CLIP-Gaze: Towards General Gaze Estimation via Visual-Linguistic Model

Pengwei Yin[1,2*], **Guanzhong Zeng**[1,2*], **Jingjing Wang**[1,2†], **Di Xie**[1,2]
[1]Hikvision Research Institute, Hangzhou, China
[2]Zhejiang Key Laboratory of Social Security Big Data, China
{yinpengwei,zengguanzhong,wangjingjing9,xiedi}@hikvision.com

Abstract

Gaze estimation methods often experience significant performance degradation when evaluated across different domains, due to the domain gap between the testing and training data. Existing methods try to address this issue using various domain generalization approaches, but with little success because of the limited diversity of gaze datasets, such as appearance, wearable, and image quality. To overcome these limitations, we propose a novel framework called CLIP-Gaze that utilizes a pre-trained vision-language model to leverage its transferable knowledge. Our framework is the first to leverage the vision-and-language cross-modality approach for gaze estimation task. Specifically, we extract gaze-relevant feature by pushing it away from gaze-irrelevant features which can be flexibly constructed via language descriptions. To learn more suitable prompts, we propose a personalized context optimization method for text prompt tuning. Furthermore, we utilize the relationship among gaze samples to refine the distribution of gaze-relevant features, thereby improving the generalization capability of the gaze estimation model. Extensive experiments demonstrate the excellent performance of CLIP-Gaze over existing methods on four cross-domain evaluations.

Introduction

Gaze estimation has been widely applied for human behavior and mental analysis. High-accuracy gaze estimation can provide strong support for many applications, such as human-computer interaction(Andrist et al. 2014; Moon et al. 2014), saliency prediction(Xu, Sugano, and Bulling 2016), augmented reality(Padmanaban et al. 2017) and driver monitoring systems(Mavely et al. 2017). Recently, appearance-based gaze estimation methods(Zhang et al. 2015; Krafka et al. 2016; Cheng, Lu, and Zhang 2018; Cheng et al. 2020) achieved significant results with the development of deep learning. These methods achieve promising performance in the within-domain evaluation, but suffer from dramatic degradation in the cross-domain evaluations due to the domain gap.

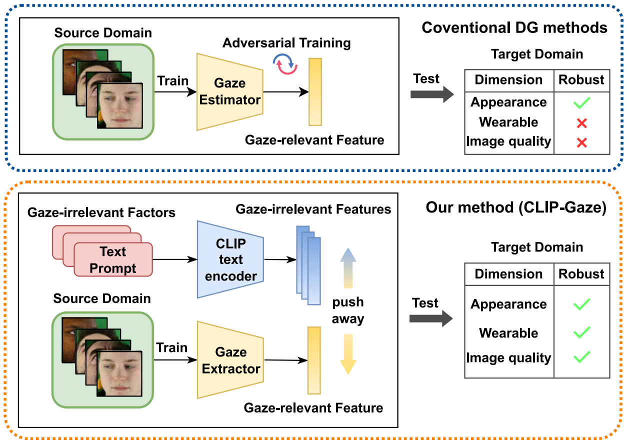

Figure 1: (1) Top subgraph: The conventional gaze generalization approach enhances the model's robustness by adversarial training, but can only mitigate a few gaze-irrelevant factors. (2) Bottom subgraph: Our method, CLIP-Gaze, constructs a text prompt from diverse language descriptions to obtain gaze-irrelevant features, and then push away the gaze-relevant feature from gaze-irrelevant features in the feature space to handle various gaze disturbing factors and achieve a robust model.

Gaze images contain rich information, but gaze labels are mainly determined by the eye direction. The eye area only occupies a small proportion of pixels in the face image(Xu, Wang, and Lu 2023), so the model prediction results are easily affected by various disturbation factors, such as hair, beard, expressions, makeup, hat, environment illumination, sensor noise and motion blur. Thus, these gaze-irrelevant factors lead to the domain gap, and pose a great challenge for the generalization of gaze model. To enhance the generalization of the model, some works purify the gaze feature with self-adversarial framework(Cheng and Bao 2022) or learn stable representation with the contrastive regression loss(Wang et al. 2022). However, their performance is limited due to the simplicity of the source domain, which fails to cover the diverse data types of the unseen target domain. To address such a problem, common domain generalization approaches(Xu, Wang, and Lu 2023; Yin et al. 2024) attempt to improve the diversity of the source dataset by data augmentation or data manipulation. However, these meth-

[*]These authors contributed equally.
[†]Corresponding author.

ods have not achieved substantial success because it is expensive to exhaustively enumerate all possible combinations of various gaze-irrelevant factors. Moreover, existing works have increased the data diversity in few factors such as identity, expression, and illumination, but they still struggle with many other important factors, which indicates that relying on a single visual modality is insufficient to handle all gaze-irrelevant factors.

On top of this, we propose a novel method named CLIP-Gaze that leverages a pretrained vision-language model(VLM) CLIP (Radford et al. 2021) to impart general transferable knowledge to the gaze estimation model and enhance the generalization ability of extracted gaze feature. CLIP learns rich visual-linguistic correlations through large-scale and aligned image-text datasets, which endows it with general, powerful and flexible representation capabilities. Consequently, CLIP-Gaze can exploit CLIP to flexibly handle various gaze-irrelevant factors, rather than relying on expensive models or uncontrollable adversarial methods that can only deal with limited gaze-irrelevant factors.

As shown in Fig. 1, we first employ the CLIP text encoder to generate a set of gaze-irrelevant features as gaze distractors from flexible and diverse language descriptions, which contains all gaze-irrelevant factors mentioned above. Subsequently, we maximize the distance between gaze-relevant feature and gaze-irrelevant features in the feature space via a feature separation loss function, so as to enhance the robustness of the gaze estimation model against various gaze disturbing factors. Furthermore, we develop a strategy for prompt optimizing and a loss function for feature refining, i.e., Personalized Context Optimization and Feature Rank Loss to further improve the domain generality of CLIP-Gaze for gaze estimation task. Personalized Context Optimization is a text prompt tuning (TPT) method that aims to avoid prompt engineering problems (a slight change in wording could make a huge difference in performance) and provide personalized text prompts for each individual. Feature Rank Loss refines the distribution of gaze-relevant features in gaze feature space by exploiting the relationship among samples, rather than only relying on the supervised loss of a single sample.

The main contributions can be summarized as follows:

- We design an efficient domain-generalization framework for gaze estimation, which is the first that introduces visual-linguistic modeling into gaze estimation task, and it deals with the diverse data types not seen in the source domain through a flexible manner.
- We develop a personalized text prompt tuning method to overcome prompt engineering issues and improve adaptation to the gaze estimation task. Furthermore, we propose a novel loss function based on the relationship among gaze samples to promote more reasonable feature distribution and learn robust gaze-relevant feature.
- Experimental results demonstrate that our CLIP-Gaze achieves remarkable performance improvement compared with the baseline model and also outperforms the state-of-the-art domain generalization approaches on gaze estimation tasks.

Related Works

Gaze Estimation Domain Generalization Appearance-based gaze estimation has become a hotspot(Zhang et al. 2015; Krafka et al. 2016; Cheng, Lu, and Zhang 2018; Cheng et al. 2020), but still has many challenges on cross-domain evaluation due to the domain gap caused by various gaze-irrelevant factors. The common approaches typically collect diverse and extensive gaze datasets(Zhang et al. 2020; Kellnhofer et al. 2019) to train a model with robust generalization capabilities, but collecting gaze data is often costly, and the diversity of the available data remains limited. This implies that we need to enhance models by using domain generalization (DG) methods, which can generalize to unseen distributions and improve cross-domain performance. However, most domain generalization methods are designed for classification tasks rather than regression tasks, and there are only a few studies on the generalization of gaze estimation. PureGaze(Cheng and Bao 2022) proposes a self-adversarial framework to alleviate the gaze disturbation by eliminating gaze-irrelevant feature and purifing gaze-relevant feature. Xu *et al.*(Xu, Wang, and Lu 2023) disturb training data against gaze-irrelevant factors by using adversarial attack and data augmentation, but they neglect many other gaze-irrelevant factors that still challenge gaze estimation.

Vision-Language Model Recently, many works have taken advantage of CLIP's flexible text manipulation and visual alignment capabilities to enhance the open detection or generalization performance of specific tasks, such as DetCLIP(Yao et al. 2022), DenseCLIP(Rao et al. 2022), CLIP-Gap(Vidit, Engilberge, and Salzmann 2023), Ordinalclip(Li et al. 2022), CLIP-Cluster(Shen et al. 2023) and so on. Furthermore, to improve the performance of vision-language models on downstream tasks, a more effective approach is to learn continuous text prompts by text prompt tuning(Zhou et al. 2022b,a). In this paper, we extend the usage of CLIP to gaze estimation, by utilizing its rich domain information to enhance the generalization ability of gaze estimation models, since it is trained on a large scale of data.

Method

Preliminaries

A generalized gaze representation learning can be formulated as:

$$\min_{G} E_{x,y}\ell(G(x,y)) + \gamma\ell_{reg}$$

where (x, y) is the input image and label, G is the gaze estimator, γ is the trade-off parameter and ℓ_{reg} denotes some regularization term.

Many methods attempt to learn a generalizable gaze feature by proposing various ℓ_{reg}, and their goals are to eliminate the influence of gaze-irrelevant factors on gaze estimation by explicit adversarial attack or data disturbation. However, it is difficult to cover all gaze-irrelevant factors. Although existing works have improved the robustness of models in some factors, i.e. identity, expression, and illumination, they still ignore many other important factors that challenge gaze estimation.

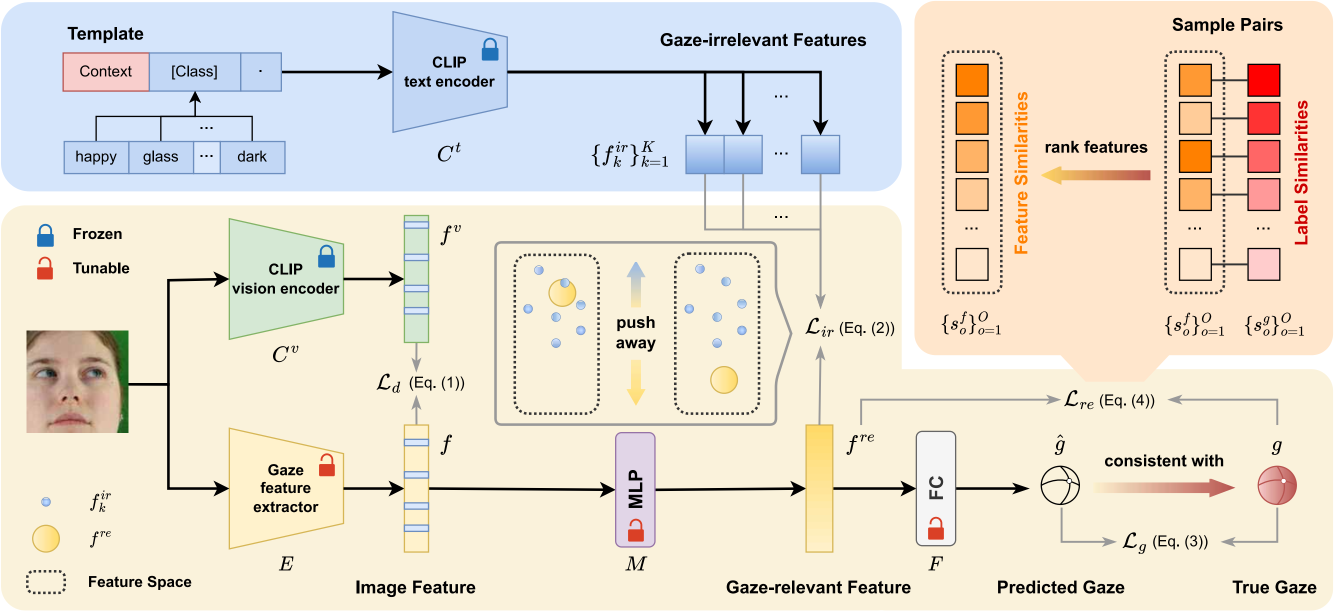

Figure 2: Overview of our CLIP-Gaze framework. We promote gaze domain generalization by introducing abundant knowledge outside the source domain to explicitly eliminate gaze-irrelevant features.

Therefore, we design a flexible and scalable method for gaze-irrelevant feature construction to cover a variety of target domains. We comprehensively define the gaze-irrelevant factors from three dimensions:

- **Appearance.** Face images contain rich but disturbing information for gaze estimation, such as identity features (e.g., face shape, eyebrows, eyes, mouth, and nose), expression variations, texture attributes (e.g., hair and beard color and style), and other characteristics (e.g., gender and age).
- **Wearable.** Besides the intrinsic factors of the face, gaze estimation is also influenced by external factors, such as wearing glasses, hats, helmets and makeup.
- **Image Quality.** Except for above image content factors, sensor noise, motion blur and environment also play a role. Therefore, we define multiple types of image clarity and illuminations(Schlett et al. 2022).

By enumerating the gaze-irrelevant factors exhaustively, we get a comprehensive set, which contains K gaze-irrelevant factors $\{\boldsymbol{c}_k\}_{k=1}^K$ in total and significantly exceeds the number of factors considered by previous methods. See the supplementary material for more details.

CLIP-Gaze Framework

In this section, we will elaborate on the proposed simple and flexible gaze generalization framework named CLIP-Gaze, which leverages the available pre-trained vision-language model to introduce rich general knowledge for gaze estimation task. Fig. 2 illustrates the whole pipeline of CLIP-Gaze that consists of two models. The first model is CLIP, which is fixed and used to generate image features $\boldsymbol{f}^v$ and construct multiple textual gaze-irrelevant features $\{\boldsymbol{f}_k^{ir}\}_{k=1}^K$ by defined prompt templates.

The second model is the gaze model, which is used to extract image feature $\boldsymbol{f}$ including gaze-relevant and gaze-irrelevant features through a convolution neural network (CNN). Then we use a multi-layer perceptron (MLP) to separate gaze-relevant features $\boldsymbol{f}^{re}$ and push it away from $\{\boldsymbol{f}_k^{ir}\}_{k=1}^K$ to learning robust gaze representation. In this way, the gaze estimator can generalize to multiple target domains. Next, we will describe more details.

Construct Gaze-irrelevant Features: Based on the flexibility of the VLM, we use the prompt template "`An image of a face with` $\{\boldsymbol{c}_k\}$`.`" to construct gaze-irrelevant features $\{\boldsymbol{f}_k^{ir}\}_{k=1}^K$ about the content of input images via CLIP text encoder C^t, where $\boldsymbol{c}_k$ is a gaze-irrelevant factor from the set $\{\boldsymbol{c}_k\}_{k=1}^K$.

Distill to CLIP Feature Space: For a given input image x, the whole gaze estimation process can be formulated as $\hat{\boldsymbol{g}} = F(M(E(x)))$. Specifically, we regard ResNet-18(He et al. 2016) as our backbone E to extract image feature $\boldsymbol{f}$, then use a MLP as feature filter M to separate gaze-relevant feature $\boldsymbol{f}^{re}$, and lastly we predict the gaze direction through a fully connected (FC) layer F.

The gaze-irrelevant features $\{\boldsymbol{f}_k^{ir}\}_{k=1}^K$ constructed above are fixed and located in the CLIP feature space. To remove gaze-irrelevant information from image features $\boldsymbol{f}$ and remain gaze-relevant features $\boldsymbol{f}^{re}$, we first align gaze image feature to the CLIP feature space by the following loss function:

$$\mathcal{L}_d\left(\boldsymbol{f}, \boldsymbol{f}^v\right) = 1 - \left(\frac{\boldsymbol{f} \cdot \boldsymbol{f}^v}{\|\boldsymbol{f}\| \, \|\boldsymbol{f}^v\|} + 1\right) * 0.5 \qquad (1)$$

Where $\boldsymbol{f}^v$ is the feature extracted by CLIP vision encoder

C^v. The loss value is normalized to the range of 0 to 1.

Separate Gaze-relevant Feature: To force the feature filter M to extract gaze-relevant feature $\boldsymbol{f}^{re}$, we minimize the similarity between $\boldsymbol{f}^{re}$ and $\boldsymbol{f}_k^{ir}$ where k from 1 to K. Note that the gaze-irrelevant factors corresponding to each sample may be different, we define the $\tilde{w}_k$ to summarize the correlation between one sample and k-th irrelevant factor. $\tilde{w}_k$ is used to alleviate the influence of gaze-irrelevant factors that are not involved in this image.

We use $\tilde{W}$ to represent the set of $\tilde{w}_k$, and $\tilde{W} = softmax(w_1, \ldots, w_K)$, where w_k is the degree of correlation between current sample and k-th irrelevant factor and can be described as:

$$w_k = \frac{\boldsymbol{f} \cdot \boldsymbol{f}_k^{ir}}{\|\boldsymbol{f}\| \, \|\boldsymbol{f}_k^{ir}\|}$$

Each sample has K gaze-irrelevant features elimination loss values, which will be re-weight by $\tilde{w}_k$ and the irrelevant loss is formally expressed as:

$$\mathcal{L}_{ir}\left(\boldsymbol{f}^{re}, \{\boldsymbol{f}_k^{ir}\}_{k=1}^{K}\right) = \sum_{k=1}^{K} \tilde{w}_k \frac{\boldsymbol{f}^{re} \cdot \boldsymbol{f}_k^{ir}}{\|\boldsymbol{f}^{re}\| \, \|\boldsymbol{f}_k^{ir}\|} \qquad (2)$$

Gaze Estimation: Lastly, we map the gaze-relevant feature $\boldsymbol{f}^{re}$ to gaze direction $\hat{\boldsymbol{g}}$, the gaze loss function is defined as:

$$\mathcal{L}_g\left(\hat{\boldsymbol{g}}, \boldsymbol{g}\right) = \arccos\left(\frac{\hat{\boldsymbol{g}} \cdot \boldsymbol{g}}{\|\hat{\boldsymbol{g}}\|\|\boldsymbol{g}\|}\right) \qquad (3)$$

where g is the gaze label.

Potential Issues and Improvement Schemes: Employing a suitable prompt template for CLIP-Gaze is non-trivial, as it requires prior knowledge about the gaze estimation task and proficiency in the language model's underlying mechanism(Radford et al. 2021). Hence, we propose a personalized text prompt tuning method to generate $\{\boldsymbol{f}_k^{ir}\}_{k=1}^{K}$ for each person.

Moreover, it only learns robust features from individual samples without further exploring the relationships between samples, which is suboptimal for the regression task. According to widely accepted understanding(Wang et al. 2022), the label and feature relationships among samples should exhibit a strong correlation. To tackle this limitation, we propose a novel loss function that explores the relationship among gaze samples and constructs a reasonable gaze-relevant feature distribution.

Personalized Context Optimization

To avoid the prompt engineering issues, a common approach is to learn prompts by prompt tuning. However, existing text prompt tuning methods may not be suitable for gaze estimation. CoOp(Zhou et al. 2022b) learns only one prompt for each class, which may fail to fully capture the gaze-irrelevant features, since each individual has a personalized facial property. CoCoOp(Zhou et al. 2022a) learns a prompt conditioned on each input image, by learning a lightweight neural network to impose the image content into the prompt. However, directly using the whole image content to learn the prompt may introduce some detrimental information into the prompt, since the image content also contain gaze-related information, and we want to use the prompt to extract only gaze-irrelevant feature. To this end, we propose a novel personalized context optimization (PCO) method. Fig. 3 illustrates our approach.

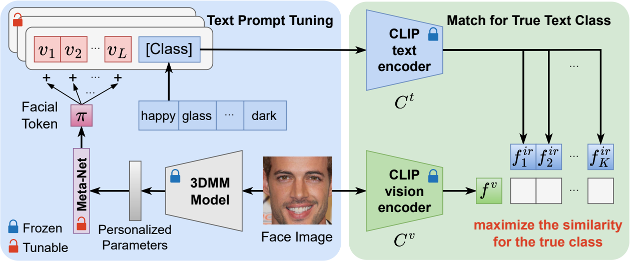

Figure 3: Our method, Personalized Context Optimization (PCO), has two learnable components: a context vector set and a lightweight neural network (Meta-Net) that produces a facial token for each identity, while the vision encoder, text encoder and 3DMM model are froze during training.

Learn Personalizing Text Prompt: First, our PCO utilizes face attribute classification as the proxy task to optimize the prompts. Specifically, we perform text prompt tuning on CelebA(Liu et al. 2015) , a large-scale and diverse face attributes dataset, to obtain a more robust gaze-irrelevant features. We extend its attribute number and leverage the CLIP model to get the pseudo labels as the attribute label for prompt tuning, inspired by previous works(Abdelfattah et al. 2023; Schlett et al. 2022). See the supplementary material for more details about how to generate pseudo labels. We follow CoOp(Zhou et al. 2022b) and introduce L learnable context vectors, $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_L\}$, each with the same dimension as the word embedding. The prompt for the i-th class, denoted by $\boldsymbol{t}_k$, is defined as $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_L, \boldsymbol{c}_k\}$ where $\boldsymbol{c}_k$ is the word embedding corresponding to one class in the class name set.

Next, to generate personalized text prompts for each individual and reduce gaze-related information, we use a pre-trained 3D Morphable Model (3DMM)(Tran and Liu 2018) to extract the personalization parameter $\boldsymbol{f}^m$ from the most frontal face image of each individual. Specifically, we use the 3DMM coefficients corresponding to identity as $\boldsymbol{f}^m$ to capture only the identity features of the face. Then, we feed $\boldsymbol{f}^m$ to Meta-Net to obtain an input-conditional token $\boldsymbol{\pi}$. Now, each personalizing context token is obtained by $\boldsymbol{v}_i(\boldsymbol{f}^m) = \boldsymbol{v}_i + \boldsymbol{\pi}$. The prompt for the i-th class is thus conditioned on the input, i.e., $\boldsymbol{t}_k(\boldsymbol{f}^m) = \{\boldsymbol{v}_1(\boldsymbol{f}^m), \boldsymbol{v}_2(\boldsymbol{f}^m), \ldots, \boldsymbol{v}_L(\boldsymbol{f}^m), \boldsymbol{c}_k\}$.

Furthermore, we feed $\boldsymbol{t}_k(\boldsymbol{f}^m)$ into text encoder $\boldsymbol{C}^t$ of CLIP to obtain text feature $\boldsymbol{f}_k^t$, and obtain face image feature $\boldsymbol{f}^v$ via vision encoder $\boldsymbol{C}^v$ of CLIP. During training, we specify the negative words for each category, such as “`not happy`” for the factor “`happy`”, and we could obtain negative text features $\overline{\boldsymbol{f}_k^t}$ via aforementioned operations. Then, we could maximize the prediction probability p_k for each gaze-irrelevant language description via a classification loss function, which is formulated as:

$$p_k = \frac{\exp(\mathrm{sim}(\boldsymbol{f}^v, \boldsymbol{f}_k^t)/\tau)}{\exp(\mathrm{sim}(\boldsymbol{f}^v, \boldsymbol{f}_k^t)/\tau) + \exp(\mathrm{sim}(\boldsymbol{f}^v, \overline{\boldsymbol{f}_k^t})/\tau)}$$

where τ is a temperature parameter learned by CLIP and sim(·,·) denotes cosine similarity.

Construct Personalizing Gaze-irrelevant Features: After finishing text prompt tuning, we first leverage the identity labels provided by ETH-XGaze and Gaze360 to select a frontal face image for each subject, and input this image and gaze-irrelevant factors $\{\boldsymbol{c}_k\}_{k=1}^K$ to our PCO module to construct K gaze-irrelevant features $\{\boldsymbol{f}_k^{ir}\}_{k=1}^K$ for each identity in the gaze training dataset.

Rank Gaze-Relevant Features

In this part, we refine the extracted gaze-relevant features through re-thinking the relationship of samples.

Intuitively, the gaze-relevant features with similar gaze directions should be close, which is suitable for multiple gaze domains. Derived from this idea, a contrastive regression loss (Wang et al. 2022) called CRLoss was proposed. CRLoss sets a threshold to push features with gaze angular difference larger than the threshold apart, and pull features with gaze angular difference less than the threshold together. However, it is hard to determine the threshold and it discards the finegrained relationships among gaze features, since gaze estimation is a continuous regression task. To deal with this deficiency, we explore the relationship among gaze-relevant features from a novel perspective.

We know that multiple gaze samples can be paired in pairs, and each pair of samples can be used to calculate a label similarity s^g and a feature similarity s^f. As shown in the upper right box of Fig. 2, for multiple sample pairs, the color intensity represents the similarity level, with darker color indicating higher similarity. Then we rank all pairs from high to low according to the label similarities of each pair, and impose penalties to force the feature similarities sequence to maintain the same order as the label similarities. At last, the distribution of gaze-relevant features will be more reasonable, and the gaze feature also can be more robust.

Specifically, based on the gaze label $\boldsymbol{g}$ and the gaze-relevant feature $\boldsymbol{f}^{re}$, a pair is composed of sample i and sample j, we calculate a label similarity s_{ij}^g and a feature similarity s_{ij}^f as:

$$s_{ij}^g = \frac{\boldsymbol{g}_i \cdot \boldsymbol{g}_j}{\|\boldsymbol{g}_i\| \, \|\boldsymbol{g}_j\|}, \quad s_{ij}^f = \frac{\boldsymbol{f}_i^{re} \cdot \boldsymbol{f}_j^{re}}{\|\boldsymbol{f}_i^{re}\| \, \|\boldsymbol{f}_j^{re}\|}$$

For pair p_1 and pair p_2 we use $\mathcal{L}_{re}$ to re-rank gaze-relevant features:

$$\mathcal{L}_{re} = \max\left(0, -S_{12} * \left(s_1^f - s_2^f\right)\right) \tag{4}$$

Where S_{12} evaluates to 1 if $s_1^g > s_2^g$, while evaluating to -1 when $s_1^g < s_2^g$.

For all samples in a mini-batch of size B, we construct $O = \frac{(B*(B-1))}{2}$ pairs of samples, then randomly choose two pairs O times to calculate the total rank loss.

Total Loss Function

In summary, the total loss function applied in our method is:

$$\mathcal{L} = \mathcal{L}_g + \lambda_1\mathcal{L}_d + \lambda_2\mathcal{L}_{ir} + \lambda_3\mathcal{L}_{re}$$

$\lambda_1, \lambda_2, \lambda_3$ are hyper-parameters, and we empirically set $\lambda_1 = \lambda_2 = \lambda_3 = 1.0$.

Experiments

Experiment Details

Gaze Data Details To verify the performance of our method in the gaze estimation task, we use ETH-XGaze(Zhang et al. 2020) and Gaze360(Kellnhofer et al. 2019) as training set, and test the gaze model on MPIIFaceGaze(Zhang et al. 2017) and Eye-Diap(Funes Mora, Monay, and Odobez 2014). Thus, we totally evaluate on four cross-domain task, and denote them as $\mathcal{D}_{\mathrm{E}}$ (ETH-XGaze)→$\mathcal{D}_{\mathrm{M}}$(MPIIFaceGaze), $\mathcal{D}_{\mathrm{E}}\rightarrow\mathcal{D}_{\mathrm{D}}$(EyeDiap), $\mathcal{D}_{\mathrm{G}}$(Gaze360)→$\mathcal{D}_{\mathrm{M}}$, $\mathcal{D}_{\mathrm{G}}\rightarrow\mathcal{D}_{\mathrm{D}}$. See Supplementary Material for more details on the data pre-processing.

Comparison Methods For Baseline, we only use $\mathcal{L}_g$ as training loss. For DG methods, we choose CDG(Wang et al. 2022), PureGaze and Xu *et al.*'s method(Xu, Wang, and Lu 2023) for comparison and use the results report by the author. Additionally, the results of SOTA UDA methods, including PnP-GA(Liu et al. 2021), RUDA(Bao et al. 2022), CRGA(Wang et al. 2022), LatentGaze(Lee et al. 2022), Liu *et al.*'s work(Liu et al. 2022) and UnReGA(Cai et al. 2023) as a reference.

Implementation Details We conduct the experiments on a single Tesla V100 GPU. We resize and normalize all the images to 224×224 and [0, 1]. We set the batch size to 128 and train the model for 30 epochs on ETH-XGaze and Gaze360. See Supplementary Materials for more details on the network, training setting, CLIP setting and others.

Performance Comparison with SOTA Methods

Quantitative result of four cross-domain gaze estimation tasks are shown in Tab. 1. The second row shows the comparison between our method and SOTA domain generalization (DG) methods. The CLIP-Gaze$^-$ is the plain CLIP-Gaze framework without PCO module and feature rank loss. It improves the generalizable capablity of Baseline and achieves the comparable performance with Xu *et al.*'s method using ResNet-18 as the backbone. The complete CLIP-Gaze achieves the best overall performance. It shows the state-of-the-art performance on three cross-domain evaluation tasks and achieves similar performance to the best method for $\mathcal{D}_{\mathrm{E}}\rightarrow\mathcal{D}_{\mathrm{D}}$, which proves the effectiveness of our proposed PCO module and feature rank loss.

Besides, we provide the comparison results with SOTA unsupervised domain adaption (UDA) methods in the third row of Tab. 1. Note that UDA methods require a small number of unlabeled target domain samples. It can be observed that our method demonstrates advanced performance with no access to target domain data. Specifically, we surpass LatentGaze on task $\mathcal{D}_{\mathrm{E}}\rightarrow\mathcal{D}_{\mathrm{D}}$, achieve better performance than Liu *et al.*.'s work on task $\mathcal{D}_{\mathrm{G}}\rightarrow\mathcal{D}_{\mathrm{M}}$ and outperform PnP-GA

Task	Methods	$\mid \mathcal{D}_t \mid$	$\mathcal{D}_\mathrm{E} \rightarrow \mathcal{D}_\mathrm{M}$	$\mathcal{D}_\mathrm{E} \rightarrow \mathcal{D}_\mathrm{D}$	$\mathcal{D}_\mathrm{G} \rightarrow \mathcal{D}_\mathrm{M}$	$\mathcal{D}_\mathrm{G} \rightarrow \mathcal{D}_\mathrm{D}$	Avg
DG	Baseline	0	8.35	9.66	7.58	9.01	8.65
	PureGaze	0	7.08	7.48	9.28	9.32	8.29
	CDG ‡	0	6.73	7.95	7.03	7.27	7.25
	Xu *et al.*	0	6.50	**7.44**	7.55	9.03	7.63
	CLIP-Gaze$^-$	0	7.04	8.51	7.55	7.73	7.71
	CLIP-Gaze	0	**6.41**	7.51	**6.89**	**7.06**	**6.97**
UDA	PnP-GA *	10	5.53	5.87	6.18	7.92	6.38
	RUDA	100	5.70	6.29	6.20	5.86	6.01
	CRGA	> 0	5.48	5.66	5.89	6.49	5.88
	LatentGaze	100	5.21	7.81	-	-	6.51
	Liu *et al.*	100	5.35	6.62	7.18	8.61	6.94
	UnReGA	100	5.11	5.70	5.42	5.80	5.51
SDA	Baseline *	100	4.63	5.86	5.67	6.26	5.61
	CLIP-Gaze*	100	**4.45**	**5.27**	**4.94**	**5.60**	**5.07**

Table 1: Comparison with SOTA methods. Results are reported by angular error in degrees, bold and underline denotes the best and the second best result among each column on one specific task. ‡ expresses the model employs ResNet-50 as backbone, * indicates that experimental settings are different, * denotes model is fine-tuned on target-domain.

and Liu *et al.*.'s method on task $\mathcal{D}_\mathrm{G} \rightarrow \mathcal{D}_\mathrm{D}$. It demonstrates the strength of our proposed method since it does not need any target domain information.

To further demonstrate the improvement of the proposed method, we randomly choose 100 target samples with labels for fine-tuning on our baseline and CLIP-Gaze model. The evaluation results in the last row of Tab. 1 show that our fine-tuned model consistently outperforms the baseline after fine-tuning. The details of the fine-tuning experiments can be found in the supplementary materials.

Ablation Study

Ablation Study of Text Prompt Tuning methods In this section, we compare different TPT methods in Tab. 2, where "Baseline" is the same as in Tab. 1. We denote CLIP-Gaze without text prompt tuning as w/o TPT, which achieved a significant improvement over the Baseline. In our TPT experiments, CoOp(Zhou et al. 2022b) learns only one prompt for each class, which only improves the performance slightly. CoCoOp(Zhou et al. 2022a) introduces instance-conditional text embedding for prompt tuning, but the instance containing gaze-relevant feature is detrimental to CLIP-Gaze, thus resulting in worse performance. CoCoOp* only uses one user face image embedding as a conditional embedding for prompt tuning, but this image embedding may also contain impure identity features that may affect prompt tuning. Our PCO method achieves the best performance, which demonstrates using the 3DMM model to extract personalizing features and introducing the identity-conditional text embedding can learn more suitable prompts.

Methods	$\mathcal{D}_\mathrm{E} \rightarrow \mathcal{D}_\mathrm{M}$	$\mathcal{D}_\mathrm{E} \rightarrow \mathcal{D}_\mathrm{D}$	$\mathcal{D}_\mathrm{G} \rightarrow \mathcal{D}_\mathrm{M}$	$\mathcal{D}_\mathrm{G} \rightarrow \mathcal{D}_\mathrm{D}$	Avg
Baseline	8.35	9.66	7.58	9.01	8.65
w/o TPT	6.72	8.16	7.07	7.64	7.40
CoOp	7.44	**7.42**	7.41	7.15	7.36
CoCoOp	7.56	8.03	7.52	8.12	7.81
CoCoOp *	7.36	7.77	7.31	8.46	7.73
PCO	**6.41**	7.51	**6.89**	**7.06**	**6.97**

Table 2: Comparison of different text prompt tuning methods for gaze model cross-domain evaluation in four tasks. Bold indicates the best results in each column, and underline denote the second best result results in each column.

Methods	$\mathcal{D}_\mathrm{E} \rightarrow \mathcal{D}_\mathrm{M}$	$\mathcal{D}_\mathrm{E} \rightarrow \mathcal{D}_\mathrm{D}$	$\mathcal{D}_\mathrm{G} \rightarrow \mathcal{D}_\mathrm{M}$	$\mathcal{D}_\mathrm{G} \rightarrow \mathcal{D}_\mathrm{D}$	Avg
Baseline	8.35	9.66	7.58	9.01	8.65
Appearance($\mathcal{A}$)	7.32	**7.09**	7.49	7.54	7.36
Wearable($\mathcal{W}$)	7.33	7.95	7.57	7.45	7.58
Quality($\mathcal{Q}$)	7.30	7.36	7.24	7.49	7.35
$\mathcal{W} + \mathcal{Q}$	6.75	7.87	6.99	7.95	7.39
$\mathcal{A} + \mathcal{Q}$	6.91	7.23	**6.85**	7.42	7.10
$\mathcal{A} + \mathcal{W}$	7.18	7.32	7.23	7.66	7.35
$\mathcal{A} + \mathcal{W} + \mathcal{Q}$	**6.41**	7.51	6.89	**7.06**	**6.97**

Table 3: Comparison of different gaze-irrelevant combinations for gaze model generalization in four cross-domain tasks. Bold indicates the best results in each column.

Ablation Study of Attributes and Conditions To investigate the effects of different gaze-irrelevant factors on the gaze model, we conduct different factor combinations experiments, as shown in Tab. 3. We can draw the following three conclusions: (1) The combination of multiple group of gaze-irrelevant factors outperforms the single group factors individually, possibly because filtering out more gaze-irrelevant features can make the model more generalizable; (2) Appearance($\mathcal{A}$) and image quality($\mathcal{Q}$) have larger impacts on the generalization ability of the gaze model than other factors, possibly because they account for the major domain gap; (3) The full combination of all factors($\mathcal{A}+\mathcal{W}+\mathcal{Q}$) achieves the best performance.

Methods	$\mathcal{D}_{\mathrm{E}} \rightarrow \mathcal{D}_{\mathrm{M}}$	$\mathcal{D}_{\mathrm{E}} \rightarrow \mathcal{D}_{\mathrm{D}}$	$\mathcal{D}_{\mathrm{G}} \rightarrow \mathcal{D}_{\mathrm{M}}$	$\mathcal{D}_{\mathrm{G}} \rightarrow \mathcal{D}_{\mathrm{D}}$	Avg
Baseline	8.35	9.66	7.58	9.01	8.65
$+\mathcal{L}_d$	7.82	9.32	7.24	9.14	8.38
$+\mathcal{L}_d + \mathcal{L}_{ir}$	<u>6.54</u>	7.94	7.44	7.57	7.37
$+\mathcal{L}_d + \mathcal{L}_{ir} + \mathcal{L}_{CR}$	6.96	7.89	7.24	7.28	7.34
$+\mathcal{L}_d + \mathcal{L}_{ir} + \mathcal{L}_{re}^{L1}$	6.81	7.91	7.07	**6.84**	7.16
$+\mathcal{L}_d + \mathcal{L}_{ir} + \mathcal{L}_{re}^{L2}$	6.79	<u>7.59</u>	<u>6.92</u>	7.21	<u>7.13</u>
$+\mathcal{L}_d + \mathcal{L}_{ir} + \mathcal{L}_{re}^{KL}$	6.96	7.89	6.97	7.08	7.23
$+\mathcal{L}_d + \mathcal{L}_{ir} + \mathcal{L}_{re}$	**6.41**	**7.51**	**6.89**	<u>7.06</u>	**6.97**

Table 4: Ablation study on loss functions. Results are reported by angular error in degrees.

Ablation Study of Loss Functions Compared with our baseline, we propose three loss functions to achieve the goal of extracting robust gaze-relevant features for gaze estimation. To investigate the effectiveness of $\mathcal{L}_d$, $\mathcal{L}_{ir}$ and $\mathcal{L}_{re}$, we conduct experiments to prove their effects by gradually increasing the loss term in baseline model training. For $\mathcal{L}_{re}$, we compare the CRLoss $\mathcal{L}_{CR}$ and our proposed rank loss and its variants, such as$\mathcal{L}_{re}^{L1}$, $\mathcal{L}_{re}^{L2}$, and $\mathcal{L}_{re}^{KL}$, which compute the L1, L2, and KL losses between the feature similarity and label similarity sequences of sample pairs, respectively.

Based on the results shown in in the third row of Tab. 4, we can see distilling gaze features to CLIP feature space enhances the average cross-domain performance, and there is a significant performance improvement after eliminating gaze-irrelevant features, this proves our proposed framework is efficient and superior. On this basis, we compare different loss forms of the relationship among sample features, it can be observed in the last row, the improvement brought by $\mathcal{L}_{CR}$ is slight and even worse than these variants of $\mathcal{L}_{re}$ that directly align the feature similarity to label similarity. Instead of learning the absolute values of feature similarities about sample pairs, our $\mathcal{L}_{re}$ constrains their relative magnitudes and achieve the best overall performance, these results above demonstrate the effectiveness of our framework and proposed loss functions.

Visualization of Extracted Features

To compare and analyze the extracted features $\boldsymbol{f}^{re}$ of different models, we follow the manner in (Wang et al. 2022) to visualize the distribution of feature on the task $\mathcal{D}_{\mathrm{G}} \rightarrow \mathcal{D}_{\mathrm{D}}$ with t-SNE(Van der Maaten and Hinton 2008). Fig. 4 displays the visualization results from four different models in Tab. 4, where feature points with similar gaze directions share similar colors.

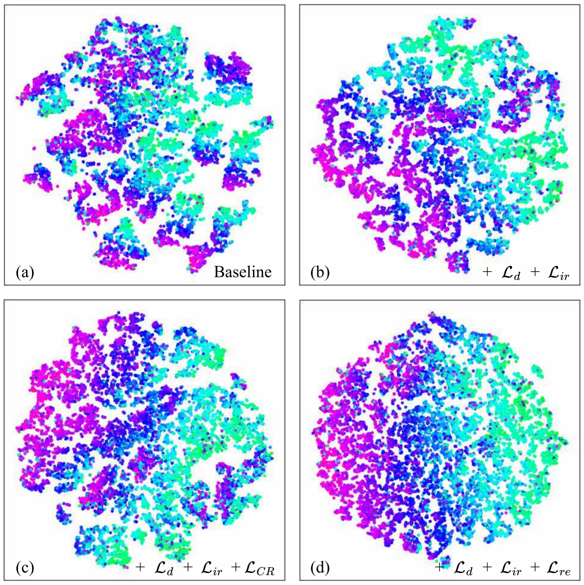

Figure 4: Visualization of the feature distribution. Different colors denotes different gaze directions and close gaze directions share similar colors. (Best viewed in color).

For the Baseline model, the features with different gaze directions are mixed together and the feature cluster is quite dispersed which is not sensible for regression task. After eliminating the gaze-irrelevant parts from extracted feature, the model of Fig. 4 (b) which omit the $\mathcal{L}_{re}$ in CLIP-Gaze shows the overall feature distribution becoming ordered on gaze directions, and the similar colors are close in the feature space. However, there is an unreasonably purple features cluster appears in the green area in the upper right of the box. Similarly, as shown in Fig. 4 (c), the model with additional $\mathcal{L}_{CR}$ does not improve the feature distribution compared with the model of Fig. 4 (b). This confirms that simply pushing features away or pulling features together by a fixed threshold is not optimal. In general, CLIP-Gaze has the most reasonable feature distribution and the visualization is shown in Fig. 4 (d), the gaze direction similarities and feature similarities have a strong correlation, this means our proposed feature rank loss $\mathcal{L}_{re}$ is effective. More visualizations provided in the supplementary materials.

Conclusion

In this paper, we propose a domain-generalization framework for gaze estimation models, which leverages visual-linguistic models to handle diverse target domains. Specifically, we define the gaze-irrelevant factors, such as face appearance, wearable, and image quality, and construct gaze-irrelevant features using language descriptions. Then, we decouple the gaze features from the CLIP feature space to enhance the model's generalization ability. To enchance the performance, a personalized context optimization method is proposed for text prompt tuning, and a rank loss is designed to learn a more reasonable gaze feature distribution. Our proposed framework achieves state-of-the-art performance on domain generalization for gaze estimation tasks.

Acknowledgements

This work was sponsored by National Key R&D Program of China (2023YFE0204200).

References

Abdelfattah, R.; Guo, Q.; Li, X.; Wang, X.; and Wang, S. 2023. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1348–1357.

Andrist, S.; Tan, X. Z.; Gleicher, M.; and Mutlu, B. 2014. Conversational Gaze Aversion for Humanlike Robots. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 25–32.

Bao, Y.; Liu, Y.; Wang, H.; and Lu, F. 2022. Generalizing Gaze Estimation With Rotation Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4207–4216.

Cai, X.; Zeng, J.; Shan, S.; and Chen, X. 2023. Source-Free Adaptive Gaze Estimation by Uncertainty Reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22035–22045.

Cheng, Y.; and Bao, Y. 2022. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 436–443.

Cheng, Y.; Lu, F.; and Zhang, X. 2018. Appearance-Based Gaze Estimation via Evaluation-Guided Asymmetric Regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Cheng, Y.; Zhang, X.; Lu, F.; and Sato, Y. 2020. Gaze Estimation by Exploring Two-Eye Asymmetry. *IEEE Transactions on Image Processing*, 29: 5259–5272.

Funes Mora, K. A.; Monay, F.; and Odobez, J.-M. 2014. 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*, 255–258.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; and Torralba, A. 2019. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6912–6921.

Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S. M.; Matusik, W.; and Torralba, A. 2016. Eye Tracking for Everyone. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2176–2184.

Lee, I.; Yun, J.-S.; Kim, H. H.; Na, Y.; and Yoo, S. B. 2022. LatentGaze: Cross-Domain Gaze Estimation through Gaze-Aware Analytic Latent Code Manipulation. In *Proceedings of the Asian Conference on Computer Vision*, 3379–3395.

Li, W.; Huang, X.; Zhu, Z.; Tang, Y.; Li, X.; Zhou, J.; and Lu, J. 2022. Ordinalclip: Learning rank prompts for language-guided ordinal regression. *Advances in Neural Information Processing Systems*, 35: 35313–35325.

Liu, R.; Bao, Y.; Xu, M.; Wang, H.; Liu, Y.; and Lu, F. 2022. Jitter Does Matter: Adapting Gaze Estimation to New Domains. *arXiv preprint arXiv:2210.02082*.

Liu, Y.; Liu, R.; Wang, H.; and Lu, F. 2021. Generalizing Gaze Estimation with Outlier-guided Collaborative Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Mavely, A. G.; Judith, J. E.; Sahal, P. A.; and Kuruvilla, S. A. 2017. Eye gaze tracking based driver monitoring system. In *2017 IEEE International Conference on Circuits and Systems (ICCS)*, 364–367.

Moon, A. J.; Troniak, D. M.; Gleeson, B.; Pan, M. K.; Zheng, M.; Blumer, B. A.; MacLean, K.; and Crof, E. A. 2014. Meet Me where I'm Gazing: How Shared Attention Gaze Affects Human-Robot Handover Timing. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 334–341.

Padmanaban, N.; Konrad, R.; Cooper, E. A.; and Wetzstein, G. 2017. Optimizing VR for All Users through Adaptive Focus Displays. In *ACM SIGGRAPH 2017 Talks*, SIGGRAPH '17. New York, NY, USA: Association for Computing Machinery. ISBN 9781450350082.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18082–18091.

Schlett, T.; Rathgeb, C.; Henniger, O.; Galbally, J.; Fierrez, J.; and Busch, C. 2022. Face image quality assessment: A literature survey. *ACM Computing Surveys (CSUR)*, 54(10s): 1–49.

Shen, S.; Li, W.; Wang, X.; Zhang, D.; Jin, Z.; Zhou, J.; and Lu, J. 2023. CLIP-Cluster: CLIP-Guided Attribute Hallucination for Face Clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20786–20795.

Tran, L.; and Liu, X. 2018. Nonlinear 3D Face Morphable Model. In *IEEE Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vidit, V.; Engilberge, M.; and Salzmann, M. 2023. CLIP the Gap: A Single Domain Generalization Approach for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3219–3229.

Wang, Y.; Jiang, Y.; Li, J.; Ni, B.; Dai, W.; Li, C.; Xiong, H.; and Li, T. 2022. Contrastive Regression for Domain Adaptation on Gaze Estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19354–19363.

Xu, M.; Wang, H.; and Lu, F. 2023. Learning a Generalized Gaze Estimator from Gaze-Consistent Feature. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3): 3027–3035.

Xu, P.; Sugano, Y.; and Bulling, A. 2016. Spatio-temporal modeling and prediction of visual attention in graphical user interfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3299–3310.

Yao, L.; Han, J.; Wen, Y.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; XU, C.; and Xu, H. 2022. DetCLIP: Dictionary-Enriched Visual-Concept Paralleled Pre-training for Open-world Detection. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 9125–9138. Curran Associates, Inc.

Yin, P.; Wang, J.; Dai, J.; and Wu, X. 2024. NeRF-Gaze: A Head-Eye Redirection Parametric Model for Gaze Estimation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; and Hilliges, O. 2020. ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, 365–381. Springer.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2015. Appearance-based gaze estimation in the wild. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4511–4520.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1): 162–175.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.