Gaze Target Estimation Anywhere with Concepts

Xu Cao[1], Houze Yang[1]*, Vipin Gunda[1], Zhongyi Zhou[2], Tianyu Xu[2], Adarsh Kowdle[2], Inki Kim[1] , James M. Rehg[1†]
[1] University of Illinois Urbana-Champaign [2] Google
{xucao2,jrehg}@illinois.edu

Abstract

Estimating human gaze targets from images in-the-wild is an important and formidable task. Existing approaches primarily employ brittle, multi-stage pipelines that require explicit inputs, like head bounding boxes and human pose, in order to identify the subject of gaze analysis. As a result, detection errors can cascade and lead to failure. Moreover, these prior works lack the flexibility of specifying the gaze analysis task via natural language prompting, an approach which has been shown to have significant benefits in convenience and scalability for other image analysis tasks. To overcome these limitations, we introduce the ***Promptable Gaze Target Estimation (PGE)*** *task, a new end-to-end, concept-driven paradigm for gaze analysis. PGE conditions gaze prediction on flexible user text or visual prompts (e.g., "the boy in the red shirt" or "person in point [0.52, 0.48]") to identify a specific subject for gaze analysis. This approach integrates subject localization with gaze estimation, and eliminates the rigid dependency on intermediate analysis stages. We develop a scalable data engine to generate* ***Gaze-Co*** *(Gaze Estimation with Concepts), a dataset and benchmark of 120K high-quality, prompt-annotated image pairs. We also propose* ***GazeAnywhere****, the first model designed for PGE. GazeAnywhere uses a transformer-based detector to fuse features from frozen encoders and simultaneously solves subject localization, in/out-of-frame presence, and gaze target heatmap estimation. GazeAnywhere achieves state-of-the-art performance on multiple PGE benchmarks, setting a strong baseline for this new problem even on a difficult out-of-domain, real-world clinical dataset. GazeAnywhere is open-sourced in github.com/IrohXu/GazeAnywhere.*

1. Introduction

Human gaze is a fundamental non-verbal cue, conveying a wealth of social and cognitive information [28, 51]. It is integral to human interaction, used to initiate social contact, signal attention and interest, manage conversational turn-taking, and regulate intimacy. As a direct proxy for cognition, gaze can also reveal a person's intentions, preferences, and emotional states [54, 62]. Consequently, its study has attracted significant interest across diverse fields, including psychology, human-computer interaction, and clinical research on conditions such as autism spectrum disorder (ASD) [1, 2, 22].

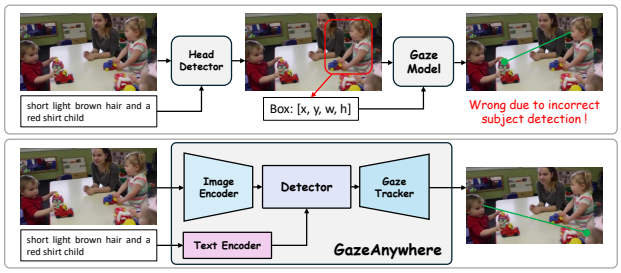


Figure 1. Gaze target estimation in in-the-wild environments. Prior methods such as Sharingan and ViTGaze have to rely on an Open-Vocabulary Detector (OVD) to produce auxiliary head boxes via dynamic human prompts, introducing a sequential dependency that becomes a major bottleneck.

Despite its importance, accurately estimating a person's gaze target in unconstrained, "in-the-wild" settings via image analysis remains a formidable challenge. Current approaches typically require explicit prior information (such as human head and face bounding boxes, pose estimations, and depth) not only for training but also as critical inputs during inference [58, 60]. Consequently, the most common architecture is a multi-stage pipeline where intermediate inputs are generated by a series of pre-processing steps prior to gaze estimation (Figure 1). This process typically requires first detecting and tracking a person and then precisely localizing their head or face with a bounding box before the gaze direction can be computed [12, 79]. This sequential dependency creates a critical bottleneck, as inaccuracies in these initial stages (common in crowded scenes, poor lighting, or with challenging cases like detecting children's faces) can

*Co-first author
†Corresponding author

cascade, leading to a failure of the entire system.

To overcome this limitation and provide greater convenience and flexibility in specifying gaze analysis tasks, we propose a paradigm shift toward an end-to-end, concept-based framework for gaze target estimation. This approach is inspired by recent advancements in vision foundation models, such as the Open-vocabulary Detectors (OVDs) [25, 81], and the Segment Anything Model (SAM) series [9], which demonstrate remarkable abilities to detect and segment objects based on semantic-level text or visual concept prompts rather than explicit localization cues via bounding boxes. We extend this core idea to the domain of human gaze understanding, designing a promptable model that can directly identify the gaze target of a specified person within an image. By conditioning on a concept of the subject (e.g., "the boy in the red shirt"), our approach eliminates the dependencies on intermediate head bounding boxes or pose keypoints. This allows the model to infer gaze targets from a semantic understanding of the scene in a user-friendly end-to-end fashion, paving the way for more robust and versatile systems. Our main contributions can be summarized as:

(1) We define the **Promptable Gaze Target Estimation (PGE)** task, which extends the traditional gaze target estimation problem to an unconstrained, text promptable end-to-end paradigm.

(2) We design a scalable data engine to generate 120K high quality PGE training annotations consisting of subject text to gaze alignment data pairs.

(3) We introduce **GazeAnywhere**, the first promptable concept-driven gaze target estimation model. Our model achieves state-of-the-art performance in several benchmarks including an out-of-domain private dataset for autism children's gaze target estimation.

2. Related Work

Human Gaze Target Estimation. Interpreting gaze is a crucial component of human behavior understanding [21, 77]. This has motivated the "gaze-following" task, introduced by datasets like GazeFollow [54, 55], VideoAttentionTarget [15], GOO [64] and ChildPlay [61], which requires a model to predict the scene location a person is looking at. Dominant approaches have thus far employed multi-branch, fusion-based architectures. These models separately process explicit cues such as head position [58, 60, 72, 73], pose [4], text-based directions [46, 69], facial expressions [37], and depth [29, 33, 45, 65], subsequently combining these features to predict gaze points [11, 30, 32, 41, 45, 68, 76, 80]. In addition, existing end-to-end models cannot specify the subject person during inference either [18, 66, 68]. While effective, these strategies are dependent on the availability and accuracy of these intermediate representations. Concurrently, other studies have expanded the task's scope, such as multi-view gaze target estimation [47] and GazeHOI [62] for open-vocabulary targets. Despite these advances, a common limitation still persists: a dependency on auxiliary information, such as precise head bounding boxes or pose estimations. This reliance poses significant challenges in realistic settings, where such priors are often unreliable or unavailable.

Promptable and Interactive Object Detection. Recent advances in promptable and open-vocabulary perception are enabling models to generalize beyond fixed label sets, especially the referring expression comprehension tasks [8, 31, 49, 50, 74, 78]. OVD methods, for instance, leverage large-scale vision-language encoders like CLIP to detect arbitrary concepts specified by text at inference time, even for categories unseen during training [25, 42, 48, 83]. In parallel, interactive segmentation frameworks demonstrate how models can respond to flexible text and visual prompts [9, 36]. Together, these advances suggest the new paradigm: visual perception systems can be effectively guided by conceptual cues, rather than rigidly predefined supervision. Building on this paradigm, we formulate gaze target estimation as a concept-conditioned reasoning task, where both the subject and their attended region are inferred from semantic prompts rather than explicit localization inputs.

Vision Foundation Models. Vision Foundation Models (VFMs) have become a dominant approach in computer vision, entailing a significant shift of models trained on massive web-scale datasets. VFMs primarily include two branches: (1) weakly-supervised models like CLIP [52], SigLIP [67, 75], and MetaCLIP [6, 16, 70], which learn powerful representations from image-text pairs using contrastive losses, and (2) self-supervised learning (SSL) models like DINO series [34, 34, 59], which learn robust visual features from unlabeled images. These powerful, pre-trained encoders now serve as general-purpose backbones for a wide range of downstream tasks. This trend has also influenced gaze estimation, where systems like ViTGaze [60] and GazeLLE [58] have successfully adapted VFM architectures and leveraged their pre-trained features to improve performance, demonstrating the value of these models for fine-grained, human-centric tasks.

3. Method

3.1. Promptable Gaze Target Estimation

We define the PGE task as estimating the gaze target location of a specific subject within an image or video, identified by a user-provided prompt. This prompt can be of two types: text prompting via a short, natural language query; and visual one by prompting a spatial coordinate, such as the center point of a head bounding box. In text prompting, our goal is to support any simple, visually-groundable noun phrase as a text prompt. However, this introduces intrinsic

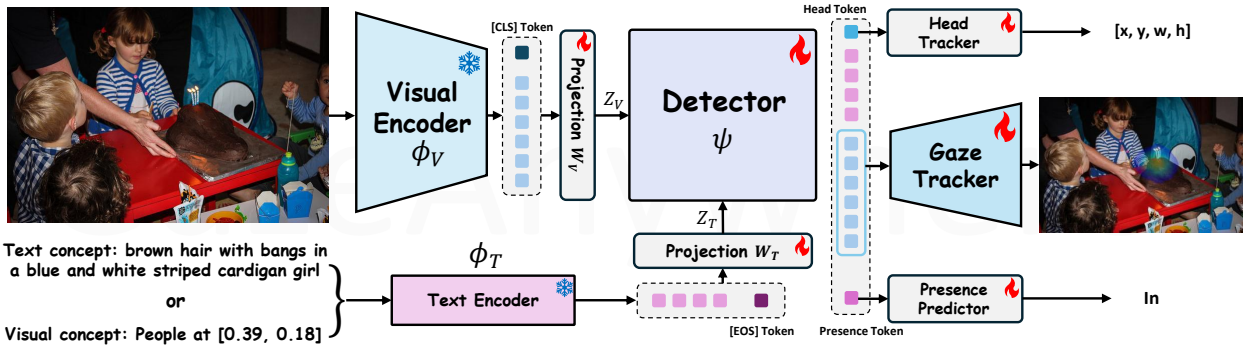


Figure 2. An overview of the GazeAnywhere end-to-end framework for PGE. The model uses frozen visual (DINOv3) and text (dino.txt) encoders to provide features to a trainable transformer-based detector. This detector utilizes multiple decoders to simultaneously predict a subject's head bounding box , in-frame presence , and gaze target. GazeAnywhere is conditioned on flexible user prompts, such as a natural language text description or visual (coordinate-based) cues.

ambiguity (e.g., "the person in the back"). To mitigate this and enable unambiguous identification, we structure text prompts around four main categories. A user can combine descriptions from these categories to specify a subject: (1) Appearance: Noun phrases describing a person, consisting of an identity (e.g., woman, man, child) and optional modifiers (e.g., hair type/color, clothes, glasses, hat). (2) Location: The subject's spatial position in the image (e.g., center, left, top-right). (3) Pose: The subject's static posture. (4) Action: Verb phrases describing what the subject is doing.

More formally, given an input RGB image $I \in \mathbb{R}^{3\times H\times W}$ and a prompt P (either text T or a visual cue), the goal of PGE is to produce a gaze heatmap $\hat{H} \in \mathbb{R}^{H_{\text{out}}\times W_{\text{out}}}$. Each element $\hat{H}(i,j)$ represents the probability that the subject specified by P is gazing at the spatial location (i,j). Unlike classic gaze estimation methods, PGE demands that the model solve the task end-to-end, directly linking a flexible, high-level query to a final gaze heatmap. This formulation is substantially more challenging as it precludes the use of auxiliary inputs common in traditional pipelines (e.g., subject bounding boxes, pose keypoints, or depth maps). Our model must implicitly learn to perform subject identification, localization, and gaze estimation jointly with text input, rather than relying on the explicit outputs of separate, specialized models like open-vocabulary detectors or pose estimators.

3.2. GazeAnywhere Architecture

Figure 2 shows the overall architecture of GazeAnywhere. The model consists of a frozen image encoder, a frozen text encoder to proceed visual modality and text modality. A transformer-based detector is used to learn joint representations and map text prompt into the main gaze target estimation and auxiliary tasks.

Image Encoder. We use a frozen ViT, denoted $\phi_V(\cdot)$, as our image encoder to extract general visual features. Consistent with PGE's problem definition, we do not employ any auxiliary models for dedicated depth or pose feature extraction. The image encoder processes an input image $I \in \mathbb{R}^{3\times H\times W}$ by dividing it into a sequence of N_V patch tokens, to which a learnable `[CLS]` token is prepended. The resulting output sequence from ϕ_V is $\phi_V(I) = [c, s_1, s_2, \ldots, s_{N_V}] \in \mathbb{R}^{(1+N_V)\times D_V}$. Here, D_V is the embedding dimension, $c \in \mathbb{R}^{D_V}$ is the final `[CLS]` token embedding, and $s_i \in \mathbb{R}^{D_V}$ is the output embedding for the i-th patch.

Text Encoder. We employ a frozen text encoder, $\phi_T(\cdot)$, which consists of a series of transformer blocks and a final linear layer. The linear layer maps the output `[EOS]` token's feature to the image embedding space. To prepare the input, a tokenizer first converts the text T into a sequence of token IDs. These IDs are then mapped to initial text embeddings T_E via an embedding layer, and the sequence is padded to a fixed context length, L_T. The encoder ϕ_T processes this embedding sequence T_E, producing the final output sequence: $\phi_T(T_E) = [t_1, t_2, \ldots, t_{N_T}, t_{eos}, \ldots, t_{pad}] \in \mathbb{R}^{L_T\times D_T}$. Here, D_T is the output embedding dimension, $t_i \in \mathbb{R}^{D_T}$ is the output embedding for the i-th content token, and $t_{eos}, t_{pad} \in \mathbb{R}^{D_T}$ are the special tokens representing the end of the input and padding, respectively.

Projection Layers. The image encoder ϕ_V and text encoder ϕ_T output features with potentially different dimensions, D_V and D_T, respectively. To map these features into a unified, shared space, we introduce two trainable linear projection layers, W_V and W_T. These layers project the high-dimensional features into a common, lower-dimension D, where $D < \min(D_V, D_T)$:

$$Z_{\text{V}} = W_V \cdot \phi_V(I) \tag{1}$$

$$Z_{\text{T}} = W_T \cdot \phi_T(T_E) \tag{2}$$

Here, W_V and W_T are the learnable projection matrices, and the operation $\cdot$ denotes a per-token linear transformation.

This process results in a sequence of visual tokens $Z_{\text{V}} \in \mathbb{R}^{(1+N_V)\times D}$ and text tokens $Z_{\text{T}} \in \mathbb{R}^{L_T \times D}$, which now share the same embedding dimension.

Task-Specific Embeddings. Beyond the primary cross-modal feature alignment, we introduce two specialized, learnable embeddings to explicitly model key sub-problems: a **head token** and a **gaze presence token**.

1. **Head Token:** This is a learnable embedding designed to explicitly predict the head localization of the prompted subject, serving as the image-text alignment objective in our task. It is initialized using the embedding of the text `[EOS]` token.
2. **Target Presence Token:** This token is introduced to address the *in/out-of-frame* gaze target boolean prediction objective. The rationale for this is that the in/out decision relies on global contextual cues from the entire image, which conflicts with the inherently local nature of the target localization objective. Forcing a single query or mechanism to handle both can be counterproductive. Therefore, we decouple the localization and in/out prediction tasks. This dedicated, learnable global token is responsible for the in/out prediction and is initialized using the embedding of the visual `[CLS]` token.

Detector Transformer. After extracting and projecting the visual and text features, we introduce a **Detector Transformer**, $\psi(\cdot)$, to fuse these representations and refine them for the gaze target estimation task. $\psi(\cdot)$ apply the same Transfomrer block in DINOv3 [59]. The input to ψ is a single sequence F constructed by concatenating the projected features and our specialized task tokens.

First, we define the **head token** $t_h \in \mathbb{R}^D$ and **target presence token** $t_p \in \mathbb{R}^D$. These are formed by combining the projected global tokens (c' from vision, t'_{eos} from text) with dedicated learnable embeddings, $\text{E}_{\text{presence}}$:

$$t_h = t'_{eos} \tag{3}$$

$$t_p = c' + \text{E}_{\text{presence}} \tag{4}$$

Let $\mathbf{s}' = [s'_1, \ldots, s'_{N_V}]$ be the sequence of projected visual patch tokens from Z_V (excluding c') and $\mathbf{t}' = [t'_1, \ldots, t'_{N_T}]$ be the projected text content tokens from Z_T (excluding t'_{eos} and padding). The full input sequence F is then assembled as:

$$F = [t_h, \mathbf{t}', \mathbf{s}', t_p] \in \mathbb{R}^{(N_T+N_V+2)\times D} \tag{5}$$

We inject positional information by adding 1D sinusoidal position embeddings to the text tokens $\mathbf{t}'$ and 2D sinusoidal position embeddings to the visual tokens $\mathbf{s}'$ [20]. The Detector transformer ψ is a stack of k standard transformer blocks; k is a hyperparameter ablated in our experiments. ψ processes F and outputs a refined sequence of the same dimension, $\psi(F) \in \mathbb{R}^{(N_T+N_V+2)\times D}$. Specific tokens from this output are then passed to dedicated decoders.

Decoders. The Detector transformer ψ outputs a refined sequence of tokens. We attach three distinct prediction heads to specific tokens from this sequence to produce the final outputs.

- **Gaze Tracker (Heatmap Decoder):** The refined visual patch tokens $\hat{\mathbf{s}} \in \mathbb{R}^{N_V \times D}$ are first re-assembled from their 1D sequence form back into a 2D spatial grid. This feature map is then fed through a convolutional decoder, consisting of two transposed convolutional layers, which upsamples the features to the output heatmap $\hat{H} \in \mathbb{R}^{H_{out} \times W_{out}}$. In our experiments, we set $H_{out} = W_{out} = 64$.
- **Head Tracker (Box Decoder):** We use the refined head token $\hat{t}_h \in \mathbb{R}^D$ for an auxiliary head localization task. The token is passed through a 3-layer feed-forward network (FFN) with ReLU activations and a hidden dimension of D. This head regresses a 4-dimensional vector $[x, y, w, h]$ representing the normalized center coordinates, width, and height of the subject's head box.
- **Presence Predictor (In/Out Decoder):** The refined gaze presence token $\hat{t}_p \in \mathbb{R}^D$ is used to predict whether the gaze target is in or out of the frame. It is processed by a 2-layer FFN (with one hidden layer of dimension D and ReLU activation) that outputs a single logit for the binary classification.

3.3. Learning Objective

We train our model end-to-end with a joint multi-task objective. The total loss $\mathcal{L}_{\text{total}}$ is a weighted linear combination of three loss terms: one for the gaze heatmap, one for the gaze presence, and one for the auxiliary head localization task.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gaze}} + \mathcal{L}_{\text{presence}} + \mathcal{L}_{\text{head}} \tag{6}$$

The gaze heatmap loss $\mathcal{L}_{\text{gaze}}$ is a pixel-wise binary cross-entropy (BCE) loss. The supervisory target is a heatmap Y, constructed by placing a 2D Gaussian ($\sigma = 3$) at the ground-truth gaze target location. Let $\hat{Y}$ be the predicted heatmap. The loss is defined as:

$$\mathcal{L}_{\text{gaze}} = -\frac{1}{N}\sum_{p=1}^{N} [y_p \log(\hat{y}_p) + (1 - y_p)\log(1 - \hat{y}_p)] \tag{7}$$

where $N = H_{out} \times W_{out}$ is the total number of pixels, and y_p and $\hat{y}_p$ are the ground-truth and predicted values for a single pixel p, respectively.

The gaze presence loss $\mathcal{L}_{\text{presence}}$ is a Focal Loss supervised with a binary label $Y_{\text{presence}} \in \{0, 1\}$. Let $\hat{Y}_{\text{presence}} \in [0, 1]$ be the model's predicted probability that the target is present ($Y_{\text{presence}} = 1$). The loss is defined as:

$$\mathcal{L}_{\text{presence}} = \mathcal{L}_{\text{focal}}(Y_{\text{presence}}, \hat{Y}_{\text{presence}}) \tag{8}$$

where the hyperparameter of the focal loss is the default value from [40].

The subject head bounding box loss $\mathcal{L}_{\text{head}}$ is a linear combination of the $\mathcal{L}_1$ loss and the generalized IoU loss. which is widely used by object detection tasks. It defined as:

$$\mathcal{L}_{\text{head}} = \lambda_{l_1} ||b - \hat{b}||_1 + \lambda_{\text{iou}} \mathcal{L}_{\text{iou}}(b, \hat{b}) \tag{9}$$

where b and $\hat{b}$ is the ground truth head box and predicted head box. $\mathcal{L}_{\text{IoU}}$ is the GIoU loss [8, 57]. λ_{l_1} and λ_{iou} are head object detection hyperparameters. We followed DETR [8] and OWLViT [48] to set $\lambda_{l_1} = 5$ and $\lambda_{\text{iou}} = 2$.

4. Gaze with Concept (Gaze-Co) Dataset

Training GazeAnywhere for the PGE task requires a large and diverse dataset annotated with concepts, a resource that no existing gaze dataset provides. To address this, we developed a scalable data engine that generates annotations via a human-in-the-loop feedback process. This engine worked in tandem with two human annotators (co-authors) to perform several key functions: aligning heterogeneous annotations, filtering low-quality frames, generating concise concept phrases, and facilitating human verification. After three rounds of iteration, we created Gaze-Co, the first large-scale dataset for PGE, containing 120K samples sourced from the training set of GazeFollow, VisualAttentionTarget (VAT) and ChildPlay. To establish a comprehensive benchmark, we also converted the test sets of these well-known gaze datasets to the PGE format, creating GazeFollow-Concept, VAT-Concept, and ChildPlay-Concept. We further conducted experiments on a private, Institutional Review Board (IRB)-approved, out-of-domain (OOD) evaluation set with several frames in 40 child social communication (Child-SC) videos.

4.1. Data Engine

Figure 3 illustrates the workflow of the data engine. We can divide the process into three stages: (1) data alignment and filter; (2) concept generation; (3) verification.

Data Alignment and Filter. The source datasets differ in coordinate conventions, split policies, and metadata. We therefore adopt a unified schema with explicit pixel coordinates for the head box $(x_{min}, y_{min}, x_{max}, y_{max})$, and a normalized gaze point $(g_x/W, g_y/H)$. Then, to ensure reliable person-scale learning signals, we apply geometric and sharpness filters. Annotations are retained only if the head box width ≥ 30 px, height ≥ 40 px, area $\geq 2500\,\text{px}^2$, and the box-to-image ratio $\in [0.008,\ 0.3]$, with sufficient Tenengrad focus. These thresholds remove extremely small, oversized, or blurry instances while preserving diverse valid samples.

Concept Generation. For each retained annotation, we produce a short, lowercase concept phrase comprising attribute,

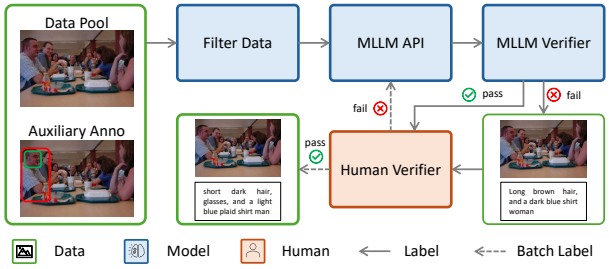


Figure 3. Overview of the GazeAnywhere data engine.

position, action, and pose, together with a coarse count of visible people. Concept generation is executed with a production Vision Language Model (VLM) accessed through API (Gemini 2.5 Pro [17, 26]), using batch processing with fixed prompts and rates. The attribute phase privileges stable visual cues (hair, glasses, beard, colors, and patterns) and the final token is constrained to one of man, woman, boy, girl, infant, child as an apparent (perceived) age/sex presentation label, used solely as a visual category cue rather than a verified identity attribute; when indeterminate, we write "adult" or "child. The position uses brief canvas references (e.g., "bottom left corner"). Action and pose are explicitly non-overlapping: action describes ongoing interaction or motion with object or direction when visible, while pose captures static body configuration and facing direction. When a field is indeterminate, we write "none."

Verification. We adopt an Multi-modal Large Language Model (MLLM)-first, human-in-the-loop verification workflow. The Gemini 2.5 Pro reviews all generated concepts and flags each as pass or fail. Human annotators then spot-check a random subset of the MLLM passed cases and evaluate the batch success rate. During review, both the MLLM and human annotators check whether each concept correctly matches its designated head box (consistency), whether all four fields are present and non-conflicting (completeness), and whether the text contains no sensitive or identifying information (privacy). If the human verifier finds the batch success rate is low, the data engine will return to the concept generation stage. The human verifier then adjusts the prompts and rules and re-runs the concept generation and verification until the observed error rate is kept low ($\leq 1\%$). For the private Child-SC dataset with IRB restriction, all concept annotations are generated manually by authorized human annotators without being sent to the MLLM.

4.2. Gaze-Co Dataset and Benchmarks

Gaze-Co is the first large-scale dataset for promptable gaze target estimation, unifying GazeFollow, VAT, and ChildPlay under a shared schema with concept-level annotations for both training and evaluation.

Training Data. The Gaze-Co 120K training set contains about 120K images from the official training splits of the

Gaze Model	Head Detection	#Total Param	Time per sample[ms]↓	GazeFollow-Concept			VAT-Concept		ChildPlay-Concept		Child-SC (OOD)	
				AUC ↑	Avg L2 ↓	Min L2 ↓	L2 ↓	AP ↑	L2 ↓	AP ↑	L2 ↓	AP ↑
ViTGaze [60]	Ground-Truth	-	-	0.955	0.097	0.045	0.098	0.879	0.113	0.905	-	-
	GroundingDINO-B [42]	255M	116	0.913	0.192	0.116	0.232	0.792	0.170	0.863	0.139	0.795
	LLMDet-L [25]	365M	339	0.897	0.189	0.113	0.235	0.763	0.158	0.869	0.165	0.825
	OWLv2-L [49]	460M	179	0.927	0.170	0.094	0.241	0.749	0.151	0.862	0.168	0.781
	RexSeek [31]	3B	1160	0.945	0.119	0.063	0.123	0.841	0.117	0.903	0.136	0.818
Sharingan [63]	Ground-Truth	-	-	0.944	0.114	0.059	0.109	0.852	0.118	0.854	-	-
	GroundingDINO-B [42]	338M	119	0.860	0.259	0.178	0.352	0.720	0.216	0.888	0.349	0.771
	LLMDet-L [25]	449M	342	0.865	0.241	0.160	0.281	0.785	0.213	0.874	0.347	0.832
	OWLv2-L [49]	544M	182	0.867	0.255	0.173	0.291	0.745	0.185	0.893	0.263	0.847
	RexSeek [31]	3B	1159	0.888	0.207	0.142	0.273	0.601	0.178	0.863	0.166	0.810
Gaze-LLE [58]	Ground-Truth	-	-	0.961	0.099	0.045	0.101	0.875	0.113	0.912	-	-
	GroundingDINO-B [42]	539M	145	0.925	0.165	0.106	0.234	0.780	0.152	0.863	0.172	0.856
	LLMDet-L [25]	650M	368	0.912	0.168	0.109	0.230	0.793	0.145	0.875	0.175	0.855
	OWLv2-L [49]	745M	208	0.941	0.146	0.089	0.229	0.792	0.127	0.893	0.161	0.830
	RexSeek [31]	3B	1183	0.954	0.108	0.054	**0.121**	0.861	0.119	0.914	0.172	0.846
GazeAnywhere-CLIP-L		430M	35	0.953	0.105	0.056	0.137	0.874	0.104	**0.915**	0.146	0.868
GazeAnywhere-DINOv3-L		870M	96	**0.958**	**0.099**	**0.050**	0.123	**0.879**	**0.098**	0.906	**0.090**	**0.902**

Table 1. PGE results on four datasets. The input is the text prompt of the subject person's appearance, position, action and pose. For baseline methods, the OVD is used to extract the bounding box with the input prompt and then feed the bounding box to the gaze models. Latency is compared with inference running speed per image in *batch size* = 1.

three source datasets. Each record includes the target head box, normalized gaze point, in/out-of-frame label, and a compact concept phrase (attribute, position, action, and pose). All samples pass image quality filters, ensuring diverse, valid instances across viewpoint, poses, scales, and interaction contexts.

Benchmark Settings. The benchmark uses the official test splits of GazeFollow, VAT, and ChildPlay, each converted into the Gaze-Co format. We evaluate concept-conditioned gaze prediction under three settings: (i) in-domain testing on the test sets; and (ii) OOD evaluation on the Child-SC dataset, a developmental sample of children's gaze behavior collected under an IRB-approved study (see the appendix for the dataset description). Every text prompt in the test set has been human-verified rather than spot-checked, ensuring accuracy and consistency. Each model receives the image, with the concept text added or altered according to the test setting. This setup provides a consistent framework for comparing models under controlled concepts-based conditions.

4.3. Metrics

We evaluate models using heatmap Area Under the Curve (AUC) in GazeFollow and pixelwise L2 in all. For heatmap AUC, the predicted heatmap is treated as a confidence map to compute an ROC curve against the binary gaze target map. Pixelwise L2 measures the Euclidean distance between the heatmap peak and the ground-truth gaze point. For GazeFollow-Concept, each image includes multiple gaze annotations directed at the same target person, so we additionally report Avg L2 (distance to the mean of all targets) and Min L2 (distance to the nearest target). For VAT-Concept, ChildPlay-Concept, and Child-SC (the IRB-approved OOD set), annotations include binary in/out labels relative to the target region; thus, we report pixelwise L2, and average precision (AP) to jointly evaluate localization and in/out binary classification.

5. Experiments

We evaluate GazeAnywhere on the PGE task, comparing its text-prompting capabilities against State-of-the-Art (SOTA) two-stage pipelines that integrate OVDs for head/human detection with a separate gaze modeling stage. We also present a series of ablation studies demonstrating the importance of the frozen encoders, validating our loss design, and analyzing the differences between visual and text prompting. Finally, we demonstrate a real-world application, the "AnyGaze Agent," a system that integrates GazeAnywhere with an Augmented Reality (AR) device and a MLLM.

5.1. Implementation Details

All models are trained for 25 epochs using the Adam optimizer and a cosine learning rate scheduler with an initial rate of 1e-3 and batch size 128, followed by an additional 5 epochs with a reduced learning rate of 1e-5. All training experiments are conducted with 4 NVIDIA H100 GPUs. The inference is running with 1 NVIDIA L40S GPU. We adopt the DigiLens ARGO smartglass as the AR platform to deploy AnyGaze Agent for real world experiments. More details are shown in the appendix.

5.2. Main Results

Table 1 compares GazeAnywhere against strong two-stage baselines, which we created by pairing three SOTA gaze methods Gaze-LLE [58], Sharingan [63], ViTGaze [60] with three leading OVDs for human detection [25, 42, 49]. De-

Model	Param	GazeFollow-Concept		VAT-Concept	
		Avg L2 ↓	Min L2 ↓	L2 ↓	AP ↑
Qwen3-VL-8B [3]	8B	0.201	0.137	0.286	0.651
Gemini 2.5 Flash [17]	-	0.216	0.156	0.292	0.661
GazeAnywhere-DINOv3-L	870M	**0.099**	**0.050**	**0.123**	**0.879**

Table 2. Compare GazeAnywhere with SOTA VLMs.

Prompt type	GazeFollow-Concept			VAT-Concept		
	AUC ↑	Avg L2 ↓	Min L2 ↓	AUC ↑	L2 ↓	AP ↑
No prompting	0.944	0.144	0.090	0.875	0.210	0.796
visual prompting	0.958	0.100	0.050	0.914	0.131	0.894
text prompting (appearance)	0.952	0.113	0.062	0.904	0.153	0.840
text prompting (position)	0.953	0.124	0.073	0.893	0.188	0.826
text prompting (action)	0.949	0.129	0.078	0.897	0.180	0.839
text prompting (pose)	0.952	0.121	0.070	0.909	0.163	0.859
text prompting (all)	0.958	0.099	0.050	0.928	0.123	0.879

Table 3. Comparison of different prompt strategies.

tails of these baselines are shown in the appendix. The encoders of GazeAnywhere can be CLIP-L [52] or DINOv3-L [59] with dino.txt [34]. On the PGE text-prompting task, GazeAnywhere achieves SOTA performance on all metrics across the three public datasets, as well as on our challenging OOD private dataset from a real-world assessment setting in which children's social communication skills are quantified by experts.

5.3. Ablation Study

We use GazeAnywhere-DINOv3-L for all following up ablation experiments in GazeFollow-Concept and VAT-Concept. More results are shown in Appendix Sec 13.

Compare GazeAnywhere with SOTA VLMs. To demonstrate the utility of GazeAnywhere in PGE, we evaluate the 0-shot performance of SOTA VLM on gaze point prediction. Table 2 shows the comparison of GazeAnywhere, Qwen3-VL-8B and Gemini 2.5 Flash. GazeAnywhere surpass all of them, highlighting the importance of building specific model for PGE.

PGE with Different Prompting. GazeAnywhere supports both visual (coordinate-based text) and text (natural language) prompts, as illustrated in Figure 2. In Table 3, we compare the performance of these different strategies. We find that text-based prompting achieves performance on par with visual prompting. Furthermore, our decomposition analysis of text prompt composition reveals that the subject's appearance and pose description are the most critical components for the PGE task.

Loss ablations. We conducted an ablation study (Table 4) on our objective function's components: gaze heatmap, presence, and head losses. The essential gaze heatmap loss was always active, while we trained models removing the presence loss, the head loss, and both. Results indicate the presence loss only supports the auxiliary in/out prediction, not help gaze estimation. The head loss, however, improves both

$\mathcal{L}_{gaze}$	$\mathcal{L}_{presence}$	$\mathcal{L}_{head}$	GazeFollow-Concept			VAT-Concept		
			AUC ↑	Avg L2 ↓	Min L2 ↓	AUC ↑	L2 ↓	AP ↑
✓	✗	✗	0.956	0.102	0.052	0.925	0.135	-
✓	✓	✗	0.955	0.103	0.055	0.924	0.136	0.863
✓	✗	✓	0.958	0.099	0.051	0.924	0.128	-
✓	✓	✓	**0.958**	**0.099**	**0.050**	**0.928**	**0.123**	**0.879**

Table 4. Ablation experiment on loss selection.

Encoder	Param	GazeFollow-Concept			VAT-Concept		
		AUC ↑	Avg L2 ↓	Min L2 ↓	AUC ↑	L2 ↓	AP ↑
CLIP-B [52]	152M	0.942	0.123	0.070	0.908	0.145	0.837
CLIP-L [52]	430M	0.953	0.105	0.056	0.913	0.137	0.874
SigLIP2-B [67]	379M	0.949	0.115	0.064	0.904	0.148	0.860
SigLIP2-L [67]	886M	0.953	0.105	0.056	0.910	0.147	0.858
DINOv3-L [34, 59]	870M	**0.958**	**0.099**	**0.050**	**0.928**	**0.123**	**0.879**
MetaCLIP2-H [16]	1.9B	0.951	0.109	0.059	0.912	0.150	0.857

Table 5. Comparison of different encoders for GazeAnywhere.

the gaze target estimation and the target presence prediction.

Comparison of Different Encoders. We conduct an ablation on the encoder backbone, comparing CLIP [52], SigLIP 2 [67], MetaCLIP 2 [16] and DINOv3 [59] (with dino.txt [34]). In all experiments, the encoders were frozen, with only the projection layer, transformer detector, and decoder heads being fine-tuned. As shown in Table 5, the DINOv3-based model achieves the best performance on nearly all metrics, highlighting its superior visual-text alignment and understanding for PGE.

5.4. Visualization

Figure 4 showcases qualitative gaze estimation results from GazeAnywhere. The input, displayed in the black boxes, is a text prompt describing only the subject's appearance, such as "light brown hair and a blue striped shirt boy". The visualizations demonstrate that GazeAnywhere performs robustly not only in simple scenarios with 2-3 people but also in complex, crowded scenes with four or more individuals. Notably, the final two examples, "long black high ponytail hair and a pink shirt girl" and "short blonde hair wearing a dark blue and light gray shirt boy", are from an OOD Child-SC video dataset. The model's successful performance on this unseen data highlights its generalization and robustness.

5.5. GazeAnywhere as Agent in AR

Previous gaze estimation models ignore the real-world application experiment. Inspired by recent tool-enhanced MLLM workflows [9, 71], we developed the GazeAnywhere Agent (Figure 5). This system uses a central MLLM (Gemini 2.5 [17]) that leverages GazeAnywhere as a specialized tool to solve advanced user queries, such as, "How many gaze shifts does this girl with the white dress present?". The workflow captures User Audio and Environment Images from an AR Glass. MLLM calls Whisper v3 [53] to transcribe the audio, followed by query reasoning and prompt rephrasing. The agent converts the high-level query into a low-level text

Figure 4. Visualization of GazeAnywhere's gaze target estimation results from several datasets. The input is the text prompt (shown in the black box) describing only the subject person's appearance and the image. GazeAnywhere detect the subject's head and track the gaze target. More qualitative comparison is shown in the appendix.

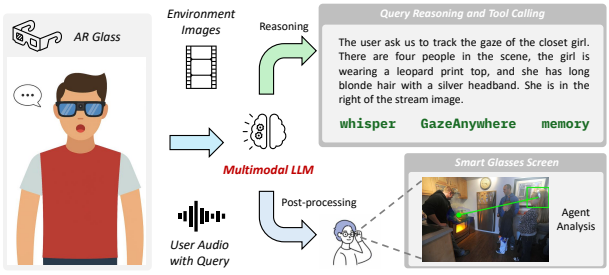


Figure 5. Workflow of MLLM-powered GazeAnywhere Agent.

MLLM	Gaze Shift MAE/min ↓	Eye Contact MAE/min ↓
Gemini 2.5 Flash (Raw)	9.247	14.763
Gemini 2.5 Flash (AnyGaze Agent)	2.337	4.756
Gemini 2.5 Pro (Raw)	6.063	10.268
Gemini 2.5 Pro (AnyGaze Agent)	2.546	6.672

Table 6. Comparison of GazeAnywhere Agent and no agent framework.

prompt (e.g., "girl with white dress") , calls the GazeAnywhere tool to generate gaze tracing, and then uses its VLM function to analyze the post-processed video, providing the user with the required analysis.

We collected 10 real-world videos with rich gaze movement using the DigiLens ARGO AR glass to test the agent's performance. The evaluation focused on two tasks: gaze shift calculation and eye contact calculation with other social partners. Using the Mean Absolute Error (MAE) per-minute, the GazeAnywhere Agent demonstrated significantly better performance than a raw, single MLLM solution. Results of GazeAnywhere Agent experiment is showed in Table 6.

6. Discussion

The practical applications of human gaze target estimation are diverse and impactful. In healthcare, for instance, this technology can significantly enhance the analysis of non-verbal communication behaviors which are implicated in the diagnosis and treatment of developmental conditions such as autism [56]. In order for AI models to be used in clinical applications, they must be sufficiently robust and easy to use by nonexperts. This works takes a significant step in that direction for the task of gaze assessment. Our concept-based approach, which allows subjects to be identified by their attributes in natural language, is a first step towards the flexible and convenient specification of a broad set of behavioral analysis tasks. In addition, by creating a unified end-to-end learnable architecture we increase robustness by eliminating brittle stage-wise approaches to identifying the subjects of gaze analysis. Our approach is beneficial even in comparison to using state-of-the-art OVD models to identify subjects, e.g. the SOTA OVD OWLv2 has only a 70% detection accurate rate in Child-SC for the child head and face detection tasks.

7. Conclusion

We present GazeAnywhere, a system that enables interactive human gaze target estimation using flexible, open-vocabulary text prompts to identify the subject. Our principal contributions include introducing the novel Promptable Gaze Target Estimation (PGE) task and Gaze-Co benchmark, proposing a tailored transformer-based detector and learning objective, and developing a human-and-AI-in-the-loop data engine to adapt existing datasets. GazeAnywhere achieves state-of-the-art results in Gaze-Co benchmark, and its robustness is further validated on a challenging out-of-domain (OOD) dataset of child social communication videos. We believe GazeAnywhere and the Gaze-Co benchmark represent important milestones, paving the way for future research and applications in social AI and human behavior understanding.

Acknowledgments

Portions of this work were supported in part by NIH R01 MH114999, the CIFAR Child and Brain Development program, and the Health Care Engineering Systems Center at University of Illinois Urbana-Champaign. Gemini API used in the project is supported by Google. This work also used Delta at the National Center for Supercomputing Applications (NCSA) through allocation CIS251391 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program [5], which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

[1] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017. 1

[2] Michael Argyle, Mark Cook, and Duncan Cramer. Gaze and mutual gaze. *The British Journal of Psychiatry*, 165(6): 848–850, 1994. 1

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 7

[4] Jun Bao, Buyu Liu, and Jun Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14126–14135, 2022. 2

[5] Timothy J Boerner, Stephen Deems, Thomas R Furlani, Shelley L Knuth, and John Towns. Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and experience in advanced research computing 2023: Computing for the common good*, pages 173–176. 2023. 9

[6] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 2

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[8] 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*, pages 213–229. Springer, 2020. 2, 5, 1

[9] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 2, 7

[10] Wenhe Chen, Hui Xu, Chao Zhu, Xiaoli Liu, Yinghua Lu, Caixia Zheng, and Jun Kong. Gaze estimation via the joint modeling of multiple cues. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1390–1402, 2021. 1

[11] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3341–3347. IEEE, 2022. 2

[12] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7509–7528, 2024. 1

[13] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M Jones, Agata Rozga, and James M Rehg. Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–20, 2017. 1

[14] 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 383–398, 2018. 1

[15] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5396–5406, 2020. 2, 1

[16] Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025. 2, 7

[17] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5, 7

[18] 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. 2, 1

[19] Bardia Doosti, Ching-Hui Chen, Raviteja Vemulapalli, Xuhui Jia, Yukun Zhu, and Bradley Green. Boosting image-based mutual gaze detection using pseudo 3d gaze. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1273–1281, 2021. 1

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[21] Nathan J Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604, 2000. 2

[22] Terje Falck-Ytter, Elisabeth Fernell, Åsa Lundholm Hedvall, Claes Von Hofsten, and Christopher Gillberg. Gaze performance in children with autism spectrum disorder when observing communicative actions. *Journal of autism and developmental disorders*, 42(10):2236–2245, 2012. 1

[23] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6460–6468, 2018. 1

[24] 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*, pages 11390–11399, 2021. 1

[25] Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14987–14997, 2025. 2, 6, 3

[26] Google DeepMind and Google. Gemini api and model card. https://ai.google.dev/gemini-api, 2025. Models: Gemini 2.5 Pro. 5

[27] Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5041–5050, 2022. 1

[28] John M Henderson. Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11):498–504, 2003. 1

[29] Nora Horanyi, Linfang Zheng, Eunji Chong, Aleš Leonardis, and Hyung Jin Chang. Where are they looking in the 3d space? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2678–2687, 2023. 2

[30] Zhengxi Hu, Kunxu Zhao, Bohan Zhou, Hang Guo, Shichao Wu, Yuxue Yang, and Jingtai Liu. Gaze target estimation inspired by interactive attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8524–8536, 2022. 2

[31] Qing Jiang, Lin Wu, Zhaoyang Zeng, Tianhe Ren, Yuda Xiong, Yihao Chen, Liu Qin, and Lei Zhang. Referring to any person. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21667–21678, 2025. 2, 6, 3

[32] Tianlei Jin, Zheyuan Lin, Shiqiang Zhu, Wen Wang, and Shunda Hu. Multi-person gaze-following with numerical coordinate regression. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021. 2

[33] Tianlei Jin, Qizhi Yu, Shiqiang Zhu, Zheyuan Lin, Jie Ren, Yuanhai Zhou, and Wei Song. Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence*, 113:104924, 2022. 2

[34] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24905–24916, 2025. 2, 7

[35] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6912–6921, 2019. 1

[36] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589, 2024. 2

[37] Yantao Lai, Rong Quan, Dong Liang, and Jie Qin. Clipgaze: Zero-shot goal-directed scanpath prediction using clip. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2

[38] Susan R Leekam, Emma Hunnisett, and Chris Moore. Targets and cues: Gaze-following in children with autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 39(7):951–962, 1998. 1

[39] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018. 1

[40] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[41] 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)*, pages 9447–9454. IEEE, 2025. 2

[42] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 2, 6, 3

[43] Manuel J Marín-Jiménez, Andrew Zisserman, and Vittorio Ferrari. " here's looking at you, kid." detecting people looking at each other in videos. 2011. 1

[44] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3477–3485, 2019. 1

[45] Qiaomu Miao, Minh Hoai, and Dimitris Samaras. Patch-level gaze distribution prediction for gaze following. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 880–889, 2023. 2

[46] Qiaomu Miao, Alexandros Graikos, Jingwei Zhang, Sounak Mondal, Minh Hoai, and Dimitris Samaras. Diffusion-refined vqa annotations for semi-supervised gaze following. In

European Conference on Computer Vision, pages 439–457. Springer, 2024. 2

[47] Qiaomu Miao, Vivek Raju Golani, Jingyi Xu, Progga Paromita Dutta, Minh Hoai, and Dimitris Samaras. Multi-view gaze target estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5371–5381, 2025. 2

[48] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer, 2022. 2, 5

[49] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023. 2, 6, 3

[50] Kai Niu, Yanyi Liu, Yuzhou Long, Yan Huang, Liang Wang, and Yanning Zhang. An overview of text-based person search: Recent advances and future directions. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):7803–7819, 2024. 2

[51] Yuko Okumura, Yasuhiro Kanakogi, Takayuki Kanda, Hiroshi Ishiguro, and Shoji Itakura. The power of human gaze on infant learning. *Cognition*, 128(2):127–133, 2013. 1

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 7

[53] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 7

[54] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *Advances in neural information processing systems*, 28, 2015. 1, 2

[55] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1435–1443, 2017. 2, 1

[56] James M Rehg, Agata Rozga, Gregory D Abowd, and Matthew S Goodwin. Behavioral imaging and autism. *IEEE Pervasive Computing*, 13(2):84–87, 2014. 8

[57] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[58] Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M Rehg. Gaze-lle: Gaze target estimation via large-scale learned encoders. 2025. 1, 2, 6, 3

[59] 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. 2, 4, 7

[60] 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):1–15, 2024. 1, 2, 6, 3

[61] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20935–20946, 2023. 2, 1

[62] Samy Tafasca, Anshul Gupta, Victor Bros, and Jean-Marc Odobez. Toward semantic gaze target detection. *Advances in Neural Information Processing Systems*, 37:121422–121448, 2024. 1, 2

[63] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Sharingan: A transformer architecture for multi-person gaze following. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2008–2017, 2024. 6, 3

[64] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3125–3133, 2021. 2

[65] 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. 2, 1

[66] 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*, pages 21860–21869, 2023. 2, 1

[67] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2, 7

[68] 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)*, pages 2192–2200. IEEE, 2022. 2, 1

[69] Jun Wang, Hao Ruan, Mingjie Wang, Chuanghui Zhang, Huachun Li, and Jun Zhou. Gazeclip: Towards enhancing gaze estimation via text guidance. *arXiv preprint arXiv:2401.00260*, 2023. 2

[70] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 2

[71] Bufang Yang, Lilin Xu, Liekang Zeng, Kaiwei Liu, Siyang Jiang, Wenrui Lu, Hongkai Chen, Xiaofan Jiang, Guoliang Xing, and Zhenyu Yan. Contextagent: Context-aware proac-

tive llm agents with open-world sensory perceptions. *arXiv preprint arXiv:2505.14668*, 2025. 7

[72] Yaokun Yang and Feng Lu. Gaze target detection based on head-local-global coordination. In *European Conference on Computer Vision*, pages 305–322. Springer, 2024. 2

[73] Yaokun Yang, Yihan Yin, and Feng Lu. Gaze target detection by merging human attention and activity cues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6585–6593, 2024. 2

[74] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022. 2

[75] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2

[76] Mingfang Zhang, Yunfei Liu, and Feng Lu. Gazeonce: Real-time multi-person gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4197–4206, 2022. 2

[77] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. Human gaze assisted artificial intelligence: A review. In *IJCAI: Proceedings of the Conference*, page 4951, 2020. 2

[78] Shizhou Zhang, De Cheng, Wenlong Luo, Yinghui Xing, Duo Long, Hao Li, Kai Niu, Guoqiang Liang, and Yanning Zhang. Text-based person search in full images via semantic-driven proposal generation. In *Proceedings of the 4th International Workshop on Human-centric Multimedia Analysis*, pages 5–14, 2023. 2

[79] Wulue Zhang, Jianbin Xiong, Xiangjun Dong, Qi Wang, and Weikun Dai. Tcnet: Gaze estimation based on temporal body-head-eyes correlation in dynamic scenes. *IEEE Sensors Journal*, 2025. 1

[80] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015. 2

[81] Zhixiong Zhang, Shuangrui Ding, Xiaoyi Dong, Songxin He, Jianfan Lin, Junsong Tang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Sec: Advancing complex video object segmentation via progressive concept construction. *arXiv preprint arXiv:2507.15852*, 2025. 2

[82] Hao Zhao, Ming Lu, Anbang Yao, Yurong Chen, and Li Zhang. Learning to draw sight lines. *International Journal of Computer Vision*, 128(5):1076–1100, 2020. 1

[83] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European conference on computer vision*, pages 350–368. Springer, 2022. 2

Gaze Target Estimation Anywhere with Concepts

Supplementary Material

8. Further Discussions & Social Impact

8.1. Toward End-to-end Gaze Target Estimation

The evolution of human gaze estimation shows a clear trend: a move away from complex auxiliary features like pose and depth towards streamlined, head box-only inputs [10, 14, 19, 23, 24, 27, 35, 39, 44, 82]. This simplification has spurred the development of end-to-end, OpenPose [7]-like, and DETR [8]-like bottom-up approaches that can detect all head box-gaze pairs within a scene [18, 65, 66, 68]. However, a critical limitation persists. These methods lack identity association; they can find the gaze of everyone but cannot identify the gaze of a specific person. This necessitates separate modules or post-processing to link a detected gaze to a particular individual. Thus, the cascaded detection error will still exist. Our work, GazeAnywhere, directly addresses this gap. We propose the first text-promptable pipeline that simultaneously resolves human identification and gaze target estimation, enabling targeted queries for a specific person's gaze.

8.2. Future Application

Joint attention, the capability of following another person's head turn and gaze direction, typically emerges in children with Autism Spectrum Disorder (ASD) years later than in typically developing children [38]. Previous research has demonstrated the strong potential of gaze target estimation models to capture these atypical joint attention behaviors, offering a promising avenue for the early screening and detection of ASD [13, 43]. The concept-based prompting flexibility of the GazeAnywhere model offers a significant evolution in this domain. In future clinical and home-based settings, this model could be deployed to continuously and non-invasively track a child's gaze behavior. Critically, GazeAnywhere and GazeAnywhere Agent can be used by pediatricians or even the patient's parents simply by describing the patient's appearance or location in the prompt (e.g., "the child in the blue shirt"), thereby omitting the complicated and labor-intensive process of manually drawing head bounding boxes for annotation. This simplified usage offers the benefit of longitudinal tracking outside of a clinical setting, enabling earlier intervention and more comprehensive developmental monitoring. In addition, user can query GazeAnywhere Agent to let MLLM post-process the target tracking video and provide high-level gaze behavior information like gaze shift.

9. GazeAnywhere Agent

In this section, we introduce the GazeAnywhere Agent, a visual agentic framework designed to process natural-language gaze estimation and post-analysis requests. Figure 6 illustrate the workflow of the agent. The system dynamically queries a MLLM to orchestrate specific tools. The initial version of the agent integrates two primary models as the tool: Whisper-large-v3 for audio-to-text conversion and our proposed GazeAnywhere model for PGE target prediction.

Given an input image or video and a user request via audio , the MLLM acts as a planner and controller. It first converts the user's audio to text, analyzes the scene context, devises a step-by-step plan, and subsequently invokes the GazeAnywhere model. After each action, the agent receives visual feedback by visualizing the gaze target within the scene. This feedback is stored in memory, enabling the agent to revise its plan and determine the next steps for analysis. This pipeline handles queries far more complex than simple noun phrases, facilitating a deeper understanding of human gaze behavior in video streams.

10. Dataset & Benchmark

10.1. Training Set

The Gaze-Co training set contains 119,525 samples in total. Each record includes the target head bounding box, normalized gaze point, an in/out-of-frame label, and a compact concept phrase (attribute, position, action, and pose). The training data are constructed from three published gaze datasets after applying the image-quality filters, MLLM-based concept generation, and human in-loop MLLM verification described in the main text. In terms of source datasets, 69.6% (83,148 samples) come from GazeFollow [54, 55], 19.6% (23,481 samples) from VideoAttentionTarget [15], and 10.8% (12,896 samples) from ChildPlay [61] (see Fig. 7a).

For the apparent subject category, 51.0% (60,983 samples) are labeled as man, 32.2% (38,508) as woman, 8.2% (9,773) as boy, 6.1% (7,337) as girl, 1.6% (1,916) as child (unspecified gender), and 0.8% (1,008) as infant (unspecified gender) (Fig. 7b). These labels reflect perceived visual categories rather than verified identity attributes. Regarding gaze location, 13.9% (16,671 samples) of annotations are out-of-frame, while 86.1% (102,854) fall within the image (Fig. 7c).

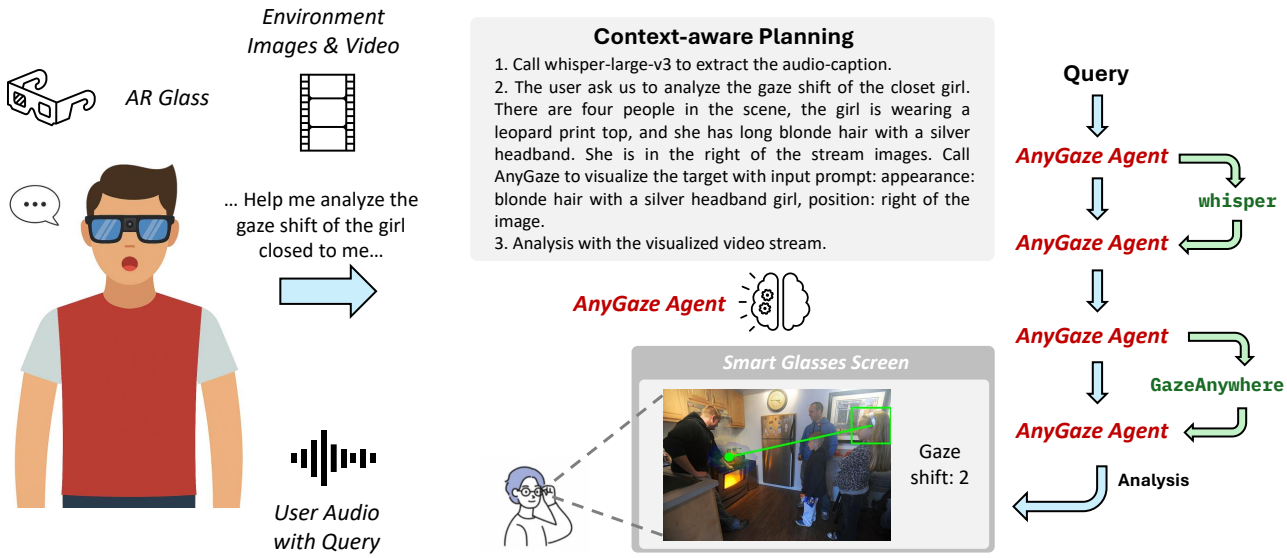
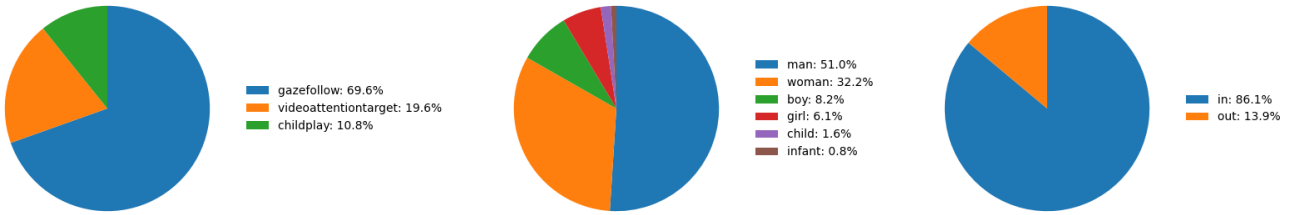


Figure 6. Step-by-step explanation of how GazeAnywhere Agent works.



(a) Source dataset distribution (b) Apparent subject category distribution (c) In-frame vs out-of-frame gaze distribution

Figure 7. Training-set statistics of the Gaze-Co dataset: (a) proportion of each source dataset, (b) distribution of apparent subject categories, and (c) proportion of in-frame vs out-of-frame gaze annotations.

10.2. Concept-based In-domain Test Set

We derive three concept-augmented test splits by converting the official test splits of GazeFollow, VAT, and ChildPlay into our unified PGE schema (image, head box, normalized gaze point, in/out-of-frame label, and concept phrase). After applying the same image-quality filters as in the training set, we obtain GazeFollow-Concept, VAT-Concept, and ChildPlay-Concept. To guarantee a high quality benchmark for both baselines and our model evaluation, all concept annotations are human verified instead of using MLLM.

GazeFollow-Concept. After filtering, GazeFollow-Concept contains 2,436 (image, head box) records. In terms of apparent subject category, 49.1% (1,197 samples) are labeled as man, 31.7% (772) as woman, 10.3% (250) as boy, 5.6% (136) as girl, 2.1% (50) as child (unspecified gender), and 1.3% (31) as infant (unspecified gender). All annotations in this split correspond to in-frame gaze targets (100%, 2,436 samples). In the dataset, each (image, head box) record is associated with multiple human gaze point annotations from the original GazeFollow dataset, which motivates the additional Avg L2 and Min L2 metrics used in the main text: Avg L2 is defined as the distance between the predicted gaze point and the mean of all human annotations, and Min L2 as the distance to the nearest human-annotated gaze point.

VAT-Concept. VAT-Concept contains 5,301 records. For apparent subject categories, 45.9% (2,435 samples) are labeled as man, 43.8% (2,324) as woman, 5.8% (310) as boy, 0.4% (20) as girl, 0.1% (5) as child (unspecified gender), and 3.9% (207) as infant (unspecified gender). Regarding gaze location, 35.5% (1,884 samples) of annotations are out-of-frame, while 64.5% (3,417) are in-frame.

ChildPlay-Concept. ChildPlay-Concept contains 1,238 records. In terms of apparent subject category, 8.5% (105 samples) are labeled as man, 32.9% (407) as woman, 36.8% (455) as boy, 12.0% (148) as girl, 4.4% (55) as child (unspec-

ified gender), and 5.5% (68) as infant (unspecified gender). For gaze location, 14.7% (182 samples) of annotations are out-of-frame, while 85.3% (1,056) are in-frame.
Across all splits, the apparent subject categories reflect perceived visual attributes rather than verified identity labels.

10.3. Concept-based Out-of-domain Test set

For out-of-domain evaluation, we utilize Child–Social Communication (Child-SC), a private dataset protected by IRB. It captures natural interactions between children and clinicians, where the clinician guides the child's attention across various targets using toys, thus eliciting frequent and structured gaze shifts. The dataset comprises 326 video clips from 40 children, sampled at 5 fps, yielding a total of 151,533 images. Due to privacy regulations, these images cannot be processed by cloud-based MLLM; consequently, all target-person concepts were manually annotated, strictly adhering to the style and protocols of our MLLM-generated concepts.

11. Baseline Details

11.1. Open-Vocabulary Detector (OVD)

As baselines, we use the OVD models to locate the target person described by a text prompt. This step supports our main task: to predict the point of view of the subject. Each OVD model takes an image and a prompt, matches text to visual regions in a shared vision–language space, and scores candidate boxes by text–image similarity. It outputs the highest-confidence bounding box for the prompted person, which we use as the subject-person localization. We also compared with the SOTA open-vocabulary human detection model RexSeek [31], which is a 3B foundation model in referring expression comprehension task.

GroundingDINO-B. GroundingDINO-B [42] is a Transformer-based detector featuring a dual-encoder single-decoder architecture that deeply fuses image and text features. It employs a language-guided query selection module to initialize object queries based on the input prompt. This mechanism produces a series of refined candidate boxes associated with prediction scores. From these outputs, we identify the target person by selecting the box with the highest confidence score for the referring phrase.

LLMDet-L. LLMDet-L [25] enhances open-vocabulary detection through multimodal co-training, where a large language model generates detailed captions to enrich feature alignment during training. At test time, with the LLM removed, the detector takes the image and prompt to generate multiple region candidates. It evaluates these regions by matching them against the text embedding, enabling us to filter the results and retrieve the top-ranked bounding box as the localized subject.

OWLv2-L. OWLv2-L [49] scales up the OWL-ViT architecture using a massive self-training strategy on over one billion weakly supervised examples. It utilizes a Vision Transformer backbone to directly predict bounding boxes and text-alignment scores from image tokens. When queried with the target person's description, the model outputs a collection of detected objects with their semantic similarity scores, from which we select the best-matching candidate to localize the person.

11.2. Gaze Model

Following the localization step, we evaluate several gaze-following models to predict the target person's point of regard. These models accept the full scene image and the localized person region as input. They output a 2D gaze heatmap (probability distribution), and we extract the coordinates of the peak value from the heatmap to represent the final predicted gaze location.

ViTGaze ViTGaze [60] is a single-modality gaze-following model that predicts a person's gaze target using RGB information only. Given the full image and the target person's head bounding box, it employs a pre-trained ViT to extract human–scene interaction cues directly from self-attention maps, eliminating the need for extra modalities. The model outputs a 2D gaze heatmap along with an in/out-of-frame score for evaluation.

Sharingan Sharingan [63] introduces a transformer-based architecture designed to capture global gaze interactions. It represents the target person via a Person Gaze Token, constructed by fusing head-crop features with normalized head-box coordinates. This token is processed with scene tokens by a ViT encoder to model human–scene dependencies. The model outputs a 2D gaze heatmap representing the spatial probability of the gaze target and an in/out-of-frame score.

Gaze-LLE Gaze-LLE [58] is a streamlined estimator built on a frozen, large-scale DINOv2 encoder, departing from traditional multi-branch head/scene architectures. Given the full image and the target person's head bounding box, it encodes the head location as a positional prompt injected into the scene features, using a lightweight transformer decoder to model head–scene relations. The model predicts a 2D gaze heatmap along with an in-/out-of-frame score.

12. Experimental Protocol

12.1. AR Device for GazeAnywhere Agent

We use DigiLens ARGO in the experiment to capture video data in real-world settings (Fig. 8). Its 48 MP camera records high-resolution video with autofocus, optical and electronic

Figure 8. DigiLens ARGO AR glasses used for video and audio capture and on-device feedback in the GazeAnywhere Agent.

stabilization, 4×4 pixel binning, and strong low-light support. For audio, a five-microphone beamforming array is designed to pick up the wearer's voice in noisy environments and provides spatial recordings suitable for analysis.

12.2. Implementation Details of GazeAnywhere-DINOv3-L

The deployed version of GazeAnywhere-DINOv3-L consists of a detector transformer with 3 layers and a dimension of $D = 256$. Both the visual and text prompts are trained jointly. For visual prompting, we apply diverse augmentation techniques during training, including head/body bounding box jittering, color jittering, random resizing and cropping, horizontal flipping, rotation, and masking of scene patches. For text prompting, as the subject position text information is fixed, we limit visual augmentation to random scene patch masking and apply text augmentation with reordering appearance, location, pose, and action attributes. During training, the input resolution is 512×512.

13. More Results

13.1. Impact of Frozen Encoder.

A key design choice for GazeAnywhere is to keep the image and text encoders frozen. We validate this approach in Table 7, which compares the default frozen model against one where the DINOv3 image encoder or the text encoder are fine-tuned. Unfreezing image or text encoders leads to a clear drop in performance. This demonstrates that DINOv3's pre-trained features are highly robust and generalizable for the PGE task, and that fine-tuning may lead to overfitting or harmful feature drift.

13.2. Impact of Detector Dimension.

We study the impact of the Detector transformer's layer dimension D in Table 8. The results indicate that performance plateaus at $D = 128$. We observed no significant performance gain from increasing D further, and thus selected

Visual	Text	Trainable Param	GazeFollow-Concept			VAT-Concept		
			AUC ↑	Avg L2 ↓	Min L2 ↓	AUC ↑	L2 ↓	AP ↑
✗	✗	870.2M	0.931	0.150	0.095	0.886	0.212	0.823
✓	✗	332.1M	0.943	0.119	0.067	0.886	0.206	0.813
✗	✓	541.7M	0.952	0.130	0.078	0.874	0.201	0.825
✓	✓	3.6M	**0.958**	**0.099**	**0.050**	**0.928**	**0.123**	**0.879**

Table 7. Comparison of the encoder frozen strategies.

D of $\psi(\cdot)$	GazeFollow-Concept			VAT-Concept		
	AUC ↑	Avg L2 ↓	Min L2 ↓	AUC ↑	L2 ↓	AP ↑
64	0.953	0.115	0.062	0.915	0.144	0.871
128	0.960	0.100	0.050	0.928	0.116	0.875
256	0.958	0.099	0.050	0.928	0.123	0.879
512	0.959	0.104	0.054	0.917	0.122	0.875

Table 8. Ablation experiment on the selection of the Detector Transformer dimension.

Layer Num of $\psi(\cdot)$	GazeFollow-Concept			VAT-Concept		
	AUC ↑	Avg L2 ↓	Min L2 ↓	AUC ↑	L2 ↓	AP ↑
1	0.949	0.126	0.076	0.882	0.179	0.846
2	0.957	0.104	0.054	0.920	0.123	0.871
3	0.958	0.099	0.050	0.928	0.123	0.879
4	0.957	0.100	0.051	0.928	0.120	0.858
5	0.959	0.095	0.047	0.929	0.121	0.888

Table 9. Ablation experiment on the selection of the transformer layer number in the Detector.

$D = 256$ as it provides the best trade-off between accuracy and computational cost.

13.3. Ablation on Detector's Transformer Layer Number

We conduct another ablation study to explore the layer number of transformer blocks in detector transformers. Results are shown in Table 9. After increasing the layer number to 3, the model shows stable performance.

14. Qualitative Analysis

In Figure 9, we qualitatively compare GazeAnywhere with the current state-of-the-art model, Gaze-LLE. Although Gaze-LLE performs well in sparse scenes with only one or two individuals, its performance degrades noticeably as crowd density increases. As shown in Figure 9, the upstream OVD module becomes unreliable in these complex settings and typically fails in two ways. First, it may localize the wrong person, causing Gaze-LLE to estimate gaze for an incorrect target. Second, it may produce an overly large bounding box that covers multiple people; even if the true target is included, Gaze-LLE cannot reliably disambiguate whom to condition on. These examples expose a key limitation of two-stage gaze estimation pipelines in real-world social scenes.

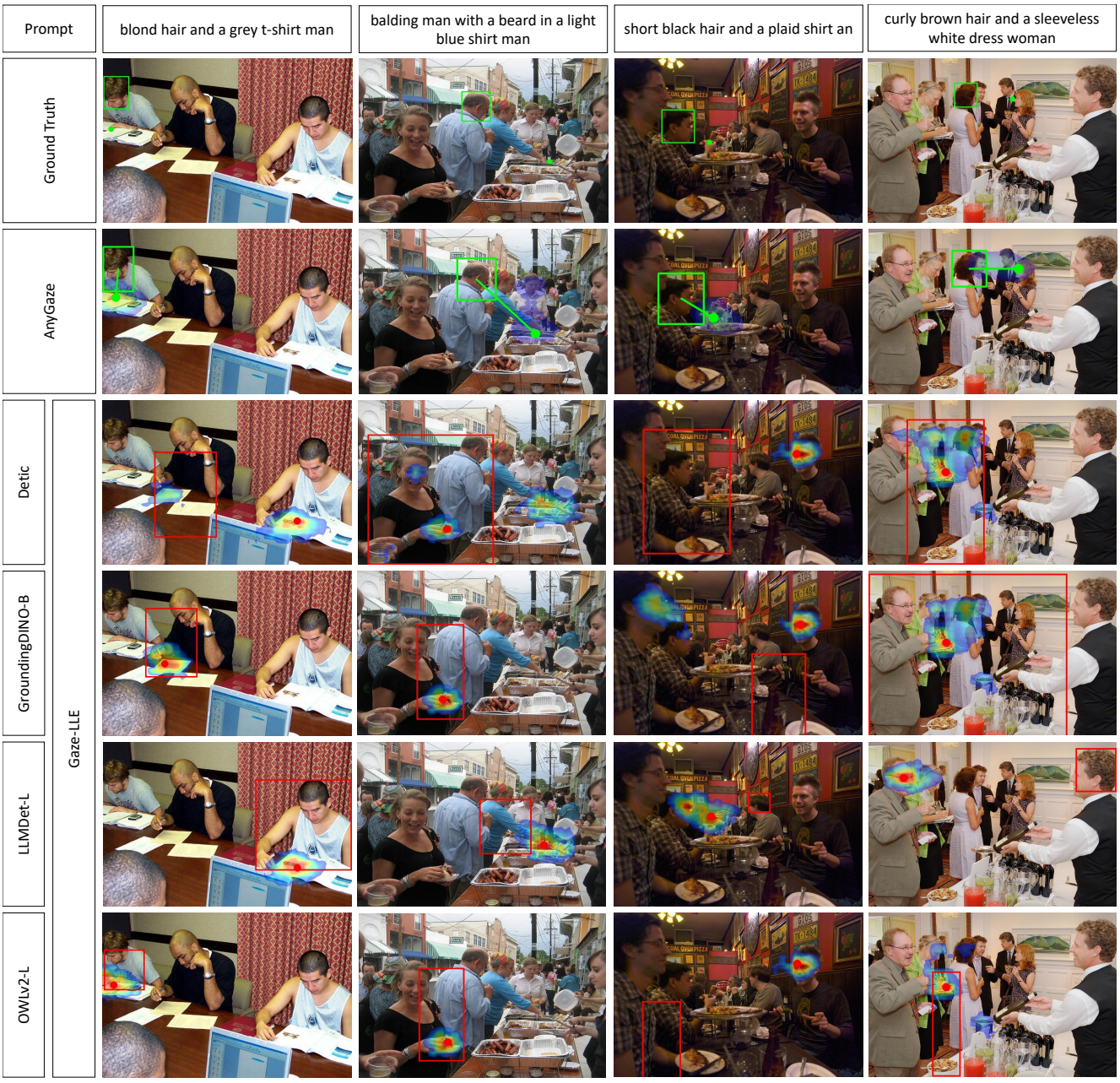


Figure 9. Qualitative comparison of gaze-target localization conditioned on appearance prompts. Each column corresponds to a different sample, and each row shows predictions from a different method. Our method produces sharper and more accurate heatmaps around the true gaze targets.

15. Related Prompts

For reproducibility, we include the exact natural-language prompts used to query the MLLM in our pipeline. These prompts support three major components: the concept-generation data engine, the MLLM-only gaze prediction baseline, and the GazeAnywhere Agent for video-based social gaze analysis. Unless otherwise noted, the prompts are shown verbatim as used in our batch API calls.

15.1. Data Engine

This section summarizes the prompts used by the data engine to construct concept-level annotations for each subject person. The attribute prompt (Fig. 10) instructs the MLLM to produce a compact description of appearance, position, action, pose, and people count for the person marked by the

green head box.

The concept verification prompt (Fig. 11) then asks the MLLM to check, field by field, whether a candidate concept matches the image and to return JSON flags for attribute, position, action, pose, and an overall pass/fail decision. Together with spot-checks from human annotators, these prompts implement the MLLM component of our human-in-the-loop data engine.

15.2. MLLM Baseline

Here we provide the prompt used for the MLLM-only gaze target prediction baselines, Gemini-2.5-flash (Fig. 12) and Qwen3-VL-8b (Fig. 13).

Given an image and a textual concept description, the model is asked to predict an gaze in/out-frame flag and a normalized 2D gaze target point, and to return the answer in a strict JSON format.

15.3. GazeAnywhere Agent

This section lists the prompts used to compare gaze-target analysis with an MLLM alone versus an MLLM assisted by the GazeAnywhere Agent on smart-glasses recordings. The raw-video prompt (Fig. 14) presents the model with the original AR recording and asks it to infer social gaze behavior directly from the unannotated video.

The GazeAnywhere-agent prompt (Fig. 15) uses the same video but with GazeAnywhere overlays (subject head box, gaze point, and out-of-frame indications), and instructs the model to count gaze shifts to social partners and overall gaze shifts.

16. Notations

We present the description all the notations in our paper in the last two pages.

Concept Generation Prompt

```
TASK
Return a description for the person with a green bounding box in head:
The description is a natural, concise attribute phrase (<30 words in total).

STYLE & CONTENT
- all lowercase
- avoid generic words:  person, people, adult; avoid starting with a/an/the
- prefer stable visible attributes, in this order:  hair style and color / hat >
glasses / beard > top garment color and pattern > pant / dress garment color and
pattern.
- the LAST word of attributes MUST be one of:  man, woman, boy, girl, infant, child
- prefer short and clear location description, such as bottom left corner.
- action:  describe ongoing interaction or movement; make it specific by adding
target/object/direction when visible.  keep “what is being done” here.  if unclear,
write "none".
- pose:  describe static body configuration and facing direction; keep “how the body
is” here (orientation, posture, limb arrangement).  if unclear, write "none".
- keep action and pose distinct and non-overlapping.
- also provide an approximate count of people visible in the scene; report a single
integer when feasible; if indeterminate, write "none".

You output format is:
<attribute> attributes of the human </attribute>
<position> position of the human in the camera </position>
<action> action of the human </action>
<pose> pose of the human </pose>
<count> estimated number of people in the scene </count>
```

Figure 10. **Concept generation prompt** used for generating concept phrases for the target person.

Concept Verification Prompt

```
TASK
You see an image with one green bounding box on a human head and a candidate
description from Prompt A, with:
<attribute>, <position>, <action>, <pose>, <count>.
Check whether the first four fields match the target human and the scene.  Ignore
<count>.


CHECKING RULES

attribute
must describe the same person in the green box.
hair / hat / glasses / beard / clothes must match.
last word must be one of:  man, woman, boy, girl, infant, child.
label as correct only if all above are satisfied.


position
must match the boxed human location (e.g., top left, bottom center, center right).
label as correct only if consistent with boxed human position.


action
must be a visible ongoing movement or interaction of this person.
if not clearly visible, the correct value should be "none".
label as correct only if supported by the image.


pose
must describe static body configuration and facing direction (orientation, posture,
limb arrangement).
must be distinct from action; if unclear, should be "none".
label as correct only if supported by the image and distinct from action.


OVERALL
overall is "pass" only if all four checks are "correct".
otherwise overall is "fail".


OUTPUT FORMAT
Output only a single JSON object with exactly these keys and values:

"attribute_check":  "correct" or "incorrect"
"position_check":  "correct" or "incorrect"
"action_check":  "correct" or "incorrect"
"pose_check":  "correct" or "incorrect"
"overall":  "pass" or "fail"
```

Figure 11. **Concept verification prompt** used to check attribute, position, action, and pose consistency for each subject.

Gaze Target Prediction Prompt of Gemini 2.5 Flash

```
You are given an image, where the top-left corner is (0, 0) and the bottom-right
corner is (1, 1).

Coordinates are normalized by the image width and height.
You are also given a description of the subject person in the image:

Attribute: {attribute}
Location: {position}
Action: {action}
Pose: {pose}
Based on this description and the image, perform the following tasks:

1. In-frame gaze flag
Indicate whether the subject person is looking at a target inside the image frame.

Output 1 if the gaze target lies within the image frame.
Output 0 if the subject person is looking outside the image frame.

2. Gaze target point
Predict the gaze target of the subject person as a point (x, y) ∈ [0, 1], with exactly
three decimal places for both x and y.

The values must be normalized by the image width and height.

Output format
Return only a valid JSON object, with no extra text, in the following format:
{
"in_frame_gaze": 0,
"gaze_target": {
"x": 0.XXX,
"y": 0.XXX
}
}
in_frame_gaze must be either 0 or 1.
x and y must be numbers in [0,1] with three decimal places.
```

Figure 12. **Gaze target prediction prompt** used for in-frame flagging and point estimation on Gemini-2.5 baseline

Gaze Target Prediction Prompt of Qwen3-VL.

```
You are given an image, where the top-left corner is (0.000 , 0.000), the bottom-right
corner is (1.000, 1.000).  The pixel point in the image is normalized to 0.000 to
1.000.  All values are rounded by 3.

Here is the description of the subject person Question

Based on the description of a subject person in the image, perform the following task:

(1) Indicate whether the subject person is looking at a target inside the image frame.

Output 1 if the gaze target lies within the image frame.  Output 0 if the subject
person is looking outside the image frame.

Provide the inside prediction between the <inside> and </inside> tags.

(2) Predict the gaze target of the subject person as a point (x, y) x and y are in
[0.000, 1.000].  If looking outside, randomly give values.

Provide the x of point between the <x> and </x> tags, y of point between the <y> and
</y> tags.
```

Figure 13. **Gaze target prediction prompt** used for in-frame flagging and point estimation on Qwen3-VL-8B.

Gaze Shift Analysis Prompt on Single MLLM Solution

```
This is a short video for human gaze target understanding.  Can you give me an
analysis of these tasks"?

The subject child is:  short black hair white dress girl

1.  "Gaze shift to social partner":  The number of gaze shifts to the nearby person
happened.

2.  "Total gaze shift count":  The number of gaze shifts happened.  (change the gaze
target to another object or out-of-frame)

Hint:  Gaze shifting is the coordinated movement of the eyes and head to look at a
new target.  Per gaze shift means changing the gaze target from one object/human to
another object/human

You should analyze the video and provide the answers to the above tasks.
```

Figure 14. **Gaze shift analysis prompt** used for counting gaze shifts and eye contact events on Single MLLM.

Gaze Shift Analysis Prompt on GazeAnywhere Agent

```
This is a short video for human gaze target understanding.  The green bounding box is
the detected subject child.  If the bounding box's color becomes blue, it indicates
the subject is looking out of the frame.  The green point is the child's gaze target,
and we overlap it with the raw video.  Can you give me an analysis of these tasks"?

The subject child is:  short black hair white dress girl

1.  "Gaze shift to social partner":  The number of gaze shifts to the nearby person
happened.

2.  "Total gaze shift count":  The number of gaze shifts happened.  (change the gaze
target to another object or out-of-frame)

Hint:  Gaze shifting is the coordinated movement of the eyes and head to look at a new
target.  Our green point in the frame can indicate the target location.  So you should
infer if the target is changed.  Per gaze shift means changing the gaze target from
one object/human to another object/human

You should analyze the video and provide the answers to the above tasks.
```

Figure 15. **Gaze shift analysis prompt** used for counting gaze shifts and eye contact events on GazeAnywhere agent.

Data and Indices	
H	Height of input image
W	Width of input image
$I \in \mathbb{R}^{3\times H\times W}$	Input RGB image
P	Prompt
T	Text
H_{out}	Height of output image
W_{out}	Width of output image
$\hat{H} \in \mathbb{R}^{H_{out}\times W_{out}}$	Gaze heatmap
Embeddings and Image Encodings	
$\phi_V(\cdot)$	Image encoder
N_V	Number of patch tokens
D_V	Visual embedding dimension
$[CLS]$	Classification token
$c \in \mathbb{R}^{D_v}$	$[CLS]$ token embedding
$s_i \in \mathbb{R}^{D_V}$	Visual output embedding token
Embeddings and Text Encodings	
$\phi_T(\cdot)$	Text encoder
$[EOS]$	End of sentence token
T_E	Initial text embeddings
L_T	Fixed context length
D_T	Text embedding dimension
t_{eos}	End of sentence token
t_{pad}	Padding token
$t_i \in \mathbb{R}^{D_T}$	Text embedding token
Projection Layers	
$W_V \in \mathbb{R}^{D_V\times D}$	Trainable visual linear projection layer
$W_T \in \mathbb{R}^{D_T\times D}$	Trainable test linear projection layer
D	Projected dimension
$Z_V \in \mathbb{R}^{(1+N_V)\times D}$	Projected visual tokens
$Z_T \in \mathbb{R}^{L_T\times D}$	Projected text tokens
Detector Transformer	
$\psi(\cdot)$	Detector transformer
$t_h \in \mathbb{R}^D$	Head token
$t_p \in \mathbb{R}^D$	Target presence token
c'	Projected global visual tokens
t'_{eos}	Projected global text tokens
E_{head}	Learnable head embeddings
$E_{presence}$	Learnable presence embeddings
$s' \in \mathbb{R}^{N_V\times D}$	Projected visual patch tokens from Z_V (excluding c')
$t' \in \mathbb{R}^{(L_T-2)\times D}$	Projected text patch tokens from Z_T (excluding t'_{eos} and padding)
$F \in \mathbb{R}^{(N_T+N_V+2)\times D}$	Full input sequence F
$\psi(F) \in \mathbb{R}^{(N_T+N_V+2)\times D}$	Output refined sequence of detector transformer
Decoder	
$\hat{s} \in \mathbb{N}_\mathbb{V} \times \mathbb{D}$	Refined visual patch tokens
$\hat{t_h} \in \mathbb{R}^D$	Refined head tokens
x	normalized center x coordinate of head tracker
y	normalized center y coordinate of head tracker

w	normalized width of head tracker
h	normalized height of head tracker
$\hat{t_p} \in \mathbb{R}^D$	Refined predict tokens
Learning Objective	
$\mathcal{L}_{total}$	Total loss
$\mathcal{L}_{gaze}$	Gaze heatmap BCE loss
σ	Standard deviation of 2D Gaussian
$\hat{Y}$	Predicted heatmap
N	Total number of pixels
p	Single pixel on the heatmap
y_p	Ground-truth of p
$\hat{y_p}$	Predicted values of p
$\mathcal{L}_{presence}$	Target presence focal loss
$Y_{presence} \in 0, 1$	Ground truth target presence
$\hat{Y_{presence}} \in [0, 1]$	Predicted target presence
$\mathcal{L}_{focal}$	Focal loss
$\mathcal{L}_{head}$	Head box loss
$\mathcal{L}_1$	Mean absolute error
$\mathcal{L}_{IoU}$	GIoU loss
b	Ground truth head box
$\hat{b}$	Predicted truth head box
λ_{l_1}	head object detection hyperparameter
λ_{IoU}	head object detection hyperparameter
Data Engine	
x_{min}	x coordinate of top-left corner of the head box
y_{min}	y coordinate of top-left corner of the head box
x_{max}	x coordinate of bottom-right corner of the head box
y_{max}	y coordinate of bottom-right corner of the head box
g_x	x coordinate of ground truth gaze point
g_y	y coordinate of ground truth gaze point