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 Practical Plug-And-Play (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].

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-
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 Practical Plug-And-Play (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 [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 generating 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 guides 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 [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 \( \beta_t \), as follows:

\[
q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I),
\]

(1)

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

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

(2)
where $\alpha_t := \prod_{s=1}^{t}(1 - \beta_s)$. 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 $t \sim \mathcal{N}(0, \mathbf{I})$ 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 $\sigma_t^2$ 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 $f_{\theta}$ 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:

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

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, $f_{\theta}$ is an image classifier that outputs $P_{\theta}(y_{target}|x_t)$, and $\mathcal{L}_{guide}$ is given by $-\log(p_{\theta}(y_{target}|x_t))$.

### 3.2. Observation

This section asks how na"{i}ve 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 noisy-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.

**Na"{i}ve external model guidance is not enough.** Here, we investigate failure cases of naive 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 naive 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 $f_{\theta_i}$, $i \in \{1, \ldots, 5\}$, where $f_{\theta_i}$ is trained on noisy inputs $x_t$, $t \in \{(i - 1) \cdot 200, \ldots, i \cdot 200\}$. We first observe that each $f_{\theta_i}$ behave differently via Grad-CAM [47]. For example, as shown in Fig. 3, $f_{\theta_1}$ and $f_{\theta_5}$ trained on cleaner images predict ‘husky’ based on distinctive canine features such as its eyes or coat pattern. Meanwhile, $f_{\theta_4}$ trained on noisy images make predictions based on overall shape.
Such behavior difference manifests when we guide diffusion using different \( f_{\phi_n} \). For example, \( f_{\phi_n} \) trained on noisy images initially generates a husky-like shape but fails to fill in finer details. On the other hand, \( f_{\phi_n} \) 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 \( N \) expert guidance model, where the \( n \)-th expert \( f_{\phi_n} \) is trained to predict the ground-truth label \( y \) given noisy data \( x_t, t \in \left\{ \frac{n-1}{N}T, \ldots, \frac{n}{N}T \right\} \). Then, during the reverse diffusion process, we assign an appropriate guidance model depending on the timestep, i.e., \( n \) for which \( t \in \left\{ \frac{n-1}{N}T + 1, \ldots, \frac{n}{N}T \right\} \).

More formally, model guidance (Eq. 4) can be rewritten as:

\[
x_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \left( x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} g(x_t, t) \right) + \sigma_t \nabla_x \mathcal{L}_{\text{guide}} (f_{\phi_n}(x_t), y),
\]

where \( n \) is such that \( t \in \left\{ \frac{n-1}{N}T + 1, \ldots, \frac{n}{N}T \right\} \). 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: Practical Plug-And-Play (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 mod-
els, 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 $f_\phi$.

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 $\{(\tilde{x}_0, y)\}$ 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 $f_\phi$ as a teacher, and use its prediction on clean data to serve as labels when training expert guidance models $f_{\phi_n^*}$. Formally, we formulate the knowledge transfer loss as:

$$ L_{KT} = \mathbb{E}_{t \sim \text{unif}[1, T \ldots, T]} [L(s_{sg}(f_\phi(\tilde{x}_0)), f_{\phi_n^*}(\tilde{x}_t))], $$

(6)

where $s_{sg}(\cdot)$ is the stop-gradient operator and $L$ 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 $L_{clsf}$ for classifiers as:

$$ L_{clsf} = D_{KL}(s_{sg}(s(f_\phi(\tilde{x}_0)/\tau)), s(f_{\phi_n^*}(\tilde{x}_t))), $$

(7)

where $s$ is the softmax operator, $\tau$ the temperature hyperparameter, 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:

- **Naive 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 tuning.
- **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 $\hat{x}_0 = \frac{x_t}{\sqrt{\alpha_t}} \cdot \sqrt{1-\alpha_t} s_{sg}(x_t, t)$.

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-
We first compare the variants of ResNet50

directly using the off-the-shelf models does not effectively
guide the diffusion model, generating irrelevant images to
given classes (Albatross and Mongoose), but leveraging
multi-experts shows more powerful guidance capabilities.

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-

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 underline 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.
Ablation study: Effect of Parameter-Efficient Tuning and data efficiency Here we analyze the effectiveness of the parameter-efficient fine-tuning strategy, in comparison 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 tuning 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 ~). 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
GLIDE consists of two diffusion models: 1) Generator: generates $64 \times 64$ images from given text inputs, and 2) Upsampler: upsamples generated $64 \times 64$ images to $256 \times 256$ 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 $64 \times 64$ 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


