# Towards Privacy-Protected Generalized Gaze Estimation Using Diffusion Models and Domain Stability Adaptation Framework

**Ziyi Wang[1], Shengcheng Ye[1], Faming Fang[1]*, Haichuan Song[1]**

[1]School of Computer Science and Technology, East China Normal University
{51265901129, 52285901002}@stu.ecnu.edu.cn, {fmfang, hcsong}@cs.ecnu.edu.cn

## Abstract

Modern gaze estimation models can accurately predict human gaze from facial images. However, due to privacy concerns and intricate data collection procedures, gaze estimation datasets are typically smaller and less diverse compared to those for other vision tasks, which directly leads to poor generalization in gaze estimation models. Common solutions, such as domain adaptation models, require additional domain-specific data, yet such data is often difficult to obtain due to privacy restrictions. Meanwhile, domain generalization models suffer from limited performance due to insufficient training data. To address these fundamental challenges—privacy and data diversity—we explore privacy-preserving gaze data generation schemes and propose a novel data-driven generalization solution. Specifically, we develop two diffusion-based generative models, DDPM-Gaze and LDM-Gaze, for synthesizing gaze data. We demonstrate that synthetic data can significantly improve generalization performance when simply used with fine-tuning-based methods. Furthermore, we introduce the Domain Stability Adaptation (DSA) framework, a simple yet effective domain generalization approach that enhances model robustness by increasing the domain uncertainty of input samples while reducing prediction uncertainty. Extensive experiments validate the effectiveness of our synthetic data and demonstrate the superiority of our data-driven generalization solution.

## Introduction

Accurate estimation of human eye gaze can provide crucial clues for other fields, such as human-computer interaction (Majaranta and Bulling 2014), affective computing (D'Mello et al. 2012) and virtual reality (Xu et al. 2018). However, compared to datasets in most other computer vision tasks, gaze datasets typically have a smaller size and exhibit less diversity. We compare the four commonly used gaze estimation datasets in Fig. 2, finding that each dataset has deficiencies in either diversity or image quality. The reasons for this characteristic are mainly twofold. First, the precise annotation of gaze requires high standards for image capture and refined experimental setup. Second, regarding privacy concerns, the requirement for facial data has introduced greater obstacles to the data collection process.

*Corresponding Author, fmfang@cs.ecnu.edu.cn


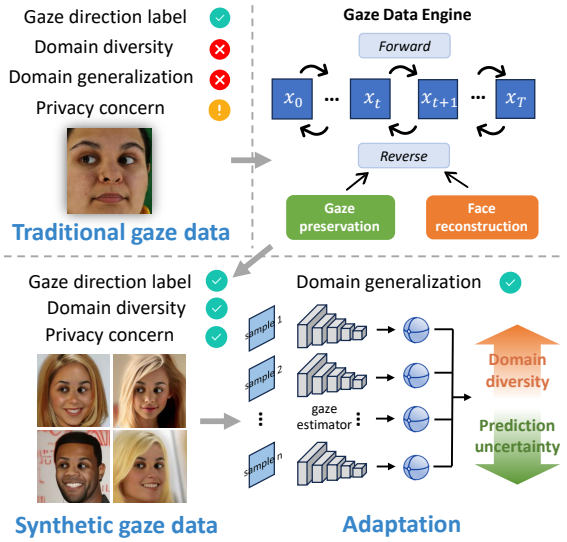


Figure 1: Proposed pipeline for tackling the generalization challenge in a fully data-driven manner.

Thus, most gaze estimation models face the generalization challenge that experience significant performance degradation in cross-domain (cross-dataset) tasks (Cheng et al. 2024). Many Domain Adaptation (DA) methods for gaze estimation have been proposed to address this issue (Liu et al. 2021; Cai et al. 2023), but they typically require supplementary source or target domain data. Due to privacy concerns, such data may well be inaccessible in practical applications. To achieve a more universal solution, Domain Generalization (DG) models are proposed to achieve good generalization performance on any unknown target domain. Existing DG methods focus on learning robust feature representations from the source domain to improve cross-domain performance (Xu, Wang, and Lu 2023; Yin et al. 2024b,a), but they also ignore the characteristics of other domains. Thus, we ponder: How can we expose the model to the characteristics of other domains without using data that poses privacy risks? Is addressing data limitations and privacy issues a new and effective approach to achieving the generalization of gaze estimation models?

Our first contribution is a diffusion-based gaze data en-

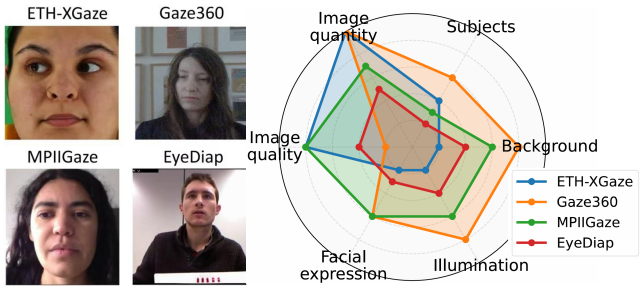


Figure 2: Compare the four commonly used datasets in the field of gaze estimation from the perspectives of image quantity, image quality, and the diversity of subjects, background, illumination, and facial expression.

gine that produces accurately labeled and highly diverse gaze data. We address the problem of insufficient diversity in conventional gaze data by introducing face-recognition data which present more variants in subjects and the environment. We employ diffusion model techniques to learn their data distribution and synthesize privacy-free face data via a controllable sampling denoising process. We also propose gaze protection diffusion process to ensure consistency between the synthetic data and the gaze direction. From an implementation perspective, the gaze generation engine includes two generation schemes, DDPM-Gaze and LDM-Gaze, built upon the classical DDPM and LDM frameworks, respectively. They generate new data in parallel, and we ultimately mix their data to obtain the final synthesized gaze data. This design stems from the complementary strengths of the two sources in key gaze-related factors, driven by their differing generative mechanisms.

We then propose a fine-tuning-based framework for generalization boosting, which is a purely data-driven solution. We employ it to verify the effectiveness of synthetic data and to demonstrate the benefits of data diversity in improving generalization. We first use gaze data from the source domain to pretrain the gaze estimator and generate synthetic data using the gaze data engine. Then, we fine-tune the model with the synthetic data. Our experiments demonstrate that diverse synthetic data significantly enhances the cross-domain performance of the gaze estimator.

Within our fine-tuning-based framework, we further introduce the Domain Stability Adaptation (DSA) framework, as its overview shown in Fig. 1. It arises from a key insight: directly fine-tuning on more diverse synthetic data can boost generalization yet also risks a new form of overfitting. Therefore, we explore a novel fine-tuning paradigm based on a new data engine. DSA first leverages the generative model to produce multiple images that share the same gaze label. Such data was impossible to obtain before the proposal of our method. With the intrinsically diverse feature of the face-recognition datasets and our two proposed generation paradigms, these label-consistent images can be regarded as multi-domain data with identical labels. DSA simultaneously predicts gaze directions for these images. In addition to constraining the predictions with their shared label, it also constrains the variance of the predicted directions, aiming to enhance prediction stability on the more diverse data, thus guiding the model to learn more generalized features. Experiments demonstrate that synthetic data combined with the DSA method effectively improves the generalization performance of gaze estimation models.

Our contributions are summarized as:

1. We formulate gaze estimation as a data-driven domain generalization problem and propose a diffusion-based gaze data engine, which consists of DDPM-Gaze and LDM-Gaze to produce accurately labeled and highly diverse gaze data.
2. We validate the reliability of the synthetic data employing a fine-tuning-based approach. We also demonstrate that fine-tuning on synthetic data effectively boosts the model's generalization performance.
3. We further propose the Domain Stability Adaptation (DSA) framework to enhance the generalization performance of gaze-estimation models by improving prediction stability on synthetic data. Our method achieves state-of-the-art performance among existing domain generalization models and even further boosts the performance of domain adaptation models.

## Related Work

### Diffusion Models and Data-driven Deep Learning

Diffusion models have emerged as a significant force in the field of generative modeling (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021). These models are inspired by the physical process of diffusion, where particles spread from an area of high concentration to an area of lower concentration over time. So far, diffusion models have been utilized across a wide range of generative modeling applications, including image generation (Zhao et al. 2023; Tang et al. 2024), image super-resolution (Gao et al. 2023; Wang et al. 2024b), image inpainting (Lugmayr et al. 2022), and mage editing (Avrahami, Lischinski, and Fried 2022). The advancements in diffusion models have provided new insights for data-driven deep learning. Training based on synthetic data or data augmentation through generative models has emerged as a novel solution (Boutros et al. 2023; Wang et al. 2024a; Fang et al. 2024; Islam et al. 2024).

### Cross-domain/dataset Gaze Estimation

The task of cross-domain/dataset gaze estimation has been a long-standing problem and absorbed adequate attention by researchers in recent years (Cheng et al. 2024). Two mainstream approaches have been proposed to solve the performance degradation problem of gaze estimation models on cross-dataset tasks, listed as Domain Adaptation (DA) methods and Domain Generalization (DG) methods. DA methods usually have a two-stage training process, including pretraining in the source domain and fine-tuning models with few labeled (Supervised Domain Adaption, SDA) or unlabeled (Unsupervised Domain Adaption, UDA) target domain data. Classic methods used in SDA include meta-learning (Park et al. 2019), gaze decomposition (Chen and

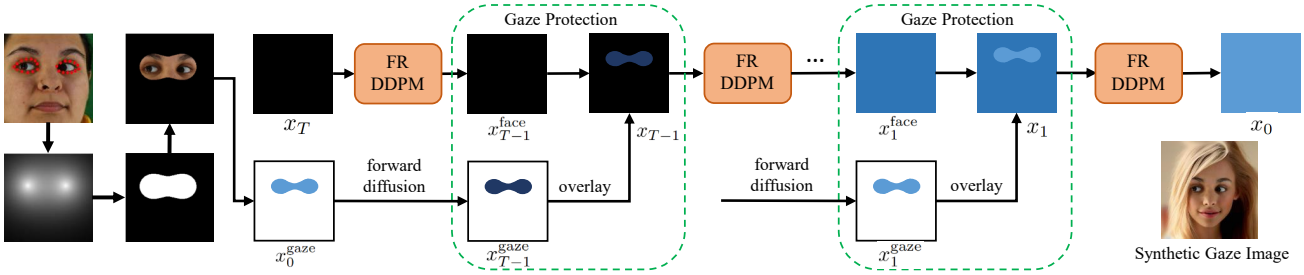


Figure 3: Illustration of our proposed DDPM-Gaze architecture.

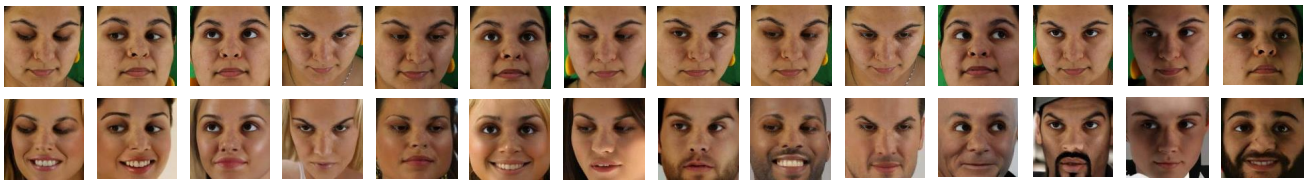

Figure 4: Synthetic gaze images with gaze labels generated by DDPM-gaze. The first row of images are from the ETH-XGaze dataset, and the second row consists of synthetic images.

Shi 2020) and differential approach for gaze estimation (Liu et al. 2018, 2019). UDA methods usually employ representation learning (Guo et al. 2020; Wang et al. 2022), rotation consistency (Bao et al. 2022), outlier guidance (Liu et al. 2021) and uncertainty reduction (Cai et al. 2023). Out of privacy concern, DG methods have gained much attention recently (Cheng, Bao, and Lu 2022; Xu, Wang, and Lu 2023; Yin et al. 2024b,a). DG gaze estimation models do not allow access to any information from the target domain when training the model on the source domain, which is more flexible in practical applications.

## Methods

### Gaze Data Engine

The Gaze Data Engine comprises two generative models: one based on DDPM (DDPM-Gaze) and the other on the latent diffusion model (LDM-Gaze). DDPM-Gaze produces precise yet relatively conservative gaze data, whereas LDM-Gaze generates more diverse data.

**DDPM-Gaze** is a training-free solution using a labeled gaze image as conditions during the sampling process to controllably generate new gaze data.

We employ an unconditional DDPM pre-trained on a face-recognition dataset (Face-DDPM) as the generative prior in pixel space. The reverse process of Face-DDPM is modeled by a neural network that predicts the parameters $\mu_\theta(x_t, t)$ and $\sigma_\theta(x_t, t)$ of a Gaussian distribution,

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)). \quad (1)$$

We progressively guide the sampling process of Face-DDPM using the crucial pixel areas of the eyes extracted from the authentic image, as shown in Fig. 3. Thanks to the characteristic of DDPM that maintains consistency in both the denoised feature space and the original image dimensions, we propose a novel gaze-protected sampling scheme for DDPM-Gaze. We denote the authentic gaze image from a gaze dataset as $x^{\text{gaze}}$, and $x_t^{face}$ represents the output of sampling at step $t$. A mask $M^{\text{eyes}}$ is used to extract pixels around the eyes area. Its acquisition method involves using facial landmark detection, creating an attention map via a Gaussian mixture model, and applying threshold-based binarization. $M^{\text{eyes}} \odot x$ represents the eyes area in an image and $(1 - M^{\text{eyes}}) \odot x$ represents the other facial parts. At every reverse step, we replace the pixels of eyes $M^{\text{eyes}} \odot x_{t-1}^{\text{face}}$ with a corresponding version of $M^{\text{eyes}} \odot x_{t-1}^{\text{gaze}}$, where $x_{t-1}^{\text{gaze}}$ keep the correct properties of the corresponding distribution,

$$x_{t-1}^{\text{gaze}} \sim \mathcal{N}(\sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t)\mathbf{I}). \quad (2)$$

$x_{t-1}^{\text{gaze}}$ can be directly obtained through the forward diffusion process of the diffusion model. Thus, we achieve the following expression for one reverse step in our approach,

$$x_{t-1} = M^{\text{eyes}} \odot x_{t-1}^{\text{gaze}} + (1 - M^{\text{eyes}}) \odot x_{t-1}^{\text{face}} \quad (3)$$

In practice, additional meticulous steps are required to ensure precise alignment between the gaze and face images, thereby enabling the generation of natural-looking facial images. We provide further details in the Supplementary Material. Fig. 4 shows a selection of synthetic image examples. Synthetic images exhibit stronger diversity, with a richer array of expressions, diverse lighting variations on the face, and a more varied range of facial features.

**LDM-Gaze** is built upon a latent diffusion model that shifts the diffusion process into the latent space. We use image features as conditional inputs to guide the generation process, because text features struggle to fully control the geometric details of gaze. The diffusion model in the latent space is designed as a U-net based DDPM model with cross attention mechanism, consistent with (Rombach et al. 2022). We propose a two-stage training method to progressively refine LDM-Gaze's control over facial attributes and gaze

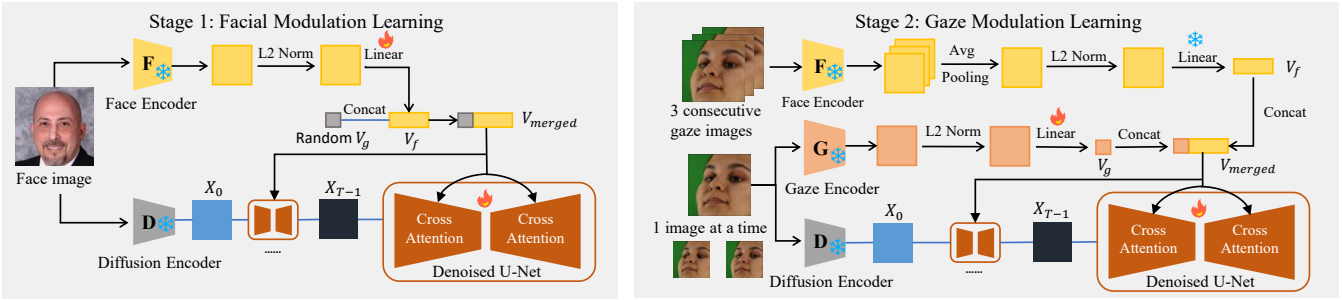


Figure 5: Illustration of our proposed LDM-Gaze architecture.

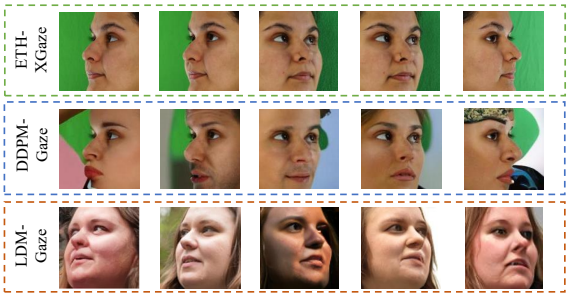


Figure 6: LDM-Gaze data exhibit greater diversity but also suffer from head-pose-related limitations.

direction attributes. The overview of LDM-Gaze's training workflow is shown in Fig. 5.

In **Facial Modulation Learning**, the model is trained to learn a stable and robust latent space distribution under the guidance of face-context information. Face-context information is obtained through an image encoder of a face recognition model (FAE), and the training data $X_i \in \mathbb{R}^{H\times W\times 3}$ is sourced from a face recognition dataset. During the FML training stage, we input images into FAE to obtain the face-context feature $Z_i$. After that, $Z_i$ passes through a linear layer to yield the face-context vector $V_f \in \mathbb{R}^m$. We then concatenate $V_f$ with a randomly initialized gaze-context vector $V_g \in \mathbb{R}^n$, before injecting it into LDM using a cross-attention mechanism. Due to the differences in scale between the face-context and gaze-context in the facial image, we do not explicitly add gaze conditions in FML. Instead, we use random vectors to retain the generative potential for diversity within the LDM. In **Gaze Modulation Learning**, we freeze the parameters for the face-context branch. Unlike FML, three consecutive images $\{Y_i, Y_{i+1}, Y_{i+2}\}$ from a gaze estimation dataset of the same subject are sent to FAE sequentially and the outputs are three homogeneous face-context features $\{Z_i, Z_{i+1}, Z_{i+2}\}$. Subsequently, these features are combined through channel average pooling. Meanwhile, gaze-context information is obtained with similar pipeline but encoded by an image encoder of a gaze estimation model (GAE). It is worth noting that as $\{Y_i, Y_{i+1}, Y_{i+2}\}$ are sent to train LDM, they share the same $V_f$ in three reverse diffusion processes. This mechanism helps prevent the model from overfitting, because a single subject contains a large amount of data in a gaze estimation dataset, and since most images are collected sequentially, the majority of facial attributes are shared in several consecutive images.

After training, synthetic images can be sampled from latent space controlled by gaze-context and face-context input from gaze data and face-recognition data respectively.

## Analysis of the synthetic gaze data

In this section, we discuss the rationale for including both DDPM-Gaze and LDM-Gaze data in the gaze data engine. DDPM-Gaze, by preserving the key periocular pixels of the source domain, generates data with high label consistency and also helps retain head-pose information. Preserving head-pose information in gaze datasets is of particular importance for gaze estimation: in natural situations the head rotates in concert with gaze direction, yet during model training this can lead a regression network to mistakenly regard head pose as a decisive cue. To address this issue, gaze datasets are deliberately collected so that gaze direction is decoupled from head orientation, facilitating more robust model training. As face-recognition datasets lack this decoupling, LDM-Gaze data - whose facial content is largely drawn from such datasets - exhibit stronger consistency between face identity and gaze direction (see Fig. 6). Meanwhile, face-recognition datasets are dominated by near-frontal faces, causing LDM-Gaze to lack samples with large head poses as well (see Fig. 6). Moreover, the accuracy of LDM-Gaze data needs to be validated; we provide the corresponding results in the Experiment Section. Nevertheless, LDM-Gaze remains essential, as it compensates for the diversity limitations that DDPM-Gaze inevitably incurs by preserving source-domain pixel information.

## Domain Stability Adaptation (DSA)

We first introduce our proposed fine-tuning-based framework. Aligned with traditional training scheme, the source-domain gaze data is used to pretrain the gaze model. Meanwhile, it is also leveraged by the gaze data engine to generate synthetic samples. Before cross-domain testing, the pretrained model will be fine-tuned on the synthesized data. We employ this method as a purely data-driven solution to verify the effectiveness of synthetic data. Compared with training from scratch on synthetic data, this method effectively reduces the required number of synthetic images.

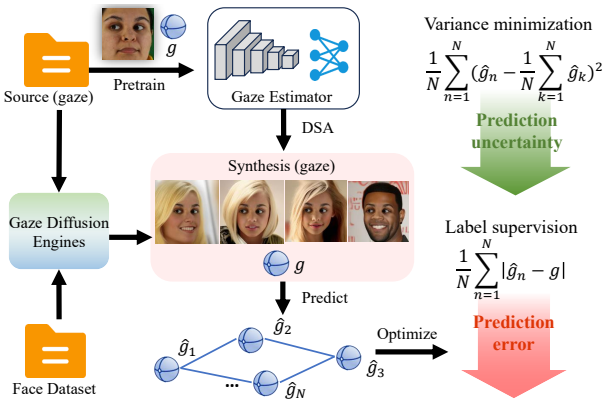


Figure 7: Illustration of the DSA roadmap.

Although the purely data-driven method can verify the effectiveness and plausibility of the synthetic data, relying solely on standard fine-tuning is suboptimal for generalization. Due to the limited scale of fine-tuning, it is prone to a new form of overfitting. To leverage synthetic data more effectively and improve the efficiency of fine-tuning, we further propose the DSA framework. DSA is inspired by a classic domain adaptation method based on collaborative learning (Cai et al. 2023; Liu et al. 2021). They employ the temporal average of a set of gaze estimators over time as pseudo-label supervision for the domain adaptation process. In comparison, DSA only adapts a specific gaze estimator with a set of labeled synthetic images. DSA enhances model's stability by increasing the samples' domain diversity and reduces model's prediction uncertainty as well as prediction error of diverse images with shared label, which can be obtained with our gaze data engine.

Let $I = \{I_1, I_2, ..., I_N\}$ denote a group of synthetic images, which includes $N$ images and they share a same gaze label $g$. $\hat{g} = \{\hat{g_1}, \hat{g_2}, ..., \hat{g_N}\}$ denotes the gaze prediction from a gaze estimator pretrained on source domain. For prediction uncertainty reduction, we update the parameters for adaptation as:

$$\ell_{uncertainty} = \frac{1}{N}\sum_{n=1}^{N}\left(\hat{g}_n - \frac{1}{N}\sum_{k=1}^{N}\hat{g}_k\right)^2. \quad (4)$$

For prediction error reduction, we continue to formulate the gaze estimation loss between the predicted vector $\hat{g}_n$ and the ground truth $g$ with the $L_1$ loss as:

$$\ell_{error} = \frac{1}{N}\sum_{n=1}^{N}|\hat{g}_n - g|. \quad (5)$$

The final loss function of DSA is defined as $\ell_{ada} = \ell_{uncertainty} + \lambda\ell_{error}$, where $\lambda$ is the weight parameter to balance two losses.

# Experiment

## Implementation Details

**Cross-domain gaze estimation.** We opt for Gaze360 (Kellnhofer et al. 2019) and ETH-XGaze (Zhang et al. 2020) as training set (source domain) for their varied gaze ranges, and a spectrum of head poses. To evaluate the efficiency of our model, we conduct tests across two well-established datasets: MPIIGaze (Zhang et al. 2015) and EyeDiap (Funes Mora, Monay, and Odobez 2014). In total, we undertake four cross-dataset experiments, which denoted as follows: $\mathcal{D}_E$ (ETH-XGaze) to $\mathcal{D}_M$ (MPIIGaze), $\mathcal{D}_E$ to $\mathcal{D}_D$ (DyeDiap), $\mathcal{D}_G$ (Gaze360) to $\mathcal{D}_M$, and $\mathcal{D}_G$ to $\mathcal{D}_D$. For preprocessing, we use the code provided in (Cheng et al. 2024).

| Model | $\mathcal{D}_E \rightarrow \mathcal{D}_M$ | $\mathcal{D}_E \rightarrow \mathcal{D}_D$ | $\mathcal{D}_E \rightarrow$ DG | $\mathcal{D}_E \rightarrow$ LG |
|---|---|---|---|---|
| CNN Baseline (ResNet-18) | 8.47 | 9.32 | 9.13 | 9.36 |
| PureGaze (ResNet-18) | 7.08 | 7.48 | 7.25 | 7.38 |
| CNN Baseline (ResNet-50) | 7.95 | 9.11 | 8.37 | 8.56 |
| PureGaze (ResNet-50) | 6.74 | 7.32 | 7.15 | 7.21 |

Table 1: Angular gaze errors performance (°) of cross-domain experiments on synthetic data. DG and LG denote the data from DDPM-Gaze and LDM-Gaze, respectively.

**Gaze data engine.** We implement the Face-DDPM in DDPM-Gaze with the guided diffusion model (Dhariwal and Nichol 2021) pretrained on CelebA-HQ (Liu et al. 2015) and FFHQ (Karras, Laine, and Aila 2019) datasets. We respectively implement the FAE and GAE in LDM-Gaze with the encoders of an advanced face recognition model ElasticFace (Fadi et al. 2022) and an advanced gaze estimation model PureGaze (Cheng, Bao, and Lu 2022). The face-context information in LDM-Gaze comes from the FFHQ dataset. During the generation phase of the gaze data engine, we randomly select data and labels from the source gaze dataset to guide conditional generation, ensuring the label distribution of the synthesized data is diverse.

The DSA framework is implemented on a NVIDIA GeForce RTX 3090 GPU using Pytorch framework. Images from face recognition datasets are all resized to a size of 224×224, and gaze images are all cropped to the same size without data augmentation. The adaptation in DSA is trained for 50 epochs using a Cosineannealing LR scheduler (Loshchilov and Hutter 2016) with a 5-epoch warm-up. We use the SGD optimizer with Nesterov momentum, a learning rate of 0.0001 for the parameter. More implementation details are included in the Supplementary Material.

## Reliability and characteristics of synthetic data

We first validate the synthetic data from the gaze data engine, focusing on evaluating both accuracy and diversity.

As for the accuracy of the LDM-Gaze, due to the generalization problem in gaze-estimation models, direct error quantification is infeasible. In Tab. 1, we design comparative experiments, which presents cross-domain results on both gaze datasets and synthetic data, using PureGaze (Cheng, Bao, and Lu 2022) and ResNet baselines to represent a domain-generalized model and a standard model. The performance on LDM-Gaze closely matches accuracy on the gaze dataset, and also matches DDPM-Gaze. By this method, we confirmed the accuracy of the LDM-Gaze data.

| Fine-tuning | $\mathcal{D}_E \rightarrow \mathcal{D}_M$ | $\mathcal{D}_E \rightarrow \mathcal{D}_D$ | $\mathcal{D}_G \rightarrow \mathcal{D}_M$ | $\mathcal{D}_G \rightarrow \mathcal{D}_D$ | Avg |
|---|---|---|---|---|---|
| Baseline | 7.08 | 7.48 | 9.28 | 9.32 | 8.29 |
| $\times$ 1000 ($\mathcal{D}$) | 6.92 | 7.33 | 8.43 | 8.38 | 7.77 |
| $\times$ 5000 ($\mathcal{D}$) | 6.65 | 7.30 | 7.81 | 8.09 | 7.46 |
| $\times$ 10000 ($\mathcal{D}$) | 6.47 | 7.22 | 7.46 | 7.88 | 7.26 |
| $\times$ 1000 ($\mathcal{L}$) | 6.85 | 7.32 | 8.57 | 8.44 | 7.80 |
| $\times$ 5000 ($\mathcal{L}$) | 6.50 | 7.25 | 7.93 | 8.05 | 7.43 |
| $\times$ 10000 ($\mathcal{L}$) | 6.42 | 7.15 | 7.56 | 7.96 | 7.28 |
| $\times$ 1000 ($\mathcal{D}+\mathcal{L}$) | 6.68 | 7.25 | 8.20 | 8.27 | 7.60 |
| $\times$ 5000 ($\mathcal{D}+\mathcal{L}$) | 6.42 | 7.13 | 7.25 | 7.83 | 7.33 |
| $\times$ 10000 ($\mathcal{D}+\mathcal{L}$) | 6.31 | 7.05 | 6.93 | 7.44 | 6.93 |

Table 2: Angular gaze errors (°) performance result of only using the fine-tuning-based framework.

| Task | Methods | $\lvert\mathcal{D}_t\rvert$ | $\mathcal{D}_E \rightarrow \mathcal{D}_M$ | $\mathcal{D}_E \rightarrow \mathcal{D}_D$ | $\mathcal{D}_G \rightarrow \mathcal{D}_M$ | $\mathcal{D}_G \rightarrow \mathcal{D}_D$ | Avg |
|---|---|---|---|---|---|---|---|
| DG | 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 | 6.41 | 7.51 | 6.89 | 7.06 | 6.97 |
| | LG-Gaze | 0 | 6.45 | 7.22 | 6.83 | 6.86 | 6.84 |
| | AGG | 0 | 7.10 | 7.07 | 7.87 | 7.93 | 7.49 |
| | GFAL | 0 | **5.72** | 6.97 | 7.18 | 7.38 | 6.81 |
| | Our DSA | 0 | 6.12 | **6.83** | **6.56** | **6.70** | **6.55** |
| 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 |

Table 3: Comparing DSA with SOTA methods with angular gaze errors performance in four cross-dataset tasks. The units of the data in the table are all degrees (°).

To demonstrate that synthetic data enhances generalization, we conducted experiments using our fine-tuning-based framework. In Tab. 2, we again use PureGaze as the pretrained baseline model, fine-tuning it with varying amounts of synthetic data and then evaluating on four cross-dataset tasks. $\mathcal{D}$ and $\mathcal{L}$ denote the use of data from DDPM-Gaze and LDM-Gaze, respectively; $\times$ indicates the number of images used; $\mathcal{D}+\mathcal{L}$ represents the mixed use of data from both sources with an equal split between them. Experimental results confirm the generalization benefits of synthetic data and demonstrate that jointly using both types of synthetic data further boosts performance. We also report the model's average performance within intervals defined by head-pose and gaze discrepancies in Fig. 8 to delve deeper into the properties of synthetic images. It can be observed that the model fine-tuned exclusively on LDM-Gaze performs markedly better on data where head pose and gaze directions are similar, while yielding larger prediction errors when these directions differ substantially—consistent with our earlier findings. Fine-tuning with a mixed data source is more effective at boosting overall model performance.

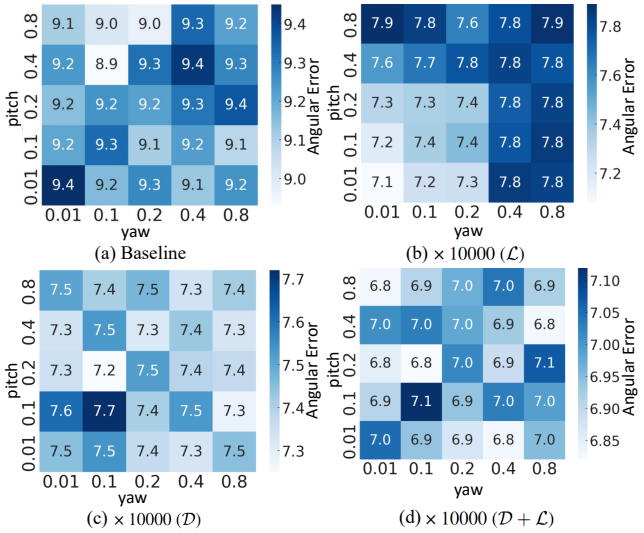


Figure 8: Generalization test results of the gaze model (PureGaze) after fine-tuning on different data. In each plot, points closer to the lower-left corner correspond to test samples whose head pose and gaze angles are more similar.

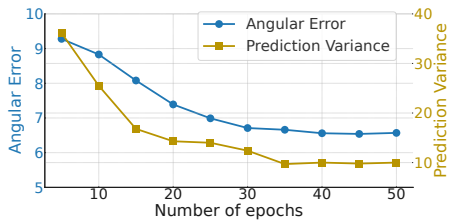


Figure 9: Training with DSA framework using PureGaze as pretrained model on task $\mathcal{D}_G \rightarrow \mathcal{D}_M$.

## Performance Comparison with SOTA Methods

In this section, we verify the advancement of the proposed DSA. We first introduce the main parameter settings of DSA. We use the generalized visual model PureGaze (Cheng, Bao, and Lu 2022) as our backbone, just as in the previous experiments, due to its open-source nature and its concise and elegant training process. Let $S = \{S_1, ..., S_K\}$ denotes the synthetic image set sending into DSA, and $K$ represents the number of gaze labels. $S_k = \{S_k^1, ..., S_k^N\}$ denotes that there are $N$ synthetic images sharing the same label. In the subsequent experiments, unless otherwise specified, t he number of synthetic images used is 10,000 and the default settings is $K = 2500$ and $N = 4$. Among the images with the same labels, the images generated by DDPM-Gaze and LDM-Gaze each account for half.

**Performance comparison with domain generalized methods.** As shown in Tab. 3, we compare our method with PureGaze baseline (Cheng, Bao, and Lu 2022), CDG (Wang et al. 2022), Xu *et al.* (Xu, Wang, and Lu 2023), CLIP-Gaze (Yin et al. 2024b), LG-Gaze (Yin et al. 2024a), AGG (Bao and Lu 2024) and GFAL (Xu and Lu 2024). We highlight the best performance in bold and the second-

| Method | Task | Origin | Ours | |
|---|---|---|---|---|
| | | | Fine-tuning | DSA |
| Gaze360 | $\mathcal{D}_G \rightarrow \mathcal{D}_M$ | 7.00 | 6.83 ↓2.4% | 6.31 ↓9.9% |
| | $\mathcal{D}_G \rightarrow \mathcal{D}_D$ | 8.77 | 8.38 ↓4.4% | 7.77 ↓11.4% |
| | $\mathcal{D}_G \rightarrow \mathcal{D}_M \rightarrow \mathcal{D}_G$ | 14.43 | 14.34 ↓0.6% | 14.11 ↓2.2% |
| | $\mathcal{D}_G \rightarrow \mathcal{D}_D \rightarrow \mathcal{D}_G$ | 14.75 | 14.58 ↓1.2% | 14.18 ↓3.9% |
| PnP-GA | $\mathcal{D}_G \rightarrow \mathcal{D}_M$ | 6.18 | 6.15 ↓0.5% | 5.92 ↓4.2% |
| | $\mathcal{D}_G \rightarrow \mathcal{D}_D$ | 7.92 | 7.58 ↓4.3% | 6.87 ↓13.3% |
| | $\mathcal{D}_G \rightarrow \mathcal{D}_M \rightarrow \mathcal{D}_G$ | 14.22 | 14.15 ↓0.5% | 14.00 ↓1.6% |
| | $\mathcal{D}_G \rightarrow \mathcal{D}_D \rightarrow \mathcal{D}_G$ | 14.38 | 14.21 ↓1.2% | 14.05 ↓2.2% |
| UnReGA- | $\mathcal{D}_G \rightarrow \mathcal{D}_M$ | 5.58 | 5.58 ↓0.0% | 5.13 ↓8.1% |
| | $\mathcal{D}_G \rightarrow \mathcal{D}_D$ | 5.84 | 5.79 ↓0.9% | 5.38 ↓7.9% |
| | $\mathcal{D}_G \rightarrow \mathcal{D}_M \rightarrow \mathcal{D}_G$ | 13.64 | 13.58 ↓0.4% | 13.11 ↓3.9% |
| | $\mathcal{D}_G \rightarrow \mathcal{D}_D \rightarrow \mathcal{D}_G$ | 13.95 | 13.84 ↓0.8% | 13.25 ↓5.0% |

Table 4: Performance comparison before and after applying synthetic data and DSA framework to UDA methods.

best with an underline. Our method achieves the best performance on three tasks and in terms of average performance, and is only slightly inferior to GFAL on one task. In Fig. 9, we present the relationship between the model's prediction error and the proposed prediction variance during training.

**Performance comparison with unsupervised domain adaptation (UDA) methods.** Compared to our method, they require images from the target domain for training. $|\mathcal{D}_t|$ represents the number of images required. These methods include PnP-GA (Liu et al. 2021), RUDA (Bao et al. 2022), CRGA (Wang et al. 2022), LatentGaze (Lee et al. 2022), Liu *et al.* (Liu et al. 2022) and UnReGA (Cai et al. 2023). Our method has generated a certain level of competitiveness with UDA methods, further narrowing the gap between DG and UDA methods while preserving privacy.

In Tab. 4, we use three open-source UDA methods as the pretrained models of our framework to verify the superiority of the DSA architecture. The three UDA models used are Gaze360 (Kellnhofer et al. 2019), PnP-GA, UnReGA-. UnReGA- is the open-source version of UnReGA without face enhancement; for more details, see (Cai et al. 2023). Specifically, both the fine-tuning-based framework and the DSA framework serve as post-training methods for the UDA method. We use the $L1$ loss as the training objective function for the fine-tuning-based framework. For each UDA model, we not only test the two cross-domain tasks but also test the performance on the source domain after domain adaptation. The experimental results show that our synthetic gaze data and DSA framework can further enhance the cross-domain performance of UDA methods while also improving the model's performance on the source domain.

## Parameter study of DSA

In this section, we conduct experiments to investigate the key parameters of DSA. In the performance comparison shown in Fig. 10, we keep $N = 4$ fixed and vary both the number of synthetic images and their source proportions. It can be seen that employing more synthetic images consistently benefits the model's generalization, while judiciously mixing data sources also plays a crucial role in improving its overall performance. In the experiments shown in Tab. 5, we fix the total number of synthetic images at 10 000, with DDPM-Gaze and LDM-Gaze images contributing equally, and vary the value of N. Special cases occur when $N = 1$, where the DSA model reduces to the standard fine-tuning method, and when $N = 3$, where 9 999 synthetic images are used instead. The results demonstrate that as N increases, the model imposes stricter requirements on prediction stability and achieves better generalization, further substantiating the effectiveness of the DSA approach.

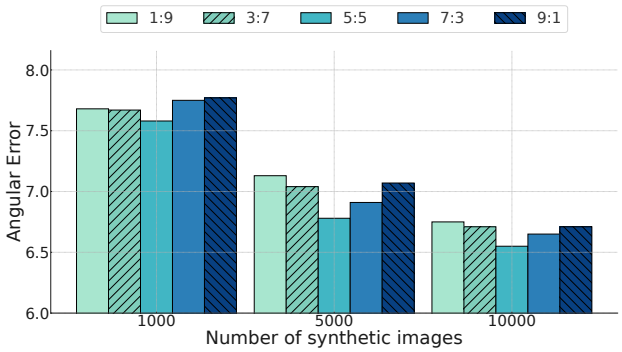


Figure 10: Average performance of DSA across four cross-domain tasks when varying the total number of synthetic images and their source proportions. The horizontal axis indicates the total number of synthetic images used, and a:b denotes the ratio of DDPM-Gaze to LDM-Gaze images.

| $N$ | $\mathcal{D}_E \rightarrow \mathcal{D}_M$ | $\mathcal{D}_E \rightarrow \mathcal{D}_D$ | $\mathcal{D}_G \rightarrow \mathcal{D}_M$ | $\mathcal{D}_G \rightarrow \mathcal{D}_D$ | Avg |
|---|---|---|---|---|---|
| $N = 1$ | 6.31 | 7.05 | 6.93 | 7.44 | 6.93 |
| $N = 2$ | 6.22 | 6.94 | 6.75 | 7.20 | 6.78 |
| $N = 3$ | 6.18 | 6.89 | 6.70 | 6.85 | 6.66 |
| $N = 4$ | 6.12 | 6.83 | 6.56 | 6.70 | 6.55 |

Table 5: Angular gaze errors (°) performance result of different values of $N$.

## Discussion and Conclusion

In this paper, we present the first synthetic-data-based solution aimed at improving the generalization of gaze estimation tasks. During the data-synthesis phase, we devised distinct mechanisms for DDPM-Gaze and LDM-Gaze to ensure both the accuracy and diversity of the generated data, and we conducted extensive experiments to verify their reliability. DSA is a simple yet effective framework for improving generalization based on synthetic data. Owing to its data-driven nature, both the selectable models and the optimizable tasks are highly flexible. We believe that synthetic data will open up new avenues for tackling a wide range of vision tasks. Moreover, as illustrated by the DSA framework in this paper, more flexible data structures also hold the potential to inspire even more effective training paradigms. In the future, we will continue to explore the potential of synthetic data in visual generalization tasks.

## Acknowledgements

This work was supported by the National Key R&D Program of China (2022ZD0161800), the National Natural Science Foundation of China under Grant 62271203, AI-Empowered Research Paradigm Reform and Discipline Leap Plan under Grant 2024AI01012 and the Open Research Fund of KLATASDS-MOE, ECNU.

## References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 18208–18218.

Bao, Y.; Liu, Y.; Wang, H.; and Lu, F. 2022. Generalizing gaze estimation with rotation consistency. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 4207–4216.

Bao, Y.; and Lu, F. 2024. From feature to gaze: A generalizable replacement of linear layer for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1409–1418.

Boutros, F.; Grebe, J. H.; Kuijper, A.; and Damer, N. 2023. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Int. Conf. Comput. Vis.*, 19650–19661.

Cai, X.; Zeng, J.; Shan, S.; and Chen, X. 2023. Source-free adaptive gaze estimation by uncertainty reduction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 22035–22045.

Chen, Z.; and Shi, B. 2020. Offset calibration for appearance-based gaze estimation via gaze decomposition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 270–279.

Cheng, Y.; Bao, Y.; and Lu, F. 2022. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *AAAI*, volume 36, 436–443.

Cheng, Y.; Wang, H.; Bao, Y.; and Lu, F. 2024. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Adv. Neural Inform. Process. Syst.*, 34: 8780–8794.

D'Mello, S.; Olney, A.; Williams, C.; and Hays, P. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies*, 70(5): 377–398.

Fadi, B.; Naser, D.; Florian, K.; and Arjan, K. 2022. Elasticface: Elastic margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 1578–1587.

Fang, H.; Han, B.; Zhang, S.; Zhou, S.; Hu, C.; and Ye, W.-M. 2024. Data augmentation for object detection via controllable diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1257–1266.

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.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023. Implicit diffusion models for continuous super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 10021–10030.

Guo, Z.; Yuan, Z.; Zhang, C.; Chi, W.; Ling, Y.; and Zhang, S. 2020. Domain adaptation gaze estimation by embedding with prediction consistency. In *ACCV*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Adv. Neural Inform. Process. Syst.*, 33: 6840–6851.

Islam, K.; Zaheer, M. Z.; Mahmood, A.; and Nandakumar, K. 2024. DiffuseMix: Label-Preserving Data Augmentation with Diffusion Models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 27621–27630.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 4401–4410.

Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; and Torralba, A. 2019. Gaze360: Physically unconstrained gaze estimation in the wild. In *Int. Conf. Comput. Vis.*, 6912–6921.

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.

Liu, G.; Yu, Y.; Mora, K. A. F.; and Odobez, J.-M. 2018. A differential approach for gaze estimation with calibration. In *Brit. Mach. Vis. Conf.*, volume 2, 6.

Liu, G.; Yu, Y.; Mora, K. A. F.; and Odobez, J.-M. 2019. A differential approach for gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(3): 1092–1099.

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 *Int. Conf. Comput. Vis.*, 3835–3844.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Int. Conf. Comput. Vis.*, 3730–3738.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 11461–11471.

Majaranta, P.; and Bulling, A. 2014. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*, 39–65. Springer.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *Int. Conf. Mach. Learn.*, 8162–8171. PMLR.

Park, S.; Mello, S. D.; Molchanov, P.; Iqbal, U.; Hilliges, O.; and Kautz, J. 2019. Few-shot adaptive gaze estimation. In *Int. Conf. Comput. Vis.*, 9368–9377.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 10684–10695.

Tang, J.; Nie, Y.; Markhasin, L.; Dai, A.; Thies, J.; and Nießner, M. 2024. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 20507–20518.

Wang, Y.; Gao, R.; Chen, K.; Zhou, K.; Cai, Y.; Hong, L.; Li, Z.; Jiang, L.; Yeung, D.-Y.; Xu, Q.; et al. 2024a. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 7246–7255.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 19376–19385.

Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.-P.; Liu, Z.; Qiao, Y.; Kot, A. C.; and Wen, B. 2024b. SinSR: diffusion-based image super-resolution in a single step. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 25796–25805.

Xu, M.; and Lu, F. 2024. Gaze from origin: Learning for generalized gaze estimation by embedding the gaze frontalization process. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6333–6341.

Xu, M.; Wang, H.; and Lu, F. 2023. Learning a generalized gaze estimator from gaze-consistent feature. In *AAAI*, volume 37, 3027–3035.

Xu, Y.; Dong, Y.; Wu, J.; Sun, Z.; Shi, Z.; Yu, J.; and Gao, S. 2018. Gaze prediction in dynamic 360 immersive videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 5333–5342.

Yin, P.; Wang, J.; Zeng, G.; Xie, D.; and Zhu, J. 2024a. Lg-gaze: Learning geometry-aware continuous prompts for language-guided gaze estimation. In *Eur. Conf. Comput. Vis.*, 1–17. Springer.

Yin, P.; Zeng, G.; Wang, J.; and Xie, D. 2024b. CLIP-gaze: towards general gaze estimation via visual-linguistic model. In *AAAI*, volume 38, 6729–6737.

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 *Eur. Conf. Comput. Vis.*, 365–381. Springer.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2015. Appearance-based gaze estimation in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 4511–4520.

Zhao, W.; Rao, Y.; Liu, Z.; Liu, B.; Zhou, J.; and Lu, J. 2023. Unleashing text-to-image diffusion models for visual perception. In *Int. Conf. Comput. Vis.*, 5729–5739.