Towards Practical Plug-and-Play Diffusion Models

Hyojun Go* Yunsung Lee* JinYoung Kim* Seunghyun Lee
Myeongho Jeong Hyun Seung Lee Seungtaek Choi†
Riiid AI Research
{hyojun.go, yunsung.lee, jinyoung.kim, seunghyun.lee
myeongho.jeong, hyunseung.lee, seungtaek.choi}@riiid.co

Abstract

*Diffusion-based generative models have achieved remarkable success in image generation. Their guidance formulation allows an external model to plug-and-play control the generation process for various tasks without fine-tuning the diffusion model. However, the direct use of publicly available off-the-shelf models for guidance fails due to their poor performance on noisy inputs. For that, the existing practice is to fine-tune the guidance models with labeled data corrupted with noises. In this paper, we argue that this practice has limitations in two aspects: (1) performing on inputs with extremely various noises is too hard for a single guidance model; (2) collecting labeled datasets hinders scaling up for various tasks. To tackle the limitations, we propose a novel strategy that leverages multiple experts where each expert is specialized in a particular noise range and guides the reverse process of the diffusion at its corresponding timesteps. However, as it is infeasible to manage multiple networks and utilize labeled data, we present a practical guidance framework termed **P**ractical **P**lug-**A**nd-**P**lay (**PPAP**), which leverages parameter-efficient fine-tuning and data-free knowledge transfer. We exhaustively conduct ImageNet class conditional generation experiments to show that our method can successfully guide diffusion with small trainable parameters and no labeled data. Finally, we show that image classifiers, depth estimators, and semantic segmentation models can guide publicly available GLIDE through our framework in a plug-and-play manner. Our code is available at https://github.com/riiid/PPAP.*

1. Introduction

Recently, diffusion-based generative models [49] have shown great success in various domains, including image generation [14, 44, 45], text-to-speech [21, 40], and text generation [32]. Specifically, for image generation, recent works have shown that diffusion models are capable of generating high-quality images comparable to those generated by GANs [8, 12], while not suffering from mode collapse or training instabilities [38].

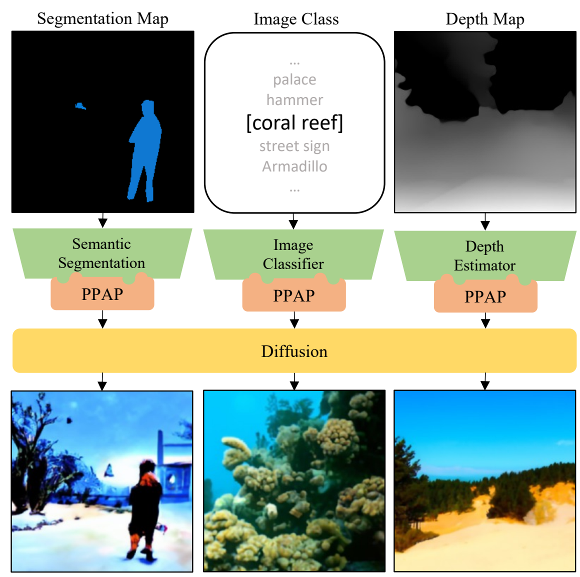


Figure 1. Overview of our framework. Practical Plug-And-Play (PPAP) enables the diffusion model to be guided by leveraging off-the-shelf models. Images shown below are generated by guiding the unconditional GLIDE [37] with DeepLabV3 [4], ResNet50 [15], and MiDaS [43] in a *plug-and-play* manner.

In addition to these advantages, their formulation allows the external model guidance [8, 49, 53], which guides the generation process of diffusion models towards the desired condition. Since guided diffusion leverages external guidance models and does not require further fine-tuning of the diffusion model, it holds the potential for cheap and controllable generation in a *plug-and-play* manner. For example, previous approaches use an image classifier for class-conditional image generation [8, 53], a fashion understanding model for fashion image editing [28], and a vision-

*Co-first autor.
†Corresponding author.

language model for text-based image generation [1, 37]. From these, if the publicly available off-the-shelf model can be used for guidance, one can easily apply one diffusion to various generation tasks.

For this purpose, an existing practice is to fine-tune the external off-the-shelf model on a noisy version of the training dataset [8, 12], to adapt the model on the noisy latent images encountered during the diffusion process. However, we argue that such a practice has two challenges for *plug-and-play* generation: (1) A single guidance model is insufficient to make predictions on inputs corrupted with varying degrees of noise, namely a too difficult task; and (2) It requires a labeled training dataset, which becomes a major hurdle whenever leveraging the off-the-shelf model.

In this paper, we first investigate the behaviors of classifiers by varying degrees of noise to understand the first challenge. On one hand, guidance models trained on corrupted images with heavy noise categorize images based on coarse structures. As a result, such a model would guide the diffusion model to generate essential skeletal features. Meanwhile, guidance models trained on cleaner images capture finer details in the images, guiding the diffusion model to work on finishing touches.

Based on these key observations, we propose a novel multi-experts strategy that uses multiple guidance models, each fine-tuned to specialize in a specific noise region. Despite the effectiveness of the multi-experts strategy, it should manage multiple networks and utilize the labeled data whenever applying new off-the-shelf models for various generation tasks.

For more practical *plug-and-play* guidance of the diffusion model with multi-experts strategy, we introduce the framework called **P**ractical **P**lug-**A**nd-**P**lay (PPAP). First, to prevent the size of guidance models from growing prohibitively large due to the multi-experts strategy, we leverage a parameter-efficient fine-tuning scheme that can adapt off-the-shelf models to noisy images while preserving the number of parameters. Second, we transfer the knowledge of the off-the-shelf model on clean diffusion-generated data to the expert guidance models, thereby circumventing the need for collecting labeled datasets.

Our empirical results validate that our method significantly improves performance on conditional image generation with off-the-shelf models with only small trainable parameters and no labeled data. We also showcase various applications with the publicly available diffusion model, GLIDE [37], by leveraging off-the-shelf image classifiers, depth estimators, and semantic segmentation models in a *plug-and-play* manner.

2. Related Work

Diffusion models Diffusion models [8, 18, 26, 38, 49] and score-based models [51, 53] are families of the generative model that generate samples from a given distribution by gradually removing noise. Unlike other likelihood-based methods such as VAEs [27] or flow-based models [9, 10], diffusion models have shown superior generation capabilities comparable to GANs [3, 12, 23]. Although diffusion models suffer from slow generation, previous works such as DDIM [50], A-DDIM [2], PNDM [33], and DEIS [55] have achieved significant acceleration in the generation process.

For conditional generation in diffusion models, classifier guidance [8, 53] and classifier-free guidance [19] are widely applied to various tasks [17,25,29,37,44]. Classifier guidance uses gradients of the external classifier, whereas classifier-free guidance interpolates between predictions from a diffusion model with and without labels. However, for classifier-free guidance, diffusion models should be learned as labeled data because it requires the prediction of labels. In this paper, we focus on the classifier guidance that freezes the unconditional diffusion model and guides it with the external model to conduct various conditional generations without labeled data in *plug-and-play* manner.

Plug-and-play generation Following [36], we use the term *plug-and-play* to refer to the capability of generating images at test time based on a condition given by a replaceable condition network without training it and generative model jointly. There have been various attempts for plug-and-play conditional generation in both image generation [11, 22, 24, 36, 52] and text generation [6, 34, 48], by binding constraints to the unconditional models, such as GAN [12], VAE [27]. These methods allow the single unconditional generative model to perform various tasks by changing the constraint model.

Most similar work to ours, Graikos *et al*. [13] attempted plug-and-play on diffusion models for various tasks by directly optimizing latent images with the off-the-shelf model. However, it fails to generate meaningful images in complex distribution as ImageNet. Contrary to this, our method successfully guidance in complex datasets by introducing small parameters into the off-the-shelf model and making it suitable for the latent.

3. Motivation

3.1. Preliminaries

Diffusion models Diffusion models [14, 44, 45, 49] are a class of generative models that sample data by gradually denoising a random noise. The diffusion model comprises two stages, namely, forward and backward processes. Forward diffusion process q gradually adds noise to the data $x_0 \sim q(x_0)$ with some variance schedule β_t, as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}). \quad (1)$$

We repeat the forward process until reaching the maximum timestep T. Given x_0, we can directly sample x_t as:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (2)$$

where $\alpha_t := \prod_{s=1}(1-\beta_s)$. Note that $\sqrt{\alpha_t}$ decreases as t grows, such that $\sqrt{\alpha_T} \approx 0$.

Reverse diffusion process starts with a random noise $x_T \sim \mathcal{N}(0, \mathbf{I})$, and produces gradually denoised samples $x_{T-1}, x_{T-2}, \ldots$, until reaching the final sample x_0. With noise predictor model $\epsilon_\theta(x_t, t)$, reverse process iteratively denoises x_t with $z \sim \mathcal{N}(0, \mathbf{I})$ as follows:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)) + \sigma_t z, \quad (3)$$

where σ_t^2 is the variance of reverse process.

Guided diffusion with external models Guided diffusion steers sample generation of the diffusion model by leveraging an external model [8, 49, 53]. To elaborate, suppose that we have some external guidance model f_ϕ that predicts certain traits of the input, e.g., a classifier that predicts the image class. At each timestep t of reverse diffusion, we use the guidance model to calculate the gradient on x_t in the direction that increases the probability of x_t having a certain desired trait y_{target}. We then update x_t to take a step in that direction, in addition to the usual denoising update. More formally, the reverse diffusion process (Eq. 3) is modified as follows:

$$\begin{aligned} x_{t-1} = &\frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)) \\ &+ \sigma_t z - s\sigma_t \nabla_{x_t} \mathcal{L}_{guide}(f_\phi(x_t), y_{target}), \end{aligned} \quad (4)$$

where $\mathcal{L}_{guide}$ and s denote guidance loss and strength, respectively. This formulation enables external models to guide the diffusion for various tasks of interest. For instance, for class-conditional image generation, f_ϕ is an image classifier that outputs $P_\phi(y_{target}|x_t)$, and $\mathcal{L}_{guide}$ is given by $-\log(p_\phi(y_{target}|x_t))$.

3.2. Observation

This section asks how naïve diffusion guidance schemes fail. Specifically, we show that when used for guided diffusion, off-the-shelf models fail due to low-confidence prediction, while models trained on data corrupted with a vast range of noise fail. Then, we report our major observation that classifiers trained on input corrupted with different noise levels exhibit different behaviors. We show that this directly affects diffusion guidance, i.e., having an expert guidance model specialized in different noise regions is crucial for successful guidance.

Setup. Our observational study involves a diffusion model and various classifiers fine-tuned on noise-corrupted data. For guidance classifiers, we used ResNet50 [15] pre-trained on ImageNet and fine-tuned them when necessary. We use the diffusion model trained on ImageNet 256 × 256 with max timesteps $T = 1000$ in [8]. To generate noisy versions of the data, we perform a forward diffusion process, i.e., given an input x_0, we obtain x_t (Eq. 2) for $t = 1, \ldots, T$. We use the DDIM sampler [50] with 25 steps, using the classifier under consideration to guide the diffusion model. More details are provided in Appendix. B.

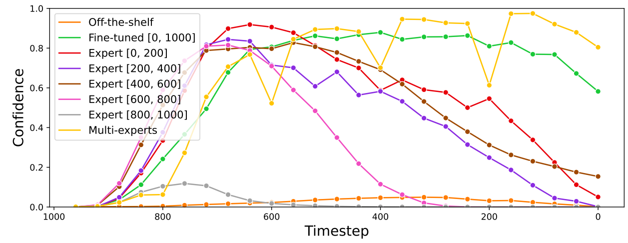
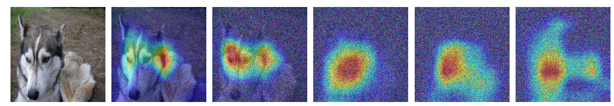


Figure 2. Classifier confidence during the reverse process. The off-the-shelf model does not increase confidence, showing that it cannot guide the diffusion to its confident region.



(a) Clean (b) f_{ϕ_1} (c) f_{ϕ_2} (d) f_{ϕ_3} (e) f_{ϕ_4} (f) f_{ϕ_5}

Figure 3. Grad-CAM visualization of each expert on corrupted images by forward process. Experts trained on larger and smaller timestep tend to focus on coarse and fine features, respectively.

Naïve external model guidance is not enough. Here, we investigate failure cases of naïve diffusion guidance. First, we try diffusion guidance with an off-the-shelf ResNet50 classifier toward an ImageNet class label. As we can see in the first row of Fig. 4, the model fails to provide a meaningful gradient for diffusion guidance. This is because an off-the-shelf model outputs low-confidence, high-entropy prediction on the out-of-distribution, noisy latent encountered throughout the reverse process as in Fig. 2.

We also experimented with a ResNet50 classifier fine-tuned on data encountered throughout the forward diffusion process, i.e., x_t (Eq. 1) for all $t \in [1, \ldots, 1000]$, where x_0 corresponds to a clean image. This works better than using a naïve off-the-shelf model, as observed in the improved FID (38.74 → 30.42) and IS (33.95 → 43.05) scores (see Section 5 for more details). However, again, as seen in Fig 2, classifier confidence drops for cleaner images ($t \approx 200$), leading to failure cases as in the second row of Fig. 4.

Behavior of classifier according to learned noise To understand the failure of the single noise-aware model, we investigate the behavior of classifiers fine-tuned on specific noise level, i.e., x_t for $t \in [a, b] \subset [0, T]$ for some suitable $a > 0$ and $b < T$.

Specifically, we fine-tuned five ResNet50 classifiers f_{ϕ_i}, $i \in \{1, \ldots, 5\}$, where f_{ϕ_i} is trained on noisy inputs x_t, $t \in \{(i-1) \cdot 200, \ldots, i \cdot 200\}$. We first observe that each f_{ϕ_i} behave differently via Grad-CAM [47]. For example, as shown in Fig. 3, f_{ϕ_1} and f_{ϕ_2} trained on cleaner images predict 'husky' based on distinctive canine features such as its eyes or coat pattern. Meanwhile, f_{ϕ_4} and f_{ϕ_5} trained on noisy images make predictions based on overall shape

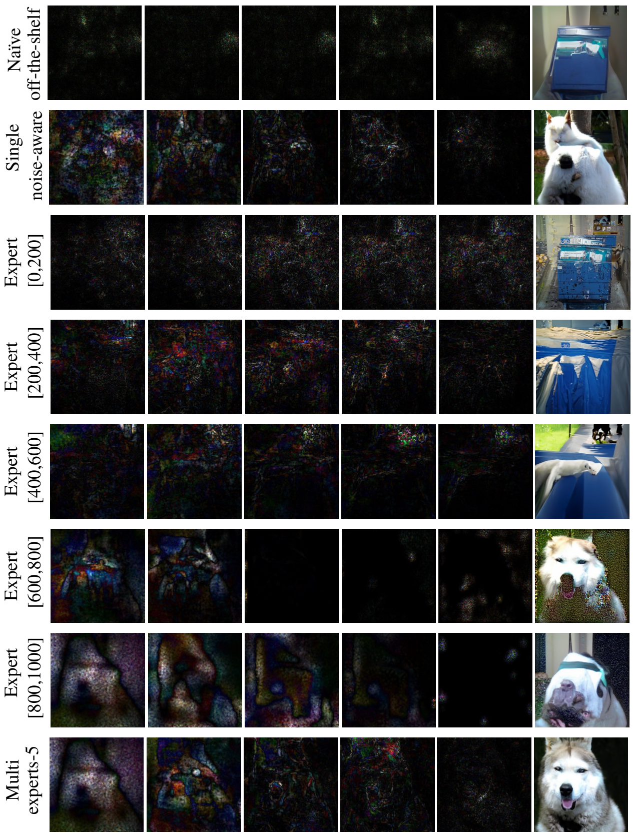


Figure 4. Gradient maps for x_t on $t \in [920, 720, 520, 320, 120]$ (left 5) and generated images (rightmost) when the reverse process is guided to husky from the same initial noise. Classifier trained on smaller noise and larger noise tends to modify finer details and coarser structure, respectively.

(albeit imperceptible to human eyes).

Such behavior difference manifests when we guide diffusion using different f_{ϕ_i}. For example, f_{ϕ_5} trained on noisy images initially generates a husky-like shape but fails to fill in finer details. On the other hand, f_{ϕ_1} trained on cleaner images seems to be focusing on generating specific details such as hairy texture but fails to generate a husky-like image due to lack of an overall structure.

These classifiers' behaviors coincide with the previous perspective; the unconditional diffusion focuses on the overall structure and finer details in larger and smaller noise, respectively [5]. Considering this, we hypothesize that the classifier can guide diffusion at the specific noise level by learning that noise level.

3.3. Multi-Experts Strategy

From the above observation, we propose a multi-experts strategy that each expert is fine-tuned to specialize in a specific noise range. Suppose we are given clean dataset $\{(x_0, y)\}$ and maximum diffusion timestep T. We train N

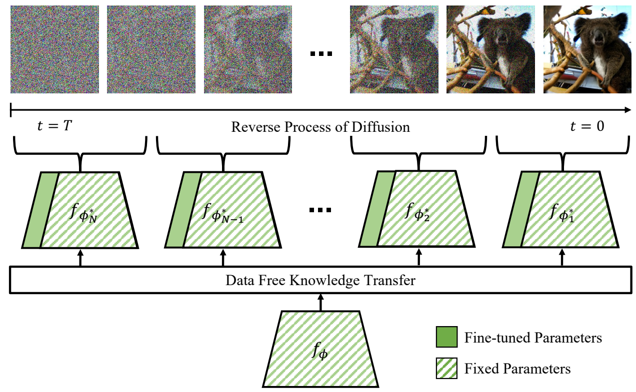


Figure 5. Overview of our method. We use parameter-efficient multi-experts that each expert is specialized in a specific noise range. We transfer the knowledge of the off-the-shelf model to each expert, thereby bypassing the need for labeled data. During the reverse process, we only need to switch the added training parameters accordingly depending on the noise region.

expert guidance model, where the n-th expert f_{ϕ_n} is trained to predict the ground-truth label y given noisy data x_t, $t \in \{\frac{n-1}{N}T, \ldots, \frac{n}{N}T\}$. Then, during the reverse diffusion process, we assign an appropriate guidance model depending on the timestep, i.e., n for which $t \in \{\frac{n-1}{N}T+1, \ldots, \frac{n}{N}T\}$. More formally, model guidance (Eq. 4) can be rewritten as:

$$\begin{aligned} x_{t-1} =& \frac{1}{\sqrt{1-\beta_t}}\big(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)\big) \\ &+ \sigma_t z - s\sigma_t \nabla_{x_t} \mathcal{L}_{guide}\big(f_{\phi_n}(x_t), y\big), \end{aligned} \tag{5}$$

where n is such that $t \in \{\frac{n-1}{N}T + 1, \ldots, \frac{n}{N}T\}$. Note that this strategy does not incur additional model inference time costs, since only one external guidance model is used depending on the reverse process timestep t.

In our observational study in Section 3.2, multi-experts guide coarse structure in larger timestep and fine-details in smaller time step, resulting in successful generation for the husky image as shown in Fig. 4.

4. Practical Plug-and-Play Diffusion

Whenever applying a new off-the-shelf model, the multi-experts strategy must utilize multiple networks and collect the labeled dataset. To deal with this impracticality, we propose a plug-and-play diffusion guidance framework: **P**ractical **P**lug-**A**nd-**P**lay (PPAP), which takes a multi-experts strategy with the following two components as shown in Fig. 5: **(1)** We introduce a parameter-efficient fine-tuning scheme based on parameter sharing in order to prevent the size of guidance models from growing prohibitively large; **(2)** We propose to use a knowledge transfer scheme that transfers an off-the-shelf model's knowledge on clean diffusion-generated data to expert guidance models, thereby bypassing the need for a labeled dataset.

4.1. Parameter Efficient Multi-Experts Strategy

One limitation of the proposed multi-experts strategy is that, as the number of guidance models increases N-fold, the number of parameters to fine-tune increases N-fold. To tackle this issue, we use a parameter-efficient strategy that only fine-tunes a small number of parameters while reusing most of the frozen off-the-shelf model. Specifically, we fine-tune bias and batch normalization, and apply LORA [20] to certain weight matrices of the off-the-shelf model. Since this method does not change architecture such as extending model depth, we do not introduce additional inference time cost. We denote n-th expert as $f_{\phi_n^*}$ to distinguish it from off-the-shelf model f_ϕ.

During the reverse process of the diffusion model, we only need to switch the added training parameters accordingly depending on the noise region, while reusing the off-the-shelf backbone model. More architectural details are provided in Appendix. C.

4.2. Data Free Knowledge Transfer

So far, we have assumed that we can access the dataset $\{(x_0, y)\}$ that was used to train the guidance model. For a practical *plug-and-play* generation, applying guidance with an off-the-shelf model should be possible without obtaining a labeled dataset suitable for each task.

We here propose to generate the clean dataset $\{\tilde{x}_0\}$ using the diffusion model, then use it to train expert guidance models. Our underlying assumption is that, by mimicking the prediction of an off-the-shelf model on a clean image, the expert can operate in the noise domain to some extent. Namely, we treat the off-the-shelf model f_ϕ as a teacher, and use its prediction on clean data to serve as labels when training expert guidance models f_{ϕ_n}. Formally, we formulate the knowledge transfer loss as:

$$\mathcal{L}_{KT} = \mathbb{E}_{t\sim \text{unif}\{\frac{n-1}{N}T,\dots,\frac{n}{N}T\}}\left[\mathcal{L}\big(\mathtt{sg}(f_\phi(\tilde{x}_0)), f_{\phi_n^*}(\tilde{x}_t)\big)\right], \tag{6}$$

where $\mathtt{sg}(\cdot)$ is the stop-gradient operator and $\mathcal{L}$ is a task-specific loss function. With this formulation, we can easily adapt our method to various tasks of interest, including image classification, monocular depth estimation, and semantic segmentation, by just using different loss functions. Due to space limitations, here we describe how we plug and play an image classifier only. Further details for other tasks can be found in Appendix. D.

Image classification An image classifier takes an image as input and outputs a logit vector of the form $f(x) \in \mathbb{R}^C$, where C is the number of image classes. We formulate knowledge transfer loss $\mathcal{L}_{clsf}$ for classifiers as:

$$\mathcal{L}_{clsf} = D_{KL}\big(\mathtt{sg}\big(\mathtt{s}(f_\phi(\tilde{x}_0)/\tau)\big), \mathtt{s}(f_{\phi_n^*}(\tilde{x}_t))\big), \tag{7}$$

where $\mathtt{s}$ is the softmax operator, τ the temperature hyperparmeter, and $D_{KL}(\cdot)$ the Kullback-Leibler divergence.

5. Experiments

In this section, we validate the effectiveness of our framework PPAP by showing that it can guide unconditional diffusion without collecting labeled data. Specifically, we first conducted various experiments in image classifier guidance to unconditional diffusion model trained on ImageNet dataset [7]. Then, we present the applicability of the image classifier, depth estimation model, and semantic segmentation model to the unconditional version of GLIDE [37] trained on a large dataset containing various domains.

5.1. ImageNet Classifier Guidance

We conduct experiments on ImageNet class conditional generation to validate the effectiveness of our framework.

Experimental Setup Based on ImageNet pre-trained unconditional ADM [8] with 256×256 size, we used two mainstream architectures: (1) CNN-based classifier ResNet50 [15] and 2) transformer-based classifier DeiT-S [54]. For each architecture, we used the following variants to serve the guidance model:

- **Naïve off-the-shelf**: ImageNet pre-trained model is used without further training on noise.
- **Single noise aware**: The model is fine-tuned on corrupted data in the whole noise range.
- **Multi-experts-N**: We fine-tune N expert guidance models in a supervised manner without applying parameter-efficient tunning.
- **PPAP-N**: N experts are parameter-efficiently knowledge transferred with generated images.

For data-free knowledge transfer (Section 4.2), we generate 500k images, which is much less than the ImageNet dataset. We use DDPM sampler [18] with 250 steps and DDIM sampler [50] with 25 steps as in [8]. We set the guidance scale s as 7.5 since it achieves good results for most variants.

We also compare other methods which can be applied to diffusion guidance with the off-the-shelf model. We use two baselines: (1) *plug-and-play priors* [13]: starting from the random noised image, they first optimize it to close good mode and desired condition. (2) *gradients on* $\hat{x}_0$: as in [1], we estimate clean images $\hat{x}_0$ from noisy images x_t as $\hat{x}_0 = \frac{x_t}{\sqrt{\alpha_t}} - \frac{\sqrt{1-\alpha_t}\epsilon_\theta(x_t,t)}{\alpha_t}$. We calculate the gradients on $\hat{x}_0$ using it for guidance. In summary, we observe that they fail to guide the diffusion. We analyze detailed limitations of their method in Appendix A.

As in [8], if the model well guides the generation process of diffusion to ImageNet class mode, the fidelity of generated images is improved by sacrificing diversity. Therefore, FID [16] becomes lower and Inception Score (IS) [46] becomes higher than its unconditional generation. Accord-

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 (≈ 1.2M)	30.42	43.05	0.5509	0.6187
		Multi-experts-5	127.5M (500%)	ImageNet (≈ 1.2M)	**19.98**	**74.78**	**0.6476**	0.5887
		PPAP-5	7.3M (28.6%)	Data-free (≈ 0.5M)	29.65	44.23	0.5872	0.6012
		PPAP-10	14.6M (57.2%)	Data-free (≈ 0.5M)	<u>27.86</u>	<u>46.74</u>	<u>0.6079</u>	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 (≈ 1.2M)	38.15	31.29	0.5426	0.6321
		Multi-experts-5	127.5M (500%)	ImageNet (≈ 1.2M)	**16.37**	**81.47**	**0.7216**	0.5805
		PPAP-5	7.3M (28.6%)	Data-free (≈ 0.5M)	22.70	52.74	0.6338	0.6187
		PPAP-10	14.6M (57.2%)	Data-free (≈ 0.5M)	<u>21.00</u>	<u>57.38</u>	<u>0.6611</u>	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 (≈ 1.2M)	44.13	28.31	0.4708	0.6030
		Multi-experts-5	109.9M (500%)	ImageNet (≈ 1.2M)	**17.06**	**80.85**	**0.7001**	0.5810
		PPAP-5	4.6M (21.3%)	Data-free (≈ 0.5M)	25.98	48.80	0.6128	0.5984
		PPAP-10	9.3M (42.6%)	Data-free (≈ 0.5M)	<u>24.77</u>	<u>50.56</u>	<u>0.6220</u>	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 (≈ 1.2M)	36.01	31.90	0.5461	0.6479
		Multi-experts-5	109.9M (500%)	ImageNet (≈ 1.2M)	**14.95**	**83.26**	**0.7472**	0.5686
		PPAP-5	4.6M (21.3%)	Data-free (≈ 0.5M)	22.30	53.62	0.6368	0.6074
		PPAP-10	9.3M (42.6%)	Data-free (≈ 0.5M)	<u>20.07</u>	<u>60.62</u>	<u>0.6734</u>	0.5963

Table 1. Overall Results on ResNet50 and DeiT-S. Note that "No" Guidance shows the same performance as it is unconditional generation. **Bold** denotes the best performance and <u>underline</u> indicates the second best performance. Multi-experts-5 and PPAP-10, which showed the best and second-best results in all cases, are both our proposed models.

ingly, precision [30] increases and recall [30] decreases. We calculate these metrics with the same reference images in [8] for quantitative comparison. For further details on experiments and implementation, refer to Appendix. E.1.

Main results We first compare the variants of ResNet50 and DeiT-S quantitatively (Table 1). Overall, we empirically confirmed the following observations to our advantage: 1) the baselines fail to guide the diffusion process, 2) the multi-experts strategy can boost performance, and 3) our practice is effective in more realistic settings.

Specifically, 1) Naive off-the-shelf and Gradients on $\hat{x}_0$ do not significantly improve any of the metrics from the unconditional generation, even worsening with the DDPM sampler. 2) Multi-experts-5 achieves the best performance of 14.95 FID, and 83.26 IS scores with DeiT-S, which significantly outperforms the baselines. Moreover, 3) PPAP-5 and PPAP-10 show superior performance (20.07 FID and 60.62 IS with DeiT-S) than the Single noise-aware of using a single expert. These results indicate that, with our framework, off-the-shelf classifiers can effectively guide the unconditional diffusion model without ImageNet supervision.

An interesting point is, PPAP outperforms the model fine-tuned with ImageNet supervision in guidance. Furthermore, it is noteworthy that our models use only small trainable parameters (21% $\sim$ 57%) and fewer unlabeled datasets (500k $<$ 1.2M). This suggests PPAP can successfully guide the diffusion models even in real-world scenarios. We will conduct an ablation study over the varying training data size to further discuss this point.

For further analysis, we show qualitative examples in Fig. 6 and confirm that the overall results are consistent with the above observations. Specifically, we observe that directly using the off-the-shelf models does not effectively guide the diffusion model, generating irrelevant images to the given classes (Albatross and Mongoose), but leveraging multi-experts shows more powerful guidance capabilities.

We also observe that DeiT-S tends to produce better guidance from Table 1. Considering that transformer-based architectures tend to capture low-frequency components and shape-based feature [31,35,39] than CNN-based architectures, we conjecture that these properties produce better guidance by focusing on shape.

Ablation study: Effect of Multi-Experts To understand the effect of using multi-experts, we compare several vari-

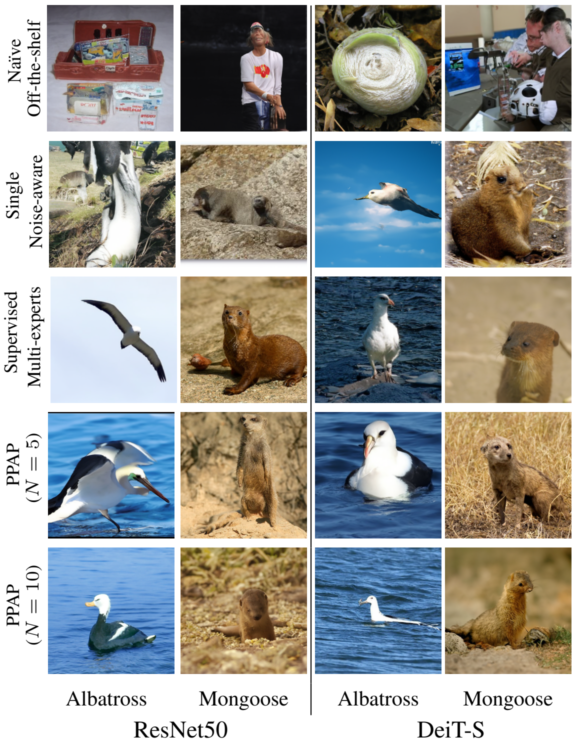
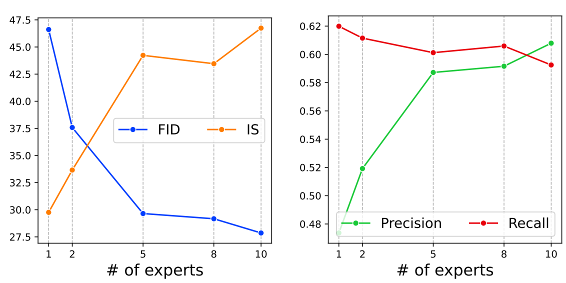


Figure 6. Qualitative results of on ImageNet class conditional generation with DDPM 250 steps. Guidance using an off-the-shelf model produces irrelevant images to given classes. On contrary, our multi-experts and PPAP generate well-guided images. More qualitative results are shown in Appendix. G.

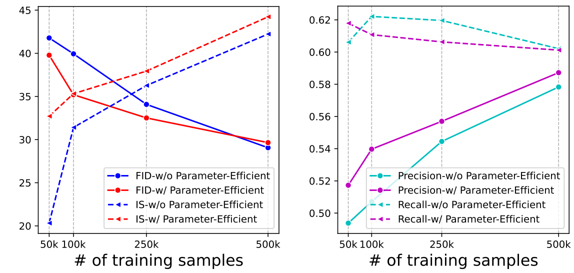


(a) FID and Inception Score (b) Precision and Recall

Figure 7. Ablation study for the number of experts. We plot quantitative results from the DDIM sampler with 25 steps by varying the number of experts. The results show leveraging more experts boosts the performance of guidance.

ants of PPAP with a varying number of parameter-efficient experts [1, 2, 5, 8, 10]. As shown in Fig. 7, leveraging more experts effectively improves diffusion guidance, and using 10 experts achieves the best performance. These results support that using multi-expert to make the noise interval more subdivided helps the guidance.

Ablation study: Effect of Parameter-Efficient Tunning and data efficiency Here we analyze the effectiveness of the parameter-efficient fine-tuning strategy, in comparison



(a) FID and Inception Score (b) Precision and Recall

Figure 8. Ablation study for parameter-efficient tunning and data efficiency. We plot quantitative results of full-finetuning and parameter-efficient tunning for five experts on the varying size of generated datasets by sampling with DDIM 25 steps. Guidance with parameter-efficient experts outperforms fully fine-tuned experts in low-data regimes. Increasing the generated data for knowledge transfer improves performance.

Figure 9. Generated images by guiding GLIDE with ResNet50. Our framework PPAP-5 succeeds in the guidance of diffusion, but our naïve off-the-shelf model fails. More qualitative results are presented in Appendix. H.1.

with full fine-tuning, on the varying sizes of the training dataset. The results are shown in Fig. 8. We observe that using the more generated data is an effective way of improving performance. Also, the parameter-efficient tunning strategy achieves comparable results to full fine-tuning with 500k generated data, showing that fine-tuning for small parameters is enough. An important finding is that the parameter-efficient experts even outperform the full fine-tuning variants in low data regimes (50k $\sim$). We posit that fully fine-tuning models with only a small dataset would ruin the representation power of the off-the-shelf models. We present more ablation studies to deeply understand the effectiveness of each component in our method in Appendix F.

5.2. Guiding GLIDE for Various Downstream Tasks

Here, we will show several practical applications by guiding GLIDE [37], which is trained on a largescale unreleased CLIP [41]-filtered dataset [37]. With our framework, we can apply an ImageNet [7] pretrained classifier [15], zero-shot depth estimator [43], and pretrained semantic segmentation model [4] as the guidance model.

Experimental Setup GLIDE [37] generates images from

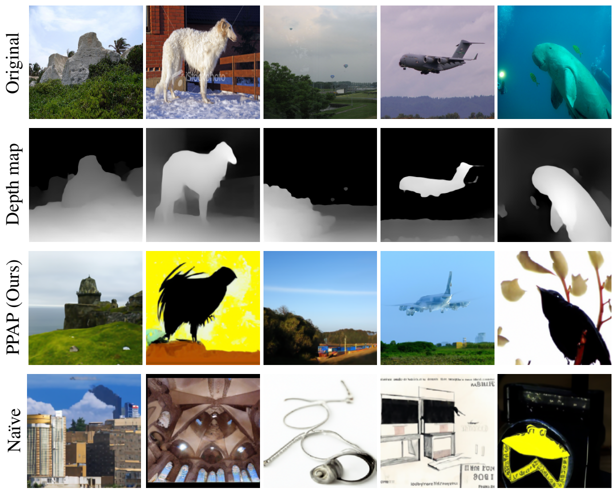


Figure 10. Generated images by guiding GLIDE with the MidaS depth estimator. Our framework PPAP succeeds in the guidance of diffusion, but our off-the-shelf model fails. More qualitative results are presented in Appendix. H.2.

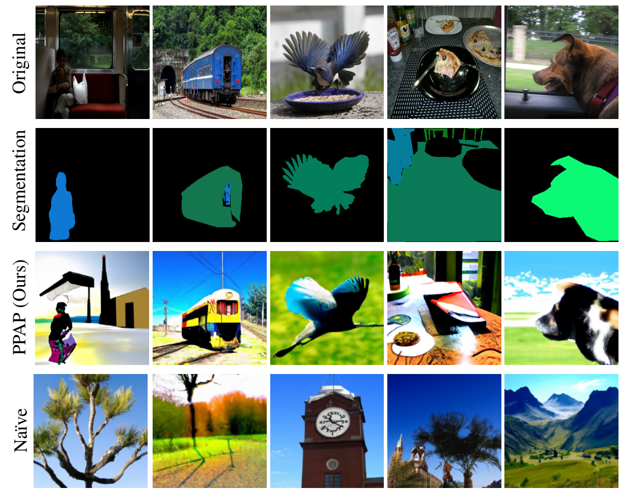


Figure 11. Generated images by guiding GLIDE with DeepLabv3 semantic segmentation. Our framework PPAP succeeds in the guidance of diffusion, but the naive off-the-shelf guidance fails. More qualitative results are shown in Appendix. H.3.

given text inputs. GLIDE consists of two diffusion models: 1) Generator: generates 64×64 images from given text inputs, and 2) Upsampler: upsamples generated 64×64 images to 256×256 images. To make GLIDE unconditional, we give the empty token as input of generator diffusion. Since the upsampler of GLIDE just changes image resolution, we aim to guide generator diffusion. All guidance models used in GLIDE experiments exploit 5 experts. For data-free knowledge transfer in our framework, we generate 500k unconditional 64×64 images from generator diffusion with the 25-step DDIM sampler. For generating guidance images, we use a DDPM sampler with 250 steps for generator diffusion and a fast27 sampler [37] for upsampler diffusion. We refer to experimental and implementation details for guiding GLIDE in Appendix. E.2.

Guiding GLIDE with ImageNet Classifier We used ImageNet [7] pre-trained ResNet50 [15] for guiding GLIDE to conduct class-conditional image generation. Figure 9 shows generated images by naive off-the-shelf guidance and PPAP ($N = 5$) guidance. The results are consistent with the results in Fig. 6 as the off-the-shelf model does not generate guided images and our method can guide GLIDE to generate well-guided images. Notably, our method can semantically guide the GLIDE with varying styles of images, such as cartoon-style images (4th image by ours), which is interesting because ResNet50 has never seen cartoon-style images in the ImageNet dataset. It shows that PPAP can obtain both the generation ability of GLIDE in various domains and the semantic understanding ability of the classifier.

Guiding GLIDE with Depth Estimator MiDaS [42,43] is a monocular depth estimation model designed for zero-shot cross-dataset transfer. We leverage their zero-shot superiority for guiding GLIDE to depth-map-to-image generation. As it is hard to provide an arbitrary depth map, we first estimate depth maps from images in the ImageNet dataset. Then, we feed the depth maps as desirable inputs for guiding the diffusion. As shown in Fig. 10, guidance with the naive off-the-shelf model generates images unrelated to the desired depth maps. On the contrary, with our proposed framework, we observe that the generated images are well-aligned with the edges in the depth map.

Guiding GLIDE with Semantic Segmentation Our experiments mentioned above have shown that the PPAP framework is capable of both semantic-level guidance from the ImageNet classifier and pixel-level guidance from the depth estimator. Based on the results, we validate whether our PPAP could apply both capabilities together and join into a semantic segmentation task guidance. Specifically, we used DeepLabv3 [4] for guidance. The results in Fig. 11 indicate that generated images from our PPAP framework tend to be aligned with the segmentation map, while generated images from naive off-the-shelf guidance fail.

6. Conclusion

In this paper, we studied how we can achieve a *practical* plug-and-play diffusion guidance. From an observation that a classifier would behave differently in varying degrees of noise, we propose the multi-expert strategy of leveraging different experts for different diffusion steps. Our experiments validate that our proposed framework makes it possible to easily utilize publicly available off-the-shelf models for guiding the diffusion process without requiring further training datasets. We deal with limitations and future work in Appendix I.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2, 5

[2] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022. 2

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 7, 8

[5] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 4

[6] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2019. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 7, 8

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 2, 3, 5, 6

[9] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 2

[10] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 2

[11] Jesse Engel, Matthew Hoffman, and Adam Roberts. Latent constraints: Learning to generate conditionally from unconditional generative models. In *International Conference on Learning Representations*, 2018. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1, 2

[13] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *arXiv preprint arXiv:2206.09012*, 2022. 2, 5

[14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 1, 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3, 5, 7, 8

[16] 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. 5

[17] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 5

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[21] Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-tts: A denoising diffusion model for text-to-speech. *arXiv preprint arXiv:2104.01409*, 2021. 1

[22] Zahra Kadkhodaie and Eero Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. *Advances in Neural Information Processing Systems*, 34:13242–13254, 2021. 2

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[24] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021. 2

[25] Heeseung Kim, Sungwon Kim, and Sungroh Yoon. Guided-tts: A diffusion model for text-to-speech via classifier guidance. In *International Conference on Machine Learning*, pages 11119–11133. PMLR, 2022. 2

[26] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[28] Chaerin Kong, DongHyeon Jeon, Ohjoon Kwon, and Nojun Kwak. Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation. *arXiv preprint arXiv:2210.05872*, 2022. 1

[29] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022. 2

[30] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[31] Yunsung Lee, Gyuseong Lee, Kwangrok Ryoo, Hyojun Go, Jihye Park, and Seungryong Kim. Towards flexible inductive bias via progressive reparameterization scheduling. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 706–720. Springer, 2023. 6

[32] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022. 1

[33] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 2

[34] Florian Mai, Nikolaos Pappas, Ivan Montero, Noah A Smith, and James Henderson. Plug and play autoencoders for conditional text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6076–6092, 2020. 2

[35] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021. 6

[36] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4467–4477, 2017. 2

[37] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2, 5, 7, 8

[38] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1, 2

[39] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 6

[40] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021. 1

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 7

[42] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 8

[43] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 7, 8

[44] 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*, pages 10684–10695, 2022. 1, 2

[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 1, 2

[46] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 5

[47] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[48] Tianxiao Shen, Jonas Mueller, Regina Barzilay, and Tommi Jaakkola. Educating text autoencoders: Latent representation guidance via denoising. In *International conference on machine learning*, pages 8719–8729. PMLR, 2020. 2

[49] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1, 2, 3

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3, 5

[51] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[52] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2021. 2

[53] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 2, 3

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 5

[55] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022. 2