Pose-Robust Calibration Strategy for Point-of-Gaze Estimation on Mobile Phones

Yujie Zhao[1,2]
zhaoyujie24s@ict.ac.cn
Jiabei Zeng[1,2]
jiabei.zeng@ict.ac.cn
Shiguang Shan[1,2]
sgshan@ict.ac.cn

[1] Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China

[2] University of Chinese Academy of Sciences
Beijing, China

Abstract

Although appearance-based point-of-gaze (PoG) estimation has improved, the estimators still struggle to generalize across individuals due to personal differences. Therefore, person-specific calibration is required for accurate PoG estimation. However, calibrated PoG estimators are often sensitive to head pose variations. To address this, we investigate the key factors influencing calibrated estimators and explore pose-robust calibration strategies. Specifically, we first construct a benchmark, MobilePoG, which includes facial images from 32 individuals focusing on designated points under either fixed or continuously changing head poses. Using this benchmark, we systematically analyze how the diversity of calibration points and head poses influences estimation accuracy. Our experiments show that introducing a wider range of head poses during calibration improves the estimator's ability to handle pose variation. Building on this insight, we propose a dynamic calibration strategy in which users fixate on calibration points while moving their phones. This strategy naturally introduces head pose variation during a user-friendly and efficient calibration process, ultimately producing a better calibrated PoG estimator that is less sensitive to head pose variations than those using conventional calibration strategies. Codes and datasets are available at our project page: https://mobile-pog.github.io.

1 Introduction

Gaze estimation has attracted growing interest in computer vision due to its ability to reveal human attention, intention, and cognitive processes. Accurate point-of-gaze (PoG) prediction enables applications in medical diagnosis [16, 17, 34], human-computer interaction [1, 2, 18, 29], VR/AR [26, 36], assisted driving [24], and so on. Powered by deep learning methods, appearance-based PoG estimation predicts 2D gaze coordinates on a designated screen from a single RGB image captured by a front-facing camera, without requiring additional hardware such as LED lights, infrared sensors, or wearable eye trackers.

Although appearance-based PoG estimation has advanced significantly with the development of large-scale datasets and data-driven learning methods, existing models still struggle

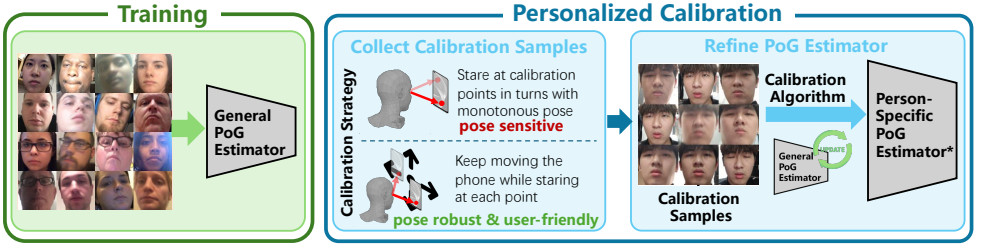

Figure 1: Pipeline of personalized calibration for PoG estimation.

to generalize across individuals due to challenging personal differences, such as variations in the anatomical structure of eyes [13, 21], that are often not discernible from RGB images alone. To customize the PoG estimator for individual users, an effective and widely used approach is to perform personalized calibration before deployment [20, 21, 22, 27, 38, 39]. As shown in Figure 1, rather than directly deploying a general PoG estimator trained on diverse data, personalized calibration typically begins by collecting a small set of user-specific samples, where the user is instructed to gaze at a few designated points (e.g., 5 or 9) while corresponding facial images are recorded. These calibration samples are then used to adapt the general model to the individual, resulting in an improved personalized estimator.

However, real-world applications show that the performance of calibrated PoG estimators is highly sensitive to head pose variations. When a user's head pose during inference differs from that during calibration, estimation accuracy often degrades significantly. This sensitivity is likely to stem from the static conventional calibration strategies, which require users to maintain a fixed head pose while gazing at predefined calibration points. Despite its practical importance, the impact of calibration strategy on gaze estimation performance remains under-explored. A key limitation is that existing mobile PoG datasets lack head pose guidance during data collection, making it difficult to simulate diverse calibration scenarios under controlled conditions. Consequently, these datasets fall short in supporting rigorous evaluation of calibration strategies and their effectiveness.

To develop a calibrated PoG estimator that remains accurate across varying head poses, we investigate key calibration factors (i.e., the diversity of calibration points and head poses) and propose a pose-robust, user-friendly calibration strategy. We introduce MobilePoG, a new phone-based PoG dataset that captures facial images of users when they are fixating on predefined points under either fixed or continuously changing head poses. MobilePoG enables the simulation of diverse calibration scenarios, serving as a benchmark for evaluating calibration strategies and analyzing influential factors systematically. Experimental analysis on MobilePoG reveals that increasing head pose diversity in calibration samples significantly enhances estimator robustness, whereas simply adding more calibration points offers limited improvement. Based on this observation, we propose a dynamic calibration strategy, where users keep moving the phone while fixating on each calibration point, instead of maintaining a static pose. This simple yet effective strategy introduces natural head pose variation and enhances the robustness of the calibrated model across different head poses, regardless of the underlying PoG estimators or calibration algorithms. The contributions of this work are summarized as follows:

- We construct MobilePoG, a mobile PoG dataset that enables the simulation of diverse calibration scenarios. It serves as a benchmark for evaluating calibration strategies and

enables systematic analysis of key influencing factors.

- We systematically analyze the impact of calibration points and head poses, and find that pose diversity in calibration samples plays a crucial role in enhancing model robustness.
- We propose a dynamic calibration strategy that naturally introduces head pose variation through a user-friendly and efficient process. Experiments on MobilePoG show that the proposed strategy effectively reduces the calibrated estimator's sensitivity to head pose.

2 Related Work

In this section, we review prior works on appearance-based gaze estimation methods, personalized gaze calibration, and point-of-gaze (PoG) datasets relevant to our approach.

Appearance-based Gaze Estimation Methods: Appearance-based methods aim to directly learn a mapping from input images to either a 3D gaze direction vector or a 2D point-of-gaze (PoG) on a screen. For 3D gaze estimation, Zhang *et al*. [40] introduced a convolutional neural network to extract gaze direction from eye images and head pose, significantly enhancing prediction accuracy. Since then, numerous studies [2, 10, 12, 31, 33] have further advanced appearance-based 3D gaze estimation. The estimated 3D gaze vector can be projected to 2D screen coordinates if the geometric relationship between the camera and screen is known. For direct estimation of PoG, Krafka *et al*. [20] and Zhang *et al*. [41] improved accuracy by incorporating spatial attention and demonstrated the advantages of using full-face images to capture head pose information. To mitigate overfitting, He *et al*. [15] simplified the architecture by using only eye images and eye corner landmarks, which also encode head pose cues. Guo *et al*. [14] introduced a knowledge distillation training scheme, while Bao *et al*. [5] and Cheng *et al*. [9] enhanced feature fusion through facial guidance and Transformer-based attention, respectively. Balim *et al*. [3] proposed an end-to-end framework that predicts gaze direction and origin separately before computing PoG. To further enhance the generalization of gaze estimators, many studies have explored unsupervised domain adaptation [6, 8, 22, 35] and domain generalization [4, 11, 37].

In spite of the improvements brought by the aforementioned studies to general gaze estimators, their generalization capability is still constrained by inter-individual differences. This paper focuses on personalized calibration for point-of-gaze estimation, which holds significant value for practical applications.

Personalized Gaze Calibration: Personalized calibration tailors the gaze estimation model to an individual user by adjusting model parameters or outputs based on user-specific data to improve accuracy. Many studies focused on the calibration algorithms for adapting gaze estimators given user-specific calibration samples [15, 20, 21, 22, 27, 38, 39]. These algorithms can be broadly categorized into three types: feature extraction-based methods, tuning-based methods, and anchor-based methods. Feature extraction-based methods utilize pre-trained models to extract features from calibration samples and then perform calibration using techniques such as support vector regression (SVR) [20], linear probe [21], or bias correction [39], without updating the model weights. In contrast, tuning-based methods [22, 27] fine-tune a subset of the model parameters under the supervision of calibration samples. Anchor-based methods [15] treat calibration samples as reference anchors, enabling the model to learn predictions by incorporating the features and labels of anchors.

However, the strategy for collecting user-specific data has received comparatively less attention. Conventional strategies typically gather calibration samples while the user fixates

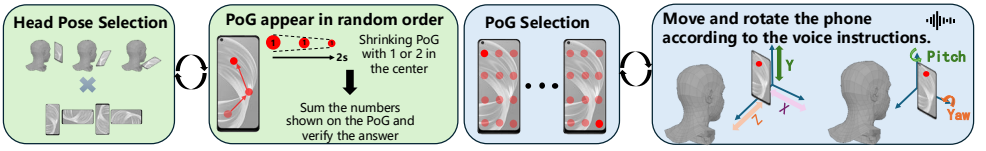

Figure 2: Dataset collection procedure of MobilePoG. (a)Static-MobilePoG: Subjects fixate on the point with fixed poses. (b)Dynamic-MobilePoG: Subjects fixate on the point while moving their phones, resulting in continuously changing poses.

on predefined calibration points with minimal variation in head pose. Some studies [15, 20, 39] have examined the influence of the number of calibration points, but often overlook head pose variation, which plays a critical role in determining the performance of calibration. In this work, we conduct a systematic analysis of the factors affecting calibration robustness and empirically demonstrate that head pose diversity during calibration is a key determinant for enhancing the generalization capability of point-of-gaze estimation.

Point-of-Gaze Dataset: Many datasets [19, 20, 25, 28, 32, 40, 41, 42] have been built for the development of gaze estimation methods. Some of them are annotated with 2D PoGs [20, 25, 28, 30, 32, 40, 41] and have various subjects, PoGs, and head poses.

Although existing datasets have significantly advanced PoG estimation research, they are limited in supporting systematic studies of personalized PoG calibration. A key limitation is the lack of control over critical calibration conditions during data collection leads to difficulty in simulating real-world calibration. To enable rigorous analysis of calibration strategies, a dataset should provide explicit guidance for both gaze target locations and head pose configurations during the data acquisition process. Such control allows for well-designed experiments where the effects of individual factors (PoG and head pose) can be independently varied and studied. Take a few examples, MPIIFaceGaze [40, 41] collected real-world data using laptops in daily-life scenarios, where participants just adopted natural head poses as if they were using a computer in their usual manner. GazeCapture [20], the largest dataset with PoG on mobile phones and tablets, simply asks participants to switch device orientation and randomly change head pose. Without precise control over head pose configurations, they are unsuitable for studying gaze calibration. TabletGaze [30] collects PoGs on mobile devices covering predefined four static poses but the limited pose diversity and high label noise make it difficult to draw reliable and generalizable conclusions for calibration research. Distinct from the previous datasets, we propose a benchmark specifically designed for personalized calibration in point-of-gaze estimation, which provides explicit guidance on head pose and point-of-gaze during data collection.

3 MobilePoG Dataset

Since existing PoG datasets were not designed to investigate calibration factors and lack the necessary conditions for such analysis, we construct a new benchmark, MobilePoG, which contains facial images of individuals when they are fixating on designated points under either fixed or continuously changing head poses.

3.1 Dataset Collection

We recruited 32 college students and recorded their facial images when they were instructed to fixate on designated points displayed on the phones. Two Android applications were developed to guide the collection procedure and form two subset of data: Static-MobilePoG with various fixed head poses, and Dynamic-MobilePoG with continuously changing poses.

Static-MobilePoG: As shown in Figure 2 (a), each participant is instructed to adopt one of three user-to-phone relative positions and to hold the phone in one of four orientations, yielding 12 head pose configurations commonly encountered during everyday smartphone use. For each pose, 55 predefined points appeared in random order. Participants are asked to fixate on each point while maintaining the same pose. Each point exists for 2 seconds and gradually shrinks in size to help guide accurate fixation. Meanwhile, a digital number (either 1 or 2) was displayed at the center of each point, and participants were required to mentally accumulate the sum throughout the sequence. The data was considered valid only if the participant reported the correct total after viewing all 55 points. The left six columns in Figure 3 show image samples with different poses in Static-MobilePoG.

Dynamic-MobilePoG: As shown in Figure 2 (b), for each of the predefined points, participants are instructed to fixate on the target point for a period of time. During this fixation period, they are guided by voice instructions to smoothly move and rotate the phone. These instructions induce head pose variations across five degrees of freedom: left-right (x), up-down (y), forward-backward (z), pitch, and yaw. Consequently, the collected images capture a comprehensive range of poses encountered during everyday mobile phone usage. The right six columns in Figure 3 show image samples with different poses in Dynamic-MobilePoG.

3.2 Dataset Characteristics

We recorded facial videos from 32 participants during both the Static-MobilePoG and Dynamic-MobilePoG procedures. For each point in Static-MobilePoG, we retained 30 frames from the middle one-second interval of the two-second duration. In Dynamic-MobilePoG, all frames for each point were retained. Facial landmarks were then detected using MediaPipe[23]. Frames with undetectable or blurry faces were excluded to ensure data quality. As a result, the MobilePoG dataset contains approximately 606,000 frames in the Static-MobilePoG subset and around 1.71 million frames in the Dynamic-MobilePoG subset. The images were captured using the front-facing camera at typical usage distances, approximately half an arm's length to an arm's length. The resolutions are either 640×360 or 640×480 pixels.

Dataset	# Subject	# PoG	# Image	# Pose	Pose Controllable
TabletGaze [30]	51	35	1,785	4 Static	✓
GazeCapture [20]	1468	Random	2,445,504	Random	✗
Static-MobilePoG	32	55	606,301	12 Static	✓
Dynamic-MobilePoG	32	12	1,707,910	Continuous	✓

Table 1: Comparison of existing mobile PoG datasets.

Table 1 presents a statistical comparison of existing mobile PoG datasets. TabletGaze [30] recorded images under only four poses for each subject. GazeCapture[20] includes a large number of participants, contributing substantial diversity to the dataset. However, each subject in GazeCapture looked at random points while maintaining self-selected, comfortable

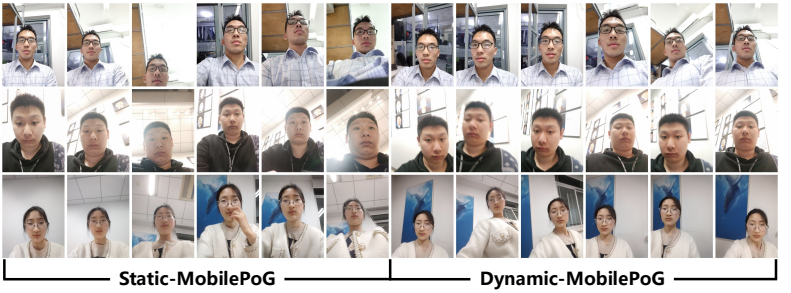

Figure 3: Examples from MobilePoG. Each subject exhibits diverse head pose variations.

poses. This randomness makes it difficult to isolate and study the effects of individual calibration factors, as multiple variables change simultaneously. In contrast, MobilePoG provides controlled yet diverse pose variations, particularly in the Dynamic-MobilePoG subset, where phone movement results in a nearly continuous range of head poses. The dataset includes images of participants fixating on the same point from different poses, as well as images captured under the same pose while fixating on different points. This design facilitates systematic analysis of how individual factors affect calibration.

4 Factor Analysis of Personalized PoG Calibration

4.1 Personalized PoG Calibration Pipeline

Figure 1 illustrates the pipeline of personalized point-of-gaze (PoG) calibration. In the training stage, a general PoG estimator is first trained on data collected from multiple subjects. Rather than using this general model directly for inference, we adapt it to individual users through personalization calibration that typically involves two steps. First, a small set of calibration samples is collected based on a chosen calibration strategy. For example, the conventional static strategy asks the user to fixate on a few designated points (e.g., 5 or 9) while corresponding facial images are captured. Second, these samples are used to refine the general estimator via a calibration algorithm, resulting in a person-specific PoG model that is then deployed for that individual.

4.2 Experimental Setup

To understand how different factors influence the performance of personalized point-of-gaze (PoG) calibration, we conducted a series of controlled experiments using the Static-MobilePoG. We focused on two key variables: the number of calibration points and the diversity of head poses during calibration.

General PoG Estimator: To ensure the generality of our analysis, we adopted two prevalent methods iTracker [20] and AFFNet [6] as the general PoG estimator. The general PoG estimator was initially pre-trained on the GazeCapture [20] dataset to endow it with stronger generalization ability and a solid foundational representation, and was subsequently trained on the training set of Static-MobilePoG. In Static-MobilePoG, we randomly selected 5 subjects as a calibration set and the remaining 27 subjects as the training set.

Calibration Samples: We investigated varied configurations of calibration points and head poses in the conventional static strategy. For each subject in the calibration set, the calibration samples are selected as the ones staring at N predefined points ($N = 1, 5, 9, 13$) with P poses ($P = 1, 2, 3, 4$). The position of points is illustrated in the supplemental materials.

Calibration Algorithm: Different calibration algorithms are adopted for calibration-algorithm-agnostic findings. The algorithms include: (1)*SVR* replaces the model's MLP head to regress the PoG. (2)*Linear Probe* trains a linear layer on top of the model. (3)*Finetune MLP* fine-tunes the MLP regression head. (4)*Full Finetune* fine-tunes the whole estimator. For the detailed implementation of algorithms, please refer to the supplementary materials.

4.3 Monotonous Pose Calibration Limits Generalization

To investigate the bottlenecks in real-world calibration scenarios, we conducted experiments using a monotonous head pose with multiple calibration points, simulating the conventional static calibration strategy. Table 2 reports the performance of calibrated estimators based on samples collected under a single head pose.

As shown in Table 2, increasing the number of calibrated points under a fixed pose leads to a substantial reduction in the same pose estimation error, with performance approaching saturation at 9 calibration points. For instance, when calibrated with 9 points using the Full Finetune strategy, AFFNet achieves a 44.1% error reduction, reaching 1.42 cm.

However, despite the improved accuracy under the same head pose, the calibrated estimators generalize poorly to test samples with different head poses, resulting in high average errors. This issue is particularly pronounced for SVR and Linear Probe, both of which train a regression head from scratch. In these cases, single-head-pose calibration samples lead to severe overfitting. Consequently, the minimum average errors for AFFNet and iTracker calibrated using SVR and Linear Probe reach 7.93 cm / 3.19 cm and 7.92 cm / 4.00 cm, respectively, which are even worse than those without any calibration. While Finetune MLP and Full Finetune exhibit slightly better robustness, their performance gains are still limited due to the narrow distribution of calibration poses. For example, calibrating AFFNet with 13 points using Full Finetune only reduces the error from 2.54 cm to 2.30 cm.

These results clearly indicate that calibration based on monotonous head pose samples fails to generalize to varying head poses. Thus, incorporating pose diversity in the calibration data is essential for achieving robust and reliable PoG estimation in real-world scenarios.

4.4 Diverse Pose Calibration Boosts Generalization

To further investigate the key factors influencing the robustness of calibrated estimators, we compared performance variations under two conditions: increasing the number of calibration points (PoGs) and increasing the diversity of calibration head poses. Figure 4 illustrates how the average error of the calibrated estimator changes with increasing numbers of calibration points and head poses in the calibration samples.

As shown in Figure 4 (a), when the head pose in calibration samples is fixed, increasing the number of PoGs yields only marginal performance gains. In some cases, adding too many PoGs even leads to overfitting. For instance, iTracker calibrated via Finetune MLP with thirteen PoGs performs worse than the uncalibrated baseline.

In contrast, Figure 4 (b) and (c) reveal that increasing head pose diversity in calibration samples significantly improves the robustness of calibrated estimators. Notably, even with

Model	Calibration Algorithm	1 calibration point			5 calibration points			9 calibration points			13 calibration points		
		Same	Diff.	Avg.	Same	Diff.	Avg.	Same	Diff.	Avg.	Same	Diff.	Avg.
iTracker	w/o calibration	2.82											
	SVR	4.17	11.17	10.60	2.30	8.93	8.42	2.21	8.47	8.02	2.27	8.31	7.92
	Linear Probe	3.82	7.08	6.81	2.40	4.13	4.00	2.17	5.30	5.07	2.29	4.58	4.43
	Finetune MLP	2.35	2.74	2.71	2.10	2.88	2.82	2.00	2.83	2.77	2.02	3.05	2.98
	Full Finetune	2.10	2.71	2.66	2.07	2.72	2.67	1.97	2.63	2.58	1.40	2.65	2.57
AFFNet	w/o calibration	2.54											
	SVR	4.18	11.17	10.60	2.18	8.40	7.93	1.69	8.91	8.40	2.03	8.60	8.17
	Linear Probe	4.18	6.84	6.62	2.02	3.29	3.19	1.97	3.69	3.57	2.05	3.55	3.45
	Finetune MLP	2.06	2.53	2.49	2.04	2.41	2.38	1.52	2.45	2.38	1.48	2.57	2.50
	Full Finetune	1.94	2.37	2.33	1.99	2.39	2.36	1.42	2.45	2.38	1.40	2.37	2.30

Table 2: Results of a single calibration head pose and different numbers of calibration points on Static-MobilePoG. "Same" and "Diff." represent the test errors of the calibrated model on test samples with the same and different head poses as the calibration head pose, respectively. The metric is the Euclidean distance (cm).

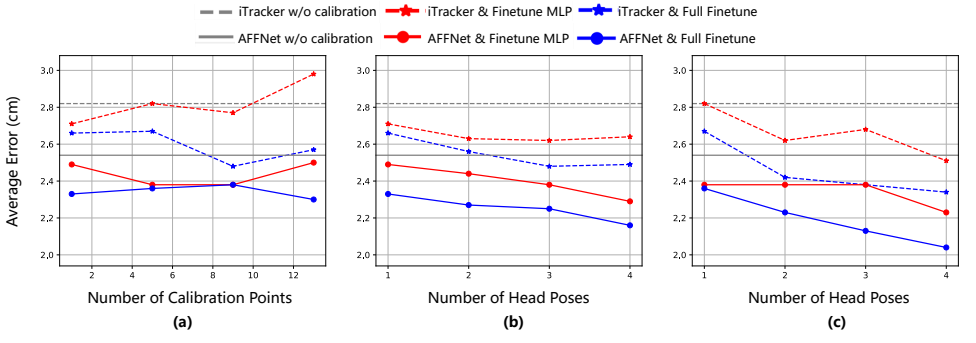

Figure 4: Comparison of increasing calibration points and head poses in calibration samples. (a) shows increasing points with a single head pose. (b) shows increasing head poses with one point. (c) shows increasing head poses with five points.

only a single point, calibration under multiple (3 or 4) head poses consistently outperforms the case with multiple (9 or 13) points under a single head pose, as shown in Figure 4 (b).

These findings strongly suggest that head pose diversity is a critical factor in personalized calibration. Incorporating varied head poses into calibration samples substantially enhances the generalization ability of calibrated estimators and should be prioritized when designing practical calibration strategies.

5 Pose-Robust Calibration Strategy

5.1 Dynamic Calibration Strategy Design

As shown in Section 4, head pose diversity in calibration samples greatly impacts the performance of person-specific estimators. However, increasing this diversity complicates data collection and demands more user cooperation. Thus, finding a user-friendly calibration strategy is crucial for practical deployment on mobile devices.

We propose a new dynamic calibration strategy: the user first selects one calibration PoG and for each fixed calibration PoG, the user just needs to rotate and translate the mobile

Model	Calibration Algorithm	1 calib. point static	1 calib. point dynamic	2 calib. points static	2 calib. points dynamic	4 calib. points static	4 calib. points dynamic	6 calib. points static	6 calib. points dynamic
iTracker	w/o calibration	2.09							
	SVR	4.18	**4.18**	3.35	**3.11**	2.26	**1.89**	2.17	**1.76**
	Linear Probe	2.86	**2.83**	2.62	**2.58**	2.03	**1.89**	1.97	**1.77**
	Finetune MLP	2.55	**1.81**	2.24	**1.89**	2.02	**1.63**	1.95	**1.53**
	Full Finetune	1.86	**1.85**	1.87	**1.74**	1.95	**1.48**	1.83	**1.36**
AFFNet	w/o calibration	1.75							
	SVR	4.18	**4.18**	3.19	**2.9**	1.60	**1.39**	1.60	**1.35**
	Linear Probe	2.77	**2.67**	2.35	2.41	1.69	**1.66**	1.58	**1.51**
	Finetune MLP	1.79	**1.57**	1.83	**1.53**	1.52	**1.33**	1.40	**1.27**
	Full Finetune	1.49	**1.43**	1.56	**1.42**	1.83	**1.27**	1.52	**1.10**

Table 3: Comparison of the static and the dynamic strategy on Dynamic-MobilePoG.The metric is the Euclidean distance (cm).

phone within a range while keeping gaze at the PoG as shown in Figure 2 (b), which not only allows for the collection of calibration samples with sufficiently diverse head poses, but is also user-friendly and easy to operate on mobile devices. As mentioned in Section 3, the collection procedure of Dynamic-MobilePoG could efficiently simulate the proposed dynamic calibration strategy as the dataset contains continuously varying head poses. We will conduct experiments on Dynamic-MobilePoG to demonstrate that our proposed dynamic strategy outperforms the conventional static strategy.

5.2 Experimental Evaluation

5.2.1 Experiment Settings

We evaluated the proposed strategy on Dynamic-MobilePoG. Five subjects were randomly selected for the calibration set, while the remaining 27 subjects were used for training. Similar to settings in Section 4, we trained the general PoG estimator by first pretraining either iTracker [20] or AFFNet [5] on the GazeCapture dataset, and subsequently trained them on the Dynamic-MobilePoG training set.

To compare the conventional static strategy and our proposed dynamic strategy, we conducted experiments on Dynamic-MobilePoG by selecting calibration samples with different settings to simulate the two strategies. For the **static strategy**, we selected consecutive frames from the sequence as calibration data and used the remaining data for testing. For the **dynamic strategy**, we sampled from the frame sequence at a fixed step to ensure that selected frames have diverse head poses and evaluated the calibrated model on the rest of the data for each person. We conducted experiments with 1, 2, 4, and 6 calibration points for both strategies, and sampled 30 frames for each calibration point. The position of points is illustrated in the supplemental materials.

Finally, to ensure the generalizability of the experimental conclusions, we also employed a range of calibration algorithms (i.e., SVR, Linear Probe, Finetune MLP, and Full Finetune) to construct personalized PoG estimators.

5.2.2 Experimental Results

Calibration Performance of the Static and the Dynamic Strategy. Table 3 reports the errors of the calibrated estimators using various configurations (i.e., two models, four calibration algorithms, and four different numbers of calibration points) under the static and the

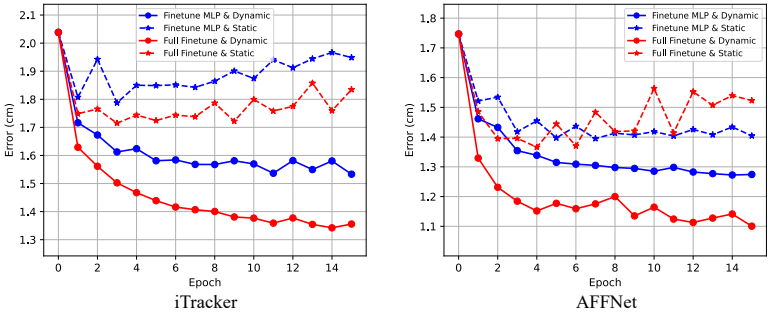

Figure 5: Calibration stability of the static and the dynamic strategy. All the models are calibrated with six calibration points.

dynamic strategy. As can be seen, the dynamic strategy outperforms the static strategy in almost all configurations except for the situation of AFFNet calibrated by Linear Probe with two calibration points. We can also see that the dynamic strategy can even surpass the performance of the static strategy, which requires a large number of calibration PoGs, using only a small number of calibration PoGs. For example, the test error of iTracker calibrated by Finetune MLP with the dynamic strategy and only one PoG achieves 1.81 cm, while the test error of that with the static strategy and six PoGs is 1.95 cm. The above experimental results demonstrate that the dynamic strategy, which is capable of collecting calibration samples with rich pose diversity, significantly enhances the robustness of the model after calibration.

Calibration Stability of the Static and the Dynamic Strategy. To further demonstrate the efficiency of our proposed dynamic strategy, we evaluated the model performance stability of different estimators and calibration algorithms under two different calibration strategies. Figure 5 shows the performance of iTracker and AFFNet during calibration using Finetune MLP and Full Finetune across different epochs. We can observe that models calibrated using the dynamic strategy gradually converge to an optimal performance and remain stable, whereas models calibrated using the static strategy exhibit severe performance fluctuations. For models with unstable performance, achieving optimal results may require careful tuning of hyperparameters such as the learning rate or manually selecting a checkpoint from a specific epoch, which is not practical in real-world calibration applications. With our proposed dynamic strategy, the person-specific estimator exhibits not only excellent but also stable performance, enabling efficient personalized PoG calibration.

6 Conclusion

In order to bridge the gap in research on personalized PoG calibration strategies and mitigate the sensitivity of the conventional static strategy to head pose variations, we first construct a new dataset, MobilePoG, which can faithfully simulate the calibration process and thus serves as a benchmark for personalized calibration. Our experiments on Static-MobilePoG demonstrate that the main bottleneck in personalized calibration lies in the diversity of head poses within the calibration samples. Finally, we propose a user-friendly dynamic calibration strategy and validate its effectiveness on Dynamic-MobilePoG, showing that it significantly improves the person-specific estimator's performance and robustness to varying head poses. Future work may focus on incorporating mechanisms for subsequent personalized calibration into the design of the general PoG estimator, particularly with respect to head pose variations, which could lead to more efficient and robust calibration.

ACKNOWLEDGEMENT: This work is supported by National Natural Science Foundation of China (No. 62176248, U2336213).

References

[1] J. San Agustin, J. Hansen, and M. Tall. Gaze-based interaction with public displays using off-the-shelf components. In *Proceedings of ACM International Conference Adjunct Papers on Ubiquitous Computing*. Association for Computing Machinery, 2010.

[2] D. Akkil, J. Kangas, J. Rantala, P. Isokoski, O. Spakov, and R. Raisamo. Glance awareness and gaze interaction in smartwatches. In *Proceedings of Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 2015.

[3] H. Balim, S. Park, X. Wang, X. Zhang, and O. Hilliges. Efe: End-to-end frame-to-gaze estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

[4] Y. Bao and F. Lu. 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 (CVPR)*, 2024.

[5] Y. Bao, Y. Cheng, Y. Liu, and F. Lu. Adaptive feature fusion network for gaze tracking in mobile tablets. In *International Conference on Pattern Recognition (ICPR), 2020*.

[6] Y. Bao, Y. Liu, H. Wang, and F. Lu. Generalizing gaze estimation with rotation consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] X. Cai, B. Chen, J. Zeng, J. Zhang, Y. Sun, X. Wang, Z. Ji, X. Liu, X. Chen, and S. Shan. Gaze estimation with an ensemble of four architectures, 2021.

[8] X. Cai, J. Zeng, S. Shan, and X. Chen. Source-free adaptive gaze estimation by uncertainty reduction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[9] Y Cheng and F. Lu. Gaze estimation using transformer. In *International Conference on Pattern Recognition (ICPR)*, 2022.

[10] Y. Cheng, X. Zhang, F. Lu, and Y. Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020.

[11] Y. Cheng, Y. Bao, and F. Lu. Puregaze: Purifying gaze feature for generalizable gaze estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[12] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *European Conference on Computer Vision (ECCV)*. Springer-Verlag, 2018.

[13] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6):1124–1133, 2006.

[14] T. Guo, Y. Liu, H. Zhang, X. Liu, Y. Kwak, B. Yoo, J. Han, and C. Choi. A generalized and robust method towards practical gaze estimation on smart phone. In *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019.

[15] J. He, K. Pham, N. Valliappan, P. Xu, C. Roberts, D. Lagun, and V. Navalpakkam. On-device few-shot personalization for real-time gaze estimation. In *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019.

[16] P. Holzman, L. Proctor, D. Levy, N. Yasillo, H. Meltzer, and S. Hurt. Eye-Tracking Dysfunctions in Schizophrenic Patients and Their Relatives. *Archives of General Psychiatry*, 31(2):143–151, 1974.

[17] M. Jiang and Q. Zhao. Learning visual attention to identify people with autism spectrum disorder. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[18] M. Kassner, W. Patera, and A. Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, 2014.

[19] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[20] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] G. Liu, Y. Yu, K. Mora, and J. Odobez. A differential approach for gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1092–1099, 2021.

[22] H. Liu, J. Qiand Z. Li, M. Hassanpour, Y. Wang, K. Plataniotis, and Y. Yu. Test-time personalization with meta prompt for gaze estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:3621–3629, 2024.

[23] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann. Mediapipe: A framework for building perception pipelines, 2019.

[24] A. Mavely, J. Judith, P. Sahal, and S. Kuruvilla. Eye gaze tracking based driver monitoring system. In *IEEE International Conference on Circuits and Systems (ICCS)*, 2017.

[25] K. Funes Mora, F. Monay, and J. Odobez. Eyediap: a database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Eye Tracking Research and Applications (ETRA)*, 2014.

[26] N. Padmanaban, R. Konrad, E. Cooper, and G. Wetzstein. Optimizing vr for all users through adaptive focus displays. In *ACM SIGGRAPH Talks*, 2017.

[27] S. Park, S. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz. Few-shot adaptive gaze estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[28] S. Park, E. Aksan, X. Zhang, and O. Hilliges. Towards end-to-end video-based eye-tracking. In *European Conference on Computer Vision (ECCV)*. Springer-Verlag, 2020.

[29] P. Prajod, N. Lavit, M. Malosio, and E. André. Gaze-based attention recognition for human-robot collaboration. In *Proceedings of International Conference on PErvasive Technologies Related to Assistive Environments*, 2023.

[30] H. Qiong, A. Veeraraghavan, and A. Sabharwal. Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Mach. Vision Appl.*, 28(5–6):445–461, 2017.

[31] R. Ranjan, S. De Mello, and J. Kautz. Light-weight head pose invariant gaze tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[32] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[33] Y. Sun, J. Zeng, S. Shan, and X. Chen. Cross-encoder for unsupervised gaze representation learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[34] S. Wang, M. Jiang, X. Duchesne, E. Laugeson, D. Kennedy, R. Adolphs, and Q. Zhao. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, 88(3):604–616, 2015.

[35] Y. Wang, T. Jiang, J. Li, B. Ni, W. Dai, C. Li, H. Xiong, and T. Li. Contrastive regression for domain adaptation on gaze estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[36] Y. Wei, R. Shi, D. Yu, Y. Wang, Y. Li, L. Yu, and H. Liang. Predicting gaze-based target selection in augmented reality headsets based on eye and head endpoint distributions. In *Proceedings of CHI Conference on Human Factors in Computing Systems*, 2023.

[37] M. Xu, H. Wang, and F. Lu. Learning a generalized gaze estimator from gaze-consistent feature. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[38] Y. Yu, G. Liu, and J. Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[39] Chen Z and B. Shi. Towards high performance low complexity calibration in appearance based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1174–1188, 2023.

[40] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[41] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[42] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision (ECCV)*, 2020.