实用型即插即用扩散模型

摘要

基于扩散的生成模型如今在图像生成方面取得了显著的成果。其指导公式允许外部模型以即插即用（plug-and-play）的方式控制生成过程，从而在无需对扩散模型进行微调的情况下执行各种任务。然而，直接使用公开可用的现成模型进行指导往往会失败，因为这些模型在处理噪声输入时表现较差。现有的解决方案是使用带有噪声标注数据对指导模型进行微调，但是这种做法存在两个局限性：(1) 单一的指导模型难以应对不同程度噪声的输入；(2) 带标签的数据集难以手机，因而妨碍到原本扩散模型的拓展。

为了解决这些局限性而提出了一种新策略，利用多个专家模型，每个专家专门处理特定的噪声范围，并在其对应的时间步长指导扩散的反向过程。然而，针对于多个网络难以管理，以及难以收集标签数据的问题，提出了一种实用的指导框架，称为 Practical Plug-And-Play (PPAP)，该框架包括了参数高效微调和无数据知识迁移技术。该框架在 ImageNet 分类条件生成实验中进行了广泛验证，结果表明此方法能够在仅使用少量可训练参数且无需标签数据的情况下成功地指导扩散。

关键词：扩散模型；即插即用；指导模型

1 引言

近年来，基于扩散的生成模型在多个领域取得了巨大成功，包括图像生成、文本到语音转换以及文本生成等领域。特别是在图像生成方面，已有研究表明，扩散模型能够生成与 GAN 生成质量相当的高质量图像，同时不会遭遇模式崩溃或训练不稳定等问题。

除了这些优势之外，扩散模型的框架还允许外部模型指导，即通过外部指导模型引导扩散模型的生成过程，使其满足特定条件。这种指导方法不需要进一步微调扩散模型，因此具备低成本、可控的生成潜力，并能够以即插即用的方式实现。例如用分类器指导相关类别图像的生成，利用语义编辑器进行 text2image 的生成等等。

在这些研究的基础上，如果公开的现成模型能够用于指导扩散过程，那么我们就可以很容易地将一种扩散模型应用于各种生成任务。然而，现有方法面临着两个主要挑战：

1. 现成模型在处理噪声输入时表现不佳：扩散模型的逆过程涉及对噪声图像进行逐步去噪，而现成的模型在这种噪声输入下往往无法提供可靠的指导信息。

2. 单一指导模型的局限性：单一的指导模型难以适应不同程度的噪声输入，这是因为在扩散逆过程的不同时间步长中，噪声的复杂程度各不相同。

为了解决这些问题，提出了一种多专家策略，其中每个专家模型专门处理特定的噪声范围，并在对应的时间步长指导扩散逆过程。这种策略可以在较大噪声时指导生成粗略结构，在

较小噪声时进一步补充细节。然而，多专家策略面临管理多个网络和依赖大量带标签数据的挑战，因而难以在实际场景中应用。

针对于上述问题，又提出了 Practical Plug-And-Play（PPAP）框架，旨在实现实用的扩散模型指导。PPAP 包括以下两个核心组件：

1. 参数高效微调：通过参数共享和高效微调，防止多专家模型规模过大。

2. 无数据知识迁移：通过将现成模型在干净扩散生成数据上的知识转移到专家模型，避免了对标签数据的依赖。

使用该框架在 ImageNet 条件图像生成任务上进行了广泛实验，结果表明，PPAP 可以在仅需少量可训练参数且无需标签数据的情况下成功指导扩散模型。

2 相关工作

2.1 传统扩散模型

针对于扩散模型而言，从 2020 年发表在 Machine Learning 的 Denoising Diffusion Probability Model（DDPM）[3] 开始，扩散模型广泛的进入了大家的视野之中。在避免了 GAN [1] 模型的诸多问题的同时，DM 表现出了相当优越的成绩。

针对于 DDPM 生成速度较慢的问题，在后来产生了大量的方法对其做了改进，如 DDIM [2]、A-DDIM、PNDM [4] 和 DEIS 等，显著加速了生成过程。

在条件生成任务中，扩散模型主要使用分类器引导和无分类器引导两种方法。分类器引导利用外部分类器的梯度来引导扩散过程，而无分类器引导则通过插值预测有标签和无标签条件下的结果。然而，对于无分类器引导方法，需要在带标签数据上训练扩散模型，因为其依赖于标签的预测结果。本工作重点关注的是分类器引导，即冻结无条件扩散模型，通过外部模型指导生成过程，从而无需标签数据，能够实现即插即用的灵活生成。

2.2 基于即插即用

针对于即插即用生成而言，即插即用（Plug-and-Play）生成指的是在测试时利用可替换的条件网络引导生成模型，而无需对条件网络与生成模型进行联合训练。在图像生成和文本生成领域，已有大量研究通过对无条件模型施加约束，实现条件生成。这些方法使得单个无条件生成模型能够通过更改约束模型来执行不同的任务。

3 本文方法

3.1 本文方法概述

此利用多个专家模型，每个专家专门处理特定的噪声范围，并在其对应的时间步长指导扩散的反向过程。然而，针对于多个网络难以管理，以及难以收集标签数据的问题，提出了一种实用的指导框架，称为 Practical Plug-And-Play (PPAP)，该框架包括了参数高效微调和无数据知识迁移技术。该框架在 ImageNet 分类条件生成实验中进行了广泛验证，结果表明此方法能够在仅使用少量可训练参数且无需标签数据的情况下成功地指导扩散。

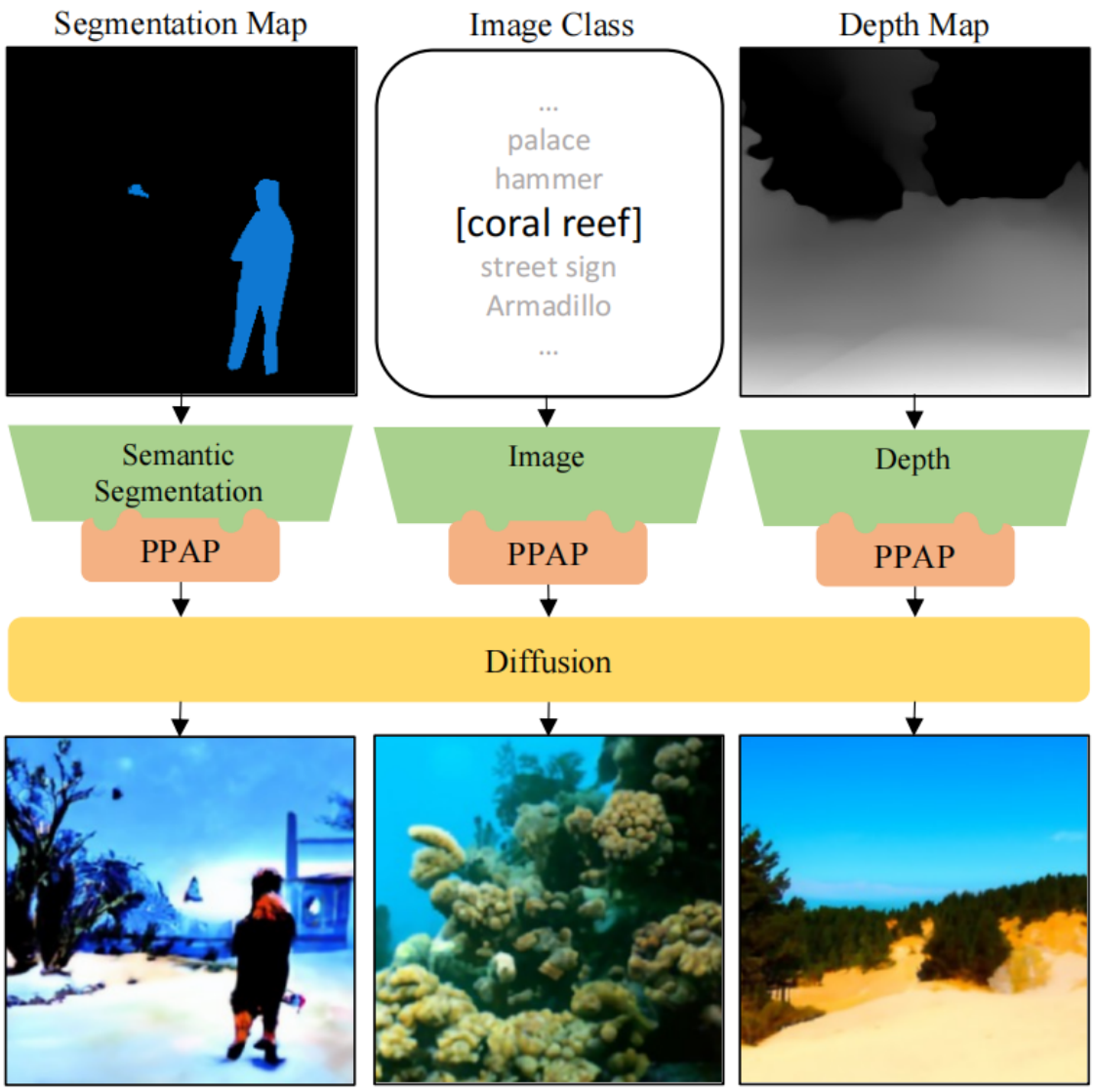

图 1. 方法示意图

3.2 对原本扩散模型引导方案的分析

在过往的引导方案中，现成的模型在引导扩散模型时往往表现出低置信度的预测，而针对整个噪声范围训练的单一模型也存在一定的局限性。对此可以设计一些实验来具体研究其表现情况。

在实验设置中，使用了一个扩散模型以及多个在噪声数据上微调过的分类器，并将分类器作为引导分类器。分类器使用了在 ImageNet 上预训练的 ResNet50，并在需要时进行微调。扩散模型基于 ImageNet256*256 数据集训练，最大时间步长 T=1000。为了使用带噪声的数据，使用了前向扩散过程，即给定输入图像 x0，通过以下公式生成带噪声数据 xt：

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon, \quad \epsilon \sim N(0, I)$$

图 2. x02xt 公式

首先使用了现成的 ResNet50 分类器对扩散模型进行引导，目标是生成特定类别的 Ima-

geNet 图像。然而结果显示，该模型未能提供有效的梯度，导致引导失败。这是因为现成模型在遇到噪声输入（即扩散过程中的潜在数据）时，输出的预测置信度较低，熵值较高。这一现象在分类器置信度的可视化结果中表现得尤为明显。

为了进一步理解单一噪声感知模型的失败原因，实验研究了在特定噪声水平上微调的分类器的行为。具体而言，分别在不同的噪声范围上（即 t [a,b] [0,T]）对 ResNet50 进行微调，得到五个专家分类器。每个分裂期专注于特定噪声水平，如第一个和第二个专注于低噪声下的数据，第四个和第五个专注于较高噪声的数据。得到的可视化结果如图所示：

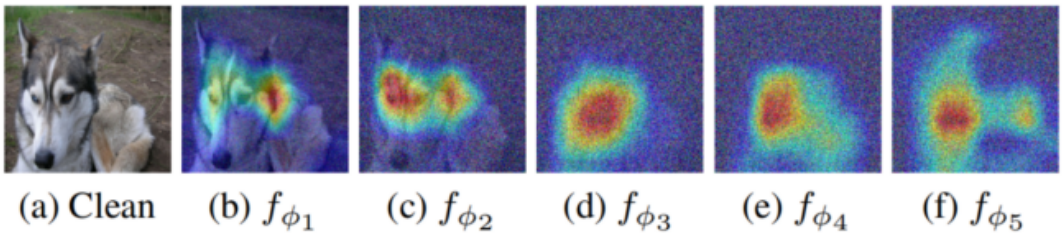

图 3. 分类器行为

根据可视化结果，可以观察到不同噪声水平下分类器的行为差异：

在较低噪声数据上训练的分类器（如第一个，第二个）能够捕捉到图像的细节特征，如物体的纹理和局部结构。在较高噪声数据上训练的分类器（如第四个，第五个）主要依赖于图像的整体结构，如物体的大致轮廓和形状。

这种行为差异在引导扩散过程时表现的尤为明显：在较高噪声水平下，分类器倾向于引导扩散模型生成大致的物体轮廓。在较低噪声水平下，分类器则帮助扩散模型补充细节，完善物体结构。可以在下图中有所体现：

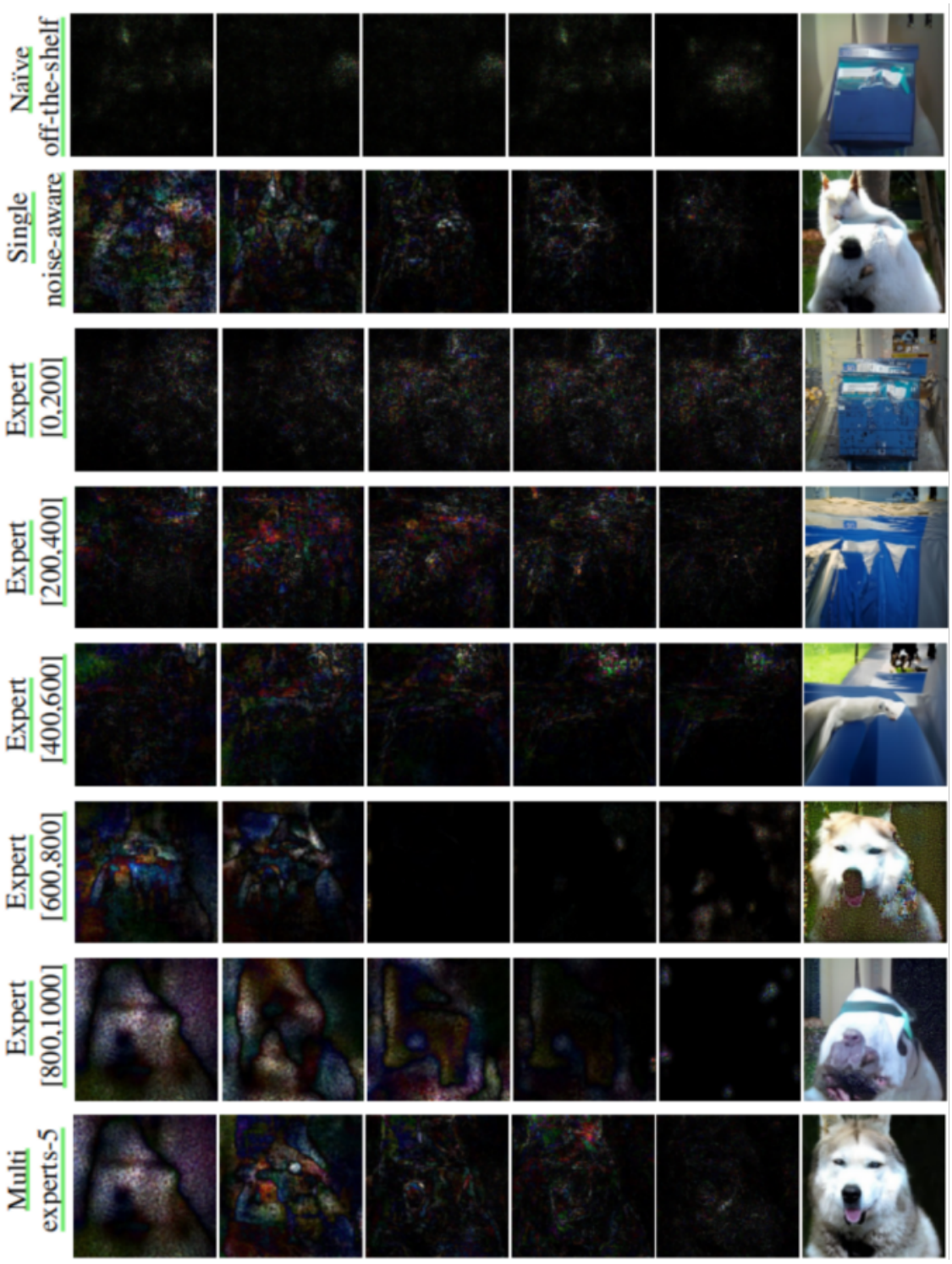

图 4. 不同分类器作用范围

在实验中，第五个分类器在高噪声水平下成功生成了一个类似于哈士奇的轮廓，但未能补充细节。而第一个分类器在低噪声水平下能够引导生成精细的毛发纹理，但由于缺少整体轮廓而未能生成完整的图像。这一观察结果与无条件扩散模型在去噪过程中的行为一致：在较大噪声下，扩散模型更关注整体结构的重建；在较小噪声下，扩散模型则更注重细节的完善。

因此，可以得出结论，分类器能够通过学习特定噪声水平的特征，从而在对应的噪声范围内有效地引导扩散模型。

3.3 多专家策略

通过前文的实验，分类器在不同噪声水平下表现出不同的行为，会直接影响扩散引导的效果。因此，为了解决单一指导模型在不同噪声水平下无法有效引导扩散模型的问题，提出了一种多专家策略，即每个专家模型专门针对一个特定的噪声范围进行优化，并在扩散逆过程的不同时间步长中进行引导。

具体来说，假设存在一个干净数据集 (x0,y) 和最大扩散时间步 T，将整个噪声区间划分为 N 个子区间，对于第 n 个子区间，对应的专家指导模型经过微调，专门学习在该噪声范围内的指导任务。噪声区间的划分和对应的指导模型分配规则如下所示：

$$t \in \left[\frac{(n-1)T}{N}, \frac{nT}{N} \right]$$

在扩散模型的反向去噪过程中，便可以修改指导公式：

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z - s\sigma_t \nabla_{x_t} L_{\text{guide}}(f_{\phi_n}(x_t), y)$$

其中，f 表示用于指导的专家模型，Lguide 表示指导损失函数，s 为指导强度。

4 复现细节

4.1 与已有开源代码对比

在本次复现过程中,参考了公开的扩散模型代码,代码出处:https://github.com/riiid/PPAP。该代码为扩散模型的基本框架和训练过程提供了基础支持。我的工作在此基础上进行了些许创新和改进，主要体现在以下几个方面：

1. **添加了 VAE 模块**：在原有的扩散模型结构中，加入了变分自编码器（VAE）模块。这一改动使得模型能够更有效地学习潜在空间，从而提高生成图像的质量和多样性。

2. 调整了噪声加入方式：通过优化噪声的加入方式，改进了模型对噪声的处理能力，从而增强了生成图像在噪声环境下的表现和稳定性。

3. 调整了采样过程：对采样过程进行了调整，提升了生成效率，并且优化了生成图像的质量，减少了训练过程中的不稳定因素。

4. 尝试修改了噪声加入强度：通过调整噪声加入的强度，确保噪声对生成过程的影响在合适的范围内，从而使得模型能够生成更清晰、细节更丰富的图像。

这些创新和改进体现了我在扩散模型研究中的技术贡献，确保了本工作在现有模型基础上具有些许差异与尝试性改进。

5 实验结果分析

5.1 实验设置

整体实验基于 ImageNet 上预训练的 256*256 的 ADM 模型，应用了两种不同的架构：1. 基于 CNN 的 ResNet50 分类器。2. 基于 transformer 的 DeiT-S 分类器。对于每一个架构而言，均设置了以下变量，来作为指导模型。

1. Naive off-the-shelf：即未经微调的 ImageNet 预训练模型。
2. Single noise aware：将指导模型在数据集上做微调，并用其指导整个生成过程。
3. Multi-experts-N：微调 N 个专家指导模型，且未应用参数微调策略等。
4. PPAP-N：微调 N 个专家指导模型，并且使用了 PPAP 框架。

5.2 实验结果

针对于无数据知识迁移策略，生成 500 000 张图片，并用其作为微调的数据集。整体的实验结果如下表所示：

Architecture	Sampler	Guidance	Trainable Parameters	Supervision	FID (↓)	IS (↑)	Precision (↑)	Recall
ResNet50	DDIM (25 Steps)	No	-	None	40.24	34.53	0.5437	0.6063
		Naïve off-the-shelf	-	None	38.74	33.95	0.5192	0.6152
		Gradients on $\hat{x}_0$	-	None	38.14	33.77	0.5277	0.6252
		Single noise aware	25.5M (100%)	ImageNet ($\approx$ 1.2M)	30.42	43.05	0.5509	0.6187
		Multi-experts-5	127.5M (500%)	ImageNet ($\approx$ 1.2M)	**19.98**	**74.78**	**0.6476**	0.5887
		PPAP-5	7.3M (28.6%)	Data-free ($\approx$ 0.5M)	29.65	44.23	0.5872	0.6012
		PPAP-10	14.6M (57.2%)	Data-free ($\approx$ 0.5M)	27.86	46.74	0.6079	0.5925
	DDPM (250 Steps)	No	-	None	28.97	40.34	0.6039	0.6445
		Naïve off-the-shelf	-	None	29.03	39.79	0.6042	0.6474
		Gradients on $\hat{x}_0$	-	None	28.81	39.80	0.6095	0.6475
		Single noise aware	25.5M (100%)	ImageNet ($\approx$ 1.2M)	38.15	31.29	0.5426	0.6321
		Multi-experts-5	127.5M (500%)	ImageNet ($\approx$ 1.2M)	**16.37**	**81.47**	**0.7216**	0.5805
		PPAP-5	7.3M (28.6%)	Data-free ($\approx$ 0.5M)	22.70	52.74	0.6338	0.6187
		PPAP-10	14.6M (57.2%)	Data-free ($\approx$ 0.5M)	21.00	57.38	0.6611	0.5996
DeiT-S	DDIM (25 Steps)	No	-	None	40.24	34.53	0.5437	0.6063
		Naïve off-the-shelf	-	None	37.51	33.74	0.5293	0.6186
		Gradients on $\hat{x}_0$	-	None	38.10	33.75	0.5288	0.6212
		Single noise aware	21.9M (100%)	ImageNet ($\approx$ 1.2M)	44.13	28.31	0.4708	0.6030
		Multi-experts-5	109.9M (500%)	ImageNet ($\approx$ 1.2M)	**17.06**	**80.85**	**0.7001**	0.5810
		PPAP-5	4.6M (21.3%)	Data-free ($\approx$ 0.5M)	25.98	48.80	0.6128	0.5984
		PPAP-10	9.3M (42.6%)	Data-free ($\approx$ 0.5M)	24.77	50.56	0.6220	0.5990
	DDPM (250 Steps)	No	-	None	28.97	40.34	0.6039	0.6445
		Naïve off-the-shelf	-	None	29.41	39.55	0.6032	0.6320
		Gradients on $\hat{x}_0$	-	None	30.26	37.75	0.6043	0.6407
		Single noise aware	21.9M (100%)	ImageNet ($\approx$ 1.2M)	36.01	31.90	0.5461	0.6479
		Multi-experts-5	109.9M (500%)	ImageNet ($\approx$ 1.2M)	**14.95**	**83.26**	**0.7472**	0.5686
		PPAP-5	4.6M (21.3%)	Data-free ($\approx$ 0.5M)	22.30	53.62	0.6368	0.6074
		PPAP-10	9.3M (42.6%)	Data-free ($\approx$ 0.5M)	20.07	60.62	0.6734	0.5963

图 5. 实验结果

实验结果表明，多专家策略对于整体的生成效果有显著的提升，但是出于网络结构过大，以及较难收集数据的原因，应用了 PPAP 框架再做实验。整体而言，在训练数据与训练参数数量大幅度减小的情况下，其训练效果并未有太多降低，表现优秀。

5.3 深度估计器生成测试

在本地实验的过程中，同样使用其框架做的些许的可视化尝试。其效果如下图：

图 6. 生成测试

最左边的图片为初始图片，随后应用了 MaDas 深度估计器对其做了深度预测，得到了对应的深度图像，再根据深度图像生成对应的图片，其效果整体而言较为优秀。随后又加入了语义编辑器，可以使用文本作为 prompt 调整生成图像的内容，整体的实用性较强。

5.4 尝试性改进

5.4.1 改进噪声的采样策略

对逆向生成图像的采样过程公式，做如下修正：

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)\right) + \sigma z + z_2\sigma$$

由于生成图像结果的随机性，其评估成绩会有波动，因此多次实验取平均值，其结果如下表所示：

	FID↓	IS↑	precesion	recall
one_z beta[0.0001-0.02]	51.62	30.87	0.49	0.57
	52.59	29.95	0.5	0.57
	52.6	29.94	0.48	0.56
	52.27	**30.25333333**	**0.49**	**0.566666667**
two_z beta[0.0001-0.02]	54.1	29.03	0.48	0.56
	53.85	30.85	0.49	0.56
	53.22	29.09	0.48	0.55
	53.72333333	**29.65666667**	**0.483333333**	**0.556666667**

图 7. 测试结果

5.4.2 修改采样过程中的 beta 参数

其结果如下图所示：

	FID↓	IS↑	precesion	recall
one_z beta[0.0001-0.02]	55.30	30.26	0.49	0.57
one_z beta[0.0002-0.02]	58.22	29.53	0.47	0.54
one_z beta[0.0001-0.04]	218.54	8.08	0.08	0.21

图 8. 测试结果

实验结果表明并未有较大幅度的提升。

6 总结与展望

本文拟研究复现了如何实现一个实用的即插即用扩散引导。通过观察分类器在不同噪声程度下的表现差异，本文提出了利用不同专家处理不同扩散步骤的多专家策略。然而对于文本的指导，项目仍然存在一些瑕疵，容易出现令人不太满意的情况。另外，在采样过程中的速度方面仍然需要改进，目前 DDIM 的采样虽有加速，但整体仍然能偏慢，可以在后续做改进。

参考文献

[1] Mehdi Mirza Bing Xu David Warde-Farley Sherjil Ozair Aaron Courville Yoshua Bengio Ian J. Goodfellow, Jean Pouget-Abadie. Generative Adversarial Networks. *Machine Learning*, 2014.

[2] Stefano Ermon Jiaming Song, Chenlin Meng. Denoising Diffusion Implicit Models. *Machine Learning*, 2020.

[3] Pieter Abbee Jonathan Ho, Ajay Jain. Denoising Diffusion Probabilistic Models. *Machine Learning*, 2020.

[4] Zhijie Lin Zhou Zhao Luping Liu, Yi Ren. Pseudo Numerical Methods for Diffusion Models on Manifolds. *Computer Vision and Pattern Recognition*, 2022.