TextGaze: Prompting Gaze Target Estimation with Textual Scene Cues

Junhui She[1,3], **Fei Wang**[2,3,*], **Kun Li**[5], **Yiqi Nie**[3,4], **Yuxin Liu**[3,4], **Zhangling Duan**[3] and **Xun Yang**[1,*]
[1]University of Science and Technology of China
[2]Hefei University of Technology
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[4] Anhui University
[5] United Arab Emirates University
{shejunhui323, ifei17.hfut, kunli.hfut, nieyiqi5, yuxinliu221}@gmail.com,
duanzl1024@iai.ustc.edu.cn, xyang21@ustc.edu.cn

Abstract

Gaze target estimation aims to infer the position of a person's gaze within a scene. Within mainstream design logic, multi-branch methods require extra supervision and annotations, while streamlined designs prioritize low-level visual saliency over true gaze intent. The former leads to a high annotation burden and hinders domain transfer, whereas the latter causes misalignment between predicted attention and actual gaze targets. To address this issue, we propose TextGaze, a unified cross-modal architecture that leverages a Large Vision-Language Model (LVLM) as scalable semantic guidance to balance the two design paradigms. The model extracts visual features from a frozen encoder and utilizes an LVLM to obtain gaze-aligned textual cues. We design a transformer-based fusion module with hierarchical text supervision to preserve task semantics. Lightweight decoding heads enable the joint prediction of gaze heatmaps and in-/out-of-frame status. We evaluate our method on four mainstream datasets, and the results show competitive performance across key metrics with robust cross-dataset generalisation without extra fine-tuning. Overall, we provide a streamlined alternative to traditional designs and highlight the potential of LVLMs as accessible auxiliary guidance for gaze estimation. All contents are available at: https://github.com/idremo/TextGaze-IJCAI2026.

1 Introduction

Gaze target estimation [Recasens *et al.*, 2017; Chong *et al.*, 2018; Miao *et al.*, 2023; Ryan *et al.*, 2025; Liu *et al.*, 2024] is a task in computer vision that aims to predict a person's gaze direction in a scene or determine if the gaze target lies within the image. As human gaze is a fundamental non-verbal cue reflecting attention allocation, intent, and cognitive state [Eckstein *et al.*, 2017], it is essential for human behavior understanding [Li *et al.*, 2023b; Li *et al.*, 2025; Li *et al.*, 2026; Wang *et al.*, 2024b], with broad applications in human-computer interaction [Katsini *et al.*, 2020; Wang *et al.*, 2026; Chen *et al.*, 2025], extended reality [Sitzmann *et al.*, 2018], and education [Wang *et al.*, 2025].

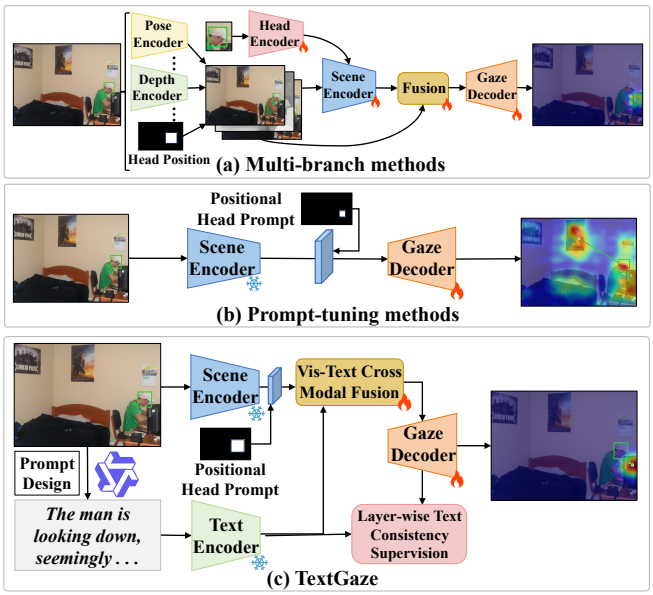

Figure 1: Pipelines for gaze target estimation. (a) Multi-branch methods [Chen *et al.*, 2021] suffer from *slow convergence* due to increased structural complexity. (b) Prompt-tuning method [Ryan *et al.*, 2025] encounter *semantic defocus* as they struggle to isolate task-relevant saliency via positional prompts. By contrast, our method (c) leverages accessible scene-level text to achieve precise semantic focus on the gaze target.

In multi-branch methods, the classical architecture is the dual-stream appearance-based design [Chong *et al.*, 2018; Lian *et al.*, 2018; Chong *et al.*, 2020; Wang *et al.*, 2024a]. In these designs, the head stream localizes the subject, and the scene stream encodes contextual visual information. Because purely visual information is insufficient to enhance effects, this paradigm has evolved into multi-branch frame-

*Corresponding authors

works that incorporate auxiliary cues. These include head features [Chen *et al.*, 2021], scene context [Saran *et al.*, 2018], depth [Bao *et al.*, 2022], and pose [Gupta *et al.*, 2022].

Although these designs enrich visual representations, they introduced the slow convergence problem [Ryan *et al.*, 2025] as adding an extra branch requires additional task-specific supervision beyond gaze annotations. As the number of branches increases, it not only increases the annotation burden and subjectivity but also complicates deployment and hinders domain transfer. More fundamentally, these systems function as post-processing geometric filters. They are capable of eliminating illogical regions but cannot reason about which objects constitute valid gaze targets. Consequently, the problem of semantic under-specification remains unresolved.

To simplify such complex architectures, recent work has introduced prompt-tuning methods based on foundation models. Gazelle [Ryan *et al.*, 2025] replaces person-specific scene re-encoding and handcrafted fusion with a unified decoder conditioned on a head position prompt, thereby addressing the need to integrate head information with global scene cues while avoiding multi-branch designs. This strategy substantially reduces architectural complexity and improves convergence behavior. However, the performance of such methods is heavily influenced by the base scene encoder and remains constrained by geometric conditions. Head position prompts also fail to provide effective information regarding task semantics. Therefore, the reasoning process is still largely influenced by low-level appearance heuristics (such as spatial proximity and visual salience), leading to semantic defocus in Figure 1, where predicted attention contradicts the subject's actual gaze intention.

Gaze following is inherently an intent-level reasoning problem rather than a purely geometric regression task. Cognitive studies [Mayrand *et al.*, 2024] show that human gaze integrates spatial cues with semantic judgments of meaningful targets and mental-state reasoning about target visibility. Yet existing models emphasize geometric signals and largely neglect task-level semantics, highlighting the need for accessible semantic priors that can guide gaze inference without reintroducing multi-branch complexity.

With the rise of large language models and LVLMs [Bai *et al.*, 2025; Wang *et al.*, 2024c], the capabilities of LVLMs are being recognized across various fields. LVLMs offer robust zero-shot capabilities for extracting contextual visual cues [Gupta *et al.*, 2024] and alleviate vision-dominated semantic underspecification by framing gaze estimation as a top-down, semantics-guided attention task. Compared to specialized, multi-branch architectures, LVLMs offer a more scalable and streamlined approach to acquiring auxiliary visual information. Their text tokens encode abstract concepts, such as objects and actions, serving as differentiable, high-level priors [Bi *et al.*, 2025]. These priors are guided by carefully designed visual prompts that specify task objectives. They constrain the solution space, improve adaptation to complex scenes, and eliminate the need for redundant branches and extra annotations. This mitigates the imbalance in the multi-branch semantic compensation cost.

Building on these insights, we introduce TextGaze, a unified trainable decoder that integrates LVLMs, scene encoding, and visual-text fusion. The model adopts a dual-stream design, shifting gaze estimation from direct geometric regression (pixels to gaze coordinates) to an intent-recognition process via a semantic layer (pixels to semantic intent to gaze target). One stream uses an LVLM with a BERT encoder to generate textual descriptions of subject behavior and plausible targets, yielding semantically enriched visual-text embeddings; the other extracts fine-grained scene cues via a frozen scene encoder and utilizes positional head prompting to enhance character positioning and facial cues. These complementary streams are fused by a Visual-Text Fusion Transformer for context-aware, cross-modal gaze estimation. To enhance model capability and robustness, we devise multiple lightweight and effective optimization strategies. Based on our design implementations, TextGaze strikes a balance between multi-branch methods and streamlined designs. And TextGaze achieves the best performance on the majority of evaluation metrics, validating the effectiveness of our fusion strategy.

In summary, the main contributions of this paper are:

- We introduce textual scene cues into gaze target estimation, providing auxiliary information to mitigate the issue of semantic defocus in the gaze decoder.
- We strengthen the GazeFollow and VideoAttentionTarget benchmarks by integrating scene-target-person attention fields with LVLM-oriented prompts; this approach generates robust contextual cues that effectively improve the accuracy of text-aided gaze estimation.
- We design a unified dual-stream cross-modal fusion framework that combines textual embeddings with fine-grained visual representations, improving spatial reasoning, scene adaptability, and mimicking human top-down attention.

2 Related Work

2.1 Gaze Target Estimation

Gaze target estimation [Recasens *et al.*, 2015; Recasens *et al.*, 2017; Lin *et al.*, 2025; Liu *et al.*, 2024; Ryan *et al.*, 2025] aims to predict the location a person is looking at in a scene image [Recasens *et al.*, 2015] or video frame [Recasens *et al.*, 2017]. The straightforward pipeline is the dual-stream approach, where one stream focuses on scene understanding and the other learns head features from cropped regions via off-the-shelf detectors [Lin *et al.*, 2025]. This design addresses the core requirements of locating individuals and learning visual scene features, but falls short of outstanding performance due to its sole reliance on visual appearance. Subsequently, multi-branch architectures are proposed to capture diverse auxiliary cues (*e.g.*, facial or head features [Chen *et al.*, 2021], scene context [Saran *et al.*, 2018], depth structures [Bao *et al.*, 2022], pose information [Gupta *et al.*, 2022], and eye [Fang *et al.*, 2021]), with multi-modal fusion becoming key for performance gains. However, complex multi-branch designs suffer from diminishing returns and high data acquisition costs, driving efforts to streamline architectures. Recently, the prompt-tuning method Gaze-LLE [Ryan *et al.*, 2025] proposed to leverage visual foundation models for gaze

prediction via positional head prompt. However, it still faces challenges in explicit visual modeling and task-relevant semantic focus. To address these issues, we propose TextGaze, a framework that leverages accessible, interpretable auxiliary semantic cues, enabling high-performance gaze target estimation through effective cross-modal fusion.

2.2 Large Vision Language Models

Large Vision Language Models (LVLMs) [Radford *et al.*, 2021; Li *et al.*, 2023a; Yang *et al.*, 2023] have become a central paradigm for multimodal visual understanding [Li *et al.*, 2023c; Li *et al.*, 2023d]. Models such as CLIP [Radford *et al.*, 2021], Claude [Priyanshu *et al.*, 2024], and GPT-4V [Yang *et al.*, 2023] exhibit strong multimodal reasoning capabilities and achieve competitive or superior performance to classical vision models in zero-shot classification. Their architectures increasingly rely on pre-trained large language models to align visual and textual representations [Grattafiori *et al.*, 2024], supported by components such as vision encoders, text encoders, decoders, and cross-attention modules. Recent advances further strengthen cross-modal fusion by unifying modalities into tokenized streams [Wang *et al.*, 2024c] or employing transfusive mechanisms [Zhou *et al.*, 2024]. In our work, LVLMs are used not as task solvers but as generators of structured visual auxiliary information for gaze target estimation. The LVLM-derived descriptions are embedded together with images into a joint representation space. A fused transformer with hierarchical fine-grained supervision is then applied to integrate these signals. This design significantly enhances cross-modal fusion and yields strong performance on gaze target estimation benchmarks.

3 Methodology

3.1 Problem Formulation and Notation

Gaze target estimation [Recasens *et al.*, 2017; Ryan *et al.*, 2025] aims to predict a heatmap that indicates the per-pixel probability of the gaze target. Formally, given an RGB image $\boldsymbol{I}_{img} \in \mathbb{R}^{H_{in} \times W_{in} \times 3}$ and the bounding box of the target person's head $\boldsymbol{x}_{bbox}$, represented by length-4 coordinates, the model predicts a 2D gaze heatmap $\boldsymbol{I}_{heat} \in [0,1]^{H \times W}$. If an in-/out-of-frame estimation task (*e.g.*, VideoAttentionTarget and ChildPlay benchmarks) is included, the model should predict a scalar probability $p_y \in [0,1]$.

3.2 Model Architecture

An architecture overview of the proposed method is shown in Figure 2. It consists of three core components: **Textual Scene Cues (§3.3)**, which extracts attention-aware textual cues from a frozen LVLM. **Feature token sequence processing (§3.4)**, responsible for enhancing visual features, embedding positional information, and constructing token sequences for fusion. **Cross-modal fusion and prediction (§3.5)**, which integrates the visual and textual sequences through fusion transformer blocks and produces a heatmap and in/out predictions. Below, we detail each component.

3.3 LVLM-based Textual Scene Cues Module

Obtaining semantically rich textual cues that align with visual focus is a key component of our design. We incorporate a frozen LVLM $\mathcal{L}$ into our design to accomplish this task. Specifically, we adopt Qwen2.5-VL-7B-Instruct in our experiments, while the framework is backbone-agnostic and compatible with any multimodal model supporting joint image-text prompting. Leveraging the zero-shot contextual reasoning and visual grounding ability of modern LVLMs, we design task-specific prompts to guide $\mathcal{L}$ in producing structured descriptions of the subject's appearance, behavior, and coarse gaze orientation. Instead of directly localizing gaze targets, which is unreliable in out-of-frame cases, this design generates context-aware auxiliary cues that remain effective across diverse scenarios.

The natural language outputs of $\mathcal{L}$ are encoded by a frozen text encoder, BERT-base-uncased in our implementation, into dense embeddings $\boldsymbol{S}_t \in \mathbb{R}^{T_t \times d_t}$, where T_t denotes the sequence length and d_t the embedding dimension. These embeddings bridge high-level linguistic context and low-level visual features, enabling effective cross-modal fusion. Overall, the prompt design and LVLM-based cues module provide a flexible and robust solution for gaze target estimation in complex settings.

3.4 Feature Token Sequence Processing Module

This module prepares text and visual features for bimodal fusion. For text processing, we augment the text feature sequence $\boldsymbol{F}_t \in \mathbb{R}^{L_t \times d}$ with learnable positional embeddings $\mathbf{PE}_{text} \in \mathbb{R}^d$, obtaining $\boldsymbol{Z}_t = \boldsymbol{F}_t + \mathbf{PE}_{text}$. For visual feature processing, we use a frozen DINOv2 encoder to extract a visual feature map $\boldsymbol{F}_v \in \mathbb{R}^{H \times W \times d}$. To enhance sensitivity to the target person's head, we incorporate head-region semantic cues using a learnable head token embedding. Given the head bounding box, we generate a downsampled binary mask $\boldsymbol{M} \in \{0,1\}^{H \times W}$ for the head region. We then add a learnable head embedding $\mathbf{E}_{head}$ to the visual features via masked modulation: $\boldsymbol{Z}_v = \boldsymbol{F}_v + \boldsymbol{M} \cdot \mathbf{E}_{head}$, where $\cdot$ denotes broadcasted element-wise multiplication. The resulting feature map is flattened into a sequence $\boldsymbol{S}_v$, and absolute 2D sinusoidal positional embeddings $\mathbf{PE}_v$ (fixed for the visual feature size) are added to form $\boldsymbol{S}_{ve} = \boldsymbol{S}_v + \mathbf{PE}_v$.

For fusion, we concatenate the text and visual sequences to construct the bimodal fusion sequence $\boldsymbol{S}_f = \mathrm{Concat}(\boldsymbol{S}_{te}, \boldsymbol{S}_{ve})$. The concatenation is operated along the token dimension. We record index boundaries to extract the visual segment for subsequent prediction. For the in-/out-of-frame classification task, a learnable token $\boldsymbol{z}_{in/out} \in \mathbb{R}^d$ is prepended, yielding $\boldsymbol{Z}_{joint} \in \mathbb{R}^{(1+HW+L_t) \times d}$:

$$\boldsymbol{Z}_{joint} = [\boldsymbol{z}_{in/out}, \underbrace{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_{HW}}_{\boldsymbol{Z}_v}, \underbrace{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_{L_t}}_{\boldsymbol{Z}_t}]. \quad (1)$$

3.5 Cross-Modal Fusion and Prediction Module

During the design of cross-modal fusion modules, we observed inherent challenges in vision-text fusion based on transformers. Such models lack explicit hierarchical constraints, leading to an excessive focus on global token interactions, resulting in representation drift across layers [Clark *et al.*, 2019; Raganato and Tiedemann, 2018]. In vision-language settings, this manifests as token-level misalignment between modalities [Lu *et al.*, 2019]. During deep fusion,

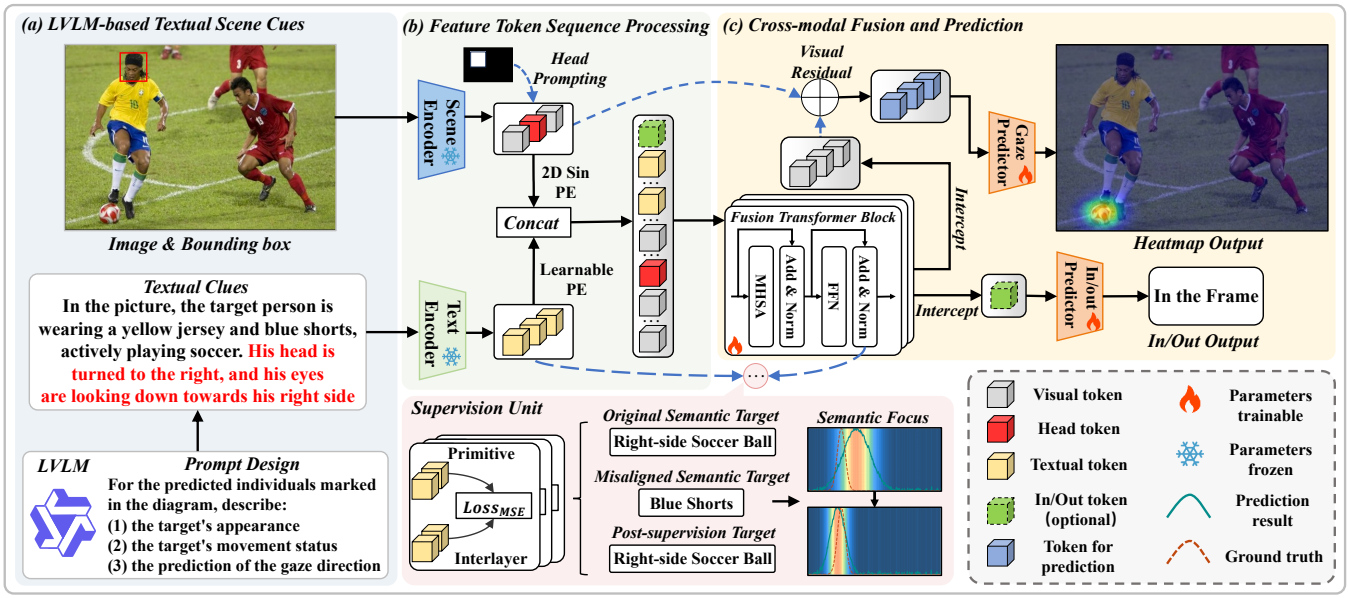

Figure 2: Overview of the proposed TextGaze. The diagram illustrates the end-to-end pipeline, including visual feature extraction via frozen DINOv2, text feature processing with pre-trained BERT, cross-modal fusion through a 3-layer fusion transformer, and final prediction heads for gaze heatmap and in/out status.

gaze-related textual semantics are gradually diluted. In the absence of intermediate supervision, this discrepancy arising from visual gaze cues and the a priori textual intent accumulates layer by layer, ultimately leading to misalignment.

To address these issues, we design a three-layer stacked fusion transformer with post-normalization. It is tailored for unified attention construction, global cross-modal interaction, and hierarchical text semantic supervision, enabling precise vision-text alignment while retaining task-critical text semantics.

The query $\boldsymbol{Q}$ and key $\boldsymbol{K}$ are derived by adding positional embeddings to the input sequence, while the value $\boldsymbol{V}$ retains the original features:

$$\boldsymbol{Q} = \boldsymbol{K} = \boldsymbol{X}^{l} + \mathbf{E}_{\text{pos}}, \quad \boldsymbol{V} = \boldsymbol{X}^{l}. \tag{2}$$

Multi-head self-attention (MHSA) follows the standard scaled dot-product formulation. Only padding masks are applied to remove invalid tokens, allowing unrestricted interaction between text and visual tokens.

Each fusion block contains two sublayers: MHSA and a feed-forward network (FFN), both equipped with residual connections and layer normalization. The MHSA sublayer output is:

$$\boldsymbol{X}'^{l} = \text{LN}\big(\boldsymbol{X}^{l} + \text{Dropout}(\text{MHSA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}))\big), \tag{3}$$

followed by the FFN sublayer:

$$\boldsymbol{X}^{l+1} = \text{LN}\big(\boldsymbol{X}'^{l} + \text{Dropout}\big(\text{FFN}(\boldsymbol{X}'^{l})\big)\big). \tag{4}$$

To safeguard task-relevant textual semantics from degradation during cross-modal fusion, we propose a hierarchical text feature consistency supervision mechanism that enforces semantic alignment across all transformer layers. For each layer $l \in \{1, 2, \ldots, L\}$, the text segment of $\boldsymbol{X}^{l}$ is extracted and aligned with pre-computed text encoder features.

Let $\boldsymbol{X}^{l}_{text} \in \mathbb{R}^{L_t \times d}$ denote the text token subset, where L_t is the number of text tokens. After permuting to $L_t \times d$, it is projected to the text hidden dimension d_{text} via a layer-specific linear projection (weight matrix $\boldsymbol{W}_l \in \mathbb{R}^{d_{text} \times d}$, bias term $\mathbf{b}_l \in \mathbb{R}^{d_{text}}$):

$$\hat{\boldsymbol{S}}^{l}_{t} = \boldsymbol{X}^{l}_{text}\boldsymbol{W}_{l}^{\top} + \mathbf{b}_l, \tag{5}$$

where $\hat{\boldsymbol{S}}^{l}_{t} \in \mathbb{R}^{L_t \times d_{text}}$ are the reconstructed text features.

The layer-wise text supervision loss is the mean squared error between $\hat{\boldsymbol{S}}^{l}_{t}$ and target text features $\boldsymbol{S}_t$:

$$\text{MSE}\Big(\hat{\boldsymbol{S}}^{l}_{t}, \boldsymbol{S}_t\Big) = \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{\boldsymbol{S}}^{l}_{t}[i] - \boldsymbol{S}_t[i]\Big)^2, \tag{6}$$

where $N = L_t \cdot d_{text}$ is the total number of elements after vectorizing the text features. The hierarchical text loss is:

$$\mathcal{L}_{text} = \frac{1}{L}\sum_{l=1}^{L}\mathcal{L}^{l}_{text}. \tag{7}$$

Following the fusion module, the visual token segment is reshaped into $\boldsymbol{F} \in \mathbb{R}^{H \times W \times d}$, where H and W denote spatial dimensions. A two-layer convolutional decoder outputs the gaze probability heatmap $\boldsymbol{I}_{heat}$, supervised by pixel-wise BCE loss with Gaussian targets (σ=3). For in-/out-of-frame classification, the task token $\boldsymbol{z}_{in/out}$ is fed to a two-layer MLP $D_{in/out}$ to predict a Bernoulli probability.

The overall objective is:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{heat}} + \lambda_1\mathcal{L}_{in/out} + \lambda_2\mathcal{L}_{text}, \tag{8}$$

where λ_1 and λ_2 balance the auxiliary losses.

Method	Input	GazeFollow			VideoAttentionTarget		
		AUC ↑	Avg L2 ↓	Min L2 ↓	AUC ↑	L2 ↓	$AP_{in/out}$ ↑
One Human	-	*0.924*	*0.096*	*0.040*	*0.921*	*0.051*	*0.093*
Recasens *et al*. [Recasens *et al.*, 2017]	I	0.878	0.190	0.113	-	-	-
Chong *et al*. [Chong *et al.*, 2018]	I	0.896	0.187	0.112	0.833	0.171	0.712
Lian *et al*. [Lian *et al.*, 2018]	I	0.906	0.145	0.081	-	-	-
Chong *et al*. [Chong *et al.*, 2020]	I	0.921	0.137	0.077	0.860	0.134	0.853
Chen *et al*. [Chen *et al.*, 2021]	I	0.908	0.136	0.074	-	-	-
Fang *et al*. [Fang *et al.*, 2021]	I+D+E	0.922	0.124	0.067	0.905	0.108	0.896
Bao *et al*. [Bao *et al.*, 2022]	I+D+P	0.928	0.122	-	0.885	0.120	0.869
Jin *et al*. [Jin *et al.*, 2022]	I+D+P	0.920	0.118	0.063	0.900	0.104	0.895
Tonini *et al*. [Tonini *et al.*, 2022]	I+D	0.927	0.141	-	0.862	0.125	0.742
Hu *et al*. [Hu *et al.*, 2022]	I+D+O	0.923	0.128	0.069	0.880	0.118	0.881
Gupta *et al*. [Gupta *et al.*, 2022]	I+D+P	0.943	0.114	0.056	0.914	0.110	0.879
Horanyi *et al*. [Horanyi *et al.*, 2023]	I+D	0.896	0.196	0.127	0.832	0.199	0.800
Miao *et al*. [Miao *et al.*, 2023]	I+D	0.934	0.123	0.065	0.917	0.109	<u>0.908</u>
Liu *et al*. [Liu *et al.*, 2024]	I+D	0.936	0.121	0.061	0.918	0.108	0.882
Tafasca *et al*. [Tafasca *et al.*, 2023]	I+D	0.939	0.122	0.062	0.914	0.109	0.834
Tafasca *et al*. [Tafasca *et al.*, 2024]	I	0.944	0.113	0.057	-	0.107	0.891
Ryan *et al*. (DINOv2 ViT-B)* [Ryan *et al.*, 2025]	I	0.955	0.106	0.048	0.928	0.110	0.877
Ryan *et al*. (DINOv2 ViT-L)* [Ryan *et al.*, 2025]	I	**0.957**	<u>0.101</u>	**0.043**	<u>0.937</u>	<u>0.103</u>	0.903
Ours (DINOv2 ViT-B)	I+T	0.953	0.105	<u>0.047</u>	0.932	0.113	0.879
Ours (DINOv2 ViT-L)	I+T	<u>0.956</u>	**0.099**	**0.043**	**0.941**	**0.102**	**0.911**

Table 1: Gaze target estimation results on GazeFollow and VideoAttentionTarget. Inputs: I (image), D (depth), E (eyes), P (pose), O (objects), and T (text). The best results are in **bold** and the second-best are <u>underlined</u>. *Results reproduced by us from the official code releases.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on four benchmark datasets: GazeFollow [Recasens *et al.*, 2015], VideoAttentionTarget [Chong *et al.*, 2020], ChildPlay [Tafasca *et al.*, 2023], and GOO-Real [Tomas *et al.*, 2021]. GazeFollow contains 117,727 training images with 124,095 annotated target persons and 4,782 test images. VideoAttentionTarget consists of 58,507 training frames with 132,563 target persons and 13,127 test frames with 31,978 target persons. ChildPlay and GOO-Real are used to assess generalization without fine-tuning, focusing on children's gaze behaviors and everyday shopping scenarios, respectively. To construct visual descriptive cues, we employ prompt engineering to guide Qwen2.5-VL-7B [Bai *et al.*, 2025] to generate appearance attributes for each image. Specifically, these attributes include (1) appearance descriptions of the target person for multi-person disambiguation; (2) action descriptions to better capture gaze behavior; and (3) gaze direction estimates without naming the target object, avoiding hallucination when the target is out of frame. Detailed implementation is given in the appendix of the supplementary material.

Evaluation Protocols. Following existing gaze target estimation settings [Chong *et al.*, 2018; Tonini *et al.*, 2022; Miao *et al.*, 2023; Ryan *et al.*, 2025], we use heatmap AUC and pixel-wise L2 distance as the primary metrics. For heatmap AUC, each pixel in the predicted heatmap is treated as a confidence score to compute the ROC curve. The pixel-wise L2 distance is defined as the Euclidean distance between the argmax location of the predicted heatmap and the ground-truth gaze target. For the GazeFollow dataset, as it has multiple annotation points, we compute the distance to the mean of annotations (Avg L2) and the distance to the nearest annotation (Min L2). For the VideoAttentionTarget and ChildPlay datasets, we also evaluate the Average Precision (AP) for the in-/out-of-frame prediction task to assess whether the gaze target lies within the visual field.

Implementation Details. The input image is resized to 448×448, while the predicted gaze heatmap has a resolution of 64×64. The scene encoder uses frozen DINOv2 [Oquab *et al.*, 2023] optimized by AdamW [Loshchilov and Hutter, 2017] optimizer with the initial learning rate of 1e-4 and weight decay of 1e-2, including ViT-B and ViT-L versions. The text encoder utilizes a frozen BERT-base-uncased model. Default loss weights are $\lambda_1 = 1$ and $\lambda_2 = 0.1$.

4.2 Performance Comparison

Comparison to State-of-the-Art. Our method's performance on the GazeFollow and VideoAttentionTarget datasets compared with existing leading approaches is summarized in Table 1. It is easy to notice that our model differs from most prior methods, which rely on additional modalities such as depth or pose; instead, we take image and text as inputs. This design originates from our thinking on balancing the acquisition of auxiliary cues and architectural complexity.

On GazeFollow, our method achieves competitive results across all metrics. With the DINOv2 ViT-L backbone, we obtain the lowest average $\mathcal{L}_2$ error and match the best minimum $\mathcal{L}_2$, while delivering a high AUC that is comparable to the state-of-the-art. Even with the lighter ViT-B backbone, our model maintains top-tier performance on key metrics. These results validate our hypothesis that the text-guided semantic priors provided by LVLMs effectively supply auxiliary cues, thereby enhancing the performance of gaze target estimation models. On VideoAttentionTarget, our DINOv2 ViT-L variant outperforms existing approaches across all three evalua-

Figure 3: Qualitative results on the GazeFollow, VideoAttention-Target (fine-tuned), and ChildPlay/GOO-Real (without fine-tuning) datasets. The input image on the left and the predicted heatmaps on the right. Our method shows consistent accuracy on different scenes.

Method	AUC ↑	L2 ↓
Chong *et al*. [Chong *et al.*, 2020]	0.670	0.334
Tonini *et al*. [Tonini *et al.*, 2022]	0.840	0.238
Miao *et al*. [Miao *et al.*, 2023]	0.869	0.202
Ryan *et al*. (DINOv2 ViT-B) [Ryan *et al.*, 2025]	0.901	<u>0.174</u>
Ryan *et al*. (DINOv2 ViT-L) [Ryan *et al.*, 2025]	0.898	0.175
Ours (DINOv2 ViT-B)	<u>0.906</u>	0.179
Ours (DINOv2 ViT-L)	**0.918**	**0.159**

Table 2: Cross-dataset results on the GOO-Real dataset. The best results are in **bold** and the second-best are underlined.

Method	AUC ↑	L2 ↓	AP ↑
Chong *et al*. [Chong *et al.*, 2020]	0.912	0.121	-
Miao *et al*. [Miao *et al.*, 2023]	0.933	0.113	-
Gupta *et al*. [Gupta *et al.*, 2024]	0.923	0.142	0.694
Tafasca *et al*. [Tafasca *et al.*, 2024]	0.932	0.115	0.600
Ryan *et al*. (ViT-B) [Ryan *et al.*, 2025]	0.941	0.119	0.993
Ryan *et al*. (ViT-L) [Ryan *et al.*, 2025]	**0.951**	**0.103**	<u>0.994</u>
Ours (DINOv2 ViT-B)	0.945	0.112	0.992
Ours (DINOv2 ViT-L)	<u>0.950</u>	<u>0.111</u>	**0.996**

Table 3: Cross-dataset results on the ChildPlay dataset. The best results are in **bold** and the second-best are underlined.

tion metrics (AUC, $\mathcal{L}_2$ error, and $\text{AP}_{\text{in/out}}$), demonstrating the value of integrating text cues for reliable gaze estimation. Our method particularly excels in the in-/out-of-frame task, highlighting its ability to determine whether a target lies within the field of view. This conclusion is further validated by cross-dataset comparison results.

It should be noted that fair comparisons with the state-of-the-art method by Ryan *et al*. [Ryan *et al.*, 2025] are conducted under matched experimental settings using DINOv2 ViT-B and DINOv2 ViT-L backbones, with their results reproduced from official code. Qualitative examples under the DINOv2 ViT-B configuration are provided in Section 4.4 to further illustrate our approach's effectiveness.

Cross-dataset Results. To evaluate the generalization capability of our approach, we conduct cross-dataset evaluation on the GOO-Real and ChildPlay datasets without modifying LVLM prompts or performing training, fine-tuning, or data augmentation. ❶ The test scenarios in the GOO-Real dataset involve gaze targets within the field of view. Table 2 compares our results with existing methods, where our approach achieves state-of-the-art performance on both AUC and L2 metrics, demonstrating strong generalization in real-world scenarios. ❷ The ChildPlay dataset includes predictions of gaze target presence within the visual field of view. Table 3 compares our results with existing methods: although our approach slightly lags behind the state of the art in AUC and L2, it outperforms all baselines on the AP metric. Combined with results from the VideoAttentionTarget, this further highlights the contribution of LVLM-derived visual descriptive cues to AP (in/out) classification. Figure 3 presents qualitative results across four datasets: GazeFollow (trained with DINOv2 ViT-B), VideoAttentionTarget (fine-tuned), and ChildPlay/GOO-Real (without fine-tuning). The left panel shows raw inputs, while the right overlays predicted heatmaps.

4.3 Ablation Study

Ablation Results of Bimodal Fusion. We explore multiple representative fusion approaches that align with our design principles to formulate our fusion strategy: additive fusion, cross-attention fusion, and the transformer-based fusion adopted in our model. Their schematic structures are illustrated in Fig. 4. Additive fusion enforces element-wise alignment between modalities and thus struggles to handle heterogeneous visual-textual representations. By comparison, cross-attention mechanisms enable directional interactions, yet remain constrained by asymmetric information flow and lack holistic cross-modal reasoning capabilities. The transformer-based fusion performs unified attention across all visual and textual tokens, allows globally symmetric interactions, and progressively optimizes gaze-related semantic information. The relevant experimental results in Table 4 demonstrate the effectiveness of this fusion strategy within our design. The ablation studies in this section are all conducted using the GazeFollow dataset. All tested models use the same DINOv2 ViT-B scene encoder and textual cues data.

Ablation Results of Key Components. Building upon the established fusion strategy, we design key components to enhance models' performance in gaze target estimation tasks. We first focus on the critical semantic guidance capability of textual cues within tasks, noting that gaze-related textual

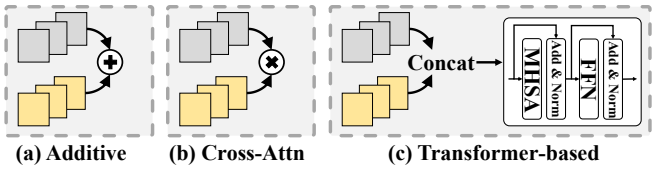

Figure 4: Schematic of fusion strategy under ablation experiments. The grey and yellow blocks represent visual tokens and textual tokens, respectively.

Bimodal Fusion Strategy	AUC ↑	Avg L2 ↓	Min L2 ↓
Additive Fusion	0.950	0.118	0.056
Cross-Attention Fusion	0.952	0.119	0.055
Transformer-based Fusion	**0.953**	**0.105**	**0.047**

Table 4: Ablation study on bimodal fusion strategies. The best results are in **bold**.

Txt Sup.	Vis Res.	Vis 2D PE	AUC ↑	Avg L2 ↓	Min L2 ↓
×	×	×	0.950	0.119	0.057
✓	×	×	0.951	0.117	0.057
✓	✓	×	0.952	0.113	0.053
✓	×	✓	0.952	0.112	0.052
✓	✓	✓	**0.953**	**0.105**	**0.047**

Table 5: Ablation study on key components under transformer-based fusion strategy. Txt Sup., Vis 2D PE, Vis 2D PE denote text supervision, visual residual, and mixed positional embedding, respectively. The best results are in **bold**.

cues may undergo semantic drift during multi-layer fusion processes. Consequently, we design an inter-layer text supervision mechanism as the foundational component and test its efficacy. We also wish to achieve spatial fidelity preservation concisely. So, we introduce pre-prediction visual residual connections to preserve the fine-scale spatial details crucial for gaze localization. Regarding positional embedding strategies for feature sequences, we attempt to incorporate explicit spatial inductive biases to model gaze-specific localization patterns. We test two approaches: modality-specific positional embedding and learnable full-sequence positional embedding. The former utilizes 2D sinusoidal positional embedding for visual features and learnable positional embedding for textual features; the latter applies learnable positional embedding to all features. Each component we designed effectively enhances the overall performance of the architecture, with relevant results shown in Table 5. Stepwise optimization experiments further validate the rationality and effectiveness of our design approach. The results further indicate that simultaneously activating these three components yields optimal performance, confirming their complementary roles in our framework design for the gaze target estimation task. We ultimately adopted the simultaneous activation scheme for our architectural implementation.

4.4 Qualitative Results

Figure 5 presents three representative scenarios from GazeFollow that illustrate the qualitative advantages of TextGaze. In the scene depicted in the first row of the image, multiple

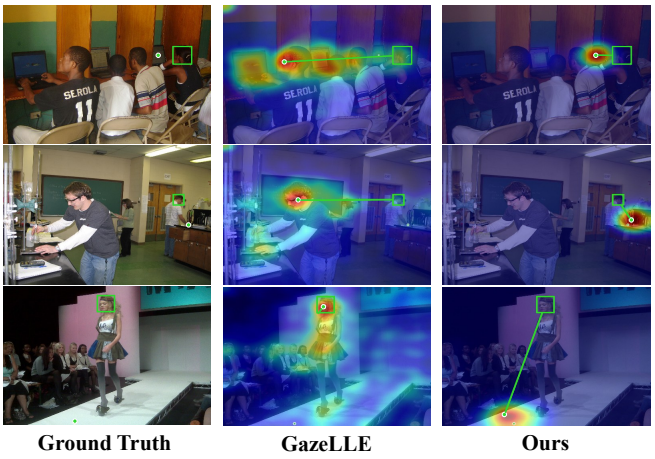

Figure 5: Qualitative comparison on the GazeFollow dataset. Compared to GazeLLE, our method demonstrates its superiority in three representative scenarios.

targets and ambiguous candidate regions are present. Purely vision-guided methods are disrupted by visually salient areas due to the absence of explicit task-oriented semantics. By incorporating textual cues from LVLM to assign subject identity and gaze intent, our model is able to avoid misjudgments. In the second typical scenario, facial cues are rendered unreliable by occlusion or low resolution. Consequently, low-level visual cues are insufficient to support the precise predictions of purely vision-guided methods. However, our approach remains reliable through LVLM-based behavioural descriptions that compensate for missing details with high-level semantic priors. Finally, the target was misclassified as being within the bounding box, a frequent error arising from its reliance on spatial proximity. Our approach prevents such errors by integrating scene context with semantic constraints. These cases validate the efficacy of our approach. Implementations that rely excessively on purely bottom-up visual understanding lack task-oriented semantics, resulting in semantic defocusing. By integrating prior knowledge guided by LVLM, we are able to shift gaze estimation from geometric regression towards a semantically driven decision process. This enables us to achieve accurate localization in complex scenes.

5 Conclusion

We address two issues in gaze target estimation: multi-branch designs require extensive supervision and annotations, causing cost imbalance and slow convergence; streamlined designs over-rely on low-level visual saliency rather than true gaze intent, leading to semantic defocus. We propose TextGaze, a unified cross-modal framework guided by an LVLM. Exploiting zero-shot capability, we produce text prompts and fuse them with frozen visual features through a dual-stream structure, enabling effective fusion while keeping the pipeline simple. We augment four datasets with text cues for future research. Experiments show strong performance and generalization, verifying our design and demonstrating LVLMs' ability to connect vision with high-level intentions.

Acknowledgements

This work was supported by the Key Science & Technology Project of Anhui Province (202304a05020068), the Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project under Grant (2025CSJZN01600), and the New Generation of Artificial Intelligence National Science and Technology Major Project (2025ZD0123303).

References

[Bai *et al.*, 2025] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

[Bi *et al.*, 2025] Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach. In *CVPR*, pages 4135–4144, 2025.

[Chen *et al.*, 2021] Wenhe Chen, Hui Xu, Chao Zhu, Xiaoli Liu, Yinghua Lu, Caixia Zheng, and Jun Kong. Gaze estimation via the joint modeling of multiple cues. *TCSVT*, 32(3):1390–1402, 2021.

[Chen *et al.*, 2025] Junjie Chen, Fei Wang, Zhihao Huang, Qing Zhou, Kun Li, Dan Guo, Linfeng Zhang, and Xun Yang. Towards seamless interaction: Causal turn-level modeling of interactive 3d conversational head dynamics. *arXiv preprint arXiv:2512.15340*, 2025.

[Chong *et al.*, 2018] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *ECCV*, pages 383–398, 2018.

[Chong *et al.*, 2020] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *CVPR*, pages 5396–5406, 2020.

[Clark *et al.*, 2019] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[Eckstein *et al.*, 2017] Maria K Eckstein, Belén Guerra-Carrillo, Alison T Miller Singley, and Silvia A Bunge. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental cognitive neuroscience*, 25:69–91, 2017.

[Fang *et al.*, 2021] 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.

[Grattafiori *et al.*, 2024] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Gupta *et al.*, 2022] Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *CVPR*, pages 5041–5050, 2022.

[Gupta *et al.*, 2024] Anshul Gupta, Pierre Vuillecard, Arya Farkhondeh, and Jean-Marc Odobez. Exploring the zero-shot capabilities of vision-language models for improving gaze following. In *CVPR*, pages 615–624, 2024.

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

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

[Jin *et al.*, 2022] Tianlei Jin, Qizhi Yu, Shiqiang Zhu, Zheyuan Lin, Jie Ren, Yuanhai Zhou, and Wei Song. Depth-aware gaze-following via auxiliary networks for robotics. *EAAI*, 113:104924, 2022.

[Katsini *et al.*, 2020] Christina Katsini, Yasmeen Abdrabou, George E Raptis, Mohamed Khamis, and Florian Alt. The role of eye gaze in security and privacy applications: Survey and future hci research directions. In *CHI*, pages 1–21, 2020.

[Li *et al.*, 2023a] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023.

[Li *et al.*, 2023b] Kun Li, Dan Guo, Guoliang Chen, Feiyang Liu, and Meng Wang. Data augmentation for human behavior analysis in multi-person conversations. In *ACM MM*, pages 9516–9520, 2023.

[Li *et al.*, 2023c] Kun Li, Dan Guo, and Meng Wang. Vigt: proposal-free video grounding with a learnable token in the transformer. *Science China Information Sciences*, 66(10):202102, 2023.

[Li *et al.*, 2023d] Kun Li, Jiaxiu Li, Dan Guo, Xun Yang, and Meng Wang. Transformer-based visual grounding with cross-modality interaction. *ACM ToMM*, 19(6):1–19, 2023.

[Li *et al.*, 2025] Kun Li, Dan Guo, Guoliang Chen, Chunxiao Fan, Jingyuan Xu, Zhiliang Wu, Hehe Fan, and Meng Wang. Prototypical calibrating ambiguous samples for micro-action recognition. In *AAAI*, volume 39, pages 4815–4823, 2025.

[Li *et al.*, 2026] Kun Li, Jihao Gu, Fei Wang, Zhiliang Wu, Hehe Fan, and Dan Guo. Ma-bench: Towards fine-grained micro-action understanding. *arXiv preprint arXiv:2603.26586*, 2026.

[Lian *et al.*, 2018] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *ACCV*, pages 35–50, 2018.

[Lin *et al.*, 2025] Zhi-Yi Lin, Jouh Yeong Chew, Jan van Gemert, and Xucong Zhang. Gazehta: End-to-end gaze

target detection with head-target association. In *ICRA*, pages 9447–9454, 2025.

[Liu *et al.*, 2024] 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 ToMM*, 20(11):1–24, 2024.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32, 2019.

[Mayrand *et al.*, 2024] Florence Mayrand, Francesca Capozzi, and Jelena Ristic. Gaze communicates both cue direction and agent mental states. *Frontiers in Psychology*, 15:1472538, 2024.

[Miao *et al.*, 2023] Qiaomu Miao, Minh Hoai, and Dimitris Samaras. Patch-level gaze distribution prediction for gaze following. In *WACV*, pages 880–889, 2023.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[Priyanshu *et al.*, 2024] Aman Priyanshu, Yash Maurya, and Zuofei Hong. Ai governance and accountability: An analysis of anthropic's claude. *arXiv preprint arXiv:2407.01557*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[Raganato and Tiedemann, 2018] Alessandro Raganato and Jörg Tiedemann. An analysis of encoder representations in transformer-based machine translation. In *EMNLP workshop*, pages 287–297, 2018.

[Recasens *et al.*, 2015] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *NeurIPS*, 28, 2015.

[Recasens *et al.*, 2017] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *ICCV*, pages 1435–1443, 2017.

[Ryan *et al.*, 2025] 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.

[Saran *et al.*, 2018] Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Human gaze following for human-robot interaction. In *IROS*, pages 8615–8621, 2018.

[Sitzmann *et al.*, 2018] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *TVCG*, 24(4):1633–1642, 2018.

[Tafasca *et al.*, 2023] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *ICCV*, pages 20935–20946, 2023.

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

[Tomas *et al.*, 2021] Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto. Goo: A dataset for gaze object prediction in retail environments. In *CVPR*, pages 3125–3133, 2021.

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

[Wang *et al.*, 2024a] Fei Wang, Dan Guo, Kun Li, and Meng Wang. Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer. In *AAAI*, volume 38, pages 5345–5353, 2024.

[Wang *et al.*, 2024b] Fei Wang, Dan Guo, Kun Li, Zhun Zhong, and Meng Wang. Frequency decoupling for motion magnification via multi-level isomorphic architecture. In *CVPR*, pages 18984–18994, 2024.

[Wang *et al.*, 2024c] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[Wang *et al.*, 2025] Weijia Wang, Fei Wang, Zhongzhen Wang, Xianyang Shi, and Chunyan Xu. Robust s3former deep learning model for the direct diagnosis and prediction of natural organic matter (nom) from three-dimensional excitation-emission-matrix (3d-eem) data. *Water Research*, 284:123994, 2025.

[Wang *et al.*, 2026] Fei Wang, Jiangnan Yang, Junjie Chen, Yuxin Liu, Kun Li, Yanyan Wei, Dan Guo, and Meng Wang. Xinsight: Integrative stage-consistent psychological counseling support agents for digital well-being. In *ACM WWW*, pages 9297–9308, 2026.

[Yang *et al.*, 2023] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023.

[Zhou *et al.*, 2024] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.