像素下的压力：探索基础模型在高分辨率医学成像中的微调范式

Zahra TehraniNasab
McGill University
MILA-Quebec AI Institute
zahra.tehraninasab@mail.mcgill.ca

Amar Kumar
McGill University
MILA-Quebec AI Institute
amar.kumar@mail.mcgill.ca

Tal Arbel
McGill University
MILA-Quebec AI Institute
tal.arbel@mcgill.ca

Abstract

基于扩散模型的基石模型在文本到图像生成方面取得了进步，但大多数努力都局限于低分辨率设置。随着高分辨率图像合成在各种应用中变得越来越重要，特别是在医学成像领域，微调作为一项关键机制，对于将这些强大的预训练模型适应特定任务需求和数据分布至关重要。在这项工作中，我们进行了一项系统性研究，探讨了不同的微调技术对扩展到高分辨率（512×512 像素）图像生成质量的影响。我们评估了一系列多样化的微调方法，包括全面的微调策略和参数高效的微调（PEFT）。我们分析了不同微调方法如何影响关键的质量指标，包括 Fréchet Inception Distance (FID)、Vendi 分数以及提示-图像对齐度。我们还评估了在数据稀缺条件下生成的图像在下游分类任务中的实用性，展示了特定的微调策略可以提高合成图像用于分类器训练和真实图像评价时的生成保真度和下游性能。我们的代码可通过项目网站 [1] 获得。

1 介绍

文本到图像的基础模型在各种计算机视觉任务中表现出显著的成功，在标准基准上持续取得强劲的表现 [12, 4] 。这些模型通过大量自然图像语料库的训练，获得了诸如纹理、形状和复杂的空间模式等视觉表示，这些往往能有效地转移到医学成像领域。这种可转移性在临床环境中特别有价值，因为有标签或配对文本-图像医疗数据的可用性稀缺。鉴于使用小规模专业数据集从零开始训练大规模模型所面临的挑战，微调作为适应基础模型到医疗应用的实际且有效策略已经浮现 [23, 14, 18] 。微调利用预训练期间学到的丰富的视觉先验知识，使模型能够通过针对性更新而非全面重新训练来适应特定领域的任务 [25, 1] 。在这项工作中，我们关注的是 Stable Diffusion [17] v1.5，一个在大规模自然图像-文本配对上预训练的显著潜在扩散模型。尽管其潜在空间架构相比像素空间替代方案能实现更高效的高分辨率合成，但在高分辨率医疗图像上微调 Stable Diffusion 仍面临重大的计算挑战。扩展到高分辨率会迅速

[1] https://tehraninasab.github.io/fine-tuning-pressure/

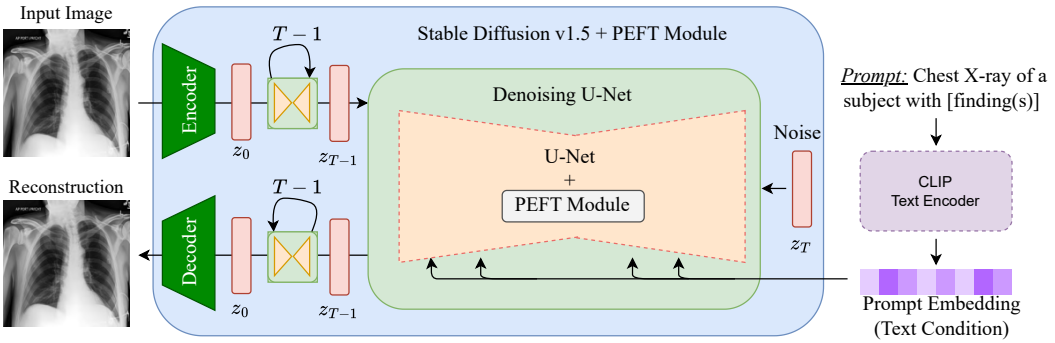

图 1: 概述我们的架构。Stable Diffusion v1.5 模型被改编用于通过不同的微调策略生成高分辨率的胸部 X 光图像。展示在 U-Net 中的 PEFT 模块仅在参数高效的微调配置（例如，LoRA, DoRA, BitFit）中使用；在完全微调设置下，直接对 U-Net（以及可选地 VAE 和文本编码器）进行微调而不使用 PEFT 模块。

增加内存和计算成本，特别是对于基于注意力的架构来说，这给部署带来了重大障碍 [25] 。这些挑战在模型容量与计算可行性之间引入了权衡，通常导致生成质量和诊断可靠性上的妥协。解决这一紧张关系对于实现文本到图像基础模型在实际医疗成像工作流程中的实用和可扩展使用至关重要 [5] 。

最近在参数高效微调（PEFT）方法，如低秩适应（LoRA）[8]、解耦秩适应（DoRA）[15]、BitFit [24] 和扩散特定 PEFT（DiffFit）[22] 方面的发展，通过选择性地更新模型组件而不是整个参数空间 [5]，为解决这些计算扩展挑战提供了有前景的解决方案。

适配器模块 [13, 2] 提供了一种替代方法，通过在冻结的预训练组件之间插入小型、可训练层，从而在不修改原始模型权重的情况下实现特定领域的适应性调整。

尽管在标准分辨率任务中展示了这些方法论上的进展及其有效性，但在高分辨率医学成像环境（512×512）中，这些参数高效方法的表现特征和扩展行为仍很大程度上未被探索，特别是关于它们在输入维度大幅增加时保持关键诊断信息的能力的同时维持计算效率方面。Dutt 等人之前的工作 [5] 分析了不同策略对在 224×224 分辨率下微调的影响，但并未关注在更高分辨率下的图像生成质量受微调影响的情况。近期，Davila 等人的工作 [3] 对医学成像中的图像分类进行了微调策略的比较，但这些图像同样处于低分辨率为 320×320 像素的状态。

本文介绍了使用预训练的基于扩散的文本到图像基础模型 Stable Diffusion 进行高分辨率图像生成微调范式的综合研究。我们的工作主要贡献如下：

- 细调策略的系统性比较—包括全细调和参数高效方法如 LoRA [8]，DoRA [15]，BitFit [24] 和 DiffFit [22]—用于 512×512 图像合成。
- 这些微调策略如何影响生成质量的深入分析，包括视觉保真度（通过 Fr'echet Inception Distance, FID [7] 测量），多样性（通过 Vendi Score [6] 评估）以及提示-图像一致性（使用基于分类器的指标进行评估）。
- 在下游评估中，分类器在合成图像上进行训练并在真实数据上进行测试，评估生成图像对现实世界诊断任务的实用性。

2 方法论

2.1 微调策略

我们评估了多种微调配置（图 1）用于高分辨率图像生成，通过一系列实验涵盖了 VAE、U-Net 和文本编码器的组合使用完整和参数高效的微调（PEFT）；参见表 1。按照 [11, 21]，我们使用模板从表格数据中生成图文对 -*带有* `[disease(s)`[2]`]` *的受试者的胸部* `X` *光片*。

表 1: 微调策略总结。✗：冻结，✓：可训练

#模型	变分自编码器	文本编码器	U-Net	描述 / 可训练
A. 完整的微调策略				
1	✗	✗	✓	U-Net only
2	✗	✓	✗	Text Encoder only
3	✓	✗	✗	VAE only
4	✗	✓	✓	Text encoder + U-Net
5	✓	✗	✓	VAE + U-Net
6	✓	✓	✗	VAE + Text Encoder
7	✓	✓	✓	U-Net + VAE + Text Encoder
B. 参数高效的 U-Net 微调				
8	✗	✗	✓	LoRA [8]
9	✗	✗	✓	DoRA [15]
10	✗	✗	✓	BitFit: Only bias terms updated [24]
11	✗	✗	✓	DiffFit: Diffusion-specific method [22]

全组件微调对于完整组件微调实验（模型 1 – 7），我们探讨了选择性训练 VAE、文本编码器和 U-Net 模块的不同组合的效果。这种方法使我们能够孤立每个组件对扩散模型整体性能的贡献。一些模型专注于单独调整一个组件，同时冻结其他组件以评估其独立影响。例如，模型 1 仅微调 U-Net，而模型 2 和 3 分别只关注文本编码器和 VAE。其他模型则涉及两个组件（模型 4 – 6）或所有三个组件（模型 7）的组合。

参数高效微调为了提高效率和部署灵活性，我们探索了四种显著的参数高效微调策略：

- *低秩适应（LoRA）*[8]: 使用低秩适应将可训练的低秩矩阵插入到网络层中，显著减少可训练参数的同时保持其适应性。
- *解耦秩适应（DoRA）*[15]: 扩展了 LoRA 框架，通过将低秩适应分解为独立的方向和缩放组件，从而允许更灵活地控制特征调制。这种方法增强了表达能力同时保持参数效率，在计算预算受限的情况下提高了性能。
- *模型 10（位适配）*[24]：将训练仅限制在每一层的偏置项上，极端地减少了参数数量。该模型强调了极小调整的惊人有效性。
- *模型 11（差拟合）*[22]：采用特定于扩散的 PEFT 策略，利用针对生成扩散模型量身定制的架构见解。

我们评估合成的医学图像的视觉质量和其在下游临床应用中的实际效用。

[2]这些疾病包括：*未发现异常, 扩大的心胸腺区, 心脏扩大, 肺不透明, 肺部病变, 水肿, 巩固, 肺炎, 肺不张, 气胸, 胸腔积液, 胸膜 其他, 断裂和支持设备*。

2.2 评估合成图像

图像生成质量 类似于先前的工作 [19, 16]，我们使用涵盖视觉保真度和分布相似性的标准指标来评估图像质量。我们采用 Fréchet Inception Distance (FID) [7] 来评估合成图像集与真实图像集之间的分布相似性。Vendi Score [6] 用于衡量生成样本的多样性，补充了以逼真度为导向的 FID 评价。

此外，使用预训练的胸部 X 光多头 Efficient-Net [20] 分类器在类别级别上评估图像提示对齐情况，以评估合成图像是否准确反映了其预期诊断标签。对于每种疾病状况，生成一组 5000 张基于相应文本提示的图像，并通过预训练分类器进行处理。通过对正确分类到各自提示类别的生成图像比例进行测量来量化对齐程度，作为语义忠实度的一个代理指标。这种基于分类器的评估方法通过明确测试是否在生成输出中保留和正确表达了特定疾病的视觉特征，补充了分布度量，并提供了有针对性的临床相关性和提示遵循性的评估。

合成图像的实用性 我们通过仅使用合成医学数据训练分类器，并在多个疾病类别的真实临床数据集上进行测试，来评估这些图像的实际效用。这提供了临床相关性的直接证据，其通过准确性等标准分类指标来衡量。

3 实验与结果

3.1 数据集和实现细节

我们在公开可用的 CheXpert 数据集上进行了实验 [10]。表 1 中的所有策略都在表 2 中提到的训练集上进行微调。需要注意的是，保留的测试集有意设置得比训练集大，以便更严格地评估基于合成图像训练的分类器的泛化性能。为了确保公平比较，所有方法都在四块 80GB H100 GPU 上进行了微调。

3.1.1 指标计算

为了计算 FID，我们使用了一个在胸部 X 光片上预训练的 DenseNet-121 [9] 特征提取器（通过 TorchXRayVision），应用于调整大小后的 224×224 灰度图像。对于 Vendi 评分，从相同的 DenseNet-121 模型中提取了 1024 维的潜在特征。评估是在一个固定的子集上进行的，每个条件最多包含 5,000 个真实样本和 5,000 个合成样本。当给定条件下可用的真实样本少于 5,000 时，相应地匹配合成样本的数量。这导致总共使用了 25,133 个真实图像和 25,133 个合成图像来计算全局 FID 和 Vendi 评分，这些图像是从保留的测试集中的六个目标条件中抽取的。

表 2: 观察的六种疾病在 CheXpert 中的训练和测试分割总结。注意：单个图像可以反映多种并发疾病的出现。

类	训练	验证	测试
Cardiomegaly	3173	1195	4515
Lung Opacity	10269	4075	14658
Edema	6447	2584	9210
No Finding	1801	722	2591
Pneumothorax	2196	827	3027
Pleural Effusion	9001	3505	12972

表 3: 评估合成图像的质量和一致性。Ca: 心脏扩大, Lo: 肺部不透明, Ed: 水肿, Nf: 无异常发现, Pn: 气胸, Pe: 胸腔积液

模型	FID↓	卖出↑	类一致性↑					
Trainable Component			钙	Lo	Ed	Nf	Pn	Pe
U-Net	3.42	5.65	24.1	10.7	25.7	91.6	15.7	21.6
U-Net + Text Encoder	6.57	2.79	11.0	3.1	8.2	96.9	9.99	5.8
U-Net + VAE	7.46	2.59	12.1	1.0	0.7	98.1	48.0	2.0
LoRA [8]	5.65	2.64	5.0	0.5	0.5	98.9	1.0	1.8
DoRA [15]	5.72	3.18	8.3	0.8	0.7	98.8	0.9	1.3
BitFit [24]	5.05	5.89	4.8	12.6	9.8	86.4	13.4	5.0

表 4: 不同模型配置在各类疾病中的准确性。注意：所有模型均使用合成数据进行训练和验证，并在真实图像上进行测试。

模型	钙	丢失	编辑	Nf	Pn	肽
U-Net	37.4	48.8	67.8	90.9	89.4	51.5
U-Net + Text Encoder	77.9	54.0	67.8	81.0	89.4	55.3
U-Net + Vae	15.7	48.8	67.5	90.9	89.4	54.7
LoRA [8]	35.3	47.6	60.2	75.8	86.4	53.6
DoRA [15]	80.6	47.4	52.9	75.8	89.3	52.7
BitFit [24]	77.3	50.9	67.8	90.9	76.4	54.5

3.2 结果

我们注意到，某些微调配置，例如仅涉及 VAE 或仅文本编码器的配置，以及 DiffFit（具体来说，来自表 1 的模型 # 3、5、6、7 和 11），由于始终产生质量较差的图像，而被排除在详细分析之外。这些模型通常会产生不真实或非医学图像输出，限制了它们的解释性和下游评估的实用性。

3.2.1 定性评估

我们在图 2 中对各种微调策略生成的胸部 X 射线图像进行定性比较。结果强调了完全微调的效果，特别是 Stable Diffusion 架构中 U-Net 组件的微调，它作为核心生成模块负责去噪和图像合成。如果不更新 U-Net，该模型难以内化和准确再现医学特征。因此，图 2 排除了一些省略 U-Net 微调的模型，例如仅更新 VAE 或文本编码器的模型，因为它们倾向于产生与原始未调整的 Stable Diffusion 模型相似的不真实输出。几种参数高效微调 (PEFT) 方法产生了视觉上具有竞争力的结果。LoRA 和 DoRA 都可以以高视觉保真度再现关键的病理标记，尽管它们仅训练了模型参数的一小部分。这证明了它们在以降低计算成本的方式将大型生成模型适应于医学领域的有效性。相反，BitFit 虽然极其轻量级，但通常会产生更模糊、结构上不太连贯的输出，尤其是在需要精细解剖细节的区域。此外，在某些情况下，BitFit 会产生类似 RGB 的图像而不是灰度 X 射线照片，表明适应性差，并表明该方法可能不足以用于医疗成像等高风险环境中的特定领域微调。这些定性发现表明，虽然全组件微调产生最高的视觉保真度，但 PEFT 策略在效率和图像质量之间提供了一个引人注目的权衡，尤其是在可扩展或资源受限的部署中。

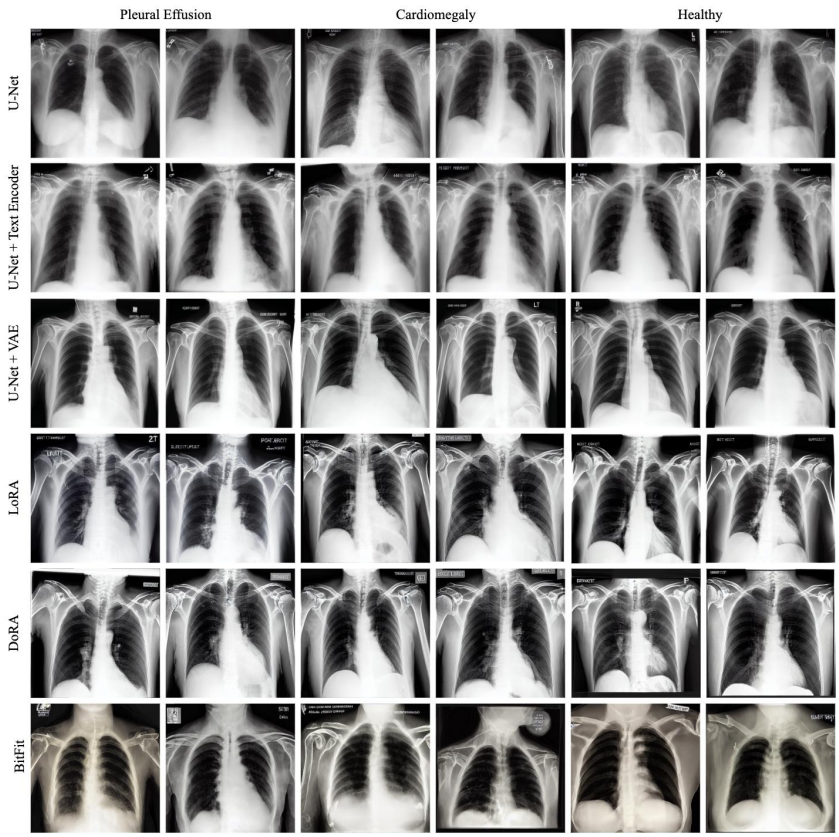

图 2: 不同微调策略下生成的胸部 X 光片在三个疾病类别（胸腔积液、心脏增大和健康）之间的定性比较。每一行展示了一种不同的微调方法（每个疾病类别两个样本）。该比较突出了各种策略之间在解剖学合理性、疾病特定特征和生成保真度方面的差异。注意：仅涉及 VAE 或仅文本编码器和 DiffFit 微调的配置因图像质量不佳而被排除。

定量评估 表 3 显示了不同微调策略在保真度、多样性和下游效用之间的权衡。U-Net 的完全微调达到了最佳的 FID 值 3.42 和强大的类别一致性得分，特别是对于未发现异常 (91.6) 和肺不透明区域 (25.7)，这确认了更新核心生成模块的重要性。在 PEFT 方法中，LoRA 和 DoRA 实现了具有竞争力的 FID 分数（分别为 5.65 和 5.18），并在诸如未发现和气胸等主要类别中保持了高分类一致性。

在合成数据训练后对真实测试图像进行评估（表 4），基于 U-Net 和文本编码器微调的模型几乎在所有疾病类别中都达到了最高精度（例如，心脏扩大为 77.9%，无发现为 94.8%），这突显了其强大的泛化能力。BitFit 在这项评估中也表现出色，表明尽管它可能在生成质量上表现不佳，但其生成的图像仍保留足够的语义结构以供下游分类使用。LoRA 和 DoRA 也在跨域转移方面展示了稳健性，特别是在高信号类别如气胸和未发现异常中，强调了它们作为高效且有效的微调替代方案的实用性。

PEFT 方法显示出计算优势，LoRA 仅需 159 万个可训练参数，而完整的 U-Net 训练方法则需要 8370 万到 9830 万个参数。尽管如此，训练时间仍然相当，PEFT 方法每个周期需要 61-70 秒，相比之下完整训练需要 77-177 秒。

4 结论

在这项工作中，我们系统评估了微调策略，以将文本到图像的基础模型 Stable Diffusion 适应于高分辨率的医学成像任务。我们将参数高效的方法（LoRA、DoRA、BitFit 和 DiffFit）与全 U-Net 训练进行了比较，使用了一个全面的评价框架来评估图像质量指标和下游任务性能。我们的结果显示，全 U-Net 训练在所有评价指标上都优于所有参数高效方法，确立了其作为高分辨率医学成像合成的最佳方法，前提是计算资源允许。虽然参数高效的方法成功解决了医学成像领域中的数据稀缺问题，但观察到的显著性能差距表明，从业者应优先考虑全 U-Net 训练以实现最大诊断准确性。开发的评估方法提供了一个稳健框架，确保技术改进转化为临床实用性。这些发现为在医学成像环境中部署基础模型提供了明确指导，并建立了未来参数高效方法的性能基准。这项工作使人们能够在资源受限的情况下明智地做出计算权衡决策，同时展示了全面网络优化对于关键医疗应用的持续优越性。

5 致谢

资助部分来自加拿大自然科学与工程研究理事会、加拿大高级研究所（CIFAR）人工智能教授项目、Mila - 魁北克人工智能研究院、Google 研究、Calcul Quebec 和加拿大数字研究联盟。

参考文献

[1] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023.

[2] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

[3] Ana Davila, Jacinto Colan, and Yasuhisa Hasegawa. Comparison of fine-tuning strategies for transfer learning in medical image classification. *Image and Vision Computing*, 146:105012, 2024.

[4] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

[5] Raman Dutt, Linus Ericsson, Pedro Sanchez, Sotirios A Tsaftaris, and Timothy Hospedales. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. *arXiv preprint arXiv:2305.08252*, 2023.

[6] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1 (2):3, 2022.

[9] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

[10] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[11] Amar Kumar, Anita Kriz, Mohammad Havaei, and Tal Arbel. Prism: High-resolution & precise counterfactual medical image generation using language-guided stable diffusion. *MIDL*, 2025.

[12] Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. Cosmicman: A text-to-image foundation model for humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6955–6965, 2024.

[13] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Cross-domain few-shot learning with task-specific adapters. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7161–7170, 2022.

[14] Gang Liu, Jinlong He, Pengfei Li, Genrong He, Zhaolin Chen, and Shenjun Zhong. Pefomed: Parameter efficient fine-tuning of multimodal large language models for medical imaging. *arXiv preprint arXiv:2401.02797*, 2024.

[15] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

[16] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, 2023.

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[18] Filippo Ruffini, Elena Mulero Ayllon, Linlin Shen, Paolo Soda, and Valerio Guarrasi. Benchmarking foundation models and parameter-efficient fine-tuning for prognosis prediction in medical imaging. *arXiv preprint arXiv:2506.18434*, 2025.

[19] Parham Saremi, Amar Kumar, Mohammed Mohammed, Zahra TehraniNasab, and Tal Arbel. Rl4med-ddpo: Reinforcement learning for controlled guidance towards diverse medical image generation using vision-language foundation models. *arXiv preprint arXiv:2503.15784*, 2025.

[20] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[21] Zahra TehraniNasab, Amar Kumar, and Tal Arbel. Language-guided trajectory traversal in disentangled stable diffusion latent space for factorized medical image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference Workshop Proceedings*, pages 4846–4851, 2025.

[22] Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4230–4239, 2023.

[23] Yeonguk Yu, Minhwan Ko, Sungho Shin, Kangmin Kim, and Kyoobin Lee. Curriculum fine-tuning of vision foundation model for medical image classification under label noise. *Advances in Neural Information Processing Systems*, 37:18205–18224, 2024.

[24] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.

[25] Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis*, 91:102996, 2024.