EDGEGAZE: EFFICIENT END-TO-END MULTI-PERSON GAZE TARGET DETECTION

Anonymous authors
Paper under double-blind review

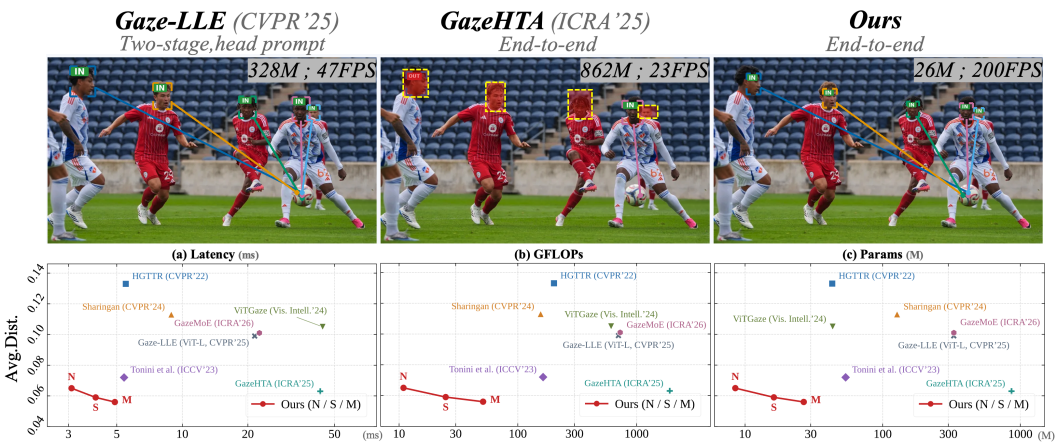

Figure 1: **Top**: qualitative comparison on an in-the-wild scene. **Bottom**: Pareto plots on GazeFollow (Recasens et al., 2015) comparing average distance with measured latency, GFLOPs, and parameter count, with log-scaled x-axes. Points closer to the bottom-left indicate a better accuracy–efficiency trade-off.

ABSTRACT

Gaze target detection on edge devices requires identifying where every person in a scene is looking under tight memory and latency budgets. Existing two-stage methods rely on external head detection and repeated per-person gaze estimation, while recent end-to-end models can grow to hundreds of millions of parameters. We introduce *EdgeGaze*, a compact end-to-end model built on EdgeCrafter that jointly detects multiple people and predicts their gaze targets in a single forward pass, with computation nearly independent of scene crowding. Gaze is strongly shaped by scene structure: people often look at other people, salient regions, and meaningful objects. We exploit this through *Scene Prior Guidance*, using saliency, object, and head priors to learn a *Prior-Guided Scene Representation* (PGSR) during training. PGSR is fused with person queries for gaze prediction, while the external prior models are discarded at inference. We further introduce a *Unified Gaze Distribution* (UGD), which places in-frame gaze locations and the out-of-frame outcome in a single normalized prediction space. EdgeGaze achieves state-of-the-art performance on GazeFollow and VideoAttentionTarget, with strong zero-shot transfer to GOO-Real and ChildPlay. On GazeFollow, EdgeGaze-M achieves 0.960 AUC and 0.798 mAP with only 26.7M parameters and 4.9 ms latency, compared with 0.929 AUC and 0.626 mAP for the 862M-parameter GazeHTA. EdgeGaze-N further reduces the model to 8.5M parameters and 3.1 ms. These results show that accurate multi-person gaze target detection can be compact, fast, and fully end-to-end, making it well suited for edge deployment. Code and models will be released; demo: https://anonymous-edgegaze.pages.dev/.

1 INTRODUCTION

Understanding a social scene requires more than knowing who is present; it also requires understanding where each person directs their attention. Gaze following predicts the target of a given person's gaze (Recasens et al., 2015), while multi-person gaze target detection must localize gaze targets for everyone in the scene and handle cases where attention falls outside the image (Chong et al., 2018; 2020). This capability is important for applications such as child-behaviour analysis,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(a) Looking at People (b) Looking at Objects (c) Across Four Datasets

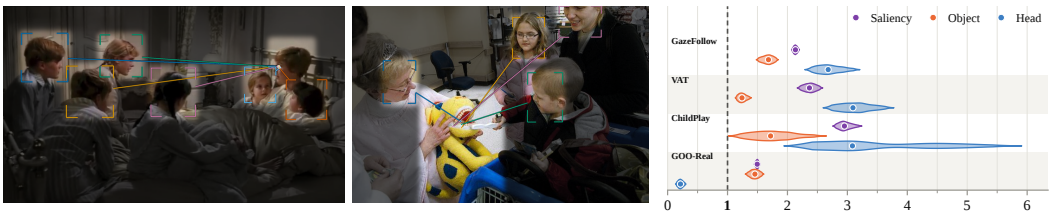

Figure 2: **Gaze targets are not arbitrary points.** (a,b) Ground-truth gaze points tend to concentrate on shared people or objects in the scene. (c) Fraction of annotated gaze targets falling within each prior region, normalized by a shuffled control matched for region area and centre bias. Prior regions include head boxes, detected non-person object boxes, and the top 20% most salient image regions. Spindles indicate 95% cluster-bootstrap confidence intervals.

retail analytics, and human–robot interaction, many of which increasingly run on embedded cameras with tight memory and latency budgets. The challenge is therefore not only to predict gaze accurately, but to do so for all people efficiently, without computation growing substantially as scenes become more crowded.

As illustrated in Figure 1, existing approaches struggle to achieve both unified inference and efficiency. Two-stage methods such as GazeLLE (Ryan et al., 2025) rely on an external head detector and retain person-dependent gaze estimation. End-to-end approaches instead predict heads and gaze targets jointly (Tu et al., 2022; Tonini et al., 2023), but recent models can be very large, as exemplified by GazeHTA (Lin et al., 2025). This leaves a clear opportunity: a compact model that detects all people and predicts their gaze targets in a single pass, with computation largely independent of scene crowding.

To this end, we build on EdgeCrafter (Liu et al., 2026), a compact set-based detector that uses a fixed set of queries to predict multiple instances jointly. We extend this formulation to gaze target detection, allowing multiple people and their gaze targets to be predicted together in a single forward pass. This removes the need for an external head detector and person-wise inference, while keeping computation nearly independent of the number of people in the scene.

We further improve gaze prediction by exploiting a simple property of human attention: gaze is strongly shaped by scene structure. People do not look at arbitrary pixels; they disproportionately attend to other people, salient regions, and semantically meaningful objects (Birmingham et al., 2008; 2009). This pattern is also evident in gaze benchmarks. In GazeFollow (Recasens et al., 2015), VideoAttentionTarget (Chong et al., 2020), and ChildPlay (Tafasca et al., 2023), gaze targets fall on another head roughly two to three times more often than under a shuffled control (Figure 2). We exploit this regularity through *Scene Prior Guidance*: saliency, object, and head priors supervise the shared image features during training to learn a *Prior-Guided Scene Representation* (PGSR). PGSR is then fused with person queries for gaze prediction, while the external prior models are discarded at inference.

We also simplify the gaze output formulation. A person's gaze has one outcome: it either falls somewhere inside the image or leaves the frame. Existing methods typically model these cases separately, using an in-frame heat map and an independent out-of-frame classifier (Chong et al., 2018; Tu et al., 2022; Tonini et al., 2023; Lin et al., 2025). We instead introduce a *Unified Gaze Distribution* (UGD), which places all in-frame locations and the out-of-frame outcome in a single normalized prediction space.

Together, these ideas form *EdgeGaze*, a compact end-to-end model for multi-person gaze target detection. EdgeCrafter provides the efficient set-based foundation, Scene Prior Guidance injects gaze-relevant scene structure without requiring external models at inference, and UGD unifies in-frame and out-of-frame gaze prediction. The resulting model jointly detects multiple people and predicts all gaze targets in one forward pass, with computation nearly independent of scene crowding.

EdgeGaze achieves state-of-the-art performance on GazeFollow (Recasens et al., 2015) and VideoAttentionTarget (Chong et al., 2020), while transferring strongly in a zero-shot setting to GOO-Real (Tomas et al., 2021) and ChildPlay (Tafasca et al., 2023). On GazeFollow, our 26.7M-parameter model reaches 0.960 AUC, compared with 0.929 for an end-to-end detector with more than $30\times$ as many parameters, while running at $4.9\,\mathrm{ms}$ per image on a single desktop GPU. Our smallest variant uses only 8.5M parameters and runs at $3.1\,\mathrm{ms}$. In Figure 1, we present Pareto frontier plots on

GazeFollow, comparing average distance with latency, GFLOPs, and parameter count. *EdgeGaze* consistently lies on the Pareto frontier, offering a substantially better accuracy–efficiency trade-off than previous approaches. These results show that accurate multi-person gaze target detection need not come at the cost of large models or inference that scales strongly with scene crowding.

Our contributions are summarized as follows:

- We present *EdgeGaze*, a compact end-to-end model that jointly detects multiple people and predicts all gaze targets in a single forward pass, with computation nearly independent of scene crowding. Despite its compact design, EdgeGaze achieves state-of-the-art performance while using substantially fewer parameters and lower latency than existing models.
- We introduce *Scene Prior Guidance*, which uses saliency, object, and head priors to learn a *Prior-Guided Scene Representation* (PGSR). PGSR directly supports gaze prediction by combining person-specific cues, scene-level structure, and their interaction, while requiring no external prior models at inference.
- We introduce a *Unified Gaze Distribution* (UGD) that places in-frame gaze locations and the out-of-frame outcome in a single normalized prediction space. This simple formulation unifies gaze localization and in/out prediction, and contributes to strong performance on GazeFollow and VideoAttentionTarget as well as zero-shot transfer to GOO-Real and ChildPlay.

An interactive demo is available at `https://anonymous-edgegaze.pages.dev/`.

2 Related Work

Two-stage gaze target detection. Gaze following was introduced with GazeFollow (Recasens et al., 2015), which predicts where a person is looking given the image and head location. Chong et al. (2018) extended the task to out-of-frame gaze, and Chong et al. (2020) later introduced the VideoAttentionTarget benchmark. Classical approaches typically combine a head crop and its location with the scene image to predict a gaze heatmap (Lian et al., 2018; Chong et al., 2020). Recent methods retain this head-conditioned formulation while improving how person–scene information is modeled. ViTGaze (Song et al., 2024) extracts interaction features from a pretrained ViT, GazeLLE (Ryan et al., 2025) uses a frozen large-scale encoder with a person-specific positional prompt, GazeMoE (Dai et al., 2026) introduces mixture-of-experts decoding, and GazeHRNet (Nan et al., 2026) incorporates head-centric spatial encoding. Sharingan (Tafasca et al., 2024) represents each person with a dedicated token, enabling multiple people to be processed in one forward pass. Other work extends the formulation to temporal and social gaze reasoning (Gupta et al., 2024) or 360-degree gaze target estimation (Dai et al., 2025). Despite these advances, these methods require head boxes as input and therefore rely on an external head detector in a complete inference system. Multi-person designs reduce redundant computation, but their gaze-estimation cost still depends on the number of people in the scene.

End-to-end gaze target detection. End-to-end methods remove the need for externally provided head boxes by jointly localizing people and predicting their gaze targets from the scene image. HGTTR (Tu et al., 2022) first formulates this problem as DETR-style set prediction (Carion et al., 2020), with each query predicting a head and its gaze target. Subsequent work extends this formulation in different directions: Tonini et al. (Tonini et al., 2023) jointly reason about heads, gaze targets, and gazed objects; GazeDETR (de Belen et al., 2025) uses separate decoders for head localization and gaze prediction; and GazeHTA (Lin et al., 2025) leverages Stable Diffusion features and explicitly models head–target associations with connection maps. More recently, GazeAnywhere (Cao et al., 2026) introduces a promptable setting that jointly localizes a specified person and predicts their gaze target, whereas our goal is to predict all people and gaze targets without prompting. Despite removing the two-stage pipeline, recent end-to-end models can be computationally heavy, with systems such as GazeHTA growing to hundreds of millions of parameters and incurring substantial inference cost. This limits their suitability for resource-constrained edge deployment, where model size, computation, and latency are all critical.

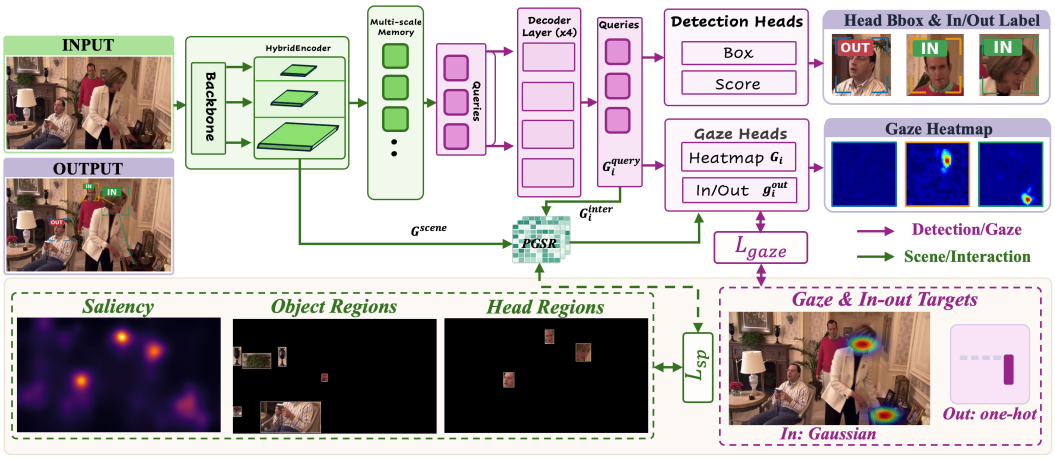

Figure 3: **Overview of EdgeGaze.** EdgeGaze uses shared image features and a fixed set of person queries to jointly detect heads and predict gaze targets in one forward pass. Scene Prior Guidance uses saliency, object, and head priors during training to learn a Prior-Guided Scene Representation (PGSR). For each person, the query term G_i^{query}, scene term G^{scene}, and person–scene interaction term G_i^{inter} are fused to form the spatial gaze logits. The Unified Gaze Distribution (UGD) then jointly models in-frame gaze locations and the out-of-frame outcome. External priors are used only during training; inference requires only the input image.

Scene semantics for gaze targets. Gaze is strongly influenced by scene semantics: people tend to look at other people and meaningful objects rather than arbitrary image locations (Birmingham et al., 2008; Mi et al., 2026). Prior work exploits such structure in two main ways. One line introduces additional scene cues at inference, including depth (Tonini et al., 2022), depth-aware saliency (Mathew et al., 2025), 3D scene geometry (Bao et al., 2022), depth-guided attention (Fang et al., 2021), and textual scene descriptions (She et al., 2026). These methods require additional modalities or auxiliary models beyond the RGB image at inference. Another line explicitly reasons about the object being viewed. Gaze object prediction estimates the location and category of the gazed object (Tu et al., 2023; Wang et al., 2024), while later work introduces pixel-level supervision from vision foundation models to better distinguish nearby or overlapping objects (Jin et al., 2024). GaTector+ (Jin et al., 2025) further removes the need for externally provided head priors at inference. Such object-centric approaches, however, rely on object-level gaze supervision that is not provided by standard gaze-following benchmarks such as GazeFollow and VideoAttentionTarget. Our approach instead uses scene semantics as *training-time guidance*. Saliency, object, and head priors supervise the learning of a Prior-Guided Scene Representation (PGSR), which is directly used for gaze prediction. The external prior models are discarded at inference, leaving an image-only end-to-end model.

3 METHOD

Figure 3 gives an overview of *EdgeGaze*. Given a single image, EdgeGaze uses shared image features and a fixed set of person queries to jointly detect multiple people and predict their gaze targets in one forward pass. Building on EdgeCrafter (Liu et al., 2026), we introduce two key ideas for gaze prediction: *Scene Prior Guidance*, which uses saliency, object, and head priors during training to learn a gaze-relevant scene representation, and the *Unified Gaze Distribution* (UGD), which jointly models in-frame gaze locations and the out-of-frame outcome.

Preliminary: EdgeCrafter. EdgeCrafter (Liu et al., 2026) is a compact dense prediction framework following the DETR-style set prediction paradigm (Carion et al., 2020). It extracts shared image features and uses a fixed set of queries to jointly predict multiple instances, with task-specific prediction heads built on top of the shared representation. EdgeGaze adopts this framework for multi-person gaze target detection: each query represents one candidate person and predicts a head bounding box, a confidence score, and the corresponding gaze target. On top of this efficient set-based foundation, Scene Prior Guidance enriches the shared scene representation, while UGD provides a unified formulation for gaze prediction.

Overall objective. For each query i, EdgeGaze predicts a head box $\hat{b}_i$, a confidence score $\hat{c}_i$, and a gaze distribution $\hat{p}_i$. The model is trained with three objectives:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{gaze}}\mathcal{L}_{\text{gaze}} + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}}, \tag{1}$$

where $\mathcal{L}_{\text{det}}$ supervises person detection, $\mathcal{L}_{\text{sp}}$ provides scene-prior guidance, and $\mathcal{L}_{\text{gaze}}$ supervises the Unified Gaze Distribution. The weights λ_{gaze} and λ_{sp} balance the gaze and scene-prior objectives. We next describe Scene Prior Guidance and UGD, followed by training and implementation details.

3.1 Scene Prior Guidance

As shown in Figure 2, gaze targets are strongly related to scene structure: people often look at other people, salient regions, and meaningful objects. We use these scene priors as auxiliary supervision to guide the learning of gaze-relevant scene features.

Scene-prior supervision. We consider three complementary priors: saliency, object regions, and head regions. During training, their target maps are generated offline by external experts. Saliency is treated as a spatial distribution and supervised with KL divergence, while object and head regions are treated as occupancy maps and supervised with binary cross-entropy and Dice losses. We write the overall scene-prior loss as

$$\mathcal{L}_{\text{sp}} = \mathcal{L}_{\text{sal}} + \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{head}}. \tag{2}$$

These priors are used only during training and introduce no external model at inference.

Prior-guided scene representation. We apply scene-prior supervision to the shared encoder features. A lightweight prediction head estimates saliency, object, and head maps, and the scene-prior loss $\mathcal{L}_{\text{sp}}$ encourages the underlying features to capture gaze-relevant scene structure. We refer to the resulting dense representation as the *Prior-Guided Scene Representation* (PGSR), denoted by G^{scene}. The external priors are used only during training; at inference, PGSR is computed directly from the image features.

Gaze prediction with scene guidance. For each person, we combine three complementary terms: a query-based gaze map G_i^{query}, the shared PGSR G^{scene}, and a person–scene interaction term G_i^{inter}. The final spatial gaze logits are

$$G_i = G_i^{\text{query}} + G^{\text{scene}} + G_i^{\text{inter}}, \qquad G_i^{\text{inter}} = \frac{1}{\sqrt{C}}\,\text{Dot}(G_i^{\text{query}}, G^{\text{scene}})\,, \tag{3}$$

where C is the feature dimension. The query term captures person-specific gaze cues, PGSR provides shared scene guidance, and the interaction term captures how the person-specific prediction relates to the scene representation. Their sum G_i forms the spatial gaze logits.

3.2 Unified Gaze Distribution

A person's gaze has one outcome: it either falls somewhere inside the image or leaves the frame. Existing methods typically model these cases separately, using a spatial gaze heatmap and an independent in/out classifier. We instead place them in a single prediction space, allowing the model to jointly decide whether the gaze is inside the image and, if so, where it is located.

For person query i, let G_i denote the fused spatial gaze logits from Eq. 3, and let g_i^{out} denote the out-of-frame logit. We concatenate them and apply a single softmax:

$$p_i = \text{softmax}\left([G_i,\, g_i^{\text{out}}]\right). \tag{4}$$

The final entry p_i^{out} gives the probability of looking outside the image, while the remaining spatial entries jointly carry the in-frame probability

$$p_i^{\text{in}} = 1 - p_i^{\text{out}}. \tag{5}$$

Thus, in-frame and out-of-frame gaze compete directly within the same normalized distribution.

Table 1: **Accuracy and efficiency of two-stage and end-to-end methods on GazeFollow (Recasens et al., 2015) and VideoAttentionTarget (Chong et al., 2020).** Params. include all components needed at inference. GFLOPs are measured in forward passes at each method's native resolution, counting a multiply-accumulate as two FLOPs; two-stage costs are $a + bK$ for K detected heads. Latency is TensorRT FP16 at batch size one on an RTX 2080 Ti, averaged over the first 100 GazeFollow test images after 50 warm-up iterations. Out. AP is the out-of-frame average precision. Best results are in bold and second-best underlined.

Method	Efficiency			GazeFollow				VideoAttentionTarget			
	Params. (M)	GFLOPs↓	Latency (ms)↓	AUC↑	Avg. L2↓	Min. L2↓	mAP↑	AUC↑	L2↓	Out. AP↑	mAP↑
Two-stage methods											
VideoAttention (Chong et al., 2020)	–	–	–	0.921	0.137	0.077	–	0.860	0.134	0.853	–
DAM (Fang et al., 2021)	–	–	–	0.922	0.124	0.067	–	0.905	0.108	0.896	–
ESCNet (Bao et al., 2022)	–	–	–	0.928	0.122	–	–	0.885	0.120	0.869	–
PDP (Miao et al., 2023)	–	–	–	0.934	0.123	0.065	–	0.917	0.109	0.908	–
Sharingan (Tafasca et al., 2024)	127.2	$86 + 12K$	8.9	0.944	0.113	0.057	–	–	0.107	0.891	–
ViTGaze (Song et al., 2024)	43.3	$50 + 93K$	44.1	0.949	0.105	0.047	–	0.938	0.102	0.905	–
Gaze-LLE (ViT-L) (Ryan et al., 2025)	328.3	$671 + 5K$	21.5	0.958	0.099	0.041	–	0.937	0.103	0.903	–
GazeHRNet (Nan et al., 2026)	110.5	–	–	0.952	0.102	0.055	–	0.929	0.103	0.909	–
TextGaze (ViT-L) (She et al., 2026)	≥8727.0	–	–	0.956	0.099	0.043	–	0.941	0.102	0.911	–
GazeMoE (Dai et al., 2026)	328.8	$671 + 9K$	22.6	0.959	0.101	–	–	0.939	0.097	0.917	–
PAGE (ViT-H+) (Ye et al., 2026)	1718.5	$1779 + 459K$	107.2	**0.966**	0.080	0.029	–	0.972	0.064	0.951	–
End-to-end methods											
HGTTR (Tu et al., 2022)	43.0	202	5.5	0.917	0.133	0.069	0.547	0.893	0.137	0.821	0.514
Tonini et al. (Tonini et al., 2023)	53.8	163	5.4	0.922	0.072	0.033	0.573	0.923	0.102	0.944	0.607
GazeHTA (Lin et al., 2025)	862.4	1903	43.0	0.929	0.063	**0.025**	0.626	0.957	0.064	**1.000**	0.772
EdgeGaze-N	**8.5**	**11**	**3.1**	0.955	0.065	0.033	0.721	0.966	0.075	0.999	0.804
EdgeGaze-S	16.2	25	4.0	0.959	0.059	0.029	0.771	**0.978**	**0.055**	0.999	0.822
EdgeGaze-M	26.7	51	4.9	0.960	**0.056**	0.027	**0.798**	0.977	0.056	**1.000**	**0.824**

Gaze supervision. We supervise the unified distribution with two complementary terms. First, an in/out loss $\mathcal{L}_{\text{in/out}}$ applies KL divergence to the two-way distribution $\left[p_i^{\text{in}}, p_i^{\text{out}}\right]$. Second, for in-frame samples, we renormalize the spatial part of p_i into a conditional gaze heatmap $\tilde{p}_i$ and match it to the ground-truth gaze heatmap using KL divergence, giving $\mathcal{L}_{\text{heat}}$. The gaze loss is

$$\mathcal{L}_{\text{gaze}} = \mathcal{L}_{\text{in/out}} + \mathcal{L}_{\text{heat}}. \tag{6}$$

This formulation couples gaze localization and in/out prediction through a single probability space: the model first allocates probability between looking inside and outside the image, while the in-frame probability is further distributed over possible gaze locations.

3.3 Training and Implementation Details

Architecture and supervision. EdgeGaze follows the backbone–encoder–decoder design of EdgeCrafter (Liu et al., 2026). All variants take 640×640 images, use $N = 20$ person queries, and predict gaze on a 64×64 grid. EdgeGaze-N/S/M share the same gaze formulation but progressively increase backbone, encoder, and decoder capacity, with 8.5M, 16.2M, and 26.7M parameters, respectively. The shared encoder features are used to form PGSR and the person queries described in Sec. 3.1. Scene-prior targets are generated offline using DeepGaze IIE (Linardos et al., 2021) for saliency, Co-DETR (Zong et al., 2023b) for objects, and annotated and pseudo-labelled heads for head occupancy. Saliency is supervised with KL divergence, while object and head occupancy use BCE and Dice losses. All external experts are used only during training and are removed at inference. Detailed architectures and prior-generation settings are provided in Sec. A of *Sup.Mat.*

Training. Following DETR-style set prediction (Carion et al., 2020), we use Hungarian matching based on head classification and localization, together with gaze and in/out terms when the corresponding annotations are available; missing labels are masked. We set $\lambda_{\text{gaze}} = \lambda_{\text{sp}} = 1$ in Eq. 1 and retain the auxiliary detection and denoising objectives of EdgeCrafter. All variants are initialized from pretrained EdgeCrafter detectors and trained on GazeFollow (Recasens et al., 2015) for 10 epochs, followed by 10 epochs of fine-tuning on VideoAttentionTarget (Chong et al., 2020). We use AdamW with a global batch size of 128, weight decay 10^{-4}, and gradient clipping at 0.1. On GazeFollow, the learning rates are 5×10^{-4} for prediction heads, 5×10^{-5} for the shared body, and 2.5×10^{-5} for the backbone; they are reduced to 1×10^{-4}, 2×10^{-5}, and 1×10^{-5}, respectively, for VideoAttentionTarget fine-tuning. Training uses four NVIDIA A40 GPUs. Further matching coefficients, schedules, EMA settings, and optimization details are given in Sec. A of *Sup.Mat.*

Table 2: **Transfer to GOO-Real (Tomas et al., 2021) and ChildPlay (Tafasca et al., 2023).** "E2E" indicates end-to-end gaze target detection, and "ZS" indicates zero-shot evaluation from a GazeFollow (Recasens et al., 2015) checkpoint. Two-stage methods use ground-truth head boxes at inference, whereas EdgeGaze detects heads and predicts gaze jointly. "Params." denotes model parameters. Bold marks the best zero-shot result. *The code for TextGaze has not been released; we therefore report a lower bound on its parameter count based on the components described in the paper, including Qwen2.5-VL (Bai et al., 2025).

Method	Params. (M)	E2E	ZS	GOO-Real		ChildPlay	
				AUC↑	L2↓	AUC↑	L2↓
TransGOP (Wang et al., 2024)	115.1	×	×	0.957	0.081	–	–
Jin et al. (Jin et al., 2024)	81.0	×	×	0.963	0.073	–	–
Gaze-LLE (ViT-L) (Ryan et al., 2025)	328.3	×	×	–	–	0.951	0.101
GazeMoE (Dai et al., 2026)	328.8	×	×	–	–	0.945	0.106
Gaze-LLE (ViT-B) (Ryan et al., 2025)	110.5	×	✓	0.901	0.174	0.946	0.114
Gaze-LLE (ViT-L) (Ryan et al., 2025)	328.3	×	✓	0.898	0.175	0.951	0.101
GazeHRNet (Nan et al., 2026)	110.5	×	✓	0.892	0.173	0.939	0.109
TextGaze (ViT-L)* (She et al., 2026)	$\geq$ 8727.0	×	✓	0.918	0.159	0.950	0.111
EdgeGaze-M	**26.7**	✓	✓	**0.947**	**0.085**	**0.965**	**0.053**

4 EXPERIMENTS

Datasets. Following the standard protocol of HGTTR (Tu et al., 2022) and GazeHTA (Lin et al., 2025), we train on GazeFollow (Recasens et al., 2015) and fine-tune on VideoAttentionTarget (Chong et al., 2020). GazeFollow annotates one person per test image, with six to ten gaze annotations for that person, but provides no in/out-of-frame label. VideoAttentionTarget provides one gaze annotation per person and labels 35.4% of its test instances as out-of-frame. For zero-shot evaluation, we directly apply the GazeFollow checkpoint to GOO-Real (Tomas et al., 2021) and ChildPlay (Tafasca et al., 2023), which contain retail and children's video scenes, respectively. Dataset splits follow HGTTR and GazeHTA; further details are provided in Sec. B of *Sup.Mat.*

Metrics. The reported distance metrics follow the annotation format of each benchmark. On GazeFollow (Recasens et al., 2015), each test person has multiple gaze annotations, so we report both the average and minimum L2 distance between the prediction and the annotated gaze points, together with AUC and mAP. On VideoAttentionTarget (Chong et al., 2020), each person has a single gaze annotation, so we report a single L2 distance, together with AUC, out-of-frame AP, and mAP. For mAP, a prediction is matched to an annotation when the head IoU is greater than 0.5 and the gaze distance is below 0.15, following GazeHTA (Lin et al., 2025). Because both datasets annotate only a subset of the heads present in each image, we follow GazeHTA and compute mAP over matched annotated instances rather than treating predictions on unannotated people as false positives. On GazeFollow, 63.8% of test images contain at least one unlabelled head; Sec. D of *Sup.Mat.* analyzes the effect of using a strict detection mAP. For zero-shot transfer, we report AUC and L2 distance following GazeLLE (Ryan et al., 2025).

Efficiency measurement. We measure parameters, GFLOPs, and latency for all methods under a unified protocol. Parameters include all components required at inference, including frozen backbones and, for two-stage methods, external modules such as head detectors or feature extractors (e.g., DINOv3 (Siméoni et al., 2025) in GazeLLE (Ryan et al., 2025)). GFLOPs are measured from actual forward passes at each method's native inference resolution, counting one multiply-accumulate as two FLOPs. For two-stage methods, we report the cost as $a+bK$, where K is the number of detected heads. Here, a denotes the computation performed once per image, including the head detector and any shared scene encoding, while b denotes the per-head gaze-estimation cost. Latency is measured with TensorRT FP16 at batch size one on a single RTX 2080 Ti. After 50 warm-up iterations, we report the average execution time over the first 100 GazeFollow (Recasens et al., 2015) test images. Two-stage methods process every detected head, totaling 599 heads across these images.

4.1 COMPARISON TO STATE-OF-THE-ART APPROACHES

Accuracy–efficiency trade-off. Table 1 compares EdgeGaze with both two-stage and end-to-end methods on GazeFollow (Recasens et al., 2015) and VideoAttentionTarget (Chong et al., 2020). EdgeGaze consistently achieves strong accuracy at substantially lower inference cost. EdgeGaze-M reaches the best or second-best result on nearly all reported metrics with only 26.7M parameters,



Figure 4: **Predictions on ChildPlay (Tafasca et al., 2023), GazeFollow (Recasens et al., 2015), GOO-Real (Tomas et al., 2021), and VideoAttentionTarget (VAT) (Chong et al., 2020).** ChildPlay and GOO-Real are evaluated zero-shot using the GazeFollow checkpoint, without fine-tuning. The red OUT tag indicates a person predicted to look outside the image.

compared with 862.4M for GazeHTA (Lin et al., 2025). At the other end of the scale, EdgeGaze-N uses only 8.5M parameters and runs at 3.1 ms per image while retaining strong performance on both benchmarks. Unlike two-stage methods, EdgeGaze also avoids external head detection and repeated per-person gaze estimation. Together, these results show that accurate multi-person gaze target detection can be achieved with a compact, single-pass model.

Zero-shot generalization. Table 2 evaluates transfer from GazeFollow to GOO-Real (Tomas et al., 2021) and ChildPlay (Tafasca et al., 2023) without target-domain training. EdgeGaze-M achieves the strongest zero-shot results on both datasets while simultaneously detecting heads and predicting gaze targets. This is a stricter setting than the compared two-stage methods, which receive ground-truth head boxes at inference. Notably, EdgeGaze achieves this transfer with only 26.7M parameters, without relying on large auxiliary vision or vision–language models. These results suggest that the learned scene and gaze representations generalize beyond the training distribution rather than overfitting to a particular benchmark.

Qualitative results. Figure 4 shows qualitative predictions from EdgeGaze-M across all four datasets. The model maintains accurate person–gaze associations across diverse scenes, including retail environments in GOO-Real (Tomas et al., 2021) and children's interactions in ChildPlay (Tafasca et al., 2023), both evaluated zero-shot. Its predictions remain focused on meaningful people and objects despite substantial changes in appearance and scene composition, consistent with the quantitative transfer results. Additional visualizations on in-the-wild images are provided in Sec. G of *Sup.Mat.* We also provide an interactive demo on an anonymous project page at `https://anonymous-edgegaze.pages.dev/`.

4.2 Ablation Study

We ablate the three key design choices of EdgeGaze: Scene Prior Guidance, person–scene gaze fusion, and the Unified Gaze Distribution. All experiments use EdgeGaze-M and the same training, matching, and evaluation protocol as the main model; only the component being ablated is modified.

Table 3: **Ablation of Scene Prior Guidance and person–scene gaze fusion on GazeFollow (Recasens et al., 2015).** All experiments use EdgeGaze-M. (a) Ablation of the three scene-prior supervision terms in $\mathcal{L}_{\text{sp}}$. (b) Ablation of the three gaze-fusion terms in Eq. 3.

Components			AUC↑	mAP↑	Avg. L2↓	Min. L2↓
(a) Scene Prior Guidance						
$\mathcal{L}_{\text{sal}}$	$\mathcal{L}_{\text{obj}}$	$\mathcal{L}_{\text{head}}$				
✗	✗	✗	0.9592	0.7857	0.0584	0.0286
✗	✓	✓	0.9597	0.7892	0.0575	0.0279
✓	✗	✓	0.9594	0.7871	0.0574	0.0277
✓	✓	✗	**0.9601**	0.7821	0.0571	0.0275
✓	✓	✓	0.9600	**0.7982**	**0.0563**	**0.0271**
(b) Person–scene gaze fusion						
G_i^{query}	G^{scene}	G_i^{inter}				
✗	✓	✓	0.9597	0.7907	0.0578	0.0277
✓	✗	✓	0.9590	0.7836	0.0577	0.0285
✓	✓	✗	**0.9602**	0.7767	0.0608	0.0305
✓	✓	✓	0.9600	**0.7982**	**0.0563**	**0.0271**

Table 4: **Ablation of Unified Gaze Distribution (UGD) on GazeFollow (Recasens et al., 2015) and VideoAttentionTarget (Chong et al., 2020).** "Separate" predicts the gaze heatmap and in/out status independently, while UGD jointly normalizes $[G_i,\ g_i^{\text{out}}]$ as in Eq. 4.

Formulation	GazeFollow				VideoAttentionTarget			
	AUC↑	Avg. L2↓	Min. L2↓	mAP↑	AUC↑	L2↓	Out. AP↑	mAP↑
Separate heatmap + in/out	0.9585	0.0584	0.0286	0.7721	0.9766	**0.0545**	0.9982	0.8125
UGD: $\text{softmax}([G_i, g_i^{\text{out}}])$	**0.9600**	**0.0563**	**0.0271**	**0.7982**	**0.9771**	0.0561	**0.9997**	**0.8240**

Scene guidance and gaze fusion. Table 3 evaluates both the training-time scene-prior losses and the three terms used for person–scene gaze fusion. In panel (a), combining $\mathcal{L}_{\text{sal}}$, $\mathcal{L}_{\text{obj}}$, and $\mathcal{L}_{\text{head}}$ gives the strongest overall performance (Sec. F.1 of *Sup.Mat.*), showing that saliency, object, and head priors provide complementary guidance for learning PGSR. In panel (b), the three fusion terms in Eq. 3 also play complementary roles: G_i^{query} provides person-specific gaze cues, G^{scene} captures shared scene structure, and G_i^{inter} links each person to relevant regions in the scene. Together, these results show that scene information improves gaze prediction both by shaping the representation through $\mathcal{L}_{\text{sp}}$ during training and by guiding person-specific prediction through gaze fusion at inference.

Unified Gaze Distribution. Table 4 compares UGD with the conventional separate heatmap and in/out predictions. Joint normalization improves all GazeFollow (Recasens et al., 2015) metrics and increases both out-of-frame AP and mAP on VideoAttentionTarget (Chong et al., 2020), with only a small trade-off in L2 distance. This suggests that treating in-frame and out-of-frame gaze as competing outcomes in a single distribution provides a cleaner and more effective prediction space.

5 Conclusion

We presented *EdgeGaze*, a compact end-to-end model for multi-person gaze target detection. Built on a set-based detector, EdgeGaze jointly detects people and predicts their gaze targets in a single forward pass, avoiding external head detection and person-wise inference while keeping computation nearly independent of scene crowding. To improve gaze prediction, we introduced *Scene Prior Guidance*, which uses saliency, object, and head priors during training to learn a *Prior-Guided Scene Representation* (PGSR) that directly supports person-specific gaze prediction. We further proposed a *Unified Gaze Distribution* (UGD), which places in-frame gaze locations and the out-of-frame outcome in a single normalized prediction space. EdgeGaze achieves state-of-the-art performance on GazeFollow and VideoAttentionTarget while remaining compact and fast, and transfers strongly in a zero-shot setting to GOO-Real and ChildPlay. These results show that accurate multi-person gaze target detection can be achieved with a lightweight, single-pass model, making end-to-end gaze understanding practical for edge deployment.

AI USE STATEMENT

Generative AI tools were used to assist with literature review, manuscript organization, language editing, interpretation and synthesis of the reported experimental results, and limited code development and debugging. Their use in result interpretation was restricted to summarizing and describing trends and comparisons already supported by the reported figures and tables. Generative AI tools were not used to generate, alter, or fabricate experimental data. All AI-assisted text and code were reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jun Bao, Buyu Liu, and Jun Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14106–14115. IEEE, 2022.

Elina Birmingham, Walter F Bischof, and Alan Kingstone. Social attention and real-world scenes: The roles of action, competition and social content. *Quarterly journal of experimental psychology*, 61(7):986–998, 2008.

Elina Birmingham, Walter F Bischof, and Alan Kingstone. Saliency does not account for fixations to eyes within social scenes. *Vision research*, 49(24):2992–3000, 2009.

Xu Cao, Houze Yang, Vipin Gunda, Zhongyi Zhou, Tianyu Xu, Adarsh Kowdle, Inki Kim, and James M Rehg. Gaze target estimation anywhere with concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31304–31315, 2026.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

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 *European Conference on Computer Vision*, pp. 397–412. Springer, 2018.

Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5395–5405. IEEE, 2020.

Zhuangzhuang Dai, Vincent Gbouna Zakka, Luis J Manso, and Chen Li. Gazetarget360: Towards gaze target estimation in 360-degree for robot perception. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 18291–18296. IEEE, 2025.

Zhuangzhuang Dai, Zhongxi Lu, Vincent G Zakka, Luis J Manso, Jose M Alcaraz Calero, and Chen Li. Gazemoe: Perception of gaze target with mixture-of-experts. *arXiv preprint arXiv:2603.06256*, 2026.

Ryan Anthony Jalova de Belen, Gelareh Mohammadi, and Arcot Sowmya. Gazedetr: Gaze detection using disentangled head and gaze representations. *arXiv preprint arXiv:2508.12966*, 2025.

Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11390–11399, 2021.

Anshul Gupta, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard, and Jean-Marc Odobez. Mtgs: A novel framework for multi-person temporal gaze following and social gaze prediction. *Advances in Neural Information Processing Systems*, 37:15646–15673, 2024.

Yang Jin, Lei Zhang, Shi Yan, Bin Fan, and Binglu Wang. Boosting gaze object prediction via pixel-level supervision from vision foundation model. In *European Conference on Computer Vision*, pp. 369–386. Springer, 2024.

Yang Jin, Guangyu Guo, and Binglu Wang. Gatector+: a unified head-free framework for gaze object and gaze following prediction. *arXiv preprint arXiv:2510.25301*, 2025.

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

Zhi-Yi Lin, Jouh Yeong Chew, Jan van Gemert, and Xucong Zhang. Gazehta: End-to-end gaze target detection with head-target association. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9447–9454. IEEE, 2025.

Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12899–12908. IEEE, 2021.

Longfei Liu, Yongjie Hou, Yang Li, Qirui Wang, Youyang Sha, Yongjun Yu, Yinzhi Wang, Peizhe Ru, Xuanlong Yu, and Xi Shen. Edgecrafter: Compact vits for edge dense prediction via task-specialized distillation. *TMLR*, 2026.

Athul M Mathew, Arshad Ali Khan, Thariq Khalid, Faroq Al-Tam, and Riad Souissi. Leveraging multi-modal saliency and fusion for gaze target detection. *arXiv preprint arXiv:2504.19271*, 2025.

Jiajie Mi, Xinyu Liu, Mengke Song, and Chenglizhao Chen. Multi-scale object-aware gaze estimation via geometric reasoning. In *European Conference on Computer Vision*, pp. 156–174. Springer, 2026.

Qiaomu Miao, Minh Hoai, and Dimitris Samaras. Patch-level gaze distribution prediction for gaze following. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 880–889. IEEE, 2023.

Tianxiang Nan, Chenglizhao Chen, Xi Chen, Zhi Li, Xiangyu Wei, and Xinyu Liu. Gazehrnet: Head-centric spatial encoding and gaze-aware feature interaction for gaze target detection. *Sensors*, 26(15):4799, 2026.

Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *Advances in neural information processing systems*, 28, 2015.

Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M Rehg. Gazelle: Gaze target estimation via large-scale learned encoders. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28874–28884. IEEE, 2025.

Junhui She, Fei Wang, Kun Li, Yiqi Nie, Yuxin Liu, Zhangling Duan, and Xun Yang. Textgaze: Prompting gaze target estimation with textual scene cues. *arXiv preprint arXiv:2607.10130*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Yuehao Song, Xinggang Wang, Jingfeng Yao, Wenyu Liu, Jinglin Zhang, and Xiangmin Xu. Vitgaze: gaze following with interaction features in vision transformers. *Visual Intelligence*, 2 (1):31, 2024.

Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20878–20889. IEEE, 2023.

Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Sharingan: A transformer architecture for multi-person gaze following. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2008–2017. IEEE, 2024.

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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3119–3127. IEEE, 2021.

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

Francesco Tonini, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Object-aware gaze target detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21860–21869, 2023.

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

Danyang Tu, Wei Shen, Wei Sun, Xiongkuo Min, and Guangtao Zhai. Joint gaze-location and gaze-object detection. *arXiv preprint arXiv:2308.13857*, 2023.

Binglu Wang, Chenxi Guo, Yang Jin, Haisheng Xia, and Nian Liu. Transgop: transformer-based gaze object prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 10180–10188, 2024.

Zhoutong Ye, Chengwen Zhang, Zhaibin Cui, Mingze Sun, Jiaqi Liu, Xiangwu Li, Qingyang Wan, Chang Liu, Xutong Wang, Huan-ang Gao, et al. Page: Towards practical human-level gaze target estimation. *arXiv preprint arXiv:2607.04860*, 2026.

Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6725–6735. IEEE, 2023a.

Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6725–6735. IEEE, 2023b.

Appendix

Table of Contents

A Implementation Details **14**
A.1 Model Variants and Training Settings 14
A.2 Inference 14

B Datasets and Evaluation Protocol **15**
B.1 Datasets 15
B.2 Evaluation Protocol 16

C Prior Generation **16**
C.1 Prior Experts 16
C.2 Cross-Dataset Prior Statistics 16

D Annotation Completeness **17**
D.1 Unlabelled Heads 17
D.2 Impact on Evaluation 17

E Efficiency Measurement **18**
E.1 Measurement Protocol 18
E.2 Efficiency Results and Parameter Breakdown 18

F Additional Ablations **19**
F.1 Scene-Prior Losses 19
F.2 Prior-Guided Scene Representation 20

G Qualitative Results **20**

A Implementation Details

A.1 Model Variants and Training Settings

Tables A1 and A2 summarize the training settings and model variants. All three variants use $N = 20$ queries, a 64×64 gaze field and the same losses, while varying the backbone and encoder widths and decoder dimensions. We initialize GazeFollow training from the released EdgeCrafter checkpoints (Liu et al., 2026). The S and M checkpoints use the released EdgeCrafter variants with COCO supervision and additional Objects365 pretraining; the N variant uses the released ECDet-N checkpoint. Each VideoAttentionTarget run starts from the GazeFollow checkpoint of the corresponding size.

EdgeGaze achieves the reported GazeFollow performance with just 10 training epochs; the M variant takes approximately 4.5 hours on four A40 GPUs. Figure A1 compares this short schedule with the 15–150 epochs reported by ten recent methods, indicating faster optimization in our training setting. The counts come from different hardware, batch sizes and initialization, and several papers release neither code nor convergence curves; the comparison is therefore about schedule length rather than controlled wall-clock speed.

A.2 Inference

At inference, EdgeGaze takes only the image. The three external prior maps for saliency, objects and heads serve only as training targets; neither the maps nor the expert predictors are needed at test time. The Prior-Guided Scene Representation (PGSR) is computed directly from image features. One forward pass yields N candidates $(\hat{b}_i, \hat{c}_i, \hat{p}_i)$, one per query. We retain candidates whose confidence $\hat{c}_i$ exceeds a threshold, report the spatial peak $\hat{g}_i$ as the in-frame gaze estimate, and read $\hat{p}_i^{\text{out}}$ from the last entry of the gaze distribution. The fixed query count sets prediction capacity, and Section E.2 measures inference cost as the number of people varies.

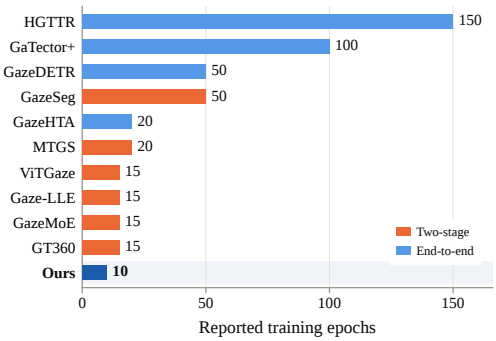

Figure A1: **Reported GazeFollow training schedules.**

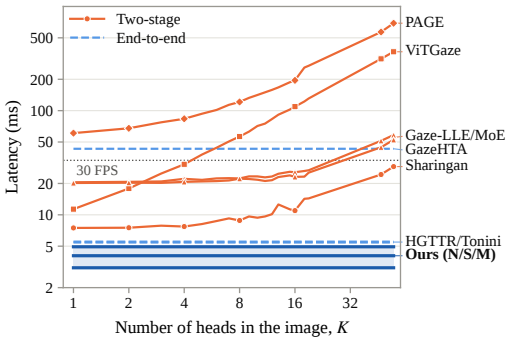

Figure A2: **Latency against the number of detected heads K.** Two-stage methods grow with K, while end-to-end methods remain nearly flat.

Table A1: **Training hyper-parameters.**

	GazeFollow	VideoAttentionTarget
Epochs	10	
Global batch size	128 (4 GPUs $\times$ 32 images/GPU)	
Input resolution	640×640	
Queries N	20	
Field grid $\lvert\Omega\rvert$	64×64	
Optimiser	AdamW, $\beta = (0.9, 0.999)$, weight decay 10^{-4}	
Gradient clipping	0.1	
Learning rate, heads	5×10^{-4}	1×10^{-4}
Learning rate, shared body	5×10^{-5}	2×10^{-5}
Learning rate, backbone	2.5×10^{-5}	1×10^{-5}
Schedule	constant	cosine to 10^{-6}
EMA decay / warm-up steps	0.9999 / 1000	0.998 / 50
Training batches per epoch	all	20%
Initialisation	EdgeCrafter detector	GazeFollow EMA weights
Matching coefficients α, training	class 2, box 5, GIoU 2, gaze 30, out 5, heatmap 0	
Matching coefficients α, evaluation	same, with gaze 100	
Gaze/scene-prior loss coefficients λ	$\lambda_{\text{gaze}} = 1$, $\lambda_{\text{sp}} = 1$, $\lambda_{\text{Dice}} = 1$	
Hardware	4$\times$ NVIDIA A40 (48 GB each)	
Wall-clock per epoch	24 min	5 min
Wall-clock per run	4.5 h	50 min

Table A2: **The three model variants.** All variants use 20 queries and a 64×64 field.

	Backbone	Embed dim	Encoder width	Decoder layers / FFN	Params (M)
EdgeGaze-N (Ours)	ViT-T	128	128	3 / 512	8.5
EdgeGaze-S (Ours)	ViT-T	192	192	4 / 512	16.2
EdgeGaze-M (Ours)	ViT-T+	256	256	4 / 1024	26.7

B Datasets and Evaluation Protocol

B.1 Datasets

GazeFollow provides 119,125 training images with 125,557 annotated people and 4,782 test images. Each test image annotates exactly one person, whose gaze point is labelled by six to ten annotators, and carries no in/out column, so we report no out-of-frame AP there. VideoAttentionTarget provides 58,507 training frames with 132,563 person-frames and 13,127 test frames with 31,978 person-frames, 35.4% of them looking out of frame. For transfer we use the dense GOO-Real test split (2,146 images; the released copy we could obtain holds ten images fewer than the one the literature reports) and the ChildPlay test split (9,934 frames, 22,649 scored person-frames). ChildPlay labels

four gaze classes that carry no in-frame point and no out-of-frame semantics; we drop those people entirely, 9.5% of the person-frames, as Dai et al. (2026) do.

B.2 Evaluation Protocol

Each benchmark defines AUC over its own tolerance region, and we follow each in turn. GazeFollow uses the pixel-level definition; VideoAttentionTarget and ChildPlay use a binary region drawn from a $\sigma = 3$ Gaussian at the annotated point; GOO-Real uses a 5×5 grid with a one-hot cell. Average and minimum distance are taken over the multiple annotators of GazeFollow. A prediction counts towards mAP when its head box reaches an IoU above 0.5 with an annotated head and its gaze point falls closer than 0.15 to the annotated point.

GazeFollow and VideoAttentionTarget annotate only a subset of the people in each scene. Under this partial annotation, strict detection mAP treats correct predictions on unlabelled people as false positives and can therefore substantially underestimate performance (Section D.2). Following Gaze-HTA (Lin et al., 2025), we perform one Hungarian assignment between predictions and available annotations before computing the reported GazeHTA-style metrics. The same assignment is shared across AUC, distance, out-of-frame AP and the GazeHTA-style mAP; unmatched predictions are excluded from the latter. This mAP column therefore measures head–gaze association on annotated instances while avoiding penalties caused solely by missing labels.

Our task-specific matching coefficients are listed in Table A1 and are fixed across all benchmark and transfer evaluations. For each training run, we select the checkpoint with the highest mAP on the corresponding test split and report all metrics from that checkpoint.

C Prior Generation

C.1 Prior Experts

The three experts label the training images once, offline. Saliency comes from DeepGaze IIE (Linardos et al., 2021) with a uniform centre bias rather than the one fitted to MIT1003, so that no other dataset's framing bias is baked into the target; the log densities are quantised per image into 8-bit PNGs that mirror the image tree. Objects come from Co-DETR ViT-L trained on the 80 COCO classes (Zong et al., 2023a), run at a single scale with no test-time augmentation, stored with their confidence and class, and read at a confidence of 0.3 during training. Heads come from the annotated boxes of each dataset together with the detector pseudo-labels distributed with HGTTR (Tu et al., 2022), read at a score of 0.7. Each expert map is attached to the image in pixel coordinates before augmentation and rasterised to the 64×64 field afterwards, so a crop or a flip moves image and target together. Saliency is area-pooled and renormalised to sum to one; the two box channels mark a cell when its centre falls inside any box. A channel whose expert map is missing for an image contributes no gradient for that image.

C.2 Cross-Dataset Prior Statistics

Figure 2(c) measures how much annotated gaze the three regions capture. For each dataset we take the share of gaze points that fall inside the head boxes, inside the detected object boxes without the person class, and inside the most salient fifth of the image, and divide it by the same share under a shuffled control that keeps the area and centre bias of each region. Heads lift gaze by 2.7–$3.1\times$ on GazeFollow, VideoAttentionTarget and ChildPlay, saliency by 2.1–$3.0\times$ and objects by 1.2–$1.7\times$. On GOO-Real, where shoppers scan shelves, the head region falls to $0.2\times$ while saliency and objects keep lifts near $1.5\times$. Spindles are 95% bootstrap intervals over source clusters: 347 for GazeFollow, 50 for VideoAttentionTarget, 21 for ChildPlay and 80 for GOO-Real. The saliency region is fixed at a fifth of the image, so its lift is comparable across datasets but not against the two box channels.



Figure A3: **Gaze heat maps with and without PGSR.** Each row is one annotated person from GazeFollow. Columns compare GazeHTA, the model without PGSR, our model, and the ground truth. Every heat map is normalised by its own peak.

D Annotation Completeness

D.1 Unlabelled Heads

Neither benchmark annotates every head. On the GazeFollow test split, 63.8% of the images contain at least one head that the benchmark leaves unlabelled; across the split the detector finds 8,844 such heads against 4,782 annotated people. Half of the training images are in the same position, and VideoAttentionTarget is only somewhat better, with unlabelled heads in 20.5% of its test frames.

D.2 Impact on Evaluation

This matters for any method that detects heads rather than receiving them. A detector that finds everybody is charged a false positive for every person the benchmark forgot. Counting every query as a prediction, our mAP on GazeFollow is 0.474. Treating the 8,844 pseudo-labelled heads as ignore regions instead, so that predictions landing on them are neither rewarded nor penalised, raises it to 0.587 at an IoU of 0.5 and 0.629 at 0.3; 12,708 of our predictions fall inside those regions. The metric reported in Table 1 is the GazeHTA-style mAP, which scores only the matched instances and is therefore insensitive to this effect. We train with the pseudo-labelled heads as auxiliary boxes, matched on geometry alone and excluded from the gaze losses, and accept the cost the strict protocol charges for it.

Table A3: **Inference cost, all measured by us.** Two-stage rows include the head detector they require. K is the number of heads in the image.

Method	Params (M)	Input resolution	GFLOPs	Latency (ms)
Two-stage Approaches				
ViTGaze (Song et al., 2024)	43.3	518^2	$50.2 + 93.5K$	44.1
Sharingan (Tafasca et al., 2024)	127.2	224^2	$86.2 + 11.6K$	8.9
Gaze-LLE (ViT-L) (Ryan et al., 2025)	328.3	448^2	$671.1 + 5.4K$	21.5
GazeMoE (Dai et al., 2026)	328.8	448^2	$671.2 + 9.4K$	22.6
PAGE (ViT-H+) (Ye et al., 2026)	1718.5	512^2	$1779.3 + 458.7K$	107.2
End-to-end Approaches				
HGTTR (Tu et al., 2022)	43.0	800×1333	201.6	5.5
Tonini et al. (Tonini et al., 2023)	53.8	800×1066	163.2	5.4
GazeHTA (Lin et al., 2025)	862.4	512^2	1902.5	43.0
EdgeGaze-N (Ours)	**8.5**	640^2	**10.9**	**3.1**
EdgeGaze-S (Ours)	16.2	640^2	24.7	4.0
EdgeGaze-M (Ours)	26.7	640^2	51.0	4.9

E Efficiency Measurement

E.1 Measurement Protocol

All three cost columns of Table A3 are measured by us under one protocol rather than copied from the papers, which report three mutually incompatible conventions: total parameters, learnable parameters only, and third-party estimates. Parameters are counted in full, including frozen backbones, and every two-stage method is charged the head detector it requires, the YOLOv5m-CrowdHuman model that Ryan et al. (2025) themselves use (21.0 M, 50.2 GFLOPs). FLOPs are counted by a real forward pass at each method's own inference resolution, and our column reports $2\times$ the multiply-accumulate count. At 224^2 our numbers land at exactly twice the two published figures that can be compared, which confirms the convention rather than a discrepancy. Latency is TensorRT FP16 at batch size one on a single 2080 Ti, with 50 warm-up iterations discarded, timing engine execution only, averaged over the first 100 GazeFollow test images; two-stage methods process every head the detector finds in each image, 599 heads in total.

E.2 Efficiency Results and Parameter Breakdown

Under this protocol all three of our variants lie at the low-cost end of Table A3: M uses 26.7 M parameters and 51.0 GFLOPs and runs in 4.9 ms, and N runs in 3.1 ms. PGSR accounts for 2.2 M of those parameters and requires no auxiliary expert at inference.

Figure A2 plots the same measurements against the number of heads. Fitting $a + bK$ per two-stage method gives slopes from 0.38 ms per head for Sharingan to 11.1 ms for PAGE. Each crosses the 30 FPS line at a different crowd size: PAGE is already past it with one person, ViTGaze at four, GazeMoE at twenty-four, Gaze-LLE at thirty and Sharingan at seventy. Across the same images every end-to-end method varies by less than 0.25 ms.

Table A4 splits the parameters of our M variant and of HGTTR by module, which is what the totals in Table A3 hide: our prediction heads are 7.0 M larger, our backbone is 23.3 M smaller, and the net difference is 16.3 M. The classification and out-of-frame heads of both models are too small to appear, together under two thousand parameters. The 862 M of GazeHTA is likewise not one model but three parts: 815.5 M of diffusion UNet, 34.2 M of VAE encoder and 12.7 M of its own modules.

Table A4: **Parameters by module** (M). EdgeGaze (Ours) is the M variant.

Module	HGTTR	EdgeGaze (Ours)
Backbone, encoder and decoder	41.5	18.2
Heat-map head	1.3	5.3
PGSR	–	2.2
Box head	0.13	0.93
Total	43.0	26.7

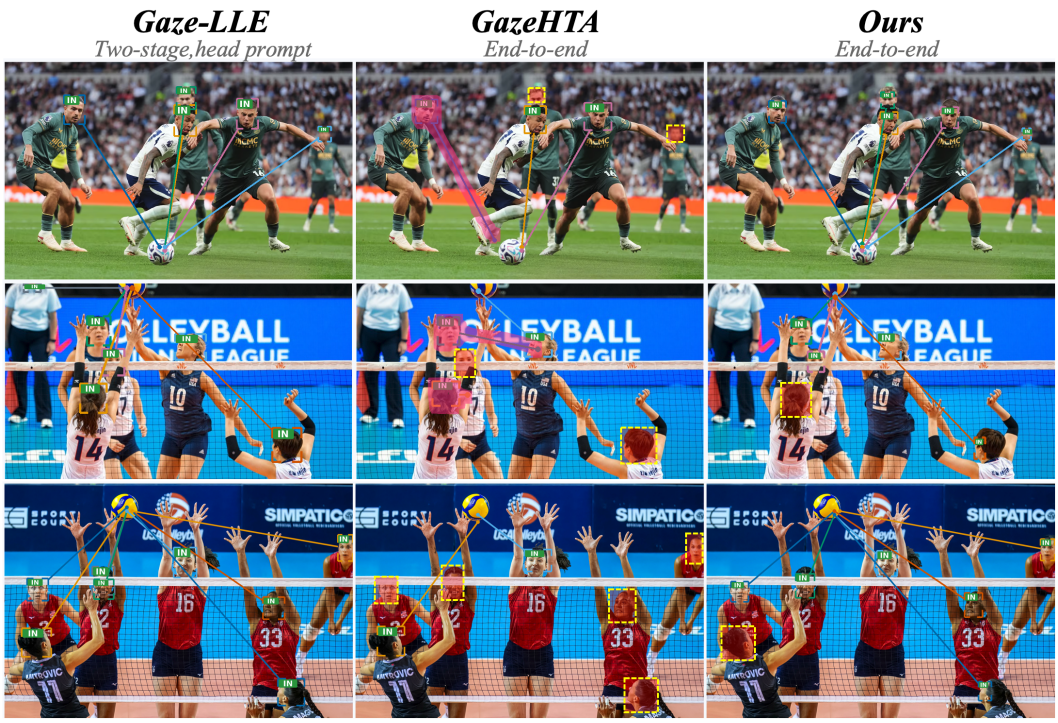

Figure A4: **Gaze target detection in the wild.** Dashed yellow boxes mark missed heads and magenta shading marks wrong gaze predictions. Gaze-LLE is prompted with detected head boxes.

Table A5: **All eight subsets of the three scene-prior losses on GazeFollow and VideoAttentionTarget.** For VideoAttentionTarget, each GazeFollow checkpoint is fine-tuned and evaluated at the final epoch.

$\mathcal{L}_{\text{sal}}$	$\mathcal{L}_{\text{obj}}$	$\mathcal{L}_{\text{head}}$	GazeFollow				VideoAttentionTarget (fine-tuned)			
			AUC↑	mAP↑	Avg. L2↓	Min. L2↓	AUC↑	mAP↑	L2↓	Out. AP↑
✗	✗	✗	0.9592	0.7857	0.0584	0.0286	0.9764	0.8102	0.0599	0.9997
✓	✗	✗	0.9597	0.7794	0.0572	0.0276	**0.9773**	0.8070	0.0575	0.9996
✗	✓	✗	**0.9601**	0.7875	0.0570	0.0278	0.9757	0.8091	0.0591	0.9996
✗	✗	✓	0.9596	0.7844	0.0566	0.0279	0.9766	0.8104	0.0574	0.9997
✓	✓	✗	**0.9601**	0.7821	0.0571	0.0275	0.9762	0.8102	0.0582	**0.9998**
✓	✗	✓	0.9594	0.7871	0.0574	0.0277	0.9771	0.8141	0.0572	0.9997
✗	✓	✓	0.9597	0.7892	0.0575	0.0279	0.9758	0.8031	0.0600	0.9997
✓	✓	✓	0.9600	**0.7982**	**0.0563**	**0.0271**	0.9763	**0.8211**	**0.0569**	0.9997

F Additional Ablations

F.1 Scene-Prior Losses

Table A5 extends Table 3(a) to all eight subsets of $\mathcal{L}_{\text{sal}}$, $\mathcal{L}_{\text{obj}}$ and $\mathcal{L}_{\text{head}}$, and fine-tunes each GazeFollow checkpoint on VideoAttentionTarget. After fine-tuning, using all three priors again gives the highest mAP, 0.8211 against 0.8031–0.8141, and the lowest L2 distance, 0.0569.

F.2 Prior-Guided Scene Representation

Figure A3 compares gaze heat maps of GazeHTA, our model without PGSR, and our full model on six annotated people from GazeFollow. The six are cases in which all three models detect the head and only the full model places its gaze estimate within 0.06 of the annotation. Without PGSR, the peak drifts to the grass in front of the frisbee player, the table in the dining scene, the torso of the man across the street or the edge of the archery target; with it, the peak sits on the ball, the target centre, the frisbee, the bowl and the two faces, and is sharper in the two rows where both models find the target.

G Qualitative Results

Figure A4 shows three in-the-wild sports scenes. GazeHTA leaves heads undetected in all three and places some gaze points away from the target. Our model recovers more heads and places its predictions on the ball.