

最近的人脸识别 (FR) 技术进步主要由在大型标注数据集上训练的深度学习模型推动。然而, 由于日益增长的隐私、伦理和法律问题, 对真实人脸身份的使用越来越受限或不可行。DataCV ICCV 人脸识别数据集构建挑战通过引入创建不包含任何现实世界身份的高质量训练数据集的任务来解决这一问题。其目标是在仅使用合成身份的情况下训练一个固定的人脸识别模型, 旨在实现与使用真实身份训练的模型在测试集上相当或更好的表现。

虽然合成数据提供了一种很有前途的替代方案, 但它引入了一些显著的挑战:

1. 身份一致性: 生成模型, 特别是基于扩散的方法, 旨在最大化多样性和真实感。然而, 这使得生成多张一致代表相同合成身份 [13, 19] 的不同图像变得困难。这种不一致性可能会混淆训练过程并降低识别准确性。
2. 类别内变化: 为了实现强大的泛化能力, FR 模型必须能够处理姿势、光照、表情和背景的变化 [5, 8, 12, 14, 16, 17]。合成数据集往往缺乏这种自然的多样性, 特别是在从单一参考图像生成样本或通过严格受限的管道生成样本的情况下。
3. 身份泄露: 数据集不得包含与公共现实世界数据集重叠的身份。这是竞赛的严格要求, 当使用预训练生成模型或外部数据源而没有明确的身份控制时, 执行起来比较困难。
4. 计算成本: 使用扩散模型进行高质量图像生成在计算上非常昂贵, 特别是当扩展到成千上万的身份, 每个身份有数十张图像时。这限制了在没有显著计算资源的情况下进行大规模合成的可行性。

为了应对这些挑战, 我们提出了一种结合了数据集清理、身份合成和基于课程的训练结构的混合方法。我们的方法首先通过精炼基线 HSFace 数据集 [18] 开始。虽然 HSFace 提供了一个强有力的起点, 但我们观察到一些身份显示出标签噪音或内部身份不一致。为了提高数据集的质量, 我们采用了一种专家混合 (MoE) [9, 10] 策略, 该策略结合了自动人脸嵌入聚类 and GPT-4o 协助 [1] 验证。对于每个身份, 我们保留最大的、一致的图像簇, 并应用数据增强以确保每个身份有 50 个样本。对于不一致性过高 (即超过 80 % 的图像被移除) 的身份, 将其丢弃。

为了填补清理后的空缺并确保多样性, 我们使用 Stable Diffusion [15] 的提示工程 [9, 11] 生成新的身份。对于每个生成的身份, 创建一张参考图像, 并使用 Vec2Face [18] 合成额外的 49 个样本, 从而获得视觉一致的图像集。然而, 这些生成的样本往往过于相似。为了减轻过拟合的风险, 我们应用课程学习 [4] 策略, 将这些容易 (低类内方差) 的身份置于训练数据集的开头, 然后加入清理后的 HSFace 集中的较难 (高类内方差) 身份 [2, 3]。这种逐步曝光提高了模型的泛化能力。

重要的是, 我们验证了所有合成身份与任何已知真实个人都不重叠, 完全满足比赛的身份不泄露要求。我们的最终数据集构建为三个规模: 10K、20K 和 100K

身份, 每个身份包含 50 张图像, 分别生成总共 500K、1M 和 5M 张合成面部图像。经过主办方的评估, 我们的提交在所有三个规模的比赛排行榜上均获得了第一名, 展示了我们的合成数据集构建策略在固定训练条件下提升人脸识别性能的有效性。

1. 方法

我们的方法旨在构建一个高质量的合成训练数据集, 使得一个固定的面部识别模型能够在不使用任何真实身份的情况下实现强大的性能。该流程包括三个主要组件: 数据集清理、身份生成和基于课程的数据结构化。

1.1. 概述

图 1 展示了我们方法的整体流程。我们首先通过使用专家混合策略 (Mixture-of-Experts, MoE) 对提供的 HSFace 数据集 [18] 进行精细化处理, 以去除噪声或不一致的身份信息。同时, 我们通过稳定扩散 (Stable Diffusion) 和 Vec2Face 生成新的合成身份, 以增加多样性和规模。结果身份被结构化, 并使用课程学习策略以提高训练效率和模型泛化能力。最后, 清理后的和生成的身份被合并成具有 10K、20K 和 100K 个身份的数据集, 每个身份包含 50 张图片。

1.2. 数据集清理

我们从组织者提供的 HSFace 数据集开始。虽然它提供了一个强健的基准, 但我们观察到, 一些身份存在标签噪声和身份内不一致性, 这可能会对训练产生负面影响, 如图 2 所示, ID:000029 包含受两个身份影响的图像, 而 ID:000786 包含受四个以上身份影响的图像。为了提高数据集的质量, 我们应用了一种包含多个专家 (MoE) 的清理策略, 其中包括:

- 使用预训练的 FR 模型进行人脸嵌入聚类, 以识别身份中的不一致图像。
- 使用语义和视觉线索的 GPT-4o 辅助验证以解决模糊的身份集群。

具体来说, 我们使用基线提供的模型 hsface300K.pth 提取每个身份的面部嵌入。这些嵌入使用 DBSCAN [6] 算法进行聚类, 采用余弦相似性作为距离度量。对于每个身份, 我们只保留最大簇, 假设这是最一致的子集。聚类后, 原始样本中剩余的不到 20 % 的身份将被丢弃, 以确保标签纯度。

对于图像质量低或视觉一致性难以评估的模糊情况, 我们采用 GPT-4o 进行多模态身份验证。我们生成一个包含给定身份的所有面部样本的合成图像, 每个样本都标注有唯一的索引。该图像与精心设计的提示一起提交给 GPT-4o, 要求模型识别哪些样本不属于大多数相同的身份。提示模板如下所示:

“你将收到一张由多个面部照片组成的图像, 这些照片以网格方式排列, 每个面孔下面都有一个数字 ID 标签。请识别所有不属于多数的那个人的面部图像。你的答案只需包含异常面孔的数字 ID, 并用逗号分隔 (例如,

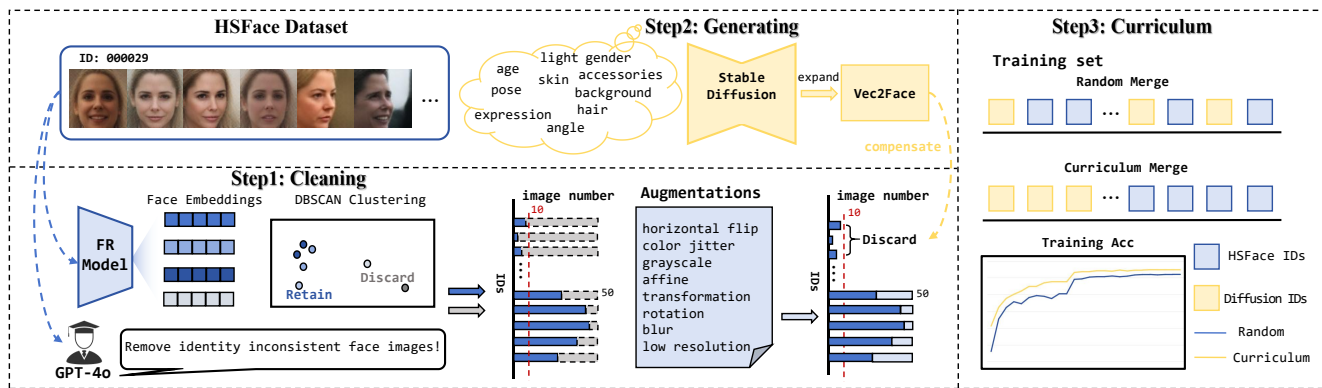


Figure 1. 我们方法的概述。我们首先使用一个人脸识别模型和 GPT-4o 验证进行数据清理（步骤 1），接着进行 DBSCAN 聚类以保留一致的人脸图像。应用数据增强技术使每个身份的图像数量恢复到 50 张。然后，我们使用 Stable Diffusion 和 Vec2Face 生成其他身份（步骤 2），引入控制的属性变化，如姿势、表情和光照。最后，在步骤 3 中，我们使用课程学习策略构建训练集，以分阶段的方式合并扩散生成的身份和 HSFace 身份，与随机混合相比，提高了收敛性和泛化能力。

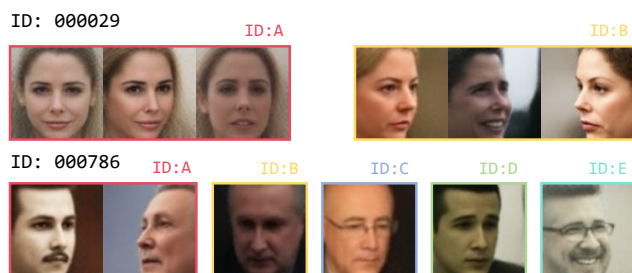


Figure 2. 在 HSFace 中存在身份不一致的图像。

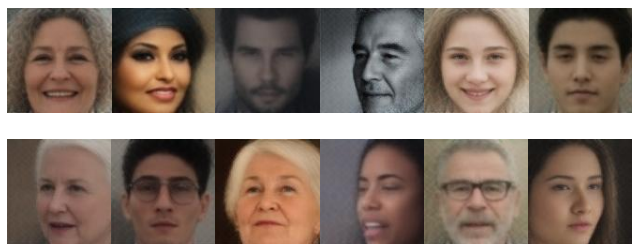


Figure 3. 由提示词通过稳定扩散生成的代表性样本。

001,005,012)。不要包含任何附加文本、标点或空格。不要返回空响应。”

为确保响应的有效性，我们应用正则表达式过滤器来验证 GPT-4o 的输出是否符合预期格式。这种多模态验证机制在解决单靠视觉相似性不足以判断的边缘情况时特别有效。

通过结合这些专家，我们能够构建出高质量的清理后 HSFace 子集，这成为我们最终训练数据集的关键组成部分之一。为了每个身份维持 50 张图像的一致数量，我们对清理后的剩余有效样本应用数据增强。增强流程的具体细节在第 1.4 节提供。

为了扩大身份空间并补偿在数据集清洗阶段移除的身份，我们使用一个两阶段的管道构建大量的合

成身份。这个过程结合了稳定扩散用于身份创建和 Vec2Face 用于身份一致的样本扩展。

我们首先从一个精心策划的包含数百个基本身份描述的列表中抽取提示。这些提示被精心设计以指定年龄、性别、种族、面部结构、发型和配饰等特征，同时确保生成的主体面部清晰可见且无遮挡。该提示被传递给一个稳定扩散 XL¹ 文本到图像模型，该模型为每一个新的合成身份合成一张高分辨率肖像图像。为了确保整个身份集在姿势、光照和表情上的多样性，我们在提示级引入了变化。每个身份提示通过一个次要描述进行增强，指定了从预定义列表中随机抽样的属性，包括头部姿态（如“面朝左”、“右倾”）、表情（如“微笑”、“严肃表情”）、光照条件（如“在自然日光下”、“由柔和灯光照亮”）、相机角度、背景和可选配饰。这种提示增强有助于最大化生成的参考图像的身份间多样性。通过这种提示方案生成的代表性样本展示在图 3 中。

一旦通过 Stable Diffusion 生成参考图像后，我们使用 insightface.app.FaceAnalysis 工具包 [7] 应用人脸检测器，以确保图像中包含清晰可见且有效的人脸。如果未检测到人脸，则丢弃该样本。否则，提取最大的人脸区域并用作标准的身份图像。为了高效地将每个身份扩展成多样的样本集，我们将裁剪后的参考图像输入到 Vec2Face 模型 [18] 中，这是一种轻量级解码器，可以快速生成 49 个身份一致的变体。这些变化包括姿势、光照和面部表情的控制变化，同时保留核心身份特征。与基于扩散的方法相比，Vec2Face 显著降低了计算成本，使其非常适合大规模数据集生成。

这个两阶段流程利用稳定扩散进行高质量的身份初始化，并使用 Vec2Face 进行快速、一致的样本生成，从而产生了一个视觉上连贯的身份类别，包含 50 张图像。它确保了高身份间的多样性和强身份内的一致性，这两者对于高效和有效的面部识别模型训练都至关重要。

¹模型提供在 <https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0/tree/main> 中。

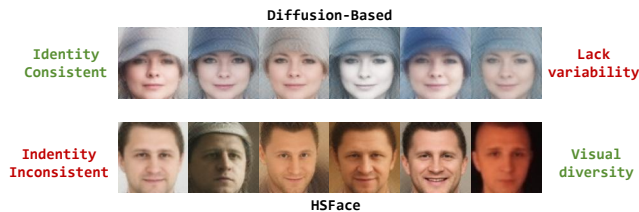


Figure 4. 扩散生成的图像往往在视觉上相似并保持身份一致性，而 HSFace 样本则表现出更大的视觉多样性，但可能包含身份不一致性。

1.3. 基于课程的数据结构

虽然生成的身份具有较强的类内一致性，但它们往往缺乏多样性，例如表情、姿势和光照的自然变化。相比之下，HSFace 数据集的清理子集包含更多视觉上多样化和真实的样本，但也引入了更大的类内变化和残留噪声，使模型更难以学习。这种对比在图 4 中有所说明，图中展示了生成的数据集和 HSFace 数据集的代表性例子。

为了缓解这种不平衡并改善泛化能力，我们采用了一种课程学习策略，根据难度递增的顺序组织训练数据。模型首先接触通过 Stable Diffusion 和 Vec2Face 生成的合成图像，这些图像提供了干净且身份一致的样本。随着训练的推进，模型会遇到来自清理后的 HSFace 数据集中更为多样和复杂的例子。这种课程安排使模型能够先从同质数据中建立稳健的身份表征，然后再适应更高的变异性。根据经验，我们发现这种渐进过程导致更快的收敛速度和更好的识别性能，尤其在训练预算受限的情况下。

在实际操作中，我们舍弃了原始 HSFace 子集中样本数量过少的身份，这些身份在数据清理过程中被识别出来。对于每个移除的身份，我们从一个合成身份库中随机选择一个替代者，每个替代者提供一整套图像（例如，50 个样本）。这确保了身份的总数量保持不变，同时保留了一致的类内结构。最终的数据集通过将所有选择的合成身份放在训练集的前面构建，随后是剩余的经过清理的 HSFace 身份。模型首先接触到高质量、身份一致的合成样本，然后再遇到更加多样化和具有挑战性的真实世界数据。身份分组在整个过程中得以保留，并且在数据集构建过程中不进行洗牌。这种简单但有效的结构促成了稳定的收敛，并显著提高了识别准确度。

1.4. 实现细节

硬件和框架。 所有实验都在单个 NVIDIA RTX 4090 GPU 上进行。实现是基于 PyTorch 的。

特征提取。 我们使用预训练模型 `hsface300.pth`² 作为数据清洗和聚类过程中的特征提取器。该模型在

²预训练模型提供在 https://huggingface.co/BooBooWu/Vec2Face/tree/main/fr_weights。

Vec2Face 基线中提供。除了比赛要求的最终固定训练外，不进行任何训练。

我们基于三个数据集构建了我们的数据集：HSFace-10K、HSFace-20K 和 HSFace-100K³，分别包含 10K、20K 和 100K 个身份。每个身份最多包含 50 张图像。

所有提示生成和描述过滤均通过 GPT-4o API 进行，该 API 能够为身份提示和字幕验证提供高质量的文本推理和属性操作。

过滤后剩余图像少于 10 的

数据清洗参数。 一个身份被丢弃。为了确保身份内的视觉一致性，我们在特征上使用余弦距离应用 DBSCAN 聚类。相似性阈值在 0.3 到 0.9 之间动态调整，以确保最大簇保留每个身份总图像的 50 % 至 80 % 之间。

对数据进行清理后，我们对所有剩余的训练图像应用了一致的数据增强流程。此步骤的主要目标是将每个身份的图像数量恢复到固定的 50 张，以补偿在过滤和聚类期间被移除的样本。除了补充数据集之外，该增强策略还引入了在姿势、光照和分辨率上的受控变化，从而在保持身份一致性的同时增强模型的鲁棒性。完整的增强序列如下：

- 以 0.5 的概率随机水平翻转
- 随机颜色扰动（亮度、对比度、饱和度、色相），概率为 0.8
- 以 0.2 的概率进行随机灰度转换
- 随机仿射变换（旋转不超过 10° ，平移不超过 5 %，缩放从 0.95 到 1.05，剪切不超过 5° ），概率为 0.5
- 以 0.5 的概率随机平面旋转 ($\pm 5^\circ$)
- 高斯模糊，核大小为 3 和 $\sigma \in [0.1, 2.0]$
- 使用定制的 `low_resolution()` 函数以 0.5 的因子降采样到低分辨率

这个增强管道是在数据清洗后离线应用的，用于生成新的图像并存储在最终训练数据集中。因此，不需要对训练代码或运行时转换进行修改。

1.5. 结果

我们的提交作品在面部识别挑战中由竞赛组织者在三种不同的训练规模下进行评估：10K、20K 和 100K 身份。每个赛道对应一个固定大小的训练集，每个身份包含 50 张图像，并使用三个官方测试子集（ACC 1, ACC 2, ACC 3）进行评估。最终排名由这三个子集的平均准确率决定。

如表 1、2 和 3 所示，我们的方法在所有规模上均稳居第一。我们在每个规模上的平均准确率都达到了最高，分别是 84.37 % (10K)、85.43 % (20K) 和 86.78 % (100K)，并且在多个单独的测试集上领先。这些结果证明了我们基于课程训练流程的有效性和可扩展性、高质量合成数据的生成以及强大的身份清理过程。

³基础数据集在 <https://huggingface.co/datasets/BooBooWu/Vec2Face/tree/main/HSFaces> 中提供。

Table 1. 使用 10K 规模的训练集进行的测试结果。最佳结果以粗体显示。

Participants	Average ACC (%)	ACC 1 (%)	ACC 2 (%)	ACC 3 (%)
nicolo.didomenico	77.26	84.84	77.67	69.26
mnogolikomus	82.83	88.21	92.52	67.77
Lin0	83.20	88.50	92.58	68.53
f10942093	83.25	88.97	93.12	67.67
GZ_Xu	83.45	88.59	92.63	69.14
XLW	83.51	91.20	91.03	68.30
anjith2006	83.66	91.26	89.42	70.30
Qile_Xu	84.31	88.69	94.35	69.91
Ours	84.37	91.53	92.62	68.97

Table 2. 20K 规模训练集的测试结果。最佳结果以粗体显示。

Participants	Average ACC (%)	ACC 1 (%)	ACC 2 (%)	ACC 3 (%)
nicolo.didomenico	76.80	81.47	78.98	69.95
mnogolikomus	83.03	88.21	92.32	68.55
f10942093	84.49	89.86	93.85	69.76
anjith2006	85.13	92.23	91.20	71.97
Ours	85.43	92.69	93.40	70.20

Table 3. 使用 10 万规模训练集的测试结果。最佳结果以粗体显示。

Participants	Average ACC (%)	ACC 1 (%)	ACC 2 (%)	ACC 3 (%)
nicolo.didomenico	77.96	83.20	80.98	69.68
anjith2006	85.74	92.19	92.88	72.14
Ours	86.78	93.96	94.17	72.23

2. 结论

本文解决了在不依赖任何真实世界身份的情况下构建高质量人脸识别合成数据集的挑战。针对 DataCV ICCV 人脸识别数据集构建挑战的要求，我们提出了一个融合数据集清理、身份合成和基于课程训练的统一流程。

我们首先通过基于嵌入的聚类和 GPT-4o 辅助验证去除噪声和不一致样本，来增强 HSFace 数据集。这提高了身份一致性，同时保留了类内多样性。为了补充清理后的数据集，我们使用 Stable Diffusion 和 Vec2Face 生成新的身份，确保在姿势、光照和表情等属性上维持视觉一致性和变化。为了指导训练过程，我们引入了一种课程学习策略，根据类内变化从低难度到高难度排列训练数据，使模型能够逐渐适应并更好地泛化。

所有生成的身份已验证与真实世界中的个体没有任何重叠，完全符合比赛的身份不泄露约束。构建的最终数据集在三个不同的规模下，均在挑战的所有赛道中获得第一名。这表明我们的合成数据构建策略在受限设定中训练竞争性人脸识别模型的效果。

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- [2] Shilong Bao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. Improved diversity-promoting collaborative metric learning for recommendation. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 9004–9022, 2024.
- [3] Shilong Bao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. Aucpro: Auc-oriented provable robustness learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.
- [4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In ICML, pages 41–48, 2009.
- [5] Fadi Boutros, Marco Huber, Patrick Siebke, Tim Rieber, and Naser Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In

- International Joint Conference on Biometrics (IJCB), pages 1–11. IEEE, 2022.
- [6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD, pages 226–231, 1996.
 - [7] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In ECCV, pages 152–168, 2020.
 - [8] Boyu Han, Qianqian Xu, Zhiyong Yang, Shilong Bao, Peisong Wen, Yangbangyan Jiang, and Qingming Huang. Aucseg: Auc-oriented pixel-level long-tail semantic segmentation. Advances in Neural Information Processing Systems, pages 126863–126907, 2024.
 - [9] Cong Hua, Qianqian Xu, Zhiyong Yang, Zitai Wang, Shilong Bao, and Qingming Huang. Openworldauc: Towards unified evaluation and optimization for open-world prompt tuning. In Forty-second International Conference on Machine Learning, 2025.
 - [10] Dmitry Lepikhin, HyounJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. In International Conference on Learning Representations, 2020.
 - [11] Feiran Li, Qianqian Xu, Shilong Bao, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. One image is worth a thousand words: A usability preservable text-image collaborative erasing framework. In Forty-second International Conference on Machine Learning, 2025.
 - [12] Yang Liu, Qianqian Xu, Peisong Wen, Siran Dai, and Qingming Huang. Not all pairs are equal: Hierarchical learning for average-precision-oriented video retrieval. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 3828–3837, 2024.
 - [13] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In ECCV, pages 241–261, 2024.
 - [14] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10880–10890, 2021.
 - [15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. CVPR, pages 10684–10695, 2022.
 - [16] Zhiwen Shao, Yong Zhou, Feiran Li, Hancheng Zhu, and Bing Liu. Joint facial action unit recognition and self-supervised optical flow estimation. Pattern Recognition Letters, pages 70–76, 2024.
 - [17] Zhiwen Shao, Yifan Cheng, Feiran Li, Yong Zhou, Xuequan Lu, Yuan Xie, and Lizhuang Ma. Mol: Joint estimation of micro-expression, optical flow, and landmark via transformer-graph-style convolution. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.
 - [18] Haiyu Wu, Jaskirat Singh, Sicong Tian, Liang Zheng, and Kevin W Bowyer. Vec2face: Scaling face dataset generation with loosely constrained vectors. In ICLR, 2025.
 - [19] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In ICCV, pages 3836–3847, 2023.