Gaze-Guided Multimodal LLMs for Social Scene Understanding

Shayan Nasiriboukani[1] **Muhammad Awais**[1,2] **Sara Atito**[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
sn01261@surrey.ac.uk,muhammad.awais@surrey.ac.uk,sara.atito@surrey.ac.uk

Abstract

Understanding where a person is looking is fundamental to human communication and social interaction. In computer vision, this task, known as gaze following, predicts an individual's point of focus within an image. Existing methods often estimate gaze using heatmaps or pixel coordinates, but these approaches fail to capture the semantic meaning of the gaze target, limiting their value for deeper scene understanding. We introduce GGVL (**G**aze **G**uided **V**ision **L**anguage), a zero shot framework for scene interpretation in static images. GGVL combines head detection, gaze estimation, and gaze conditioned vision language captioning. By leveraging the principle that gaze aligns with the most relevant elements of a scene, our framework generates more accurate and meaningful descriptions of what individuals are likely observing. It also produces holistic summaries of shared attention and overall scene activity, enabling richer social understanding. Comprehensive evaluation demonstrates the effectiveness of GGVL: it achieves state of the art performance on two benchmark datasets for gaze target prediction, while qualitative results show that it often recognizes gaze targets more accurately and meaningfully than ground truth labels. In a user study, participants consistently preferred the gaze guided captions produced by GGVL over those generated by baseline vision language models. These findings highlight the value of integrating gaze into vision language models to advance human centric scene understanding.

1 Introduction

Understanding where people are looking and what they are looking at is central to human computer interaction and scene understanding. Gaze offers a direct signal of visual attention that reveals intentions, interests, and social focus. In real world settings such as assistive robotics, driver monitoring, video surveillance, and social behaviour analysis, reliable estimation of gaze targets helps machines interpret actions in context. Social gaze prediction goes further by reasoning about groups. It asks who looks at what, whether people share a point of focus, and how attention shifts over time. This level of understanding enables systems to capture subtle social cues such as coordinated attention to an event or an object.

Prior work has mainly predicted where a person is looking by regressing a heatmap or a two dimensional coordinate from head and scene cues [10, 9, 5]. This localisation alone does not ensure that the attended object is understood. In crowded or safety critical scenes, recognising the identity and category of the target is often essential. Vocabulary guided methods add recognition but depend on a fixed list of categories and specialised training, which limits generalisation to open world scenarios [50]. Recent vision language models such as BLIP 2, LLaVA, and Gemini produce rich descriptions without task specific retraining, yet they often miss who is looking where or whether attention is shared when a scene contains multiple people. Unlike prior work that either (i) localizes

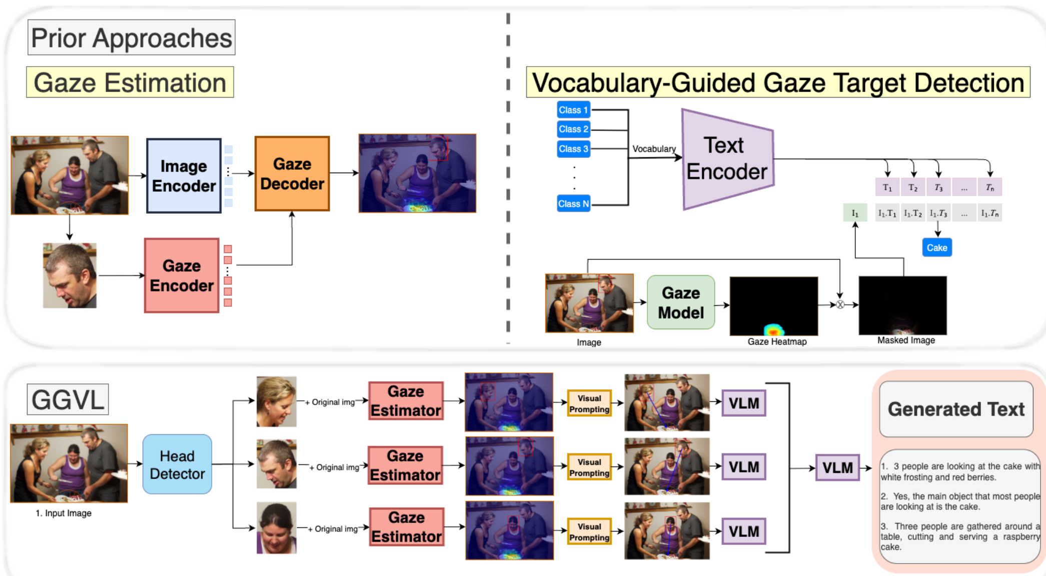

Figure 1: Comparison between prior approaches for gaze understanding at the top and the proposed GGVL framework at the bottom, which integrates gaze estimation with vision language reasoning to produce socially and contextually aware scene descriptions.

gaze with no semantics, or (ii) uses vocabulary-limited target detection, GGVL enables zero-shot, open-vocabulary gaze reasoning by explicitly conditioning VLMs on estimated gaze.

Accurate social gaze reasoning remains challenging for several reasons. First, multi person scenes introduce occlusion, small heads, and diverse viewpoints, which degrade head and eye evidence. Second, people may attend to objects that lie outside the image, or to small items that are easy to miss without explicit guidance. Third, models that rely on fixed vocabularies or heavy supervision struggle to handle the open world of objects and activities that appear in unconstrained images. These factors motivate methods that couple strong visual recognition with explicit attention cues while avoiding task specific retraining.

We address these gaps with the task of Semantic Social Gaze Understanding and with a zero shot framework called Gaze Guided Vision Language Model (GGVL). As illustrated in Fig. 1, our goal is to inject explicit gaze cues into the reasoning loop so that descriptions are grounded in who looks where and at what. The system first estimates per person gaze, then uses these signals to guide vision language inference, which enables identification of attended objects, detection of shared attention, and concise scene narratives. This design targets applications that need open vocabulary semantics without task specific retraining and is especially useful in education and assistive settings, for example when several students converge on a demonstration or when a learner's attention drifts.

Our contributions are fourfold. First, we formalise Semantic Social Gaze Understanding that jointly reasons about localisation, target identity, and shared attention in scenes with multiple people. Second, we propose a zero-shot pipeline that conditions vision language reasoning on explicit gaze cues to obtain open vocabulary targets for each person. Third, we show that grounding a vision language model with gaze improves interpretability and social awareness over point only localisation [10, 9, 5] and over vocabulary limited target detection [50]. Finally, we demonstrate that this approach preserves efficiency by reusing a single image encoding while scaling to scenes with many people, which supports practical deployment in real world environments.

2 Related Works

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 [46] multiplied a viewer-independent scene saliency with a head-conditioned gaze mask to yield person-specific maps, and it motivated many extensions on data and fusion [9, 51, 52, 27, 10, 70, 47, 25, 69]. Building on this template, 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 [2, 17, 38, 54, 23, 35]. 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 [21, 45, 66, 22]. 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 [23, 17]. 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 [71, 7, 8, 4, 6, 48, 58, 18]. 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 [25, 55, 72, 60].

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 [29, 16, 34, 59, 40, 24, 31, 43]. 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 regression with language-driven differential contrast or distill toward vision–language spaces [68, 15, 63, 62, 28, 65].

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 [13, 37]. Joint attention estimated a shared focus using interaction-aware transformers over per-person attributes together with scene-based attention maps [39, 10]. 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 [20]. 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, 53].

Vision–language models enable open-vocabulary cues and zero-shot reasoning. Foundational systems such as CLIP [44], BLIP [33], BLIP-2 [32], LLaVA [36], Gemini [19], Qwen-VL [1], CogVLM [61], GPT-4 [41], and Molmo [11] support image–text alignment, instruction following, and grounded recognition without task-specific labels. Two-stage methods integrated image–text matching or VQA with gaze transformers by injecting projected cues to modulate attention efficiently [20]. More unified approaches framed semantic gaze target detection, jointly predicting gaze location and target identity. GTR detects people and predicts per-person heatmaps, attended boxes, and categories through interacting decoders [57], while promptable gaze-following reuses a single image encoding with person-specific gaze tokens and contrastive supervision for efficient open-vocabulary grounding [50]. Promptable decoding reduces recomputation, contrastive shaping improves cross-domain robustness, and together they address efficiency and generalization. This trajectory highlights three open challenges: scalable multi-person processing, stronger cross-domain transfer, and richer gaze-grounded semantics for open-world understanding.

3 Methodology

The aim of this work is to transform human gaze into a guiding signal for open-vocabulary scene understanding. To achieve this, we introduce the Gaze Guided Vision Language (GGVL) pipeline, which operates in four consecutive stages: head detection, per-person gaze estimation, visual prompt-

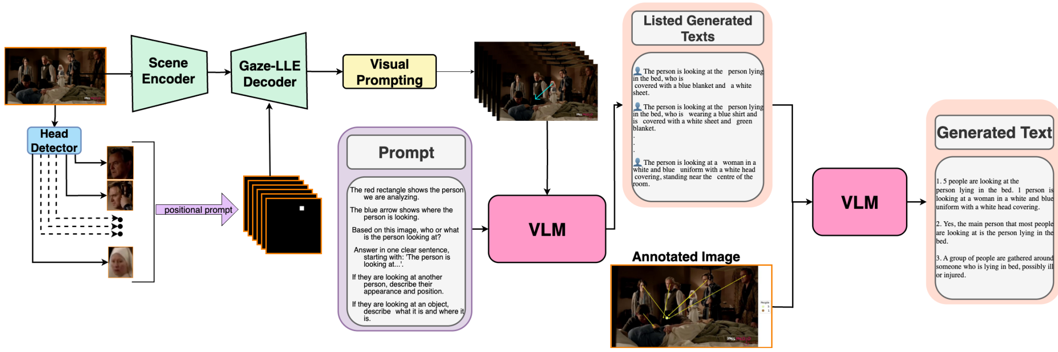

Figure 2: We introduce GGVL, a new 4 stage framework for gaze guided scene understanding. First, YOLOv8[30] detects all heads in the RGB image and extracts the corresponding crops. Next, the full image is encoded by a frozen DINOv2[42] backbone to obtain scene tokens, which are combined with the head crop features and passed to a lightweight Gaze LLE[49] decoder that predicts 2D gaze coordinates. These coordinates are overlaid on the scene to create an annotated image. Finally, the annotated image with a prompt are given to the Gemini Flash 2 [19] to produce a concise description of the scene's shared focus.

ing, and vision-language reasoning. Each stage contributes to grounding language outputs in human attentional cues, while the overall design emphasizes efficiency by encoding the scene once and reusing features across all individuals. An overview of the full pipeline is shown in Figure 2, and the complete step-by-step Algorithm is detailed in the supplementary materials.

3.1 Head Detector

The first stage of the pipeline identifies all visible heads in an input RGB image. Head localization is a critical step, as it defines the regions of interest that guide subsequent gaze prediction. We employ a pretrained YOLOv8 detector [64] without additional fine tuning. This choice reflects the robustness of modern detection architectures that generalize well to unconstrained settings where heads may be small, partially occluded, or densely clustered.

We choose a head detector rather than face detectors such as RetinaFace [12], MTCNN [67], and BlazeFace [3] because Gaze LLE [49] consumes head boxes and must handle profiles and back views where faces are not visible. In the original Gaze LLE pipeline, YOLOv5 [26] was used as the head detector. We adopt YOLOv8 [64] instead for three reasons. First, its anchor free and decoupled detection head improves recall on small and crowded heads. Second, the updated backbone and distribution based box regression provide tighter localization under occlusion. Third, the improved balance between speed and accuracy simplifies deployment in multi person scenes.

3.2 Gaze Predictor

The second stage employs a robust zero-shot gaze predictor, Gaze LLE [49], which we use without architectural changes or fine-tuning. Gaze LLE leverages a single shared scene encoding: a frozen DINOv2 encoder [42] processes the input image once to produce a feature map that is reused for all detected individuals. This design keeps computation low in multi-person scenes and supports stable zero-shot behavior. To condition the prediction on a specific subject, the detected head box is converted into a binary mask and passed to a lightweight decoder with three ViT blocks [14]. The decoder attends to the shared scene features and outputs both a fixation heatmap and an in-frame or out-of-frame decision. If the fixation lies inside the image, we convert the heatmap to a single coordinate representing the predicted point of gaze.

This stage yields one fixation per person and serves as the bridge between raw visual appearance and the higher-level gaze-aware reasoning in our pipeline. Unlike prior gaze-following methods that re-encode the scene for each subject, our approach reuses a single encoding, which improves efficiency and scalability to multi-person settings.

3.3 Visual Prompting

We translate gaze into lightweight visual prompts that a vision–language model can directly interpret. First, for each detected person we draw a blue arrow from the center of the head box to the fixation predicted by Gaze LLE [49]. For a scene with N people, we generate N prompted images $\{I_i'\}_{i=1}^{N}$, where I_i' contains only the arrow of person i. Arrows use a thin stroke and light transparency so that facial detail and small objects remain visible. If Gaze LLE predicts an out-of-frame fixation, we render a ray from the head center to the nearest image border and add a small tag labeled out-of-frame. These per-person images are later used to obtain targeted captions that specify what each subject is attending to. We adopt the blue arrow design because, as shown in the ablation study (Sec. 4.3, Tab. 3), it consistently yields the best performance across multiple vision–language models, outperforming alternative prompts such as red dots, grayscale, or overlays.

Second, we render a composite image I' that overlays all arrows to expose global attention patterns. Let $\mathcal{G} = \{\mathbf{g}_i\}$ be the set of in-frame fixation points. We cluster $\mathcal{G}$ using a radius τ and merge points within this radius. Each cluster is drawn as a small landmark at its centroid with a multiplicity badge $\times k$ that indicates how many individuals share that fixation. Arrows remain thin and landmarks small to avoid clutter while still highlighting shared attention.

These two prompting strategies provide complementary views: the per-person images support precise subject-centered reasoning, while the composite image summarizes collective focus across the scene. Unlike prior gaze-following approaches that output only heatmaps or 2D coordinates [46, 10], our method explicitly converts gaze into interpretable visual prompts. This design makes attentional cues both minimal and unambiguous, directly aligning them with the reasoning process of modern vision–language models [32, 36, 19].

3.4 Vision-Language Model

The fourth stage translates gaze-conditioned visual information into natural language descriptions. We employ a vision-language model, which receives the annotated image and produces both per-person captions and a global scene summary. The reasoning process is conducted in two passes to maximize clarity and structure.

In the first pass, the model generates exactly one sentence per detected individual, explicitly naming the object, person, or region that the subject is looking at. If the gaze falls outside the visible frame or the fixation is ambiguous, the model outputs out of frame as the description. This step ensures that each subject receives a precise and interpretable caption grounded in their attentional focus.

In the second pass, the set of individual captions is provided back to the model along with the annotated image. The model is then asked to group individuals who are attending to the same or similar targets, report the counts within each group, and compose a short summary of the overall scene. This approach provides both fine-grained and high-level understanding, capturing how attention is distributed across the scene and whether multiple people share a common focus.

Through this staged process, the GGVL pipeline achieves zero-shot gaze-aware scene understanding without any task-specific training. The design choices, namely head detection, shared encoding for gaze estimation, lightweight visual prompting, and structured vision-language reasoning, jointly enable efficient, interpretable, and grounded descriptions of complex visual environments.

4 Experiments

We evaluated GGVL through both qualitative and quantitative studies on the *GazeFollow*[46] and *GazeHOI*[50] datasets. For the qualitative analysis, we looked at visual examples showing per–person predictions, shared–attention summaries, and model explanations. We also compared our predictions with the dataset annotations. In many cases, GGVL produced labels that were more meaningful and semantically precise than the ground truth, and in some examples it even corrected mistakes in the official annotations.

Method	Learnable Params	Input	GazeFollow			Recognition		
			AUC ↑	Avg L2 ↓	Min L2 ↓	Acc@1 ↑	Acc@3 ↑	MultiAcc@1 ↑
Recasens et al.$_{[\text{NeurIPS'15}]}$ [46]	50M*	I	0.878	0.190	0.113	—	—	—
Chen et al.$_{[\text{TCSVT'21}]}$[5]	50M*	I	0.908	0.136	0.074	—	—	—
Fang et al.$_{[\text{CVPR'21}]}$[17]	68M	I+D+E	0.922	0.124	0.067	—	—	—
Bao et al.$_{[\text{CVPR'22}]}$[2]	29M*	I+D+P	0.928	0.122	—	—	—	—
Jin et al.$_{[\text{EAAI'22}]}$[27]	52M*	I+D+P	0.920	0.118	0.063	—	—	—
Hu et al.$_{[\text{TCSVT'22}]}$[25]	61M*	I+D+O	0.923	0.128	0.069	—	—	—
Gupta et al.$_{[\text{CVPR'23}]}$[23]	35M	I+D+P	<u>0.943</u>	0.114	0.056	—	—	—
Horanyi et al.$_{[\text{CVPR'22}]}$[21]	46M	I+D	0.896	0.196	0.127	—	—	—
Miao et al.$_{[\text{WACV'23}]}$[38]	61M	I+D	0.934	0.123	0.065	—	—	—
Tafasca et al.$_{[\text{ICCV'23}]}$[51]	25M*	I+D	0.939	0.122	0.062	—	—	—
Tafasca et al.$_{[\text{CVPR'24}]}$[52]	135M*	I	<u>0.944</u>	0.113	0.057	—	—	—
Tafasca et al.$_{[\text{NeurIPS'24}]}$[50]	116M	I+V	—	0.108	0.051	0.447	0.642	0.516
GGVL	0	I	**0.958**	**0.099**	**0.041**	**0.621**	**0.728**	**0.686**

Table 1: Comparison of the proposed GGVL model with state-of-the-art baselines on *GazeFollow*. The row highlighted in grey represents a model that is currently considered SOTA, while the row highlighted in green corresponds to the proposed model.

4.1 Quantitative

Our goal is to demonstrate that the proposed framework substantially improves recognition while preserving state-of-the-art localization. To ensure the highest possible localization quality, we adopt Gaze LLE [49], the current state-of-the-art gaze predictor, as the backbone for localization. Since our pipeline integrates this module without modification, all reported gains come from the recognition stage, where directional prompting and vision–language reasoning provide the improvement. Moreover, to enable a fair comparison with prior work, we evaluate our method using the same vocabulary as Tafasca et al.[50], ensuring that performance differences reflect the effectiveness of our approach rather than inconsistencies in label space.

On *GazeFollow*, refer to Table 1, the benefit of our approach is immediately visible. By keeping localization fixed and only modifying the recognition stage, GGVL lifts performance well beyond the strongest prior method of Tafasca et al. [50]. Top-1 accuracy goes up by 18%, Top-3 by almost 8%, and the multi-person metric by 17% as well. In practical terms, this means that in crowded, multi-view scenes our system can not only pinpoint where people look, but also name the correct object of attention with far higher reliability. These gains underscore the value of turning raw gaze predictions into clear, interpretable signals that a vision–language model can reason over.

On *GazeHOI*, refer to Table 2, the advantage of this design extends to a more challenging setting. In strict zero-shot evaluation, GGVL nearly doubles Top-1 accuracy compared to the baseline and improves Top-3 accuracy by more than 20 points. Interestingly, even localization accuracy goes up by about one percent, which confirms that adopting Gaze LLE as our zero-shot gaze predictor was the right design choice. Even against the fine-tuned version of Tafasca et al. [50], our method achieves slightly higher Top-1 accuracy, while being about four percents lower in Top-3 accuracy. This shows that our zero-shot approach remains competitive with task-specific training, confirming the strength of combining state-of-the-art localization with lightweight prompting and reasoning.

Method	GazeAcc ↑	Acc@1 ↑	Acc@3 ↑
Tafasca et al.$^{\dagger}_{[\text{NeurIPS'24}]}$[50]	0.652	0.306	0.463
Tafasca et al.$_{[\text{NeurIPS'24}]}$[50]	0.723	0.583	**0.706**
GGVL	**0.731**	**0.588**	0.668

Table 2: Comparison of the proposed GGVL model with the baseline method on *GazeHOI*. The baseline result Tafasca et al.[50] is shown in gray, while the proposed GGVL model is highlighted in green. Bold values indicate the best overall performance.

Taken together, these findings demonstrate that GGVL converts strong localization into genuine semantic understanding. Rather than treating gaze as a low-level coordinate, the framework elevates it into a guiding cue for recognition, delivering results that are both quantitatively stronger and qualitatively closer to how humans perceive shared attention.

4.2 Qualitative

We show that both the human evaluation (20 participants on a 1–5 scale) and the AI evaluation consistently favored GGVL over directly feeding the raw image into a VLM. The difference is evident

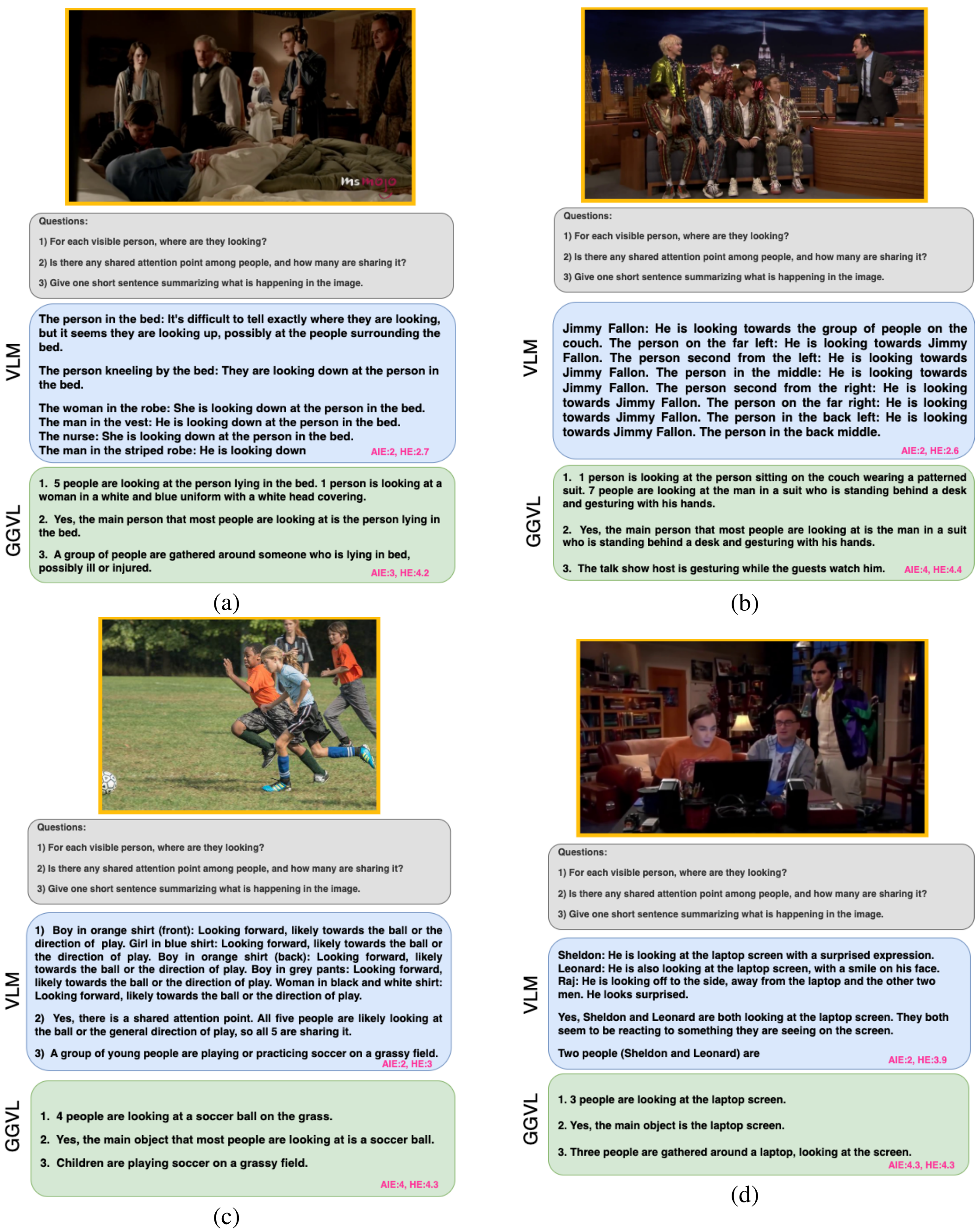

Figure 3: Qualitative evaluation sample 1. AIE: AI evaluation score; HE: Average human evaluation score (scale: 1–5, where 1=lowest quality and 5=highest quality).

in qualitative tests. For example, in Fig. 3-a and Fig. 3-b, the baseline VLM extracted fragments such as “nurse”, “bed”, or other isolated objects, but failed to form coherent or readable sentences.

GGVL, on the other hand, produced fluent, human-like descriptions with a consistent style. In some simple cases the model also preserved the style correctly, but small mistakes appeared, such as predicting five people instead of four. Fig. 4 shows that even when the ground-truth labels are not wrong, they can still be overly generic. For example, the dataset uses the label “person,” while GGVL generates more specific and natural terms such as “groom” or “bride.” Similarly, instead of “shears” the model outputs “pruning shears,” and instead of “card” it produces “credit card.” These predictions are more precise and match the way humans describe scenes.

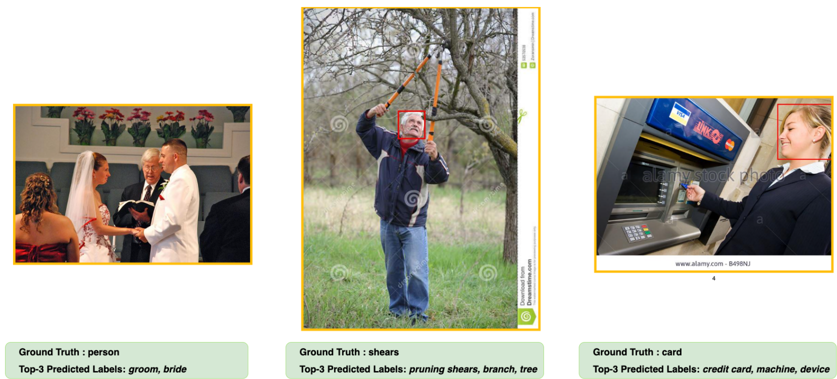

Figure 4: Examples where the ground-truth labels are technically correct but overly generic, while GGVL predictions provide more semantically precise and human-like descriptions.

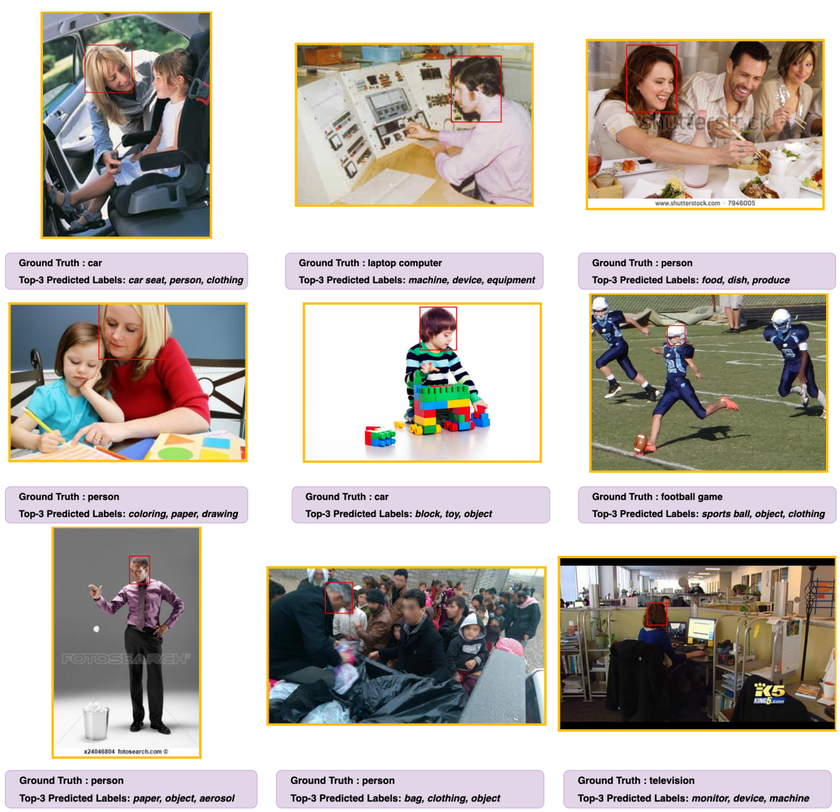

Figure 5: Examples where the ground-truth labels are incorrect, while GGVL predictions correctly capture the objects and context of the scene.

Fig. 5 highlights cases where the ground-truth labels are simply incorrect, while GGVL predictions capture the actual content of the scene. For instance, when the ground truth is "car," the model correctly identifies "car seat"; when the ground truth is "laptop computer," the model predicts "machine" or "device"; and when the ground truth is "person," the predictions include "food" or "dish," which are in fact accurate given the image. These examples demonstrate that GGVL not only improves the style of descriptions but also corrects dataset errors. To ensure a fair comparison, we used the same controlled vocabulary for both the baseline VLM and GGVL. This way, the improvements we see are due to the model's ability to generate more accurate and semantically rich descriptions, not differences in label space. Additional tests in the supplementary materials further confirm this trend, showing that GGVL consistently produces corrections that align better with human judgment than the dataset-provided labels.

4.3 Ablation Study

Table 3 presents the ablation studies on the *GazeFollow*, where we evaluated four vision language models under five different prompting strategies as shown in Fig. 6: grayscale, blur, red overlay, red dots, and blue arrow. Ablation studies on *GazeHOI* datasets is shown in the supplementary materials.

Across all models, the blue arrow emerges as the most effective and generalizable prompt. It delivers the strongest accuracy for Gemini Flash 2, Qwen2.5-VL-7B, and InternVL3-8B, and remains competitive with Molmo, which overall lags behind the other models. The arrow's advantage lies in its ability to explicitly encode both the position and the direction of gaze, while leaving the scene and attended object visible. This combination provides a clear and interpretable signal that other prompts cannot match. Other prompts perform less reliably. Grayscale occasionally improves Molmo, likely because reduced color variation simplifies its visual encoding, but it generalizes poorly across models. Red dots highlight target regions but lack directional cues, which limits their effectiveness. Blur and red overlay consistently hurt performance by obscuring important contextual information required for correct reasoning.

As for the choice of the vision language model, Gemini Flash 2 achieves the highest overall scores, which suggests that its stronger multimodal alignment particularly benefits from explicit directional information. Qwen2.5-VL-7B and InternVL3-8B also show consistent improvements with the arrow, while Molmo, despite lower absolute performance, still benefits from the arrow compared to the other prompting strategies.

VLM →	Gemini Flash 2[19]			Molmo-7B-D[11]			Qwen2.5-VL-7B[1]			InternVL3-8B[73]		
Prompt ↓	Acc@1 ↑	Acc@3 ↑	MultiAcc@1 ↑	Acc@1 ↑	Acc@3 ↑	MultiAcc@1 ↑	Acc@1 ↑	Acc@3 ↑	MultiAcc@1 ↑	Acc@1 ↑	Acc@3 ↑	MultiAcc@1 ↑
Blur	0.46	0.531	0.484	0.073	0.21	0.073	0.284	0.435	0.31	0.284	0.398	0.305
Gray	0.522	0.619	0.556	**0.21**	**0.40**	**0.246**	<u>0.375</u>	0.572	<u>0.408</u>	0.422	0.574	0.439
Red dots	<u>0.557</u>	<u>0.646</u>	<u>0.575</u>	0.12	0.33	0.16	0.272	0.564	0.302	<u>0.452</u>	<u>0.61</u>	<u>0.459</u>
Red overlay	0.522	0.592	0.544	0.14	0.326	0.163	0.354	<u>0.584</u>	0.366	0.388	0.574	0.392
Blue Arrow	**0.621**	**0.728**	**0.686**	<u>0.142</u>	<u>0.337</u>	<u>0.165</u>	**0.384**	**0.608**	**0.412**	**0.482**	**0.614**	**0.538**

Table 3: Comprehensive comparison of different visual prompts across multiple VLMs on the *GazeFollow* dataset. This unified table facilitates direct comparison of prompt effectiveness and model performance, with the highlighted row in green indicating the proposed model.

5 Conclusion

This paper presented GGVL, a zero-shot and training-free pipeline that integrates gaze estimation with vision–language reasoning to achieve socially aware scene understanding. By combining head detection, gaze localization, and gaze-guided prompting, the framework generates detailed per-person descriptions as well as high-level social scene summaries. Experiments on GazeFollow and GazeHOI showed substantial improvements over prior work, with GGVL surpassing existing methods by 18% Top-1 accuracy on GazeFollow and 28.2% on GazeHOI. Beyond numerical gains, qualitative evaluations demonstrated that the framework produces richer and more meaningful captions, even correcting annotation errors and capturing fine-grained social dynamics that baseline VLMs overlook. These findings establish gaze guidance as a powerful signal for enhancing semantic reasoning in large vision–language models, and we propose Social Scene Description as a new benchmark task for socially grounded scene understanding.

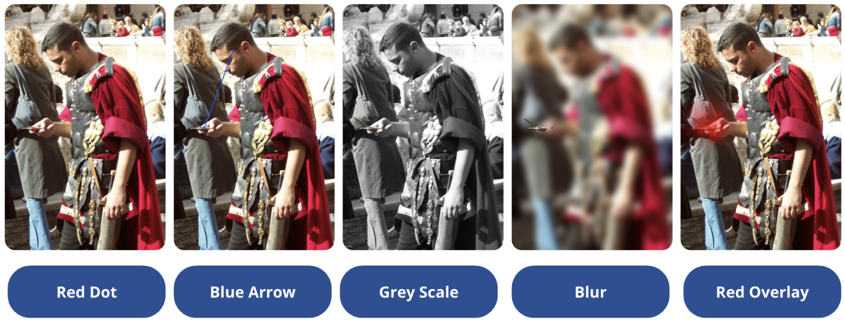

Figure 6: Examples of visual prompts: red dot for gaze target, blue arrow for gaze direction, blue grayscale for attention region, and red overlay for highlighted areas.

6 Future Work

With the GGVL pipeline in place, future work can extend its use toward building a dedicated benchmark dataset for shared attention. Such a resource would capture scenarios where multiple individuals attend to the same object or person, addressing a key limitation of existing datasets. This direction is particularly valuable, as shared gaze underpins real-world applications in social robotics, collaborative AI, and research on conditions such as autism and ADHD. Developing such a benchmark would not only enable systematic evaluation of shared attention but also accelerate progress in socially aware AI by providing models with richer training and testing environments. In addition, future extensions may focus on improving gaze localization with stronger backbones such as DINOv3 or exploring multimodal pretraining strategies, further enhancing both accuracy and generalization in complex social scenes.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

[2] Jun Bao, Buyu Liu, and Jun Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *CVPR*, pages 14126–14135, 2022.

[3] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus, 2019.

[4] Xin Cai, Boyu Chen, Jiabei Zeng, Jiajun Zhang, Yunjia Sun, Xiao Wang, Zhilong Ji, Xiao Liu, Xilin Chen, and Shiguang Shan. Gaze estimation with an ensemble of four architectures, 2021.

[5] Wenhe Chen, Hui Xu, Chao Zhu, Xiaoli Liu, Yinghua Lu, Caixia Zheng, and Jun Kong. Gaze estimation via the joint modeling of multiple cues. *IEEE TCSVT*, 32(3):1390–1402, 2022.

[6] Zhaokang Chen and Bertram E. Shi. Appearance-based gaze estimation using dilated-convolutions, 2019.

[7] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation, 2020.

[8] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE TIP*, 29:5259–5272, 2020.

[9] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency, 2018.

[10] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting attended visual targets in video, 2020.

[11] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, pages 91–104, 2025.

[12] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild, 2019.

[13] Bardia Doosti, Ching-Hui Chen, Raviteja Vemulapalli, Xuhui Jia, Yukun Zhu, and Bradley Green. Boosting image-based mutual gaze detection using pseudo 3d gaze, 2020.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[15] Lingyu Du, Xucong Zhang, and Guohao Lan. Unsupervised gaze-aware contrastive learning with subject-specific condition, 2023.

[16] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning, 2019.

[17] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *CVPR*, pages 11390–11399, 2021.

[18] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *ECCV*, 2018.

[19] Gemini Team et al. Gemini: A family of highly capable multimodal models, 2023.

[20] Anshul Gupta, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard, and Jean-Marc Odobez. A novel framework for multi-person temporal gaze following and social gaze prediction, 2024.

[21] Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings, 2023.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[23] Nora Horanyi, Linfang Zheng, Eunji Chong, Aleš Leonardis, and Hyung Jin Chang. Where are they looking in the 3d space? In *CVPRW*, pages 2678–2687, 2023.

[24] Zhengxi Hu, Dingye Yang, Shilei Cheng, Lei Zhou, Shichao Wu, and Jingtai Liu. 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.

[25] Zhengxi Hu, Kunxu Zhao, Bohan Zhou, Hang Guo, Shichao Wu, Yuxue Yang, and Jingtai Liu. Gaze target estimation inspired by interactive attention. *IEEE TCSVT*, 32(12):8524–8536, 2022.

[26] Muhammad Hussain. Yolov5, yolov8 and yolov10: The go-to detectors for real-time vision, 2024.

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

[28] Swati Jindal and Roberto Manduchi. Contrastive representation learning for gaze estimation, 2022.

[29] Swati Jindal, Mohit Yadav, and Roberto Manduchi. Spatio-temporal attention and gaussian processes for personalized video gaze estimation, 2024.

[30] Glenn Jocher, Akash Chaurasia, Jian Qian, Ryuichi Fukuda, et al. Yolov8: State-of-the-art object detection at 140 fps. `https://github.com/ultralytics/ultralytics`, 2023.

[31] Yuki Kawana, Shintaro Shiba, Quan Kong, and Norimasa Kobori. Ga3ce: Unconstrained 3d gaze estimation with gaze-aware 3d context encoding, 2025.

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[34] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at!, 2019.

[35] Feiyang Liu, Kun Li, Zhun Zhong, Wei Jia, Bin Hu, Xun Yang, Meng Wang, and Dan Guo. Depth matters: Spatial proximity-based gaze cone generation for gaze following in wild. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(11), 2024.

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[37] Manuel J. Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos, 2019.

[38] Qiaomu Miao, Minh Hoai, and Dimitris Samaras. Patch-level gaze distribution prediction for gaze following, 2022.

[39] Chihiro Nakatani, Hiroaki Kawashima, and Norimichi Ukita. Interaction-aware joint attention estimation using people attributes, 2023.

[40] Soma Nonaka, Shohei Nobuhara, and Ko Nishino. Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In *CVPR*, pages 2192–2201, 2022.

[41] OpenAI et al. Gpt-4 technical report, 2023.

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[43] Jiawei Qin, Xucong Zhang, and Yusuke Sugano. Unigaze: Towards universal gaze estimation via large-scale pre-training, 2025.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. 2021.

[45] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer, 2020.

[46] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *NeurIPS*, 2015. * indicates equal contribution.

[47] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *ICCV*, 2017.

[48] Zhonghe Ren, Fengzhou Fang, Gaofeng Hou, Zihao Li, and Rui Niu. Appearance-based gaze estimation with feature fusion of multi-level information elements. *Journal of Computational Design and Engineering*, 10(3):1080–1109, 04 2023.

[49] Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M. Rehg. Gazelle: Gaze target estimation via large-scale learned encoders. In *CVPR*, pages 28874–28884, 2025.

[50] Samy Tafasca, Anshul Gupta, Victor Bros, and Jean-Marc Odobez. Toward semantic gaze target detection. In *NeurIPS*, volume 37, pages 121422–121448. Curran Associates, Inc., 2024.

[51] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour, 2023.

[52] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Sharingan: A transformer-based architecture for gaze following, 2023.

[53] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Sharingan: A transformer architecture for multi-person gaze following. In *CVPR*, pages 2008–2017, 2024.

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

[55] Francesco Tonini, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Object-aware gaze target detection. In *ICCV*, pages 21860–21869, 2023.

[56] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. End-to-end human-gaze-target detection with transformers, 2022.

[57] Danyang Tu, Wei Shen, Wei Sun, Xiongkuo Min, and Guangtao Zhai. Joint gaze-location and gaze-object detection, 2023.

[58] Roberto Valenti, Nicu Sebe, and Theo Gevers. Combining head pose and eye location information for gaze estimation. *IEEE TIP*, 21(2):802–815, 2012.

[59] Pierre Vuillecard and Jean-Marc Odobez. Enhancing 3d gaze estimation in the wild using weak supervision with gaze following labels. In *CVPR*, pages 13508–13518, 2025.

[60] Binglu Wang, Chenxi Guo, Yang Jin, Haisheng Xia, and Nian Liu. Transgop: Transformer-based gaze object prediction. *AAAI*, 38(9):10180–10188, 2024.

[61] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2024.

[62] Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. Contrastive regression for domain adaptation on gaze estimation. In *CVPR*, pages 19376–19385, 2022.

[63] Lifan Xia, Yong Li, Xin Cai, Zhen Cui, Chunyan Xu, and Antoni B. Chan. Collaborative contrastive learning for cross-domain gaze estimation. *PR*, 161:111244, 2025.

[64] Muhammad Yaseen. What is yolov8: An in-depth exploration of the internal features of the next-generation object detector, 2024.

[65] Pengwei Yin, Guanzhong Zeng, Jingjing Wang, and Di Xie. Clip-gaze: Towards general gaze estimation via visual-linguistic model, 2024.

[66] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction, 2021.

[67] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, October 2016.

[68] Lin Zhang, Yi Tian, XiYun Wang, Wanru Xu, Yi Jin, and Yaping Huang. Differential contrastive training for gaze estimation, 2025.

[69] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 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)*, July 2017.

[70] Hang Zhao, Mengmeng Lu, Angela Yao, et al. Learning to draw sight lines. *IJCV*, 128(4):1076–1100, 2020.

[71] Yupeng Zhong and Sang Hun Lee. 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, 02 2025.

[72] Yuchen Zhou, Linkai Liu, and Chao Gou. Learning from observer gaze: Zero-shot attention prediction oriented by human-object interaction recognition. In *CVPR*, pages 28390–28400, 2024.

[73] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [NA]

   Justification: No theoretical assumptions made.

   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Code and pretrained models will be made publicly available to facilitate reproducibility and ease of use.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We use standard benchmarks with pre-defined train/val/test splits.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

   Answer: [No]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: the NeurIPS Code of Ethics are respected.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Our code builds upon publicly available repositories licensed under CC-BY 4.0. All such sources are clearly credited within the codebase wherever they are used, in accordance with the license terms.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: Code is well documented.

    Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [NA]

    Justification: Not applicable

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: LLM was only used for enhancing the writing.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.