue dot, grayscale masking, and blur masking improve performance, but remain below arrow-based prompts. This shows that explicitly encoding gaze direction is more effective. Among single-arrow prompts, the blue arrow achieves the best result with 48.80 top-1 accuracy, slightly outperforming the red arrow, while the green arrow performs lower. We further find that arrow width has a smaller but noticeable effect: a 5px blue arrow achieves the best overall performance with 49.29 top-1 accuracy and 55.79 multi-label accuracy. Overall, the 5px blue arrow provides the best balance between visibility and limited occlusion.

Retrieval-Size Ablations The retrieval-size ablation in Table 5 shows that $k = 6$ gives the best overall performance. Retrieval helps the model in two ways: it provides relevant examples for uncertain cases and reduces the effective search space from the full vocabulary of more than 300 fixed labels. This makes the reasoning process more focused and prevents the VLM from being overwhelmed by many possible object categories. Smaller values of k provide less context, while larger values may introduce less relevant examples and slightly reduce performance.

Open-Vocabulary Evaluation. We evaluate GTA in both fixed and open vocabulary settings. The open-vocabulary setting removes the predefined label constraint and evaluates whether the model can generate meaningful free-form predictions. As shown in Table 6, GTA improves performance in both fixed and open-vocabulary settings, with larger gains in the open setting. In the fixed setting, GTA improves Top-1 by 3.54 points, Top-3 by 3.03 points, and MultiAcc@1 by 4.31 points. In contrast, in the open-vocabulary setting, GTA improves Top-1 by 9.79 points, Top-3 by 10.52 points, and MultiAcc@1 by 10.48 points. This shows that agent-based reasoning and retrieval refinement are especially effective when predictions are not restricted to a predefined label set.

Table 4: Visual prompt ablation. We compare different prompt types, arrow colors, and arrow widths. The color arrows without a stated width use 4 px.

Prompt	**Acc@1** ↑	**Acc@3** ↑	**MultiAcc@1**↑
Head bbox	28.69	67.57	34.36
Blue dot	40.99	66.90	47.49
Gray scale	38.39	69.72	44.42
Blur	38.98	67.75	44.92
Red arrow	48.29	72.08	54.45
Green arrow	43.91	71.98	50.17
Yellow arrow	48.06	72.21	54.25
Blue arrow	48.80	71.70	**55.50**
3px blue arrow	48.41	72.25	54.83
5px blue arrow	**49.29**	72.81	**55.79**
6px blue arrow	49.10	**72.84**	55.40

Table 5: Retrieval-size ablation. We vary the number of retrieved examples used during retrieval-based refinement.

k	Acc@1 ↑	Acc@3 ↑	MultiAcc@1 ↑
3	48.68	72.75	55.02
4	48.56	72.69	54.89
5	49.21	**72.81**	55.58
6	**49.29**	**72.81**	**55.79**
10	49.04	72.27	55.37

Exact string matching can underestimate open-vocabulary performance because valid semantic variants may not exactly match the ground-truth label. For example, predictions such as "phone" and "cell phone" may refer to the same object, while labels such as "person" may be described as "man," "woman," "player," or "face." We also report E5 [62] semantic matching, where predictions are considered correct if their cosine similarity with the ground-truth label is at least 0.90. Open + GTA achieves 47.85 Top-1, while Fixed + GTA reaches 51.48, confirming that GTA produces semantically closer gaze-target predictions.

Computational Cost. We also analyze the computational cost of the proposed retrieval-based refinement. The GGVL baseline, which directly sends the gaze-guided visual prompt to the VLM, requires approximately 4.94×10^{12} FLOPs per sample. When uncertainty exceeds the threshold, GTA additionally activates retrieval refinement, adding approximately 9.73×10^{12} FLOPs for that sample. With the dynamic threshold, retrieval is triggered for about 45% of samples, which increases the average cost. A more efficient alternative is to use a fixed threshold of 55, which still preserves most of the performance gain while triggering retrieval for only 729/4781 samples, or approximately 15.2% of the test

Table 6: Vocabulary-setting ablation. We compare fixed and open-vocabulary evaluation for GGVL and GTA using both exact matching and E5 semantic matching.

Vocabulary	Acc@1 ↑	Acc@3 ↑	MultiAcc@1 ↑	E5 Acc@1 ↑	E5 Acc@3 ↑
Fixed + GGVL	45.75	69.78	51.48	48.43	72.40
Fixed + GTA	**49.29**	**72.81**	**55.79**	**51.48**	**76.29**
Open + GGVL	30.84	43.75	35.59	40.11	57.28
Open + GTA	40.63	54.27	46.07	47.85	64.62

set. In this setting, the average cost becomes

$$(1 - 0.152)\, 4.94{\times}10^{12} + 0.152\, (4.94 + 9.73){\times}10^{12} \approx 6.42{\times}10^{12}$$

FLOPs per sample. Among the 729 triggered samples, 520 were originally incorrect, showing that the uncertainty score effectively identifies challenging cases while avoiding unnecessary retrieval calls.

4.3 Qualitative Evaluation

In addition to quantitative results, we conduct qualitative analyses to examine the reasoning capabilities of the proposed agent. The first two analyses use our agent with the predefined vocabulary, showing that it can produce richer semantic descriptions of attended objects and correctly identify gaze targets in cases with incorrect ground truth annotations. The final analysis removes the vocabulary constraint and demonstrates that the agent can generate flexible open-vocabulary predictions.

Richer Semantic Predictions Figure 3a presents examples where the predicted labels provide more detailed semantic descriptions of the attended objects than the original dataset annotations. Rather than producing generic object categories, the agent often generates labels that better reflect the visual context of the scene. This demonstrates the ability of the framework to leverage the semantic knowledge of vision language models for more descriptive gaze target interpretations.

Ground Truth Annotation Errors Figure 3b shows cases where the dataset ground truth annotations appear incorrect. In these examples, the agent predicts object labels that better correspond to the visual gaze target. These cases highlight limitations in existing annotations and illustrate how the model's reasoning can reveal inconsistencies in the dataset.

Open-Vocabulary Predictions Finally, Figure 4 demonstrates the behavior of the agent when it is not restricted to a predefined vocabulary. Without label constraints, the model frequently generates more precise and context-aware descriptions of the attended objects. These results indicate that the agent is capable of leveraging open vocabulary reasoning to produce flexible and semantically rich predictions beyond the fixed categories used during quantitative evaluation.

(a) Examples of semantically richer gaze-target predictions generated by GTA.

(b) Examples with incorrect ground-truth labels, where GTA correctly predicts the attended object.

Fig. 3: Examples illustrate GTA's semantic reasoning and robustness to annotation errors. The ground-truth label and top-3 model predictions are shown for each image.

5 Conclusion

We presented **GTA**, a fully training-free framework for gaze-guided semantic reasoning that unifies gaze localization, object recognition, grounding, and uncertainty-aware refinement within a single pipeline. By converting gaze predictions into explicit visual prompts, the approach enables off-the-shelf vision language models to reason about attended objects in a transparent and interpretable manner. A key component of our design is a lightweight memory module that is activated only for uncertain samples, retrieving semantically similar gaze-conditioned examples to refine predictions through in-context conditioning without any parameter updates.

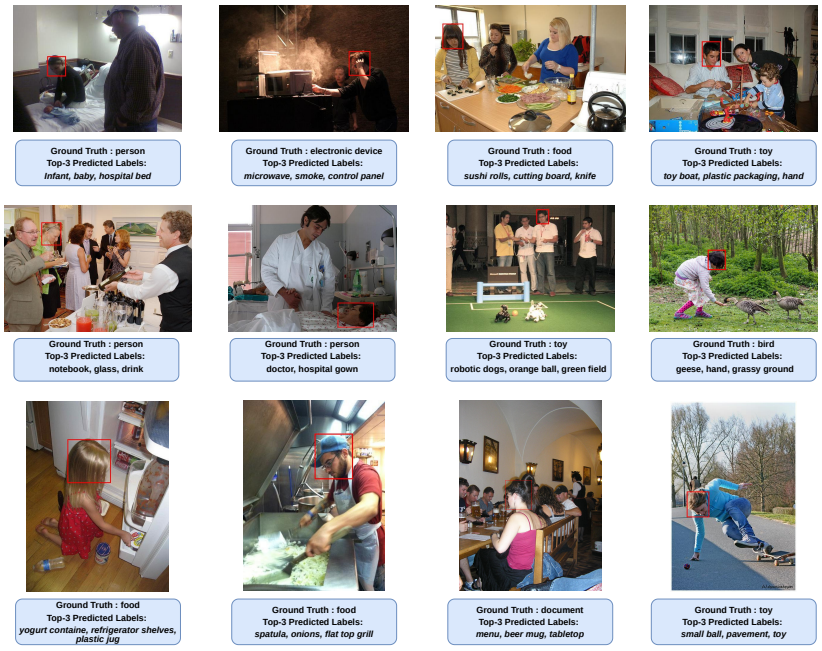

Fig. 4: Examples where GTA is not restricted to a fixed vocabulary. The predicted labels frequently provide more precise descriptions than the ground truth annotations.

Extensive experiments on *GazeFollow* and *GazeHOI* demonstrate that GTA achieves state of the art performance while remaining competitive with fine-tuned approaches. Qualitative analyses further show that GTA produces richer semantic interpretations of gaze targets, can recover correct predictions in the presence of annotation errors, and naturally supports open-vocabulary outputs when label constraints are removed. Together, these results indicate that gaze can serve as a powerful and general guiding signal for training-free visual reasoning, improving both accuracy and interpretability without additional supervision.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) through the University of Surrey Doctoral Landscape Award [grant number EP/W524463/1].

The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023] [35].

References

1. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
2. Bai, S., et al.: Qwen2.5-vl technical report (2025)
3. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-vl technical report (2025)
4. 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 (CVPR). pp. 14126–14135 (2022)
5. Birmingham, E., Bischof, W.F., Kingstone, A.: Social attention and real-world scenes: the roles of action, competition and social content. Q J Exp Psychol (Hove) **61**(7), 986–998 (2008)
6. Cai, X., Chen, B., Zeng, J., Zhang, J., Sun, Y., Wang, X., Ji, Z., Liu, X., Chen, X., Shan, S.: Gaze estimation with an ensemble of four architectures (2021)
7. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
8. Chen, X., Ma, Z., Zhang, X., Xu, S., Qian, S., Yang, J., Fouhey, D.F., Chai, J.: Multi-object hallucination in vision-language models (2024)
9. Chen, Z., Shi, B.E.: Appearance-based gaze estimation using dilated-convolutions. In: Jawahar, C., Li, H., Mori, G., Schindler, K. (eds.) Computer Vision – ACCV 2018. pp. 309–324. Springer International Publishing, Cham (2019)
10. Cheng, Y., Zhang, X., Lu, F., Sato, Y.: Gaze estimation by exploring two-eye asymmetry. IEEE Trans. Image Process. **29**, 5259–5272 (2020)
11. 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) (2018)
12. 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 (CVPR) (2020)
13. Deitke, M., Clark, C., Lee, S., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision–language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 91–104 (2025)
14. Doosti, B., Chen, C.H., Vemulapalli, R., Jia, X., Zhu, Y., Green, B.: Boosting image-based mutual gaze detection using pseudo 3d gaze. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1273–1281 (2021)
15. Du, L., Zhang, X., Lan, G.: Unsupervised gaze-aware contrastive learning with subject-specific condition (2023)
16. 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 (ICCV) (2019)
17. 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 (CVPR). pp. 11390–11399 (2021)
18. 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) (2018)

19. 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. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing systems (NeurIPS). vol. 37, pp. 15646–15673. Curran Associates, Inc. (2024)
20. 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 Workshops (CVPRW). pp. 5041–5050 (2022)
21. 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 Workshops (CVPRW). pp. 2678–2687 (2023)
22. Hu, Z., Yang, D., Cheng, S., Zhou, L., Wu, S., Liu, J.: We know where they are looking at from the rgb-d camera: Gaze following in 3d. IEEE Transactions on Instrumentation and Measurement **71**, 1–14 (2022)
23. Hu, Z., Zhao, K., Zhou, B., Guo, H., Wu, S., Yang, Y., Liu, J.: Gaze target estimation inspired by interactive attention. IEEE TCSVT **32**(12), 8524–8536 (2022)
24. 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)
25. Jindal, S., Manduchi, R.: Contrastive representation learning for gaze estimation (2022)
26. Jindal, S., Yadav, M., Manduchi, R.: Spatio-temporal attention and gaussian processes for personalized video gaze estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 604–614 (2024)
27. Kawana, Y., Shiba, S., Kong, Q., Kobori, N.: Ga3ce: Unconstrained 3d gaze estimation with gaze-aware 3d context encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3081–3090 (2025)
28. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding (2023)
29. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (ICML). pp. 19730–19742. PMLR (2023)
30. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (ICML). pp. 12888–12900. PMLR (2022)
31. Lian, D., Yu, Z., Gao, S.: Believe it or not, we know what you are looking at! In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) Computer Vision – ACCV 2018. pp. 35–50. Springer International Publishing, Cham (2019)
32. Liu, F., Li, K., Zhong, Z., Jia, W., Hu, B., Yang, X., Wang, M., Guo, D.: Depth matters: Spatial proximity-based gaze cone generation for gaze following in wild. ACM Trans. Multimedia Comput. Commun. Appl. **20**(11) (2024)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing systems (NeurIPS). vol. 36, pp. 34892–34916. Curran Associates, Inc. (2023)

34. Marin-Jimenez, M.J., Kalogeiton, V., Medina-Suarez, P., Zisserman, A.: Laeo-net: Revisiting people looking at each other in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
35. McIntosh-Smith, S., Alam, S.R., Woods, C.: Isambard-ai: a leadership class supercomputer optimised specifically for artificial intelligence (2024)
36. 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 (WACV). pp. 880–889 (2023)
37. Nakatani, C., Kawashima, H., Ukita, N.: Interaction-aware joint attention estimation using people attributes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10224–10233 (2023)
38. Nasiriboukani, S., Awais, M., Atito, S.: Gaze-guided multimodal llms for social scene understanding. In: NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning (2025)
39. Nonaka, S., Nobuhara, S., Nishino, K.: Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2192–2201 (2022)
40. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
41. Qin, J., Zhang, X., Sugano, Y.: Unigaze: Towards universal gaze estimation via large-scale pre-training. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5809–5820 (2026)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PmLR (2021)
43. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer (2020)
44. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing systems (NeurIPS). vol. 28. Curran Associates, Inc. (2015)
45. Recasens, A., Vondrick, C., Khosla, A., Torralba, A.: Following gaze in video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2017)
46. Ren, Z., Fang, F., Hou, G., Li, Z., Niu, R.: Appearance-based gaze estimation with feature fusion of multi-level information elements. Journal of Computational Design and Engineering **10**(3), 1080–1109 (2023)
47. Robinson, I., Robicheaux, P., Popov, M., Ramanan, D., Peri, N.: Rf-detr: Neural architecture search for real-time detection transformers (2026)
48. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28874–28884 (2025)
49. Sapkota, R., Cheppally, R.H., Sharda, A., Karkee, M.: Yolo26: Key architectural enhancements and performance benchmarking for real-time object detection (2026)

50. Tafasca, S., Gupta, A., Bros, V., Odobez, J.M.: Toward semantic gaze target detection. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing systems (NeurIPS). vol. 37, pp. 121422–121448. Curran Associates, Inc. (2024)
51. 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 (ICCV). pp. 20935–20946 (2023)
52. 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 (CVPR). pp. 2008–2017 (2024)
53. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
54. 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. ACM (2022)
55. 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 (ICCV). pp. 21860–21869 (2023)
56. Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., Shen, W.: End-to-end human-gaze-target detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2192–2200 (2022)
57. Tu, D., Shen, W., Sun, W., Min, X., Zhai, G.: Joint gaze-location and gaze-object detection (2023)
58. Valenti, R., Sebe, N., Gevers, T.: Combining head pose and eye location information for gaze estimation. IEEE Trans. Image Process. **21**(2), 802–815 (2012)
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need **30** (2017)
60. Vuillecard, P., Odobez, J.M.: Enhancing 3d gaze estimation in the wild using weak supervision with gaze following labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13508–13518 (2025)
61. Wang, B., Guo, C., Jin, Y., Xia, H., Liu, N.: Transgop: Transformer-based gaze object prediction. Proceedings of the AAAI Conference on Artificial Intelligence **38**(9), 10180–10188 (2024)
62. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., Wei, F.: Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533 (2022)
63. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Chen, K., Xu, B., Li, J., Dong, Y., Ding, M., Tang, J.: Cogvlm: Visual expert for pretrained language models. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing systems (NeurIPS). vol. 37, pp. 121475–121499. Curran Associates, Inc. (2024)
64. Wang, Y., Jiang, Y., Li, J., Ni, B., Dai, W., Li, C., Xiong, H., Li, T.: Contrastive regression for domain adaptation on gaze estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19376–19385 (2022)
65. Xia, L., Li, Y., Cai, X., Cui, Z., Xu, C., Chan, A.B.: Collaborative contrastive learning for cross-domain gaze estimation. PR **161**, 111244 (2025)

66. 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)
67. Yaseen, M.: What is yolov8: An in-depth exploration of the internal features of the next-generation object detector (2024)
68. Yin, P., Zeng, G., Wang, J., Xie, D.: Clip-gaze: Towards general gaze estimation via visual-linguistic model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6729–6737 (2024)
69. Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J.: Hrformer: High-resolution transformer for dense prediction (2021)
70. Zhang, L., Tian, Y., Wang, X., Xu, W., Jin, Y., Huang, Y.: Differential contrastive training for gaze estimation. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 3477–3486. MM '25, Association for Computing Machinery, New York, NY, USA (2025)
71. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)
72. Zhao, H., Lu, M., Yao, A., et al.: Learning to draw sight lines. International Journal of Computer Vision (IJCV) **128**(4), 1076–1100 (2020)
73. Zhong, Y., Lee, S.H.: Gazesymcat: A symmetric cross-attention transformer for robust gaze estimation under extreme head poses and gaze variations. Journal of Computational Design and Engineering **12**(3), 115–129 (2025)
74. Zhou, Y., Liu, L., Gou, C.: Learning from observer gaze: Zero-shot attention prediction oriented by human-object interaction recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28390–28400 (2024)
75. Zhu, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models (2025)