

用于手机上 注视点估计的姿态鲁棒校准策略

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

尽管基于外观的注视点 (PoG) 估计法有所改进,但由于个人差异,估计器仍然难以在不同个体间泛化。因此,为了获得准确的 PoG 估计,个人特定的校准是必需的。然而,经过校准的 PoG 估计器通常对头部姿态的变化非常敏感。为了解决这个问题,我们研究了影响校准估计器的关键因素,并探索姿态鲁棒的校准策略。具体而言,我们首先构建了一个基准数据集, MobilePoG, 其中包括 32 个人在固定或连续变化的头部姿态下对准指定点的面部图像。使用这个基准数据集,我们系统地分析了校准点和头部姿态的多样性如何影响估计精度。我们的实验表明,在校准过程中引入更广泛的头部姿态范围可以提高估计器处理姿态变化的能力。在这一见解的基础上,我们提出了一种动态校准策略,其中用户在移动手机的同时注视校准点。这个策略在用户友好且高效的校准过程中自然引入了头部姿态变化,最终产生了一个比使用传统校准策略更不易受到头部姿态变化影响的校准 PoG 估计器。代码和数据集可在我们的项目页面获取: <https://mobile-pog.github.io>。

1 引言

注视估计因其能够揭示人类注意力、意图和认知过程而在计算机视觉中引起了越来越多的关注。准确的看的点 (PoG) 预测能够应用于医学诊断 [16, 17, 34]、人机交互 [1, 2, 18, 29]、VR/AR [26, 36]、辅助驾驶 [24] 等领域。在深度学习方法的推动下,基于外观的 PoG 估计可以从由前置摄像头捕捉的单个 RGB 图像中预测指定屏幕上的二维注视坐标,而无需额外的硬件,如 LED 灯、红外传感器或可穿戴眼动追踪器。尽管随着大规模数据集和数据驱动学习方法的发展,基于外观的 PoG 估计取得了显著进展,但由于个体之间的不同挑战,例如眼睛的解剖结构变化 [13, 21] 等现有模型在跨个体泛化时仍面临困难,这些差异往往仅靠 RGB 图像难以分辨。为了为个别用户定制 PoG 估计器,一种有效且广泛使用的方法是在部署前执行个性化校准 [20, 21, 22, 27, 38, 39]。如图 1 所示,与其直接部署在多样化数据上训练的通用 PoG 估计器,个性化校准通常从收集一小组特定于用户的样本开始,其间要求用户注视几个指定点(例如,5 或 9 个),同时记录相应的面部图像。然后使用这些校准样本使通用模型适应个体,从而得到一个经过改进的个性化估计器。

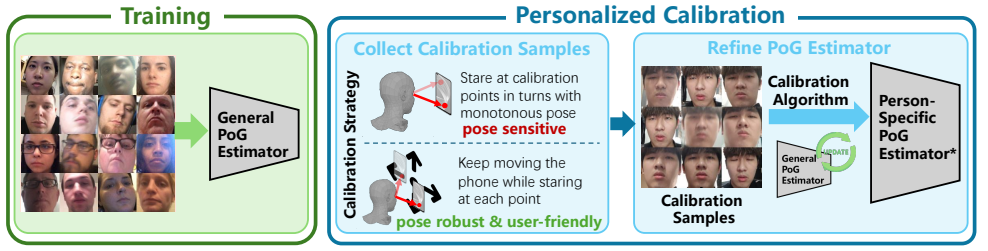


Figure 1: 个性化校准用于 PoG 估计的流程

然而，现实世界的应用表明，校准后的注视点估计器的性能对头部姿态的变化高度敏感。当用户在推理时的头部姿态与校准时不同，估计精度通常会显著下降。这种敏感性可能源于静态的传统校准策略，该策略要求用户在注视预定义的校准点时保持固定的头部姿态。尽管校准策略在实际中非常重要，但其对注视估计性能的影响仍然未被深入研究。一个主要的限制是，现有的移动注视点数据集在数据采集时缺乏头部姿态指导，这使得在受控条件下模拟多样化的校准场景变得困难。因此，这些数据集在支持校准策略的严格评估及其有效性方面表现不足。

为了开发一个在不同头部姿态下仍然准确的校准 PoG 估计器，我们研究了关键的校准因素（即，校准点和头部姿态的多样性）并提出了一种对姿态具有鲁棒性的、用户友好的校准策略。我们引入了 MobilePoG，这是一个新的基于手机的 PoG 数据集，该数据集捕捉了用户在固定或不断变化的头部姿态下注视预定义点时的面部图像。MobilePoG 能够模拟多种校准场景，为评估校准策略和系统分析影响因素提供了一个基准。在 MobilePoG 上的实验分析表明，增加校准样本中头部姿态的多样性显著增强了估计器的鲁棒性，而仅仅增加更多的校准点只能带来有限的改进。基于这一观察，我们提出了一种动态校准策略，在这种策略中，用户在注视每个校准点时持续移动手机，而不是保持固定姿态。这个简单而有效的策略引入了自然的头部姿势变化，并增强了校准模型在不同头部姿态下的鲁棒性，无论基础的 PoG 估计器或校准算法为何。这项工作的贡献总结如下：

- 我们构建了 MobilePoG，这是一个移动 PoG 数据集，可以用于模拟各种校准场景。它作为评估校准策略的基准，并能够系统地分析关键影响因素。
- 我们系统地分析了校准点和头部姿态的影响，发现校准样本中姿态的多样性在增强模型鲁棒性方面起着至关重要的作用。
- 我们提出了一种动态校准策略，通过一种用户友好且高效的过程自然引入头部姿态变化。在 MobilePoG 上的实验表明，该策略有效地降低了经过校准的估计器对头部姿态的敏感性。

2 相关工作

在本节中，我们回顾与我们的方法相关的基于外观的注视估计方法、个性化注视校准以及视线点 (PoG) 数据集的先前工作。

基于外观的视线估计方法：基于外观的方法旨在直接学习从输入图像到 3D 视线方向向量或屏幕上的 2D 注视点 (PoG) 的映射。对于 3D 视线估计，张 *et al.* [40] 引入了一个卷积神经网络来从眼睛图像和头部姿态中提取视线方向，显著提升了预测精度。从那时起，众多研究 [7, 10, 12, 31, 33] 进一步推进了基于外观的 3D 视线估

计。如果已知摄像头和屏幕之间的几何关系, 则估计的 3D 视线向量可以投射到 2D 屏幕坐标。对于直接估计 PoG, Krafka *et al.* [20] 和张 *et al.* [41] 通过结合空间注意力提高了准确性, 并展示了使用全脸图像来捕捉头部姿态信息的优势。为减轻过拟合, 何 *et al.* [15] 通过仅使用眼睛图像和眼角标志简化了结构, 这些标志也编码了头部姿态线索。郭 *et al.* [14] 引入了一种知识蒸馏训练方案, 而包 *et al.* [5] 和程 *et al.* [9] 则分别通过面部指导和基于 Transformer 的注意力增强特征融合。巴利姆 *et al.* [3] 提出了一个端到端框架, 在计算 PoG 之前分别预测视线方向和起点。为了进一步增强视线估计器的泛化能力, 许多研究探索了无监督领域适应 [6, 8, 22, 35] 和领域泛化 [4, 11, 37]。

尽管上述研究提升了通用凝视估计器的性能, 但其泛化能力仍受到个体间差异的限制。本文着重于凝视点估计的个性化校准, 这对于实际应用具有重要意义。

个性化注视校准: 个性化校准通过基于用户特定数据调整模型参数或输出来定制注视估计模型, 以提高准确性。许多研究集中于给定用户特定校准样本的情况下适应注视估计器的校准算法 [15, 20, 21, 22, 27, 38, 39]。这些算法大致可以分为三种类型: 特征提取方法、调参方法和锚点方法。特征提取方法利用预训练模型从校准样本中提取特征, 然后使用支持向量回归 (SVR) [20]、线性探针 [21] 或偏差校正 [39] 等技术进行校准, 而无需更新模型权重。相比之下, 调参方法 [22, 27] 微调模型参数的子集, 以校准样本为监督。锚点方法 [15] 将校准样本视为参考锚点, 使模型通过结合锚点的特征和标签来学习预测。然而, 收集用户特定数据的策略得到了较少关注。传统策略通常在用户注视预定义校准点时收集校准样本, 并且头部姿势变化较小。一些研究 [15, 20, 39] 探讨了校准点数量的影响, 但往往忽略了头部姿势变化, 这在决定校准性能方面起着关键作用。在这项工作中, 我们系统地分析了影响校准鲁棒性的因素, 并通过实验证明在校准过程中头部姿势的多样性是增强注视点估计泛化能力的关键决定因素。**凝视点数据集:** 许多数据集 [19, 20, 25, 28, 32, 40, 41, 42] 已经为凝视估计方法的发展而建立。其中一些数据集附有二维凝视点 [20, 25, 28, 30, 32, 40, 41], 具有各种不同的主体、凝视点和头部姿势。尽管现有的数据集在凝视点估计研究上取得了显著进展, 但在支持个性化凝视点校准的系统研究方面仍然存在局限性。一个关键的限制是在数据收集过程中缺乏对关键校准条件的控制, 导致在模拟现实世界校准时出现困难。为了能够彻底分析校准策略, 数据集应在数据获取过程中对观察目标的位置和头部姿态配置提供明确指导。这种控制允许设计精良的实验, 其中个别因素 (凝视点和头部姿态) 的影响可以独立变化和研究。举几个例子, MPIIFaceGaze [40, 41] 使用笔记本电脑在日常生活场景中收集了现实世界的的数据, 参与者只是采用自然的头部姿态, 好像他们在以通常的方式使用电脑。GazeCapture [20], 是在手机和平板电脑上拥有最多凝视点的数据集, 只是要求参与者切换设备方向并随机改变头部姿态。由于没有对头部姿态配置进行精确控制, 它们不适合研究凝视校准。TabletGaze [30] 在移动设备上收集凝视点, 涵盖了预定义的四种静态姿势, 但有限的姿势多样性和高标签噪声使得难以为校准研究得出可靠和普遍适用的结论。有别于之前的数据集, 我们提出了一个专门为凝视点估计中的个性化校准设计的基准, 提供了在数据收集过程中对头部姿态和凝视点的明确指导。

3 MobilePoG 数据集

由于现有的 PoG 数据集并未设计用于研究校准因素, 且缺乏进行此类分析的必要条件, 我们构建了一个新的基准 MobilePoG, 其中包含个体在固定或连续变化的头部姿势下凝视指定点时的面部图像。

我们招募了 32 名大学生, 并在他们被指示固定凝视在手机屏幕上的指定点时记录他们的面部图像。开发了两个安卓应用程序来指导数据收集过程, 并形成两个数

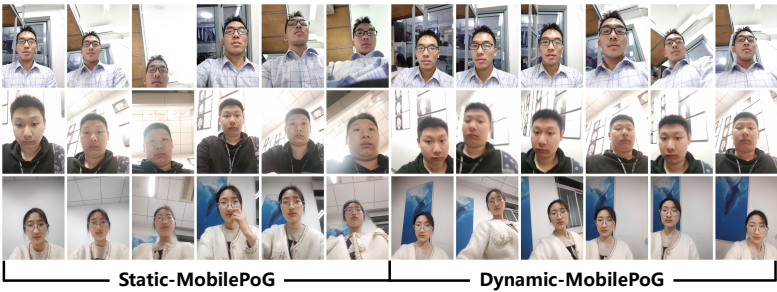


Figure 2: 来自 MobilePoG 的例子。每个受试者表现出多样的头部姿态变化。

据子集：Static-MobilePoG 子集包含各种固定的头部姿势，Dynamic-MobilePoG 子集则包含连续变化的姿势。

静态-移动 PoG：如图 ?? (a) 所示，要求每位参与者采用三种用户与手机的相对位置之一，并以四种方向之一握持手机，从而产生在日常智能手机使用中常见的十二种头部姿态配置。对于每种姿势，55 个预定义的点随机出现。参与者被要求在保持相同姿势的同时凝视每个点。每个点持续存在 2 秒，并逐渐缩小以帮助引导准确的凝视。在此期间，每个点的中心显示一个数字（1 或 2），参与者需要在序列中进行心算累加。只有当参与者在查看完所有 55 个点后报告出正确的总数，数据才被视为有效。图 2 的左六列显示了静态-移动 PoG 中不同姿势的图像样本。

Dynamic-MobilePoG：如图 ?? (b) 所示，对于每个预定义的点，参与者被指示在目标点上凝视一段时间。在这个凝视期间，他们通过语音指令指导下，平稳地移动和旋转手机。这些指令引导了在五个自由度上的头部姿势变化：左右 (x)、上下 (y)、前后 (z)、俯仰和偏航。因此，所收集的图像捕捉到了日常移动手机使用过程中遇到的全面姿势范围。图 2 右侧的六列显示了在 Dynamic-MobilePoG 中不同姿势的图像样本。

3.1 数据集特征

我们在静态-MobilePoG 和动态-MobilePoG 程序中记录了 32 名参与者的面部视频。对于静态-MobilePoG 中的每个点，我们从两秒持续时间的中间一秒间隔中保留了 30 帧。在动态-MobilePoG 中，每个点的所有帧都被保留。随后使用 MediaPipe [23] 检测面部特征点。面部无法检测到或模糊的帧被排除，以确保数据质量。因此，MobilePoG 数据集在静态-MobilePoG 子集中包含约 606,000 帧，在动态-MobilePoG 子集中包含约 171 万帧。

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: 现有移动 PoG 数据集的比较。

表 1 对现有移动 PoG 数据集进行了统计比较。TabletGaze [30] 仅在四个姿势下记录每个被试的图像。GazeCapture [20] 包含大量参与者，为数据集贡献了极大

的多样性。然而，在 GazeCapture 中，每个被试在保持自己选择的舒适姿势时观看随机点。这种随机性使得隔离和研究个别校准因素的影响变得困难，因为多个变量同时发生变化。相比之下，MobilePoG 提供了控制但又多样的姿势变化，尤其是在 Dynamic-MobilePoG 子集中，手机移动导致几乎连续的头部姿势范围。数据集包括参与者从不同姿势注视同一点的图像，以及在相同姿势下注视不同点的图像。这种设计便于系统分析个别因素如何影响校准。

4 个性化 PoG 校准的因子分析

4.1 个性化凝聚点校准流程

图 1 展示了个性化注视点 (PoG) 校准的流程。在训练阶段，首先在多个受试者的数据上训练一个通用的 PoG 估计器。为了不直接使用这个通用模型进行推理，我们通过个性化校准将其适应于个体用户，这通常涉及两个步骤。首先，根据选定的校准策略收集一小组校准样本。例如，传统的静态策略要求用户聚焦在几个指定的点上 (如 5 或 9)，同时拍摄相应的面部图像。其次，这些样本被用于通过校准算法细化通用估计器，从而得到一个个性化的 PoG 模型，该模型随后被部署给该个人使用。

4.2 实验装置

为了理解不同因素如何影响个性化注视点 (PoG) 校准的性能，我们使用静态移动 PoG 进行了一系列的控制实验。我们关注两个关键变量：校准点的数量和校准期间头部姿态的多样性。通用 PoG 估计器：为了确保我们分析的普遍性，我们采用了两种流行的方法 iTracker [20] 和 AFFNet [5] 作为通用的 PoG 估计器。通用 PoG 估计器最初在 GazeCapture [20] 数据集上进行了预训练，使其具有更强的泛化能力和稳固的基础表示，随后在静态移动 PoG 的训练集上进行了训练。在静态移动 PoG 中，我们随机选择 5 位受试者作为校准集，其余 27 位受试者作为训练集。

校准样本：我们研究了传统静态策略中校准点和头部姿态的各种配置。对于校准集中的每个受试者，校准样本被选择为以 N 个预定义点 ($N = 1, 5, 9, 13$) 和 P 个姿态 ($P = 1, 2, 3, 4$) 注视的样本。点的位置在补充材料中进行了说明。

校准算法：针对与校准算法无关的发现，采用了不同的校准算法。算法包括：(1) 使用 SVR 替换模型的 MLP 头部来回归 PoG。(2) 线性探针在模型顶部训练一个线性层。(3) Finetune MLP 微调 MLP 回归头。(4) 完全微调微调整个估计器。算法的详细实现请参考补充材料。

4.3 单调姿势校准限制泛化

为了研究真实世界校准场景中的瓶颈问题，我们进行了实验，使用单一头部姿势与多个校准点，模拟传统的静态校准策略。表 4 报告了基于单一头部姿势下收集的样本的校准估计器的性能。

如表 4 所示，在固定姿态下增加校准点的数量会大幅减少相同姿态的估计误差，且在 9 个校准点时性能趋于饱和。例如，使用全微调策略在 9 个点进行校准时，AFFNet 实现了 44.1 % 的误差减少，达到 1.42 厘米。

然而，尽管在相同的头部姿态下精度有所提高，校准的估计器在推广到具有不同头部姿态的测试样本时表现不佳，导致较高的平均误差。这个问题在 SVR 和线性探针中尤为明显，这两者都是从头开始训练回归头。在这些情况下，单一头部姿态的校准样本导致了严重的过拟合。因此，使用 SVR 和线性探针校准的 AFFNet 和 iTracker 的最小平均误差分别达到了 7.93 厘米/3.19 厘米和 7.92 厘米/4.00 厘米，甚

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	7.91
	Linear Probe	3.82	7.08	6.81	2.40	4.13	4.00	2.17	5.30	5.07	2.29	4.58	4.41
	Finetune MLP	2.35	2.74	2.71	2.10	2.88	2.82	2.00	2.83	2.77	2.02	3.05	2.99
	Full Finetune	2.10	2.71	2.66	2.07	2.72	2.67	1.97	2.63	2.58	1.40	2.65	2.59
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	3.43
	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	2.34

Table 2: 单一校准头部姿态和不同数量校准点在 Static-MobilePoG 上的结果。“Same” 和 “Diff.” 分别表示校准模型在测试样本上的测试误差，这些样本的头部姿态与校准头部姿态相同或不同。度量标准是欧氏距离（厘米）。

至比没有任何校准时还要糟糕。虽然微调 MLP 和完全微调展示了稍好的鲁棒性，但由于校准姿态分布狭窄，其性能提升仍然有限。例如，使用完全微调对 13 个点进行 AFFNet 校准仅将误差从 2.54 厘米减少到 2.30 厘米。

这些结果清楚地表明，基于单调的头部姿态样本的校准无法泛化到变化的头部姿态。因此，在校准数据中加入姿态多样性对于在现实世界场景中实现稳健和可靠的注视点估计是至关重要的。

4.4 多样化姿态校准提升泛化能力

为了进一步研究影响校准估计器稳健性的关键因素，我们比较了在两种条件下性能的变化：增加校准点 (PoGs) 的数量和增加校准头部姿态的多样性。图 ?? 展示了校准估计器的平均误差如何随着校准样本中校准点和头部姿态数量的增加而变化。

如图 ?? (a) 所示，当校准样本中的头部姿态固定时，增加 PoG 的数量仅带来边际的性能提升。在某些情况下，添加过多的 PoG 甚至会导致过拟合。例如，使用十三个 PoG 通过 Finetune MLP 校准的 iTracker 的表现比未校准的基线更差。

相比之下，图 ?? (b) 和 (c) 显示，在校准样本中增加头部姿态的多样性显著提高了校准估计器的鲁棒性。值得注意的是，即使只有一个点，在多种 (3 或 4 种) 头部姿态下进行校准仍然一致优于在单一头部姿态下用多个 (9 或 13 个) 点的情况，如图 ?? (b) 所示。

这些发现强烈表明头部姿态的多样性是个性化校准中的一个关键因素。将不同的头部姿态纳入校准样本可以显著增强校准估计器的泛化能力，在设计实际校准策略时应当优先考虑这一点。

如第 4 节所示，校准样本中的头部姿势多样性极大地影响了特定人估计器的性能。然而，增加这种多样性会使数据收集变得复杂，并需要更多用户合作。因此，寻找一种用户友好的校准策略对于在移动设备上的实际部署至关重要。

如图 ?? (b) 所示，我们提出了一种新的动态校准策略。对于每个校准的注视点，用户只需在一定范围内旋转和移动手机，同时保持注视该点，这不仅能够收集具有足够多样化头部姿势的校准样本，而且对用户友好，并且易于在移动设备上操作。正如在 3 节中提到的，Dynamic-MobilePoG 的收集过程可以有效地模拟所提出的动态校准策略，因为数据集中包含持续变化的头部姿势。我们将对 Dynamic-MobilePoG 进行实验，以证明我们提出的动态策略优于传统的静态策略。

Model	Calibration Algorithm	1 calib. point		2 calib. points		4 calib. points		6 calib. points	
		static	dynamic	static	dynamic	static	dynamic	static	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: 动态-MobilePoG 上静态策略和动态策略的比较。度量是欧几里得距离（厘米）。

4.5 实验评估

4.5.1 实验设置

我们在 Dynamic-MobilePoG 上评估了所提策略。随机选择五个受试者作为校准集，其余 27 个受试者用于训练。与 4 节中的设置相似，我们首先通过在 GazeCapture 数据集上预训练 iTracker [20] 或 AFFNet [5]，然后在 Dynamic-MobilePoG 训练集上进行训练来训练通用注视点估计器。

为了比较传统的静态策略和我们提出的动态策略，我们在 Dynamic-MobilePoG 上进行了实验，通过选择不同设置的校准样本来模拟这两种策略。对于静态策略，我们从序列中选择连续的帧作为校准数据，并使用剩余的数据进行测试。对于动态策略，我们以固定步长从帧序列中采样，确保选定的帧具有多样的头部姿态，并在每个人的其余数据上评估校准后的模型。我们对两种策略分别进行了 1、2、4 和 6 个校准点的实验，并且为每个校准点采样了 30 帧。点的位置在补充材料中有所说明。

最后，为了确保实验结论的可推广性，我们还采用了一系列校准算法（即，SVR、线性探针、MLP 微调和全微调）来构建个性化的 PoG 估计器。

4.5.2 实验结果

静态和动态策略的校准性能。表 3 报告了在静态和动态策略下使用各种配置（即两种模型、四种校准算法和四种不同数量的校准点）时校准估计器的误差。可以看出，动态策略在几乎所有配置中都比静态策略表现出色，除了 AFFNet 在使用线性探针和两个校准点的情况下的表现。我们还可以看到，动态策略甚至能在仅使用少量校准点时超过静态策略的性能，而静态策略需要大量的校准 PoGs。例如，iTracker 通过动态策略并使用 Finetune MLP 进行校准且仅使用一个 PoG 的测试误差达到了 1.81 cm，而静态策略和六个 PoG 的测试误差为 1.95 cm。上述实验结果表明，能够收集具有丰富姿态多样性的校准样本的动态策略显著提高了模型校准后的鲁棒性。

静态与动态策略的校准稳定性。为了进一步展示我们所提动态策略的效率，我们评估了不同估计器和校准算法在两种不同校准策略下的模型性能稳定性。图 3 显示了在不同的训练周期中使用微调 MLP 和完整微调对 iTracker 和 AFFNet 进行校准时的性能表现。我们可以观察到，使用动态策略校准的模型逐渐收敛到最佳性能并保持稳定，而使用静态策略校准的模型则表现出明显的性能波动。对于性能不稳定的模型，实现最佳结果可能需要细致调整超参数，例如学习率，或者从特定的训练周期中手动选择检查点，这在真实世界的校准应用中是不切实际的。通过我们提出的动态策略，个体特定的估计器不仅表现出卓越且稳定的性能，而且能够进行 PoG 校准。

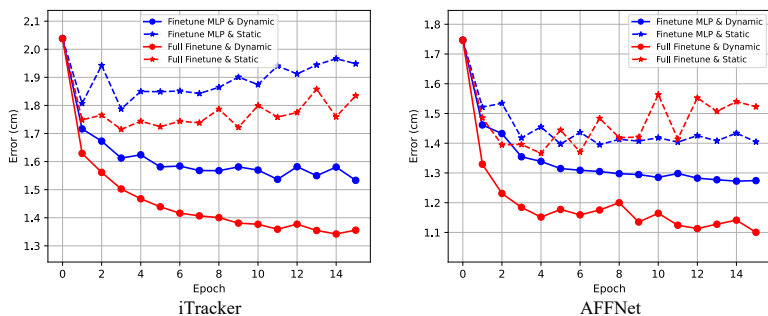


Figure 3: 静态和动态策略的校准稳定性。所有模型都使用六个校准点进行校准。

5 结论

为了弥合个性化 PoG 校准策略研究中的差距，并减轻传统静态策略对头部姿态变化的敏感性，我们首先构建了一个新数据集 MobilePoG，它能真实模拟校准过程，从而作为个性化校准的基准。我们在 Static-MobilePoG 上的实验表明，个性化校准的主要瓶颈在于校准样本中头部姿态的多样性。最后，我们提出了一种用户友好的动态校准策略，并在 Dynamic-MobilePoG 上验证了其有效性，显示其显著提高了针对个体的估计器在不同头部姿态下的性能和鲁棒性。未来的工作可能着重于将后续个性化校准的机制整合到通用 PoG 估计器的设计中，特别是在头部姿态变化的情况下，这可以导致更高效的校准和更强的校准模型鲁棒性。

致谢：本工作由国家自然科学基金（编号 62176248, U2336213）资助。

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. www.xueshuxiangzi.com

- [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. Yassillo, 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. Qian, 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.

在主论文中，我们提到传统的静态策略对头部姿势变化非常敏感，并且展示了校准样本的头部姿势多样性对特定人的凝视点（PoG）估计器有很大影响。在本节中，我们将对 PoG 估计和校准中的头部姿势进行更多讨论。

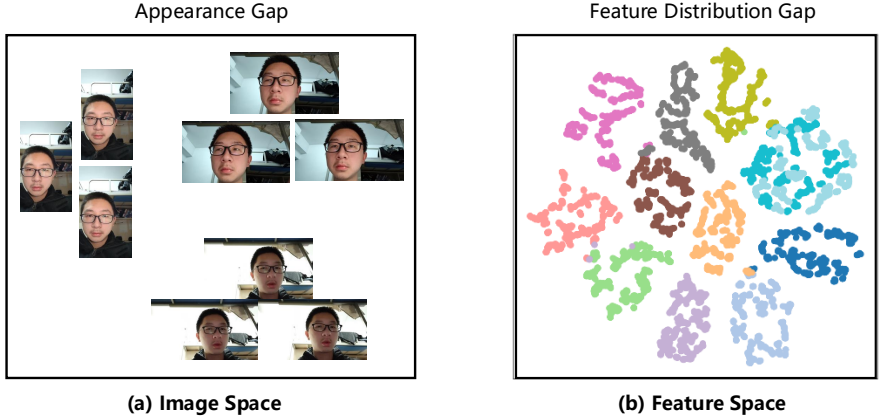


Figure 4: 不同头部姿势之间的差距。(a) 显示了图像空间中的外观差距。(b) 用 t-SNE 显示了特征空间中的分布差距。

图 4 展示了静态-移动注视点之间不同头部姿态的差距。图 4 (a) 展示了图像空间中的外观差距。我们可以观察到，不同头部姿态之间存在显著的差距。当我们使用某些校准算法，如完整微调时，具有有限姿态多样性的校准样本由于外观差距可能导致模型过拟合，使其对姿态变化的鲁棒性降低。在图 4 (b) 中，我们对一个主题由训练过的通用估计器提取的特征进行 t-SNE 分析，其中每种颜色代表静态-移动注视点中的一个头部姿态。我们可以看到，学习到的特征基于头部姿态展现了聚类模式。在个性化校准阶段，如果我们采用一些算法如 SVR 和微调 MLP，它们会冻结估计器的骨干并微调回归头。这相当于将特征作为输入提供给模型。因此，图 4 (b) 所示的聚类模式表明，单一头部姿态将导致过拟合。总体而言，头部姿态的多样性是驱动个性化校准性能的主要因素。

在本节中，我们展示了 MobilePoG 数据集的更多可视化内容。图 ?? 显示了来自 Static-MobilePoG 的一些示例，其中包含 12 种类型的静态头部姿态。图 5 显示了来自 Dynamic-MobilePoG 的一些示例，其中头部姿态在帧序列中连续变化。可以观察到，MobilePoG 数据集展示了丰富的头部姿态多样性，使其非常适合用于评估模型在不同实验设置下对头部姿态变化的鲁棒性。

A 更多实验细节

在本小节中，我们将提供有关主论文实验设置的更多细节。

估计器：我们选择 iTracker [20] 和 AFFNet [5] 作为我们的估计器，因为它们分别是 PoG 估计中的代表性基线和最新的模型。它们都由基于 CNN 的骨干网络和 MLP 回归头组成。

校准 PoG：图 6 显示了我们在实验中使用的校准点的位置。校准点的位置尽可能均匀地分布在手机屏幕上，旨在实现校准空间的更全面覆盖。

在本小节中，我们将提供有关我们实验中使用的校准算法的更多详细信息。

SVR 是一种回归方法，它应用 SVM 的原理，通过在指定误差范围内拟合一个函

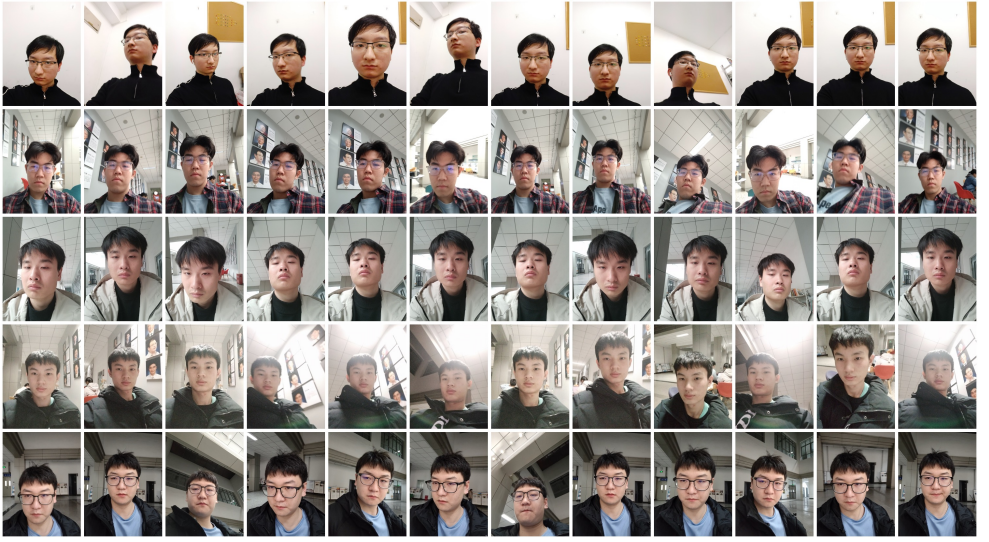


Figure 5: 动态移动 PoG 的可视化。我们可以看到每行的头部姿势在不断变化。

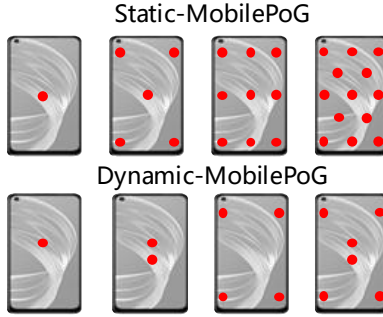


Figure 6: 校准点的位置。

数来预测连续值。为了将 SVR 应用于 PoG 校准，我们用 SVR 替换模型的 MLP 头部，以对从主干提取的特征进行回归，正如 [20] 所述。

线性探头通过在顶部训练一个简单的线性层进行预测，同时保持骨干网络冻结。Liu *et al.* [21] 提出，可以通过线性模型估计地面真值和预测之间的偏差。我们将这一结论从 3D 凝视方向推广到 2D PoG，并在校准过程中，在模型顶部添加一个线性层，同时冻结其余的参数。

微调 MLP 是在校准期间冻结特征提取骨干的参数，只学习 MLP 头部的参数。这种方法在模型性能和训练效率之间取得了平衡，是视觉模型中微调和迁移学习的常用算法。我们采用 AdamW 优化器，并在校准样本上训练 15 个周期。Static-MobilePoG 的学习率是 $1e-4$ ，Dynamic-MobilePoG 的学习率是 $5e-4$ 。

全微调是在新的数据集上训练所有的模型参数。尽管这涉及到更多的参数量，但它可以显著提高模型的迁移性能。我们采用 AdamW 优化器并在校准样本上训练 15 轮。在 Static-MobilePoG 上的学习率为 $1e-5$ ，而在 Dynamic-MobilePoG 上的学习率为 $5e-5$ 。

还有一些其他的校准算法，比如 Prompt Tuning [22] 和 Anchor Method [15]。我们没有选择这些算法，因为它们需要更复杂的流程，这使得它们难以在大规模上灵活调整校准设置。如在 Section ?? 中讨论的那样，如果我们采用传统的静态策略，不同头部姿态之间固有的差异仍然会限制这些算法的性能。

在这一小节中，我们提供了主论文中未显示的关于静态-移动 PoG 的所有实验结果。

下表与主要论文中的表 2 和图 4 直接相关。它们提供了扩展的结果和额外的细节以支持主要发现的可信性。有关实验的全面分析和讨论，请参见正文第 4 节。

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.
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
	Linear Probe	3.82	7.08	6.81	2.40	4.13	4.00	2.17	5.30	5.07	2.29	4.58
	Finetune MLP	2.35	2.74	2.71	2.10	2.88	2.82	2.00	2.83	2.77	2.02	3.05
	Full Finetune	2.10	2.71	2.66	2.07	2.72	2.67	1.97	2.63	2.58	1.40	2.65
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
	Linear Probe	4.18	6.84	6.62	2.02	3.29	3.19	1.97	3.69	3.57	2.05	3.55
	Finetune MLP	2.06	2.53	2.49	2.04	2.41	2.38	1.52	2.45	2.38	1.48	2.57
	Full Finetune	1.94	2.37	2.33	1.99	2.39	2.36	1.42	2.45	2.38	1.40	2.37

Table 4: 单一校准头部姿势和不同数量校准点在静态-移动注视点上的结果。“相同”和“不同”分别表示校准模型在与校准头部姿势相同和不同的测试样本上的测试误差。度量标准是欧几里得距离（厘米）。

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.
iTracker	w/o calibration	2.82										
	SVR	4.15	8.70	7.95	2.31	5.75	5.22	2.13	4.87	4.48	2.01	5.34
	Linear Probe	3.25	4.48	4.28	2.24	3.13	2.99	2.54	3.29	3.18	2.47	3.95
	Finetune MLP	2.53	2.65	2.63	2.24	2.69	2.62	1.84	2.75	2.62	1.76	2.64
	Full Finetune	2.46	2.58	2.56	1.70	2.56	2.42	1.69	2.53	2.41	1.76	2.45
AFFNet	w/o calibration	2.54										
	SVR	4.03	7.82	7.19	1.82	6.45	5.73	1.99	5.05	4.61	1.81	6.09
	Linear Probe	2.89	4.20	3.98	2.23	2.64	2.58	2.35	2.50	2.48	2.42	3.07
	Finetune MLP	2.07	2.51	2.44	2.00	2.45	2.38	1.62	2.41	2.29	1.62	2.65
	Full Finetune	1.71	2.38	2.27	1.70	2.32	2.23	1.60	2.30	2.20	1.68	2.36

Table 5: 关于校准头部姿态和不同数量的校准点对静态移动注视点的结果。“相同”和“不同”分别代表经过校准的模型在与校准头部姿态相同和不同的测试样本上的测试误差。度量标准为欧氏距离（厘米）。

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	5.99	5.54	2.37	4.47	3.98	1.74	4.02	3.52	1.77	4.21	3.7
	Linear Probe	2.75	3.10	3.02	2.51	3.03	2.91	2.14	2.75	2.62	2.38	3.17	3.0
	Finetune MLP	2.59	2.63	2.62	2.11	2.85	2.68	1.75	2.73	2.51	1.81	2.77	2.5
	Full Finetune	2.32	2.53	2.48	2.06	2.48	2.38	1.90	2.52	2.38	1.73	2.49	2.3
AFFNet	w/o calibration	2.54											
	SVR	3.73	5.90	5.36	2.27	4.41	3.91	2.03	4.19	3.72	1.84	4.08	3.6
	Linear Probe	2.62	2.96	2.88	2.25	2.75	2.64	2.09	2.66	2.54	2.30	2.75	2.6
	Finetune MLP	2.38	2.37	2.38	1.86	2.54	2.38	1.89	2.32	2.22	1.73	2.47	2.3
	Full Finetune	2.05	2.31	2.25	1.64	2.28	2.13	1.57	2.21	2.07	1.59	2.23	2.1

Table 6: 三个校准头部姿势和不同数量校准点在静态移动凝视点上的结果。“相同”和“不同”分别表示校准模型在测试样本上与校准头部姿势相同和不同的头部姿势的测试误差。度量标准是欧几里得距离（厘米）。

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	3.83	4.62	4.36	2.13	3.45	3.04	2.09	3.15	2.84	1.85	3.10	2.7
	Linear Probe	2.73	2.78	2.76	2.55	2.67	2.63	2.67	2.71	2.70	3.03	2.99	3.0
	Finetune MLP	2.64	2.65	2.64	2.25	2.62	2.51	2.10	2.63	2.47	2.03	2.56	2.4
	Full Finetune	2.53	2.47	2.49	2.19	2.41	2.34	2.02	2.37	2.27	1.96	2.30	2.2
AFFNet	w/o calibration	2.54											
	SVR	3.57	3.87	3.79	1.97	2.84	2.57	1.77	2.79	2.49	1.74	2.77	2.4
	Linear Probe	2.47	2.36	2.40	2.19	2.34	2.30	2.26	2.56	2.47	2.62	2.91	2.8
	Finetune MLP	2.39	2.24	2.29	2.07	2.30	2.23	1.97	2.32	2.22	1.92	2.31	2.2
	Full Finetune	2.20	2.14	2.16	1.96	2.08	2.04	1.88	2.10	2.03	1.85	2.06	2.0

Table 7: 四种校准头部姿势和不同数量校准点对静态移动 PoG 的影响结果。”相同”和”不同”分别表示校准模型在测试样本上与校准头部姿势相同和不同的头部姿势下的测试误差。衡量标准是欧几里得距离（厘米）。