

---

# OMNIGAZE: Reward-inspired Generalizable Gaze Estimation in the Wild

---

Hongyu Qu<sup>1\*</sup>, Jianan Wei<sup>2\*</sup>, Xiangbo Shu<sup>1†</sup>, Yazhou Yao<sup>1</sup>, Wenguan Wang<sup>2†</sup>, Jinhui Tang<sup>3</sup>

<sup>1</sup>Nanjing University of Science and Technology <sup>2</sup>Zhejiang University <sup>3</sup>Nanjing Forestry University

<https://github.com/quhongyu/OmniGaze>

## Abstract

Current 3D gaze estimation methods struggle to generalize across diverse data domains, primarily due to **i) the scarcity of annotated datasets**, and **ii) the insufficient diversity of labeled data**. In this work, we present OMNIGAZE, a semi-supervised framework for 3D gaze estimation, which utilizes large-scale unlabeled data collected from diverse and unconstrained real-world environments to mitigate domain bias and generalize gaze estimation in the wild. First, we build a diverse collection of unlabeled facial images, varying in facial appearances, background environments, illumination conditions, head poses, and eye occlusions. In order to leverage unlabeled data spanning a broader distribution, OMNIGAZE adopts a standard pseudo-labeling strategy and devises a reward model to assess the reliability of pseudo labels. Beyond pseudo labels as 3D direction vectors, the reward model also incorporates visual embeddings extracted by an off-the-shelf visual encoder and semantic cues from gaze perspective generated by prompting a Multimodal Large Language Model to compute confidence scores. Then, these scores are utilized to select high-quality pseudo labels and weight them for loss computation. Extensive experiments demonstrate that OMNIGAZE achieves state-of-the-art performance on five datasets under both in-domain and cross-domain settings. Furthermore, we also evaluate the efficacy of OMNIGAZE as a scalable data engine for gaze estimation, which exhibits robust zero-shot generalization on four unseen datasets.

## 1 Introduction

Eye gaze provides human with a means for evaluating an individual’s interest in their internal and external environments [1, 2], which is subtle but informative. 3D gaze estimation, as a crucial topic in the field of gaze signal analysis, aims to directly predict gaze direction from face images, which serves as the foundational representation in various applications, such as virtual reality [3, 4, 5], human-computer interaction [6, 7, 8], medical diagnosis [9, 10], and driver monitor systems [11, 12].

Due to the variants of subject appearance, background environments, image quality, shooting angle and illumination across existing datasets [13, 14, 15, 16], the performance of gaze estimation methods [17, 18] trained on a single dataset suffer from performance degradation when testing on new, unseen datasets. This limitation motivates recent research [19, 20, 21, 22, 23, 24, 25, 26] to focus on cross-domain generalization for gaze estimation, seeking to bridge inter-dataset discrepancies. Though effective, these methods are still constrained by the limited diversity of labeled training data, restricting their applicability for real-world applications. In contrast, enormous face images can be easily accessed by crawling from Internet [27] or synthetic generation using generative models [28].

---

\* Equal contribution

† Corresponding author

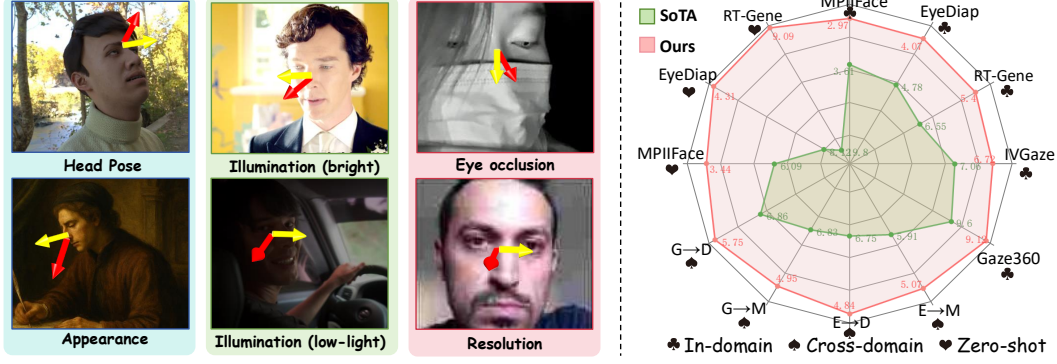


Figure 1: **Left:** By making efficient use of large-scale diverse unlabeled datasets via reward-driven pseudo label selection, our OMNIGAZE can estimate high-quality 3D gaze directions for in-the-wild images in diverse conditions, *e.g.*, extreme head poses, varying lighting conditions, and appearance, *etc.* Red and yellow arrows represent predictions from OMNIGAZE and base model. **Right:** our OMNIGAZE achieves state-of-the-art performance on five datasets under three settings, *i.e.*, in-domain, cross-domain, and zero-shot generalization.

To generalize gaze estimation with large-scale unlabeled face datasets, there are two main strands of research: weakly supervised learning and unsupervised learning. Regarding to weakly supervised learning, previous works enhance 3D gaze estimation with weak social gaze interaction labels (*e.g.*, mutual-gaze [29] and gaze-following [30]), while [31] generates gaze pseudo-annotations by leveraging 3D eye region geometry. However, their effectiveness is constrained by reliance on labeled datasets from gaze-related domains. In the context of unsupervised learning, self-supervised pre-training stands out as the leading paradigm, which endeavors to learn a robust gaze representation via well-designed pretext tasks, *e.g.*, eye-consistent image reconstruction [32, 33, 34], masked image restoration [35] and gaze redirection [36]. Nevertheless, these pretext tasks exhibit weak semantic relevance to gaze estimation, resulting in inefficient utilization of unlabeled face images.

In light of these limitations, we find that there remains a notable void for semi-supervised frameworks capable of effectively harnessing both labeled data and large-scale unlabeled datasets in the gaze estimation community. Then, we propose OMNIGAZE, a semi-supervised learning framework (*cf.* Fig. 1), which employs a pseudo-labeling strategy to generalize gaze estimation in the wild with large-scale unlabeled face datasets. Concretely, OMNIGAZE implements a standard SSL three-phase training protocol: **i)** a teacher model is trained via supervised learning on annotated datasets; **ii)** this model is utilized to generate pseudo-labels for unlabeled samples and high-quality instances are selected to enhance training data; **iii)** a generalized student model is optimized by integrating both annotated and pseudo-labeled data. However, applying this strategy to 3D gaze estimation is confronted with three critical challenges: **❶** existing threshold-based pseudo-labeling methods [37, 38, 39], specifically tailored for classification tasks, are *inapplicable for regression output*; **❷** pseudo-labels generated by a teacher model trained on labeled datasets with limited diversity suffer from domain bias [40], which leads to *difficulty in utilizing the pseudo labels*; **❸** learning robust gaze representations demands training data with rich diversity [41, 30, 35] to *capture the wide variability across individuals*.

Faced with challenge **❶**, OMNIGAZE devises a dedicated reward model that utilizes unlabeled images paired with pseudo labels to assess the reliability of these pseudo labels. To learn reliable reward scores, we propose two advancements: **i)** Each pseudo gaze label is interpolated into a 3D gaze direction vector, thereby enabling a geometry-aware representation and enhancing alignment with natural gaze behaviors; **ii)** To harness the enormous knowledge stored in large-scale pretrained language models, we extract visual embeddings of unlabeled images via an off-the-shelf visual encoder and define a prompt to guide the Multimodal Large Language Model to generate *scene-specific gaze descriptions* for unlabeled images; These linguistic descriptions are encoded via the text encoder of CLIP and combined with visual features to construct a multi-modal gaze representation. Thus, the reward model can capture the nuanced nature of gaze for robust confidence assessments.

As a response to challenge **❷**, OMNIGAZE adopt two strategies: **i)** utilize confidence scores to filter out unreliable pseudo labels and reweight the importance of different high-quality samples for loss computation; **ii)** establish a loop for mutual boosting between the student model and reward model training, enabling continuous refinement of pseudo labels to progressively enhance both gaze estimation accuracy and pseudo-label quality in OMNIGAZE.

To tackle challenge ❸ and fuel the proposed semi-supervised data engine, we curate a diverse collection of unlabeled face images from six publicly available sources, exhibiting wide variability in terms of facial appearance, lighting conditions, head poses, imaging environments, *etc* (Table 1).

Through embracing scaling up data as well as effective reward-inspired pseudo label selection, our OMNIGAZE surpasses all top-leading solutions on five datasets under both in-domain and cross-domain settings (§4.2). Furthermore, we demonstrate the efficacy of OMNIGAZE as a scalable data engine for generating reliable gaze annotations for facial images under diverse conditions. Without any fine-tuning, OMNIGAZE exhibits robust zero-shot generalization across four unseen datasets, evidencing its great potential for deployment in wild-scene applications (§4.3).

## 2 Related Work

**Appearance-based Gaze Estimation.** Appearance-based gaze estimation aims to regress 3D gaze from 2D face images captured by web cameras. Early methods develop their algorithm using scene-restricted datasets and attempt to enhance generalizability through strategies such as extracting gaze-correlated face features [42, 15, 43, 44] or integrating geometric constraints [45, 46, 47]. Though effective enough for certain subjects, they suffer from performance degradation in unconstrained environments, *e.g.*, free head motion and profile faces of subjects positioned further from the camera. To track this issue, subsequent studies endeavor to construct datasets [13, 14, 15, 16] for gaze estimation in more physically unconstrained settings. Though they employed various methods to simulate real-world scenarios, such as using panoramic cameras to record multiple participants at once [14] or multi-view photogrammetry to simulate gaze variations under extreme head poses [13], these approaches still rely on pre-defined assumptions that inherently simplify real-world complexity, and remain difficult to scale compared to web-crawled [27], crowd-sourced [15] or synthetic data [28].

**Cross-domain Gaze Estimation.** The scarcity of diverse *labeled training data* in appearance-based gaze estimation leads current fully-supervised methods to achieve strong *within-domain* performance but suffer from poor generalization in *cross-domain* scenarios. To address this challenge, recent efforts for gaze estimation can be categorized into two paradigms: domain adaptation and domain generalization. Domain adaptation approaches primarily try to minimize the domain discrepancy between source domain and known target domain via strategies, *e.g.*, adversarial learning [14, 48], collaborative learning [49], contrastive learning [22], and consistency learning [23, 24]. In contrast, recent endeavors in domain generalization address a more realistic scenario without access to target samples, focusing on learning domain-invariant features through methods, *e.g.*, self-adversarial learning to preserve gaze information [25] or data augmentation based on gaze-irrelevant factors [26]. However, given the inherent diversity of face images (*e.g.*, illumination, head orientation, and eye occlusion), these methods remain constrained by their reliance on the coverage of labeled source-domain data, which hinders their performance in the unconstrained real-world environments.

**Semi-supervised Learning.** The goal of semi-supervised learning (SSL) is to enhance model’s performance under the limited availability of labeled data by leveraging unlabeled data. The two mainstream methods are entropy minimization [37, 38] and consistency regularization [39, 50, 51, 52]. The former is proposed based on the manifold assumption or the smoothness assumption, *i.e.*, the model should output similar predictions regardless of input perturbations, which necessitates well-designed data augmentations based on the prior of specific tasks. The latter encourages the model itself to output confident predictions on unlabeled data, leading to the main problem of SSL: *how to efficiently select high-quality pseudo labels?* For classification tasks, FixMatch [53] uses a fixed confidence threshold to filter out uncertain samples, while FlexMatch [54] enhances this strategy with class-aware thresholds. Furthermore, SemiReward [55] introduces a reward score based on cosine similarity between pseudo and groundtruth labels to evaluate the quality of pseudo labels. Despite these advances, the SSL for gaze estimation still remain to be explored.

## 3 OMNIGAZE

### 3.1 Semi-supervised Training Pipeline

We first formulate the semi-supervised framework in 3D gaze estimation. Let  $\mathcal{D}_L = \{x_i^l, y_i^l\}_{i=1}^{N_L}$  and  $\hat{\mathcal{D}}_U = \{x_j^u\}_{j=1}^{N_U}$  denotes the labeled and unlabeled datasets, where  $x_i^l$  and  $x_j^u$  are the labeled and

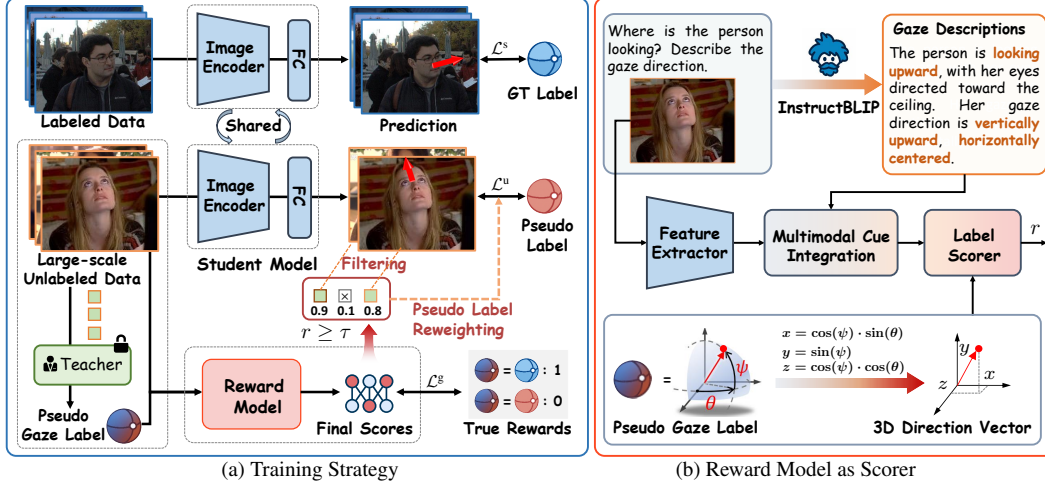


Figure 2: **Overview of the proposed semi-supervised learning framework.** (a) OMINGAZE jointly trains on both labeled data and large-scale unlabeled data, and utilizes a reward model to select and reweight high-quality pseudo labels for unlabeled data. (b) The reward model evaluates the reliability of pseudo labels by jointly reasoning over visual appearance, scene-specific gaze descriptions, and geometric gaze directions.

unlabeled face images, and  $y_i^l$  is the 3D gaze ground truth label. To make use of all training data, the training pipeline of OMINGAZE can be divided into three phases: **i) Pseudo-label generation.** Following previous self-training strategies [37, 38], a teacher model  $\theta_T$  pre-trained on  $\mathcal{D}_L$  via supervised learning is applied to generate pseudo labels  $y_j^u$  for  $\hat{\mathcal{D}}_U$ ; **ii) Reward-driven pseudo-label selection.** To measure the reliability of continuous pseudo labels, a reward model is proposed to predict confidence scores, which are then used to filter out low-quality pseudo labels and reweight the contribution of high-quality samples in the loss calculation (§3.4); **iii) Student model self-training.** These high-quality samples along with labeled data are utilized to train a robust student model  $h_S$  (§3.5). A brief illustration of training pipeline is shown in Fig. 2.

### 3.2 Learning Labeled Face Images

**Model Architecture.** Our gaze estimation model adopts a Vision Transformer (ViT) architecture [56]. Specifically, given an input image  $x$ , the model first extracts gaze representation via a transformer encoder, and then employs a lightweight MLP to regress the gaze direction  $\hat{y}$  as yaw and pitch angles.

**Supervised Loss.** Following [57, 58, 44], we adopt an angular loss to optimize our gaze estimator:

$$\mathcal{L}^s = \frac{1}{N_L} \sum_{i=1}^{N_L} \|\hat{y} - y_i^l\|_2, \quad (1)$$

where  $\hat{y} = h_T(x)$  is the estimation result.

### 3.3 Unleashing the Power of Unlabeled Face Images

**Unlabeled Face Image Collection.** Due to impoverished labeled data, current gaze estimators struggle to generalize to diverse real-world conditions, *e.g.*, different facial appearance, extreme head poses, varying lighting conditions, and broader gaze distributions. To learn robust gaze representations, we attempt to harness the power of large-scale unlabeled face images, which are widely available through online repositories and publicly curated datasets [59, 60, 28, 61, 62, 27] originally designed for facial analysis tasks. Concretely, we compile face images from six public datasets to construct a large-scale unlabeled dataset compassing over 1.4 million images, which covers diverse head poses, lighting conditions, appearance, *etc.* Table 1 provides a detailed breakdown of this dataset.

**Pseudo-label Generation.** Given an unlabeled dataset, we first developed a high-performing teacher model to automatically generate pseudo gaze labels. Specifically, we make full use of existing labeled datasets to train this teacher model  $h_T$  in a supervised manner. Then, we utilize  $h_T$  to assign pseudo gaze labels on unlabeled images:

$$\mathcal{D}_U = \{(x_j^u, y_j^u) | y_j^u = h_T(x_j^u), x_j^u \in \hat{\mathcal{D}}_U\}_{j=1}^{N_U}, \quad (2)$$

Table 1: Key characteristics of unlabeled training datasets used in OMINGAZE. In total, our OMINGAZE is trained on labeled images and **1.4M unlabeled facial images** jointly.

Dataset	Appearance Diversity	Scene	Illumination	Head Pose	Eye Occ.	Face Res.	Size
CelebA [59]	Attributes, makeup, age	Studio-like	Controlled	Mostly frontal	✗	Low	~177K
VGGFace2 [60]	Identity, age, ethnicity	Real-world	Varied	Wide range	✓	Varying	~489K
FaceSynthetics [28]	Synthetic with variation	Synthetic	Controlled	Wide range	✓	High	~86K
SFHQ-T21 [61]	Broad demographic	Synthetic	Varied	Wide range	✓	High	~120K
VFHQ [62]	High-fidelity	Real-world	Varied	Wide range	✓	High	~210K
WebFace [27]	Identity, ethnicity	Indoor	Controlled	Mostly frontal	✗	Medium	~354K

where  $\mathcal{D}_U$  is the pseudo labeled dataset. Then we combine the labeled dataset and pseudo labeled dataset as a new training dataset  $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$  to jointly train a student model  $h_S$ .

**Pseudo-label Selection.** Pseudo-labels generated by teacher models pre-trained on limited annotations are susceptible to confirmation bias. The effective utilization of *noisy pseudo-labels* during training remains a persistent challenge in self-training paradigms. Prior research [53, 54] has predominantly addressed this by filtering out low-quality pseudo-labels through dynamic or handcrafted thresholding strategies. Though effective, these strategies are oriented for classification tasks while ill-suited for regression tasks like gaze estimation, where 3D gaze labels are continuous signals. In this work, we accompany the student model  $h_S$  with an auxiliary reward model  $h_G$  that generates confidence scores to assess the reliability of pseudo-gaze labels. By learning to distinguish between reliable and unreliable pseudo labels,  $h_G$  can select high-quality samples from  $\mathcal{D}_U$ , thereby enhancing the utilization of unlabeled data and improving the efficacy of self-training for the student model.

### 3.4 Empowering Reward Model with Multimodal Cues

Our reward model  $h_G$  (cf. Fig. 2b) evaluates the reliability of pseudo labels by reasoning of multimodal cues: geometric gaze directions, visual feature, and semantic context. By leveraging these cues,  $h_G$  can capture the nuanced and context-dependent nature of gaze for robust confidence assessments.

**Multimodal Cues Integration.** To improve the generalization capability of the reward model for in-the-wild samples, we integrate multimodal cues, *e.g.*, visual and linguistic cues, into the reward model, enabling it to distinguish visually similar but semantically different gaze patterns. **First**, we extract the visual cues by encoding the input image  $x_k \in \mathcal{D}$  via the visual encoder of CLIP [63]:

$$\mathbf{f}_k^v = [\mathbf{f}_{\text{cls}}^v, \mathbf{f}_1^v, \mathbf{f}_2^v, \dots, \mathbf{f}_M^v] = \text{MLP}(\text{Encoder}_v(x_k)), \quad (3)$$

where  $M$  denotes the number of patch in the image. **Second**, we further obtain the linguistic descriptions by questioning MLLMs, *e.g.*, InstructBLIP [64], on the input image with a pre-defined prompt: *In 3D space, where is the person looking, including details about horizontal (left/right) direction, vertical (up/down) direction, and forward/backward relative to the viewer?* Then, these descriptions are converted into linguistic embeddings  $\mathbf{f}_k^l$  via the text encoder of CLIP. Finally, we adopt cross-attention to aggregate  $\mathbf{f}_k^v$  and  $\mathbf{f}_k^l$ , resulting in an semantic-aware gaze representation:

$$\hat{\mathbf{f}}_k^v = \text{AvgPool}(\text{LN}(\text{CrossAttn}(\mathbf{f}_k^v, \mathbf{f}_k^l))), \quad (4)$$

where AvgPool is the average pooling, LN is the standard layer normalization. CrossAttn denotes the standard cross-attention operation, where visual features  $\mathbf{f}_k^v$  is the query, and text features  $\mathbf{f}_k^l$  is the key and value. As such, our reward model can dynamically attend to relevant semantic cues when interpreting visual features, enhancing its ability to disambiguate subtle gaze variations. We provide more examples for generating scene-specific gaze descriptions in the Appendix.

**Reward Model as Scorer.** Given a pseudo label  $y_k = (\theta_k, \psi_k)$ , we first interpolate it into a 3D direction vector. Compared to angular representations such as yaw and pitch, 3D direction vectors provide a more expressive and continuous formulation for modeling gaze behavior, which facilitates more precise alignment with semantic-aware gaze representations. Specifically, we convert the pseudo gaze label into a 3D direction vector  $\mathbf{v}_k$  using a Spherical-to-Cartesian coordinate transformation:

$$\mathbf{v}_k = [\cos(\psi_k) \cdot \sin(\theta_k), \cos(\psi_k), \cos(\psi_k) \cdot \cos(\theta_k)]. \quad (5)$$

Leveraging the obtained semantic-aware gaze representation  $\hat{\mathbf{f}}_k^v$  and 3D gaze vector  $\mathbf{v}_k$ , the reward model predicts confidence scores via a cross-attention followed by an MLP:

$$\hat{r}_k = \text{Sigmoid}(\text{MLP}(\text{CrossAttn}(\hat{\mathbf{f}}_k^v, \mathbf{v}_k))), \quad (6)$$

where  $\hat{r}_k \in [0, 1]$  and Sigmoid denotes the Sigmoid function. Such confidence score allows the reward model to assess the consistency between visual appearance, semantic cues, and gaze labels.



To further strengthen the evaluation capability of reward model, we feed the confidence score  $\hat{r}_k$  and the cosine similarity score (between the student model prediction  $\hat{y}_k$  and the pseudo label  $y_k$ ) into a label scorer implemented via a lightweight MLP to obtain final confidence scores  $r_k \in [0, 1]$ , yielding a holistic and reliable measure of pseudo-label quality:

$$r_k = \text{Sigmoid}(\text{MLP}([\hat{r}_k, \text{sim}(\hat{y}_k, y_k)])) \quad (7)$$

These confidence scores are then utilized to filter out unreliable samples and reweight high-quality ones (see §3.5), thus enhancing the stability and effectiveness of student model self-training.

### 3.5 Training with Pseudo Labels

**Training of Reward Model  $h_G$ .** We train reward model  $h_G$  by jointly using the labeled and unlabeled dataset (*i.e.*,  $\mathcal{D}_L$  and  $\mathcal{D}_U$ ), where we treat ground-truth gaze labels of labeled data as trust pseudo-labels. Formally, given confidence scores  $\{r_k\}_{k=1}^{N_L+N_U}$  for  $\mathcal{D}_L$  and  $\mathcal{D}_U$ , the reward model  $h_G$  is supervised via a binary classification loss:

$$\mathcal{L}^g = \sum_{k=1}^{N_L+N_U} -(c_k \log(r_k) + (1 - c_k) \log(1 - r_k)), \quad (8)$$

where  $c_k \in \{0, 1\}$  is a binary observability mask indicating the label source. Specifically,  $c_k = 1$  denotes that  $y_k$  is a ground-truth label, while  $c_k = 0$  indicates that  $y_k$  is a pseudo label. By this means, the reward model gradually learns to distinguish between reliable and unreliable pseudo labels.

**Training of Student Model  $h_S$ .** Meanwhile, our student model  $h_S$  receives the confidence scores from the reward model  $h_G$  for  $\mathcal{D}_U$ , which indicate the reliability of the corresponding pseudo labels. These scores are used to modulate the learning of the student model  $h_S$  in two ways: **i)** filtering out low-confidence pseudo labels (*i.e.*,  $r_j < \tau$ ), and **ii)** adaptively reweighting the contribution of the remaining pseudo labels. In other words, the reward model  $h_G$  guides the student model to attach more attention to correct labels and ignore erroneous labels. Formally, given  $\mathcal{D}_U = \{x_j^u, y_j^u\}_{j=1}^{N_U}$  and corresponding confidence scores  $\{r_j\}_{j=1}^{N_U}$ , the unsupervised loss on unlabeled data can be defined as:

$$\mathcal{L}^u = \sum_{j=1}^{N_U} \mathbb{1}[r_j \geq \tau] \cdot r_j \cdot \|h_S(x_j^u) - y_j^u\|_2, \quad (9)$$

where  $\mathbb{1}[\cdot]$  is the indicator function for threshold filtering, and  $\tau = 0.5$  is a confidence threshold. Specifically, if  $h_G$  considers a pseudo label unreliable (*i.e.*,  $0 < r_j < 0.5$ ), the unsupervised loss  $\mathcal{L}^u$  encourages  $h_S$  to increase its prediction gap from that pseudo label. In contrast, when  $h_G$  highly trusts a pseudo label (*i.e.*,  $r_j \rightarrow 1$ ),  $\mathcal{L}^u$  enforces stronger alignment between the student prediction and pseudo label. We also explored confidence-based soft weighting  $\alpha \in [0, 1]$  for labeled samples, but it yielded limited practical benefits, leaving deeper analysis for future work. Finally, **overall training objective** of the student model  $h_S$  is an average combination of  $\mathcal{L}^s$  (Eq. 1) and  $\mathcal{L}^u$  (Eq. 9).

**Pseudo-label Update Strategy.** To ensure training stability and robustness, we adopt a periodic pseudo-label update strategy, where the frozen teacher model’s parameters are periodically refreshed with the student model’s weights every  $K$  epochs to regenerate pseudo-labels (ablation study in Table 7 of Appendix). This interval mitigates the risk of early-stage overfitting to noisy labels while allowing the student model to progressively benefit from improved predictions over time.

## 4 Experiment

### 4.1 Experimental Settings

**Training.** OMNIGAZE is trained with a batch size of 512. The training of OMNIGAZE can be divided into two stages: **i)** The teacher model is trained on labeled datasets for 50 epochs. We utilize the Adam optimizer [65] with an initial learning rate of 0.005, and a weight decay of 0.05. **ii)** We train the student model and reward model on both labeled and unlabeled data for 40 epochs with a base learning rate of 0.001 and 0.0001, respectively. Hyper-parameter  $K$  is empirically set to 10.

**Testing.** Following previous works [13, 58, 31], we use one input image scale of  $224 \times 224$  without any data augmentation for the sake of fairness. Note that, during model deployment, our OMNIGAZE does not bring any change to network architecture of the student model or additional computation cost. The reward model  $h_G$  is directly discarded after network training.

**Evaluation Metric.** Following the conventions [66, 67, 68, 44], we use the angular error for evaluation, where lower values indicate better performance.

Table 2: **Quantitative in-domain gaze estimation results** (§4.2) on five benchmarks [43, 16, 44, 14, 12].

Method	MPIIFaceGaze [43]	EyeDiap [16]	RT-Gene [44]	Gaze360 [14]	IVGaze [12]
FullFace [43] [CVPRW17]	4.93	6.53	10.00	14.99	13.67
RCNN [68] [BMVC18]	4.10	5.31	10.30	11.23	-
Gaze360 [14] [ECCV19]	4.06	5.36	7.06	11.04	8.15
RT-Gene [44] [ECCV18]	4.66	6.02	8.60	12.26	-
XGaze [13] [ECCV20]	4.80	6.50	12.00	-	7.06
CANet [58] [AAAI20]	4.27	5.27	8.27	11.20	-
GazeTR [57] [ICPR22]	4.00	5.17	6.55	10.62	7.33
AGE-Net [70] [ICIP24]	3.61	4.78	-	-	-
3DGazeNet [31] [ECCV24]	4.00	-	-	9.60	-
OMNIGAZE (Ours)	<b>2.97</b> $\pm 0.09$	<b>4.07</b> $\pm 0.15$	<b>5.40</b> $\pm 0.21$	<b>9.12</b> $\pm 0.11$	<b>6.72</b> $\pm 0.15$

## 4.2 OMNIGAZE as Generalized Gaze Estimator

We first pretrain OMNIGAZE on our curated training dataset, and then fine-tune it on downstream gaze estimation tasks to examine OMNIGAZE as the weight initialization. We investigate the effectiveness of OMNIGAZE on two settings: 1) *in-domain* gaze estimation, *i.e.*, training and testing on the same dataset, and 2) *cross-domain* gaze estimation, *i.e.*, training on one dataset, testing on an unseen one. In this part, we approach our algorithm on ViT-B encoder pre-trained on ImageNet-21K dataset [69].

**Dataset.** We curate a comprehensive training dataset by combining labeled gaze datasets with six public unlabeled face datasets [59, 60, 28, 61, 62, 27] (Table 1). Note that in *in-domain* gaze estimation, labeled gaze datasets comprises ETH-XGaze [13] along with specific evaluation training datasets; In *cross-domain* gaze estimation, we only use source datasets as labeled gaze datasets. For evaluation, beyond the test split of Gaze360 [14], we further assess OMNIGAZE on four widely used benchmarks: MPIIFaceGaze [43], EyeDiap [16], RT-Gene [44], and IVGaze [12].

**In-domain Gaze Estimation.** We first compare OMNIGAZE with top-leading solutions on five widely used gaze estimation benchmarks under the in-domain evaluation setup. As shown in Table 2, OMNIGAZE surpasses recent state-of-the-art gaze estimation algorithms by solid margins. In particular, it yields **0.64°**, **0.71°**, and **1.15°** reductions in angular error on MPIIFaceGaze [43], EyeDiap [16], and RT-Gene [44] respectively. It is particularly impressive considering the fact that the improvement is solely achieved by our semi-supervised training scheme, without any network architectural modification. Notably, we adopt the basic network design for in-domain gaze estimation, and we believe our results can be further enhanced if equipped with more advanced architectures.

### Cross-domain Gaze Estimation.

Next we investigate the cross-domain generalization of OMNIGAZE under the cross-domain setting. Like previous studies [26, 71, 72, 73] that report results for models trained on ETH-XGaze [13] or Gaze360 [14], we evaluate OMNIGAZE trained on the same labeled gaze dataset for the sake of fairness. Table 3 reports our comparison results with SOTA domain-generalization methods on ETH-XGaze [13] and Gaze360 [14]. We can observe that our training algorithm always improves generalization on any dataset. This proves the effectiveness of making use of labeled datasets and large-scale diverse unlabeled datasets.

Table 3: **Quantitative cross-domain gaze estimation results** (§4.2).  $\mathcal{D}_E$ ,  $\mathcal{D}_G$ ,  $\mathcal{D}_M$ , and  $\mathcal{D}_D$  denote ETH-XGaze [13], Gaze360 [14], MPIIFaceGaze [43] and EyeDiap [16] datasets.

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
FullFace [43] [CVPRW17]	11.13	14.42	12.35	30.15
CANet [58] [AAAI20]	-	-	27.13	31.41
PureGaze [25] [AAAI22]	7.08	7.48	9.28	9.32
RAT [23] [CVPR22]	7.40	6.91	7.69	7.08
Gaze-Consistent [26] [AAAI23]	6.50	7.44	7.55	9.03
AGG [71] [CVPR24]	5.91	6.75	7.87	7.90
CLIP-Gaze [72] [AAAI24]	6.41	7.51	6.89	7.06
LG-Gaze [73] [ECCV24]	6.45	7.22	6.83	6.86
OMNIGAZE (Ours)	<b>5.07</b> $\pm 0.18$	<b>4.84</b> $\pm 0.23$	<b>4.95</b> $\pm 0.15$	<b>5.75</b> $\pm 0.29$

## 4.3 OMNIGAZE as Automatic Data Engine

A fundamental challenge in gaze estimation lies in the limited availability of diverse and well-annotated training data. To address this, we propose to leverage OMNIGAZE as a general-purpose *automatic data engine*—a system capable of generating reliable gaze annotations for face images under diverse conditions. Thus, we comprehensively validate the zero-shot gaze estimation capability of OMNIGAZE, *i.e.*, directly estimating gaze directions on unseen datasets or in-the-wild images.

**Dataset.** We curate a comprehensive training dataset by combining two public labeled 3D gaze datasets (*i.e.*, Gaze360 [14] and ETH-XGaze [13]) with six public unlabeled face datasets (see Table 1

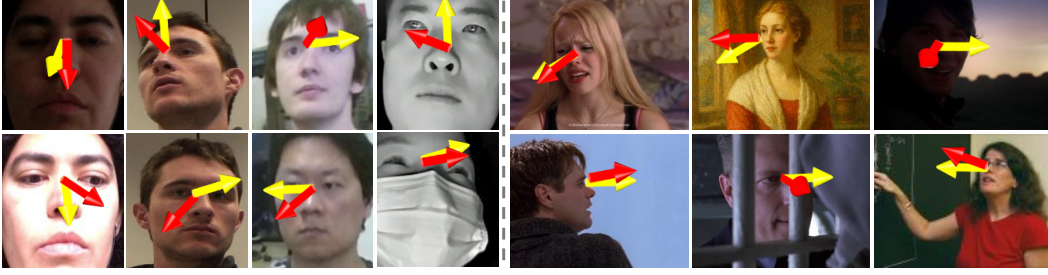


Figure 3: **Visual comparison results** (§4.3) on four unseen datasets (**left**) and in the wild (**right**). **Red** and **yellow** arrows represent gaze estimation predictions from our OMNIGAZE and base model trained only on labeled datasets. Four datasets from left to right: MPIIFaceGaze [76], EyeDiap [16], RT-Genie [44], and IVGaze [12].

Table 4: **Quantitative zero-shot generalization results** (§4.3) on MPIIFaceGaze [43], EyeDiap [16], RT-Genie [44], and IVGaze [12]. Note that all methods in the first block follow the in-domain evaluation. We provide three model scales, based on ViT-S (24.8M), ViT-B (97.5M), and ViT-L (335.3M), respectively.

Method	Zero-shot	MPIIFaceGaze [43]	EyeDiap [16]	RT-Genie [44]	IVGaze [12]
FullFace [43] [CVPRW17]	✗	4.93	6.53	10.00	13.67
Gaze360 [14] [ICCV19]	✗	4.06	5.36	7.06	8.15
CANet [58] [AAAI20]	✗	4.27	5.27	8.27	-
3DGazeNet [31] [ECCV24]	✗	4.00	-	-	-
DINO-B [74] [ICCV21]	✓	6.52	8.50	21.09	19.82
ViT-B [56] [ICLR21]	✓	6.17	8.85	20.73	18.97
FaRL-B [75] [CVPR22]	✓	6.09	8.12	19.80	18.06
OMNIGAZE-Small (Ours)	✓	3.70 $\pm$ 0.13	4.85 $\pm$ 0.21	10.02 $\pm$ 0.35	11.87 $\pm$ 0.29
OMNIGAZE-Base (Ours)	✓	3.44 $\pm$ 0.10	4.31 $\pm$ 0.12	9.09 $\pm$ 0.21	10.81 $\pm$ 0.27
OMNIGAZE-Large (Ours)	✓	3.03 $\pm$ 0.07	4.15 $\pm$ 0.14	9.01 $\pm$ 0.18	10.43 $\pm$ 0.20

for more details), comprising over 2.4M data samples in total. We assess OMNIGAZE on four unseen gaze estimation benchmarks, *i.e.*, MPIIFaceGaze [43], EyeDiap [16], RT-Genie [44], and IVGaze [12].

**Network Architecture.** In principle, our semi-supervised learning scheme can be applied into any feature encoder. In our experiments, we approach our algorithm on ViT-S, ViT-B, and ViT-L encoder.

**Quantitative Zero-shot Benchmark Evaluation.** Table 4 summarizes the zero-shot generalization comparison results on four representative unseen datasets[43, 16, 44, 12]. Note that we also adopt certain powerful ViT-based models (DINO-B [74], ViT-B [56], and FaRL-B [75]) as the gaze feature encoder to fine-tune on labeled datasets, and directly evaluate on unseen datasets. DINO-B and ViT-B have general semantic representations, while FaRL-B is designed for face analysis tasks. As seen, both with a ViT-B encoder, OMNIGAZE consistently outperforms other ViT-based models on diverse scenes, proving the high versatility of our algorithm. It is also worth noting that, though previous SOTA methods in the first block uses the corresponding training images (*not zero-shot anymore*), our OMNIGAZE is still evidently superior to them on certain datasets, *e.g.*,  $4.00^\circ \rightarrow 3.44^\circ$  on MPIIFaceGaze. This highlights the remarkable potential of our model as an automatic data engine.

**Qualitative Zero-shot Results.** We visualize predicted gaze directions by OMNIGAZE and base model (*i.e.*, train on labeled datasets and directly evaluate unseen datasets) on four unseen datasets in Fig. 3 (left). We observe that, after considering data scale-up and filtering out noisy pseudo labels, OMNIGAZE demonstrates generalization improvements with a lower angular error on each dataset.

**Qualitative Results in the Wild.** In Fig. 3 (right), we provide additional qualitative results on in-the-wild images, to resemble the practical zero-shot application in real-world conditions. Compared to the base model trained only on labeled datasets, our OMNIGAZE can predict gaze direction accurately in unseen diverse environments, *e.g.*, extreme head poses, challenging lighting conditions, and diverse appearances. We respectfully refer the reviewer to the appendix (§F) for more qualitative results.

#### 4.4 Diagnostic Analysis

For in-depth analysis, we conduct ablative studies using ViT-B encoder under zero-shot setting (§4.3).

**Key Component Analysis.** In Table 5a, we first examine the efficacy of essential components in our algorithm. The 1<sup>st</sup> row reports the result of the baseline model, which only trains on labeled datasets. For 2<sup>nd</sup> row, through jointly training on labeled datasets and large-scale diverse unlabeled data in a semi-supervised manner, we observe consistent and modest improvements against the baseline



Table 5: **A set of ablative studies** (§4.4) on multiple unseen datasets (MG: MPIIFaceGaze [76], ED: EyeDiap [16], RG: RT-Genie [44]) under the zero-shot setting.

Labeled data	Unlabeled data	Pseudo-labels selection	MG	ED	RG	Confidence score	MG	ED	RG
✓			6.17	8.45	20.73	w/o reward	4.97	5.42	13.75
✓	✓		4.97	5.42	13.75	$\hat{r}_k$ (Eq. 6)	3.71	4.68	9.96
✓	✓	✓	<b>3.44</b>	<b>4.31</b>	<b>9.09</b>	$r_k$ (Eq. 7) (Ours)	<b>3.44</b>	<b>4.31</b>	<b>9.09</b>

(a) Core components in OMNIGAZE

Evaluator Component	MG	ED	RG
BASELINE	4.52	5.19	12.01
+ Scene-specific Gaze Descriptions <i>only</i>	3.69	4.78	10.23
+ 3D Direction Vector <i>only</i>	4.03	4.93	10.56
Gaze Descriptions + 3D Direction Vector (Ours)	<b>3.44</b>	<b>4.31</b>	<b>9.09</b>

(c) Network design for reward model

(b) Confidence score

Filtering	Reweighting	MG	ED	RG
		4.97	5.42	13.75
✓		4.18	5.02	11.71
	✓	3.84	4.78	10.39
✓	✓	<b>3.44</b>	<b>4.31</b>	<b>9.09</b>

(d) Filtering strategy

Table 6: **Comparison results with the same backbone** (§4.4) on MPIIFaceGaze [76], EyeDiap [16] and RTGenie [12] under the in-domain setting.

Method	Backbone	MPIIFaceGaze [76]	EyeDiap [16]	RT-Genie [12]
FullFace [43]	AlexNet	4.93	6.53	10.00
Gaze360 [14]	ResNet-18	4.06	5.36	7.06
XGaze [13]	ResNet-50	4.80	6.50	12.00
CANet [58]	ResNet-50	4.27	5.27	8.27
GazeTR [57]	ResNet-18	4.00	5.17	6.55
AGE-Net [70]	ResNet-34	3.61	4.78	—
3DGazeNet [31]	ResNet-18	4.00	—	—
BASELINE	ResNet-18	4.48	5.56	7.45
<b>OminGaze (Ours)</b>	ResNet-18	<b>3.46</b>	<b>4.37</b>	<b>5.89</b>

on each dataset (e.g.,  $8.45^\circ \rightarrow 5.42^\circ$  on EyeDiap [16]). This supports our claim that large-scale unlabeled face images provides significantly high level of data diversity, thus enhancing zero-shot generalization of our method. Furthermore, after assessing and filtering out low-quality pseudo-labels via the reward model, the performance boosts to **3.44°** and **4.31°** on MPIIFaceGaze and EyeDiap, respectively. This suggests that scaling up datasets and further selecting high-quality samples can work in a collaborative manner, confirming the effectiveness of our overall algorithmic design.

**Network Design for Reward Model.** We investigate the impact of scene-specific gaze descriptions and gaze label interpolation (cf. Eq.5) in the reward model (§3.4), which is summarized in Table 5c. We construct a BASELINE model that directly predicts confidence scores based on the visual appearance and gaze labels. First, upon aggregating visual appearance and scene-specific gaze descriptions, all datasets observe notable improvements (e.g.,  $4.52^\circ \rightarrow 3.69^\circ$  on MPIIFaceGaze [76]). This verifies the effectiveness of learning semantic-aware gaze representation. Second, after interpolating gaze labels into 3D direction vectors, we also achieve significant angular error reductions, revealing the value of capturing the underlying geometry of gaze behaviors. Finally, our full reward model delivers the best performance across all datasets, validating the joint effectiveness of our network design.

**Pseudo-label Filtering Strategy.** We further probe the influence of different pseudo-label filtering strategies (§3.5). As outlined in Table 5d, by filtering out low-quality pseudo-labels, the method has a slight improvement of **0.79°** and **0.40°** angle error on MPIIFaceGaze [76] and EyeDiap [16], respectively. In addition, using confidence scores to reweight the importance of different samples, the model also exhibits improvements in angle error, achieving **3.84°** and **4.78°** on MPIIFaceGaze and EyeDiap, respectively. Outstandingly, OMNIGAZE achieves the highest performance on all datasets by integrating both confidence-based pseudo-label filtering and reweighting. The empirical evidence proves that our design facilitates gaze representation learning of the student model.

**Student Prediction in Confidence Evaluation.** To evaluate the contribution of student model and pseudo label prediction similarity  $\text{sim}(\hat{y}_k, y_k)$  in final scores  $r_k$  (cf. Eq. 7), we remove the similarity  $\text{sim}(\hat{y}_k, y_k)$  from the final confidence scores (cf. Eq. 6) and report the results in Table 5b. The absence of additional information about gaze estimation results in a slight performance decline, suggesting that this can serve as a complementary cue for the assessment ability of the reward model.

**Comparison with Same Backbone.** To ensure a fair comparison and address potential concerns regarding backbone capacity, we present additional experiments using the same lightweight ResNet-18 backbone as several SOTA methods [57, 70, 31] under the in-domain setting. The comparison results are summarized in Table 6. As shown, our OMNIGAZE achieves consistently better per-

formance than both the baseline and existing SOTA methods (*e.g.*, GazeTR [57], AGE-Net [70], and 3DGazeNet [31]) when adopting the similar backbone architecture. This confirms that the improvements stem from our proposed semi-supervised learning strategies rather than model scales.

## 5 Conclusion

In this work, we present OMNIGAZE, a novel semi-supervised framework to effectively generalize gaze estimation in the wild via harnessing the power of large-scale unlabeled data. To achieve this, we carefully construct a diverse collection of unlabeled face images, varying in head poses, illumination conditions, facial appearances, *etc.*, and devise a reward model to filter out noisy pseudo labels in unlabeled data. The reward model jointly reasons over visual appearance, semantic gaze context, and geometric gaze labels to predict confidence scores for accurate pseudo-label assessments. Extensive empirical analysis demonstrates that OMNIGAZE sets new SOTAs on both in-domain and cross-domain settings, and also exhibits excellent zero-shot generalization ability.

**Acknowledgement.** This work was supported by the National Natural Science Foundation of China (Grant No. U25A20442, 62222207, 62427808, 62472222), Fundamental Research Funds for the Central Universities (226-2025-00057), Zhejiang Provincial Natural Science Foundation of China (No. LD25F020001), the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2025A02), the Major Research Program of Jiangsu Province (Grant No. BG2024042), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX25\_0755), and the Natural Science Foundation of Jiangsu Province (No. BK20240080).

## References

- [1] Nathan J Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604, 2000.
- [2] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *ICCV*, pages 5724–5733, 2019.
- [3] Katerina Mania, Ann McNamara, and Andreas Polychronakis. Gaze-aware displays and interaction. In *SIGGRAPH*, pages 1–67. 2021.
- [4] Yun Suen Pai, Benjamin Tag, Benjamin Outram, Noriyasu Vontin, Kazunori Sugiura, and Kai Kunze. Gazesim: simulating foveated rendering using depth in eye gaze for vr. In *SIGGRAPH*, pages 1–2. 2016.
- [5] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions On Graphics*, 35(6):1–12, 2016.
- [6] Myungguen Choi, Daisuke Sakamoto, and Tetsuo Ono. Kuiper belt: Utilizing the “out-of-natural angle” region in the eye-gaze interaction for virtual reality. In *CHI*, pages 1–17, 2022.
- [7] Carlos Elmadjian and Carlos H Morimoto. Gazebar: Exploiting the midas touch in gaze interaction. In *CHI*, pages 1–7, 2021.
- [8] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for humanlike robots. In *CHI*, pages 25–32, 2014.
- [9] Zhuoqing Chang, J Matias Di Martino, Rachel Aiello, Jeffrey Baker, Kimberly Carpenter, Scott Compton, Naomi Davis, Brian Eichner, Steven Espinosa, Jacqueline Flowers, et al. Computational methods to measure patterns of gaze in toddlers with autism spectrum disorder. *JAMA pediatrics*, 175(8):827–836, 2021.
- [10] Sam Perochon, J Matias Di Martino, Kimberly LH Carpenter, Scott Compton, Naomi Davis, Brian Eichner, Steven Espinosa, Lauren Franz, Pradeep Raj Krishnappa Babu, Guillermo Sapiro, et al. Early detection of autism using digital behavioral phenotyping. *Nature Medicine*, 29(10):2489–2497, 2023.
- [11] Muhammad Qasim Khan and Sukhan Lee. A comprehensive survey of driving monitoring and assistance systems. *Sensors*, 19(11):2574, 2019.

- [12] Yihua Cheng, Yaning Zhu, Zongji Wang, Hongquan Hao, Yongwei Liu, Shiqing Cheng, Xi Wang, and Hyung Jin Chang. What do you see in vehicle? comprehensive vision solution for in-vehicle gaze estimation. In *CVPR*, pages 1556–1565, 2024.
- [13] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *ECCV*, pages 365–381, 2020.
- [14] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *ICCV*, pages 6912–6921, 2019.
- [15] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *CVPR*, pages 2176–2184, 2016.
- [16] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*, pages 255–258, 2014.
- [17] Hengfei Wang, Jun O Oh, Hyung Jin Chang, Jin Hee Na, Minwoo Tae, Zhongqun Zhang, and Sang-Il Choi. Gazecaps: Gaze estimation with self-attention-routed capsules. In *CVPR*, pages 2669–2677, 2023.
- [18] Jun O Oh, Hyung Jin Chang, and Sang-Il Choi. Self-attention with convolution and deconvolution for efficient eye gaze estimation from a full face image. In *CVPR*, pages 4992–5000, 2022.
- [19] Isack Lee, Jun-Seok Yun, Hee Hyeon Kim, Youngju Na, and Seok Bong Yoo. Latentgaze: Cross-domain gaze estimation through gaze-aware analytic latent code manipulation. In *ACCV*, pages 3379–3395, 2022.
- [20] Xin Cai, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Source-free adaptive gaze estimation by uncertainty reduction. In *CVPR*, pages 22035–22045, 2023.
- [21] Jiawei Qin, Takuru Shimoyama, and Yusuke Sugano. Learning-by-novel-view-synthesis for full-face appearance-based 3d gaze estimation. In *CVPR*, pages 4981–4991, 2022.
- [22] Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. Contrastive regression for domain adaptation on gaze estimation. In *CVPR*, pages 19376–19385, 2022.
- [23] Yiwei Bao, Yunfei Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with rotation consistency. In *CVPR*, pages 4207–4216, 2022.
- [24] Yoichiro Hisadome, Tianyi Wu, Jiawei Qin, and Yusuke Sugano. Rotation-constrained cross-view feature fusion for multi-view appearance-based gaze estimation. In *WACV*, pages 5985–5994, 2024.
- [25] Yihua Cheng, Yiwei Bao, and Feng Lu. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *AAAI*, volume 36, pages 436–443, 2022.
- [26] Mingjie Xu, Haofei Wang, and Feng Lu. Learning a generalized gaze estimator from gaze-consistent feature. In *AAAI*, volume 37, pages 3027–3035, 2023.
- [27] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.
- [28] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *ICCV*, pages 3681–3691, 2021.
- [29] Rakshit Kothari, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, and Jan Kautz. Weakly-supervised physically unconstrained gaze estimation. In *CVPR*, pages 9980–9989, 2021.
- [30] Pierre Vuillecard and Jean-Marc Odobez. Enhancing 3d gaze estimation in the wild using weak supervision with gaze following labels. *arXiv preprint arXiv:2502.20249*, 2025.
- [31] Evangelos Ververas, Polydefkis Gkagkos, Jiankang Deng, Michail Christos Doukas, Jia Guo, and Stefanos Zafeiriou. 3dgazenet: Generalizing 3d gaze estimation with weak-supervision from synthetic views. In *ECCV*, pages 387–404, 2024.
- [32] Yunjia Sun, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Cross-encoder for unsupervised gaze representation learning. In *ICCV*, pages 3702–3711, 2021.
- [33] Yiwei Bao and Feng Lu. Unsupervised gaze representation learning from multi-view face images. In *CVPR*, pages 1419–1428, 2024.

- [34] Yaoming Wang, Jin Li, Wenrui Dai, Bowen Shi, Xiaopeng Zhang, Chenglin Li, and Hongkai Xiong. Bootstrap autoencoders with contrastive paradigm for self-supervised gaze estimation. In *ICML*, 2024.
- [35] Jiawei Qin, Xucong Zhang, and Yusuke Sugano. Unigaze: Towards universal gaze estimation via large-scale pre-training. *arXiv preprint arXiv:2502.02307*, 2025.
- [36] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *CVPR*, pages 7314–7324, 2020.
- [37] Zhi-Hua Zhou and Ming Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24:415–439, 2010.
- [38] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020.
- [39] Jiwon Kim, Youngjo Min, Daehwan Kim, Gyuseong Lee, Junyoung Seo, Kwangrok Ryoo, and Seungryong Kim. Conmatch: Semi-supervised learning with confidence-guided consistency regularization. In *ECCV*, pages 674–690. Springer, 2022.
- [40] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.
- [41] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pages 10371–10381, 2024.
- [42] Timo Schneider, Boris Schauerte, and Rainer Stiefelhausen. Manifold alignment for person independent appearance-based gaze estimation. In *ICPR*, pages 1167–1172, 2014.
- [43] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It’s written all over your face: Full-face appearance-based gaze estimation. In *CVPRW*, pages 51–60, 2017.
- [44] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *ECCV*, pages 334–352, 2018.
- [45] Wangjiang Zhu and Haoping Deng. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *ICCV*, pages 3143–3152, 2017.
- [46] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *ECCV*, pages 721–738, 2018.
- [47] Rajeev Ranjan, Shalini De Mello, and Jan Kautz. Light-weight head pose invariant gaze tracking. In *CVPRW*, pages 2156–2164, 2018.
- [48] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. Generalizing eye tracking with bayesian adversarial learning. In *CVPR*, pages 11907–11916, 2019.
- [49] Yunfei Liu, Ruicong Liu, Haoqi Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *ICCV*, pages 3835–3844, 2021.
- [50] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.
- [51] Junbo Yin, Jin Fang, Dingfu Zhou, Liangjun Zhang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Semi-supervised 3d object detection with proficient teachers. In *ECCV*, pages 727–743. Springer, 2022.
- [52] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In *ICCV*, pages 16197–16208, 2023.
- [53] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, pages 596–608, 2020.
- [54] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, pages 18408–18419, 2021.
- [55] Siyuan Li, Weiyang Jin, Zedong Wang, Fang Wu, Zicheng Liu, Cheng Tan, and Stan Z Li. Semireward: A general reward model for semi-supervised learning. In *ICLR*, 2023.

- [56] 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. In *ICLR*, 2021.
- [57] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *ICPR*, pages 3341–3347, 2022.
- [58] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *AAAI*, pages 10623–10630, 2020.
- [59] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.
- [60] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, pages 67–74, 2018.
- [61] David Beniauguev. Synthetic faces high quality - text 2 image (sfhq-t2i) dataset, 2024.
- [62] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *CVPR*, pages 657–666, 2022.
- [63] 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 *ICML*, pages 8748–8763, 2021.
- [64] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv*, 2023.
- [65] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.
- [66] Mingjie Xu and Feng Lu. Gaze from origin: learning for generalized gaze estimation by embedding the gaze frontalization process. In *AAAI*, pages 6333–6341, 2024.
- [67] Huan Liu, Julia Qi, Zhenhao Li, Mohammad Hassanpour, Yang Wang, Konstantinos N Plataniotis, and Yuanhao Yu. Test-time personalization with meta prompt for gaze estimation. In *AAAI*, pages 3621–3629, 2024.
- [68] Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues. In *BMVC*, 2018.
- [69] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.
- [70] Yichen Shi, Feifei Zhang, Wenming Yang, Guijin Wang, and Nan Su. Agent-guided gaze estimation network by two-eye asymmetry exploration. In *ICIP*, pages 2320–2326, 2024.
- [71] Yiwei Bao and Feng Lu. From feature to gaze: A generalizable replacement of linear layer for gaze estimation. In *CVPR*, pages 1409–1418, 2024.
- [72] Pengwei Yin, Guanzhong Zeng, Jingjing Wang, and Di Xie. Clip-gaze: towards general gaze estimation via visual-linguistic model. In *AAAI*, pages 6729–6737, 2024.
- [73] Pengwei Yin, Jingjing Wang, Guanzhong Zeng, Di Xie, and Jiang Zhu. Lg-gaze: Learning geometry-aware continuous prompts for language-guided gaze estimation. In *ECCV*, pages 1–17, 2024.
- [74] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021.
- [75] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *CVPR*, pages 18697–18709, 2022.
- [76] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, pages 4511–4520, 2015.
- [77] Shreya Ghosh, Munawar Hayat, Abhinav Dhall, and Jarrod Knibbe. Mtgls: Multi-task gaze estimation with limited supervision. In *WACV*, pages 3223–3234, 2022.
- [78] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, pages 1021–1030, 2017.



- [79] MA FISCHLER AND. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.
- [80] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *ETRA*, pages 1–9, 2018.
- [81] Peng Wu, Xiankai Lu, Hao Hu, Yongqin Xian, Jianbing Shen, and Wenguan Wang. Logiczsl: Exploring logic-induced representation for compositional zero-shot learning. In *CVPR*, pages 30301–30311, 2025.
- [82] Hongyu Qu, Rui Yan, Xiangbo Shu, Hailiang Gao, Peng Huang, and Guo-Sen Xie. Mvp-shot: Multi-velocity progressive-alignment framework for few-shot action recognition. *IEEE Transactions on Multimedia*, 2025.
- [83] Hongyu Qu, Jianan Wei, Xiangbo Shu, and Wenguan Wang. Learning clustering-based prototypes for compositional zero-shot learning. In *ICLR*, 2025.
- [84] Ling Xing, Hongyu Qu, Rui Yan, Xiangbo Shu, and Jinhui Tang. Locality-aware cross-modal correspondence learning for dense audio-visual events localization. *arXiv preprint arXiv:2409.07967*, 2024.
- [85] Pengpeng Li, Xiangbo Shu, Chun-Mei Feng, Yifei Feng, Wangmeng Zuo, and Jinhui Tang. Surgical video workflow analysis via visual-language learning. *npj Health Systems*, 2(1):5, 2025.
- [86] Ling Xing, Alex Jinpeng Wang, Rui Yan, Xiangbo Shu, and Jinhui Tang. Vision-centric token compression in large language model. In *NeurIPS*, 2025.
- [87] Jianan Wei, Tianfei Zhou, Yi Yang, and Wenguan Wang. Nonverbal interaction detection. In *ECCV*, pages 277–295, 2024.
- [88] Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Hydra-sgg: Hybrid relation assignment for one-stage scene graph generation. In *ICLR*, 2025.
- [89] Guikun Chen, Jin Li, and Wenguan Wang. Scene graph generation with role-playing large language models. In *NeurIPS*, pages 132238–132266, 2024.
- [90] Liulei Li, Jianan Wei, Wenguan Wang, and Yi Yang. Neural-logic human-object interaction detection. In *NeurIPS*, pages 21158–21171, 2023.
- [91] Liulei Li, Wenguan Wang, and Yi Yang. Human-object interaction detection collaborated with large relation-driven diffusion models. In *NeurIPS*, pages 23655–23678, 2024.
- [92] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen. Cascaded human-object interaction recognition. In *CVPR*, pages 4263–4272, 2020.

## NeurIPS Paper Checklist

### 1. Claims

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

Answer: [\[Yes\]](#)

Justification: The main claims made in the abstract and introduction accurately reflect our paper's contributions and scope.

Guidelines:

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

### 2. Limitations

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

Answer: [\[Yes\]](#)

Justification: The limitations of this work is discussed and the related details can be found in Appendix [G](#).

Guidelines:

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

### 3. Theory assumptions and proofs

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

Answer: [NA]

Justification: This paper does not include theoretical results. Instead, we provide comprehensive ablation study on our provided method.

Guidelines:

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

#### 4. Experimental result reproducibility

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

Answer: [Yes]

Justification: We disclose all the information needed to reproduce the main experimental results of this paper and the data used in this paper is publicly available (see §3 and §4).

Guidelines:

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

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

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

Answer: [No]

Justification: The code of this paper will be released later while the data used in this paper is publicly available.

Guidelines:

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

## 6. Experimental setting/details

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

Answer: [Yes]

Justification: We specify all the training and test details in §4 to understand the experimental results.

Guidelines:

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

## 7. Experiment statistical significance

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

Answer: [Yes]

Justification: Please see the experimental results our method in Table 2, Table 3, and Table 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

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

Answer: [Yes]

Justification: We conduct all experiments on four NVIDIA RTX 3090 GPUs which is provided §4.

Guidelines:

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

#### 9. Code of ethics

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

Answer: [Yes]

Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>.

Guidelines:

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

#### 10. Broader impacts

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

Answer: [Yes]

Justification: We discuss the broader impacts of this work in Appendix G.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.



- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

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

Answer: [NA]

Justification: This paper poses no such risks.

Guidelines:

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

## 12. Licenses for existing assets

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

Answer: [Yes]

Justification: We have correctly and respectfully cited the original paper that produced the dataset used in this paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. **New assets**

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

Answer: [NA]

Justification: About the code and model of this paper, they will be publicly available as soon as the paper is published.

Guidelines:

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

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

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

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

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

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

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

Answer: [NA]

Justification: There is no research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: The core methodology and experimental pipeline of this study do not involve the use of any large language models (LLMs).

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## Summary of the Appendix

This supplementary document provides additional details for the main paper, titled “OMINGAZE: Reward-inspired Generalizable Gaze Estimation in the Wild”. The appendix is organized as follows:

- §A provides additional dataset analysis.
- §B presents more quantitative results.
- §C provides more implementation details of OMINGAZE.
- §D provides the pseudo-code of the reward model.
- §E shows generated scene-specific gaze descriptions along with corresponding face images.
- §F offers more qualitative results.
- §G discusses our limitations, broader impact, future work, and ethical considerations.

## A Additional Dataset Analysis

### A.1 Unlabeled Dataset Details

**CelebA** [59] is a large-scale dataset containing a variety of facial attributes, including 40 attributes such as makeup, age, and gender. This dataset consists of over 200,000 images collected from several celebrities in a screen-based gaze target setup. The head poses of these face images are mostly frontal, and there are minimal facial occlusions. To reduce redundancy and filter out samples with extreme head poses, we sub-sample about 177,000 images from CelebA.

**VGGFace2** [60] is a diverse face dataset that includes images from people of various ages, ethnicities, and identities around the world. Unlike CelebA, VGGFace2 focuses on real-world images that capture a variety of environmental conditions, backgrounds, and lighting variations. This dataset features a wide range of head poses, including frontal, profile, and other angles, providing greater challenges for gaze estimation. There are also varying levels of occlusion present in the images. To reduce redundancy while maintaining diversity, we select 55 images per identity from VGGFace2.

**FaceSynthetics** [28] is a synthetic face image dataset designed to simulate real-world variations in facial features, age, gender, and expression. This dataset includes a wide range of head poses, from frontal to profile. As the data is synthetic, various levels of occlusion are introduced for experimental purposes. Additionally, these face images are of high quality, with high resolution and rich detail.

**SFHQ-T2I** [61] is a synthetic face dataset that covers a broad demographic range, including variations in age, gender, ethnicity, and facial expressions. The dataset consists of around 120,000 images generated in various environments, including different lighting settings and background variations. The head poses are diverse, covering a wide range of angles, from frontal to profile and other perspectives. The image quality is high, with detailed and clear facial features.

**VFHQ** [62] is a high-fidelity face video dataset with a focus on generating highly realistic images. VFHQ contains various facial features and expressions, wide head poses and gaze variations. To reduce redundancy while maintaining diversity, we sub-sample every 20 frames from VFHQ.

**WebFace** [27] is a large-scale real-world face dataset collected from various online sources. It includes a diverse range of demographics, *e.g.*, different races, ages, genders, and facial expressions. Due to the data from the internet, WebFace features a wide variety of conditions, such as varying lighting, backgrounds, and facial poses. To reduce redundancy and filter out samples with extreme head poses, we sub-sample about 354,000 images from WebFace.

### A.2 Dataset Diversity

We provide additional examples from each dataset in Fig. 4. It can be observed that each dataset covers a narrow range of visual conditions, *e.g.*, specific capturing environments (*e.g.*, indoor [13], outdoor [14], synthetic [28, 61], or in-vehicle [12] scenes), controlled or studio-like illumination, specific facial appearance, and frontal head poses. Due to the domain-specific bias, gaze estimation models trained on one single dataset suffer from performance degradation when testing on new, unseen datasets. To overcome this limitation, *our OMINGAZE not only combines multiple labeled*

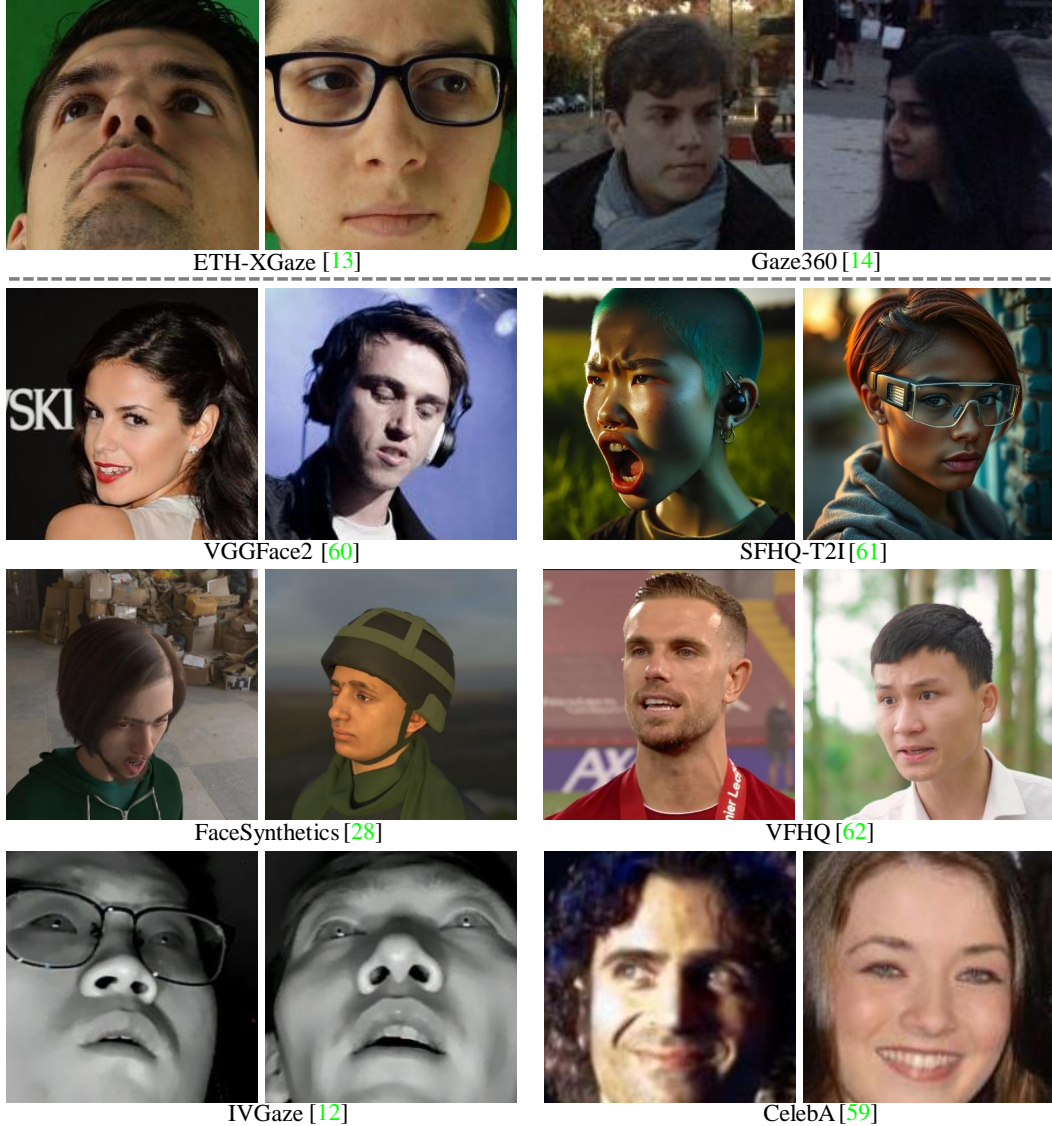


Figure 4: Additional examples of each dataset (§A.2). Our curated training dataset (*i.e.*, labeled datasets and large-scale unlabeled data) exhibits wide variability in terms of capturing environments (*e.g.*, indoor, outdoor, synthetic, or in-vehicle scenes), facial appearance, lighting conditions, head poses, *etc.*

*datasets but also incorporates a diverse collection of large-scale unlabeled datasets, varying in facial appearance, capturing environments, illumination conditions, head poses, eye occlusions, etc.* By effectively harnessing both labeled data and large-scale unlabeled datasets, OMNIGAZE mitigates domain bias and achieves robust, high-quality 3D gaze estimation in the wild.

## B More Quantitative Results

**Pseudo-label Update Strategy.** We next probe the effectiveness of the periodic pseudo-label update strategy and the choice of update interval  $K$  under the zero-shot setting, which is summarized in Table 7. The 1<sup>st</sup> row, which removes the pseudo-label update mechanism during training, results in a consistent performance drop on each dataset, confirming that noisy pseudo labels hinder zero-shot gaze estimation generalization of the student model. We further investigate the impact of the pseudo-label update interval  $K$ . As outlined in Table 7, our OMNIGAZE yields the best performance with a moderate update interval (*i.e.*,  $K = 10$ ). Too frequent updates regarding pseudo-labels (*i.e.*,



Table 7: **Analysis of pseudo-label update strategy and update interval  $K$**  (§B) on MPIIFaceGaze [43], EyeDiap [16] and RTGene [12] under the zero-shot setting. The adopted network designs are marked in red.

Update Strategy	Update Interval $K$	MPIIFaceGaze [43]	EyeDiap [16]	RTGene [12]
✗	-	3.77	4.98	10.15
✓	1	3.89	5.19	11.17
✓	5	3.52	4.38	9.27
✓	<b>10</b>	<b>3.44</b>	<b>4.31</b>	<b>9.09</b>
✓	20	3.59	4.55	9.78

$K = 1$ ) may introduce instability due to noisy early predictions, while overly sparse updates regarding pseudo-labels (*i.e.*,  $K = 20$ ) limit the model’s ability to correct its own mistakes, thus significantly increasing the training complexity. The empirical evidence proves that updating pseudo-labels every  $K = 10$  epochs mitigates the risk of early-stage overfitting to noisy labels while allowing the student model to progressively benefit from improved predictions over time.

**Unlabeled Data Size.** We study the impact of the unlabeled data size under the zero-shot setting in Fig. 5. We randomly sample subsets from each component dataset to create 25%, 50%, and 75% subsets of the full unlabeled data. Note that 0% subset means our OMNIGAZE is trained only on labeled data without any unlabeled data. When jointly training our OMNIGAZE on both labeled and unlabeled data, we observe that OMNIGAZE gains stable improvements (*e.g.*,  $6.17^\circ \rightarrow 4.32^\circ$  on MPIIFaceGaze [76] and  $20.73^\circ \rightarrow 16.98^\circ$  on RT-Gen [44]) as the size of unlabeled data grows (*e.g.*, 0% subset  $\rightarrow$  50% subset). When more than 75% subset, further increasing the unlabeled data size gives marginal performance gains. We speculate this is because our performance is primarily driven by data diversity, which appears sufficient at this scale. This study confirms our motivation to make full use of large-scale and diverse unlabeled data for predicting gaze direction accurately across various domains.

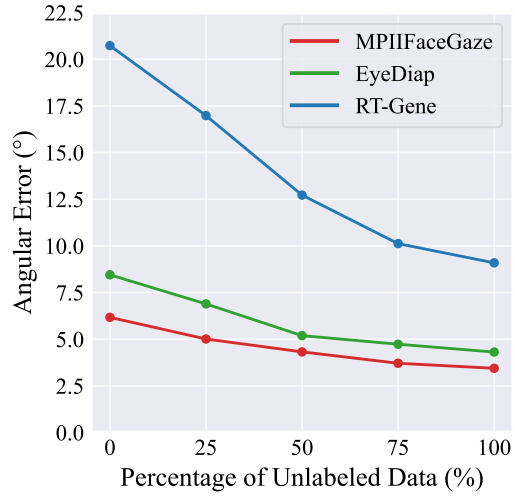


Figure 5: **The impact of unlabeled data size** (§B).

**Effect of Scene-specific Gaze Descriptions.** We investigate the effect of only using scene-specific gaze descriptions to train the reward model, without incorporating other semi-supervised strategies proposed in OMNIGAZE (*e.g.*, 3D direction vector pseudo labels, pseudo label filtering and reweighting, periodic pseudo-label updating). We report the results in Table 8. As seen, the performance improvement brought by using only scene-specific gaze descriptions is limited, *e.g.*,  $0.42^\circ$  gains on MPIIFaceGaze [76]. The empirical evidence proves that the major gains of OMNIGAZE come from the ensemble of our semi-supervised training strategies rather than the use of MLLM for calibration.

Table 8: **Effect of scene-specific gaze descriptions** (§B) on MPIIFaceGaze [76], EyeDiap [16] and RTGene [12] under the zero-shot setting.

Method	MPIIFaceGaze [76]	EyeDiap [16]	RTGene [12]
BASLINE	4.62	5.24	12.95
+scene-specific gaze descriptions	4.20	4.91	12.03
OMNIGAZE (Ours)	<b>3.44</b>	<b>4.31</b>	<b>9.09</b>

**Efficiency Analysis.** Table 9 reports the inference speed comparison between OMNIGAZE and the baseline under different backbones. Note that, OMNIGAZE introduces the reward model only during training for pseudo-label assessment and selection, and discards the reward model during the testing phase. Thus, as shown in Table 9, OMNIGAZE neither introduces additional computational overhead nor architectural modification to the base model during testing compared to the base model.

**Comparison with Methods Utilizing Large-scale Unlabeled Data.** To further validate the effectiveness of OMNIGAZE in leveraging unlabeled data, we conduct a comprehensive comparison with

Table 9: **Inference speed comparison** (§B) between baseline and OMNIGAZE under different backbones.

Method	Backbone	Inference Speed (ms)	MPIIFaceGaze [76]	EyeDiap [16]	RTGene [12]
Baseline	ResNet-18	2.7	4.48	5.56	7.45
OMNIGAZE (Ours)	ResNet-18	<b>2.7</b>	<b>3.46</b>	<b>4.37</b>	<b>5.89</b>
Baseline	ViT-B	10.9	4.62	5.24	12.95
OMNIGAZE (Ours)	ViT-B	<b>10.9</b>	<b>3.44</b>	<b>4.31</b>	<b>9.09</b>

several state-of-the-art methods that utilize large-scale unlabeled facial datasets in Table 10. As shown, OMNIGAZE consistently outperforms all the competitors trained with large-scale unlabeled data. This verifies the effectiveness of our model design and training strategy.

Table 10: **Comparison results with methods that utilize large-scale unlabeled data** (§B) on MPIIFaceGaze [76], EyeDiap [16] and Gaze360 [14] under the in-domain setting.

Method	Backbone	MPIIFaceGaze [76]	EyeDiap [16]	Gaze360 [14]
3DGazeNet [31]	ResNet-18	4.00	–	9.60
MTGLS [77]	ResNet-50	4.07	–	12.83
UniGaze [35]	ViT-B	4.75	5.52	9.64
ST-WSGE [30]	ViT-B	6.40	8.20	13.20
OMNIGAZE (Ours)	ViT-B	<b>2.97</b>	<b>4.07</b>	<b>9.12</b>

**More Cross-domain Results.** To further evaluate the generalization ability of OMNIGAZE, we train our OMNIGAZE on ground truth datasets that are more limited (*e.g.*, MPIIFaceGaze [76] or GazeCapture [15]), and test our model on testing datasets that have large diversity (*e.g.*, Gaze360 [14]). We compare our OMNIGAZE with LAEO [29] and 3DGazeNet [31], and provide the cross-domain comparison results in Table 11. As observed, OMNIGAZE consistently outperforms these methods under the cross-domain setting, demonstrating its effectiveness as a generalized gaze estimator.

Table 11: **Quantitative cross-domain gaze estimation results** (§B). Note that we train our OMNIGAZE on ground truth datasets that are more limited, and test our model on testing datasets that have large diversity.

Method	MPIIFaceGaze [76] → Gaze360 [14]	GazeCapture [15] → Gaze360 [14]
LAEO [16]	–	27.2
3DGazeNet [14]	17.6	17.6
<b>OminGaze (Ours)</b>	<b>13.8</b>	<b>14.2</b>

**Reward Model Training Strategy.** We study the impact of different reward model training strategies in Table 12. Here we pre-train the reward model by jointly using the labeled and unlabeled datasets, and then evaluate the performance of our model with the pre-trained reward model under the zero-shot setting. As seen, the performance of the pre-trained reward model is inferior to that of the online-trained counterpart. We hypothesize that this is because the pre-trained model tends to overfit to the initial distribution of pseudo labels, and thus struggles to assess the reliability of pseudo labels at different training stages.

## C Implementation Details

**Unlabeled Data Pre-processing.** We first detect facial landmarks [78] and estimate the 3D head pose through the Perspective-n-Point (PnP) algorithm [79]. Based on the estimated pose, we apply data normalization [80] to crop face images, so as to align each face image to a canonical coordinate system. Specifically, five key landmarks (*i.e.*, eye centers, nose tip, and mouth corners) are aligned to pre-defined facial templates. Such alignment procedure is crucial for reducing pose variation and improving the generalization of gaze estimation models. Moreover, we filter out samples with extreme head poses for unlabeled datasets.

**Training.** OMNIGAZE is trained with a batch size of 512. All face images are in the size of  $224 \times 224$  after the data normalization process. The training of OMNIGAZE can be divided into two stages: **i)** The teacher model is trained on labeled datasets for 50 epochs. We utilize the Adam optimizer [65] with an initial learning rate of 0.005, and a weight decay of 0.05. **ii)** We train the student model and reward model on both labeled and unlabeled data for 40 epochs with a base learning rate of 0.001 and 0.0001, respectively. Both labeled and unlabeled datasets are balanced in a minibatch to ensure each

Table 12: **Ablation studies on different reward model training strategies** (§B) on MPIIFaceGaze [76], EyeDiap [16] and RTGene [12] under the zero-shot setting.

Reward Model Training	MPIIFaceGaze [76]	EyeDiap [16]	RTGene [12]
Pre training	3.71	5.03	10.24
Online training ( <b>Ours</b> )	<b>3.44</b>	<b>4.31</b>	<b>9.09</b>

dataset accounts for an almost equal ratio. During training, we do not apply any image augmentation. The pseudo-label updating interval  $K$  is empirically set to 10.

**Reproducibility.** OMNIGAZE is implemented in PyTorch, and trained on 4 NVIDIA RTX 3090 GPUs with 24GB memory per card.

## D Pseudo Code

Algorithm S1 provides the pseudo-code of the reward model. To guarantee reproducibility, our code and pre-trained models will be made publicly available.

---

**Algorithm S1** Pseudo-code for the reward model of OMNIGAZE in a PyTorch-like style.

---

```

"""
img: input face image
gaze_des: scene-specific gaze descriptions
pseudo_label: pseudo gaze label
pre_label: student model prediction
"""

def RewardModel(img, gaze_des, pseudo_label, pre_label):

    # Encode images using CLIP
    img_feats = EncodeImage(img) # CLIP's visual encoder and MLP, Eq. 3

    # Tokenize and encode gaze descriptions via CLIP
    text_feats = EncodeText(Tokenize(gaze_des))

    # Multimodal feature integration via cross-attention
    img_enfeats = AvgPool(CrossAttn(img_feats, text_feats)) # Eq. 4

    # Convert pseudo gaze label to 3D direction vector
    theta, psi = pseudo_label
    dir_vector = [cos(psi)*sin(theta), cos(psi), cos(psi)*cos(theta)] # Eq. 5

    # Obtain initial confidence scores via the reward model
    cross_feat = CrossAttn(img_enfeats, dir_vector)
    int_score = Sigmoid(MLP(cross_feat)) # Initial confidence score in [0,1], Eq. 6

    # Compute cosine similarity between pseudo label and prediction
    sim_score = CosineSimilarity(pseudo_label, pre_label)

    # Define Final confidence scores
    final_input = Cat([int_score, sim_score])
    final_score = Sigmoid(MLP(final_input)) # Final confidence score, Eq. 7

    return final_score

```

---

## E Scene-specific Gaze Descriptions

The detailed, *scene-specific* descriptions are obtained by questioning MLLMs, *e.g.*, InstructBLIP [64], on the input image with a pre-defined prompt: *In 3D space, where is the person looking, including details about horizontal (left/right) direction, vertical (up/down) direction, and forward/backward relative to the viewer?* We provide several examples of *scene-specific* gaze descriptions in Fig. 6. As seen, *scene-specific* gaze descriptions complement low-level visual features by providing additional high-level spatial semantics, which are often ambiguous or underdetermined in raw image space.





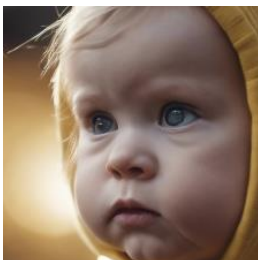

Face Image	Gaze Description	Face Image	Gaze Description
	The person is looking <b>slightly to their left and upward</b> . The person is <b>not looking directly forward</b> ; their gaze is <b>angled slightly away</b> from the viewer.		The person is looking <b>straight ahead and downward</b> , with no significant left or right deviation. The gaze is <b>angled downward</b> , so it is <b>away from</b> the viewer.
	The person is looking <b>slightly to the left and at eye - level</b> . The person maintains a <b>direct gaze</b> towards the viewer.		The person is looking <b>straight ahead</b> , with minimal deviation to the left or right. The gaze is at <b>eye - level</b> . The gaze is <b>looking forward</b> .
	The baby is looking <b>slightly to the left</b> relative to the viewer and directed <b>upwards</b> . The baby is <b>looking forward towards</b> the viewer.		The person is looking almost <b>straight ahead</b> , with minimal deviation to the right. The person is <b>looking forward</b> towards the viewer.

Figure 6: Examples of the generated scene-specific gaze descriptions along with face images (§E).

## F More Qualitative Results

We provide more qualitative results on in-the-wild images in Fig. 7, to resemble the practical zero-shot application in real-world conditions. We use a pre-trained facial landmark detector [78] to normalize input images for gaze estimation, and de-normalize the gaze direction predictions for visualization in the original image space. We observe that our OMNIGAZE can predict gaze directions accurately in unseen diverse environments, *e.g.*, extreme head poses, challenging lighting conditions, background environments, and diverse appearances, based on large-scale diverse unlabeled datasets and reward-driven pseudo label selection.

## G Discussion

**Limitation.** One limitation of our algorithm is that it needs a large amount of unlabeled data, varying in facial appearances, illumination conditions, head poses, and eye occlusions, which may be time-consuming to collect. However, in practice, enormous face images can be easily accessed by crawling from the Internet [27] or synthetic generation using generative models [28]; we compile face images from six public datasets [59, 60, 28, 61, 62, 27] to construct a large-scale unlabeled dataset encompassing over 1.4 million images, which covers significant data diversity. In the future, we will attempt to collect face images from more sources to train a more capable student model for generalizable gaze estimation in the wild. Additionally, though the reward model evaluates the reliability of pseudo labels and selects high-quality ones, OMNIGAZE still requires repeating the teacher model and reward model several times (*i.e.*, pseudo-label update strategy) to refine pseudo labels like many previous semi-supervised learning algorithms, which incurs extra computational costs. Therefore, an important consideration in future research is the balance between computational cost and the quality of pseudo labels provided by the teacher model. Moreover, though the reward



Figure 7: **Visual comparison results (§F)** on in-the-wild images. **Red** and **yellow** arrows represent gaze estimation predictions from our OMNIGAZE and base model trained only on labeled datasets.

model utilizes VLMs [63] and MLLMs [64] to offline extract visual features and generate scene-specific descriptions, respectively, it still requires extra computational budget for pseudo-label assessment during the training phase. Note that we directly discard the reward model during the testing phase without any network architectural modification or extra inference cost.

**Broader Impact.** This work introduces OMNIGAZE, a powerful semi-supervised framework for generalizable gaze estimation in the wild via harnessing both labeled data and large-scale unlabeled datasets, which overcomes the limitations of previous solutions struggling to generalize across diverse data domains due to the scarcity and insufficient diversity of annotated datasets. Like every coin has two sides, using our framework will have both positive and negative impacts. On the positive side, OMNIGAZE pushes the boundary of gaze estimation algorithms, particularly under the cross-domain and zero-shot/few-shot settings [81, 82, 83] that are common in real-world scenarios. This advancement can significantly contribute to a number of potential real-world applications, *e.g.*, virtual reality [3, 4, 5], human-computer interaction [6, 7, 8], video understanding [84, 85], and autonomous driving [11, 12]. For potential negative social impact, the reward model in OMNIGAZE relies heavily on VLMs [63] and MLLMs [64, 86] for pseudo-label assessment, thus leading to the reinforcement of biases and inequalities inherent in the data used during their large-scale pre-training stage. In addition, it is essential to ensure that gaze algorithms do not invade the privacy of people by adhering to ethical standards and legal regulations, so as to avoid potential negative societal impacts.

**Future Work.** Our OMNIGAZE aims to estimate high-quality 3D gaze directions for in-the-wild images in diverse conditions by making efficient use of large-scale unlabeled datasets and reward-driven pseudo label selection. It is also interesting to extend the idea of our algorithm to develop a scalable data engine for other visual tasks, which might improve data engineering techniques for producing reliable supervision. Moreover, the design of our reward model, which reasons over multimodal cues for pseudo-label assessments, stands for an early attempt to select high-quality pseudo-labels in gaze estimation and deserves to be further explored. In the future, we plan to generalize this framework to broader domains, *e.g.*, nonverbal communication understanding [87] and relational reasoning [88, 89, 90, 91, 92] tasks.

**Ethical Considerations.** Our research utilizes existing facial and gaze datasets, and does not generate any new face images. In accordance with ethical guidelines, we assume that these datasets are originally collected and published in compliance with relevant ethical and data protection standards. Our experimental protocols only focus on image content, ensuring that no personally identifiable information or links to other personal data are involved.