This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Unconditional Priors Matter! Improving Conditional Generation of Fine-Tuned Diffusion Models

Prin Phunyaphibarn Phillip Y. Lee Jaihoon Kim Minhyuk Sung
KAIST
{prin10517, phillip0701, jh27kim, mhsung}@kaist.ac.kr

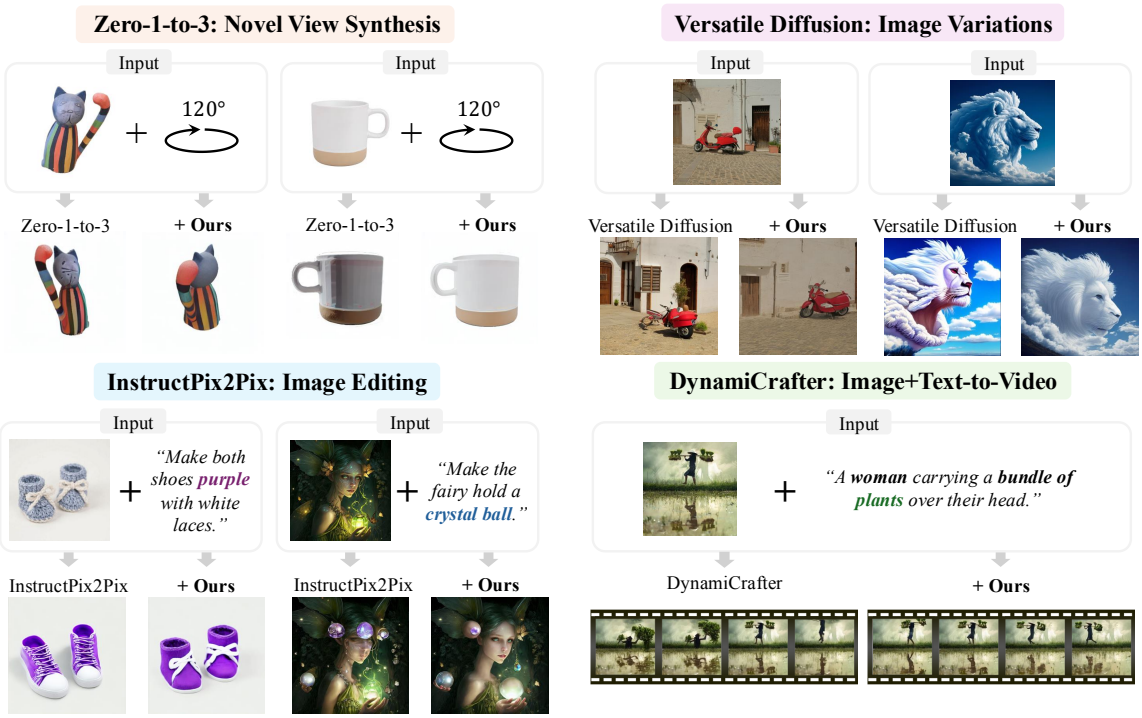


Figure 1. **Unconditional Priors Matter in CFG-Based Conditional Generation.** Fine-tuned conditional diffusion models often show drastic degradation in their unconditional priors, adversely affecting conditional generation when using techniques such as CFG [27]. We demonstrate that leveraging a diffusion model with a richer unconditional prior and combining its unconditional noise prediction with the conditional noise prediction from the fine-tuned model can lead to substantial improvements in conditional generation quality. This is demonstrated across diverse conditional diffusion models including Zero-1-to-3 [44], Versatile Diffusion [61], InstructPix2Pix [7], and DynamiCrafter [59].

Abstract

Classifier-Free Guidance (CFG) is a fundamental technique in training conditional diffusion models. The common practice for CFG-based training is to use a single network to learn both conditional and unconditional noise prediction, with a small dropout rate for conditioning. However, we observe that the joint learning of unconditional noise with limited bandwidth in training results in poor priors for the unconditional case. More importantly, these poor unconditional noise predictions become a serious reason for degrading the quality of conditional generation. Inspired by the fact that most CFG-based conditional models are trained by fine-tuning a base model with better unconditional generation, we first show that simply replacing the unconditional noise in CFG with that predicted by the base model can significantly improve conditional generation. Furthermore, we show that a diffusion model other than the one the fine-tuned model was trained on can be used for unconditional noise replacement. We experimentally verify our claim with a range of CFG-based conditional models for both image and video generation, including Zero-1-to-3, Versatile Diffusion, DiT, DynamiCrafter, and InstructPix2Pix.

1. Introduction

In recent years, diffusion models [28, 53, 56] have shown great success in generation tasks, becoming the de facto standard generative model across many data modalities such as images [48–51], video [4, 5, 29, 63], and audio [13, 32, 42]. The success of diffusion models is not only due to their high-quality results and ease of training, but also the simplicity of adapting them into *conditional* diffusion models. While previous generative models such as GANs [22] and VAEs [37] require separate training for each conditional generation task, making it costly to create various conditional generative models, diffusion models introduced a considerably more effective approach: training an unconditional model (or a conditional model with simple conditions, such as text) as a base and branching out into multiple conditional models.

At the core of the extendability of diffusion models in easily converting an unconditional (or less conditioned) base model into a conditional (or more conditioned) model is the Classifier-Free Guidance (CFG) [27] technique. CFG proposed to learn to predict both unconditional and conditional noises using a single neural network, without introducing another network, such as a classifier, as in the classifier-guidance [14] approach. CFG combines unconditional and conditional noise predictions to generate data conditioned on a given input. It has been widely adopted not only for training a conditional model from scratch but also for *fine-tuning* a base model to incorporate other conditions, by adding encoders for the conditional input. Many successful conditional generative models have been fine-tuned using CFG from a base model. For example, Zero-1-to-3 [44] and Versatile Diffusion [61] use variants of Stable Diffusion [50] (`SD`) as a base, with additional encoders to incorporate the input image as conditions, while InstructPix2Pix [7] uses `SD1.5` as a base and incorporates text editing instructions and input reference images as conditions to perform instruction-based image editing.

Despite its successes and widespread usage, fine-tuning a conditional model from a base model using the CFG technique has limitations, most notably producing lower-quality results for unconditional generation. This is because both conditional and unconditional noise are learned by the same noise prediction network, thus sharing the limited capacity of the neural network. Typically, the bandwidth allocated for the unconditional noise is even more limited by setting a 5-20% drop rate of the condition, an issue which is exacerbated when the training data is limited or the model is fine-tuned multiple times. More importantly, the low quality of unconditional noise also negatively affects the quality of conditional generation, since conditional generation is performed by combining both conditional and unconditional noise predictions in the CFG formulation.

The crucial oversight in this practice is that the base model already provides useful guidance for unconditional generation, and the quality of its generated outputs is generally much better than that of the fine-tuned model. Hence, we demonstrate that in conditional generation using a fine-tuned diffusion model and CFG, simply replacing the unconditional noise of the fine-tuned diffusion model in CFG with that of the base model leads to significant improvements. This is a *training-free* solution that requires no additional training or modifications to the neural networks. It also highlights that when fine-tuning a diffusion model with additional conditioning using CFG, the unconditional noise does not need to be learned jointly.

Surprisingly, we show that the unconditional noise does not need to come from the base unconditional model used for fine-tuning, but can also come from other pretrained diffusion models. This further eliminates the need to jointly learn both unconditional and conditional noise using the same noise prediction network when a pretrained unconditional diffusion model is available.

2. Related Works

Guidance in Diffusion Models. Classifier-Free Guidance (CFG) [27] has become the de facto guidance technique for conditional generation with diffusion models, leading to notable improvements in both condition alignment and image quality. However, recent research has highlighted some of its limitations. Kynkäänniemi et al. [39] have shown that the specific timesteps at which CFG is applied significantly impact image diversity, and proposed to restrict CFG to certain intervals.

Another line of work [1, 30] addresses the limited applicability of CFG for *text-based* conditions when using off-the-shelf diffusion models like Stable Diffusion [50]. These approaches introduce a guidance technique that extends to a broader range of generation tasks, including unconditional generation, inverse problems, and conditional generation with *non-text* conditions (*e.g.*, depth maps [64]). However, most works have not explored how the dynamics of CFG shift when a diffusion model is *fine-tuned* for a specific task [7, 44, 61]. Motivated by transfer learning, Zhong et al. [67] propose to improve the generation quality of fine-tuned DiTs [48] by using the unconditional noise from the original model. However, they consider only small class-conditional DiTs while our method applies to popular large-scale diffusion models across different architectures (even when the base and fine-tuned models have different network architectures) and in cases where the base and fine-tuned model conditions have different modalities. Based directly on empirical observations of unconditional noise degradation in large-scale diffusion models that occurs during fine-tuning, we propose a solution by combining noise predictions from various diffusion models.

Karras et al. [36] proposed Autoguidance which uses the

noise estimate from an *under-trained* version of itself, instead of unconditional noise, to improve image quality. In Autoguidance, both versions of the model are conditioned on the *same* condition to isolate the image quality improvement from the improvement in condition-alignment. In contrast, our method combines models whose conditions have *different* modalities, correcting the degradation that stems from the quality of the unconditional prior used in the CFG formulation. Furthermore, Autoguidance emphasizes the importance of designing the degradations to match the degradation of the conditional model. They show that certain degradations will degrade rather than improve the generation quality. One of our main contributions is showing that the fine-tuned unconditional degradation hurts rather than helps the quality of conditional generation.

Merging Diffusion Models. Aligned with the mixture-of-experts [8] and model merging [62] literature on foundation models, there is growing research on methods for merging diffusion models to enable effective composition of multiple conditions. Diffusion Soup [3] directly merges weights of different diffusion models, Mix-of-Show [24] combines the weights of LoRA adapters [31], and MaxFusion [46] merges intermediate model features. Notably, leveraging the iterative denoising process of diffusion models, merging their noise estimates has emerged as a simple yet powerful technique for composing conditions. By merging noise estimates from the *same* diffusion model with different input conditions, it becomes possible to generate outputs that contain a combination of these conditions [2, 16, 17, 20, 21, 65]. Interestingly, multiple studies have shown that noise estimates from *different* diffusion models [11, 19, 45] can also be merged effectively. In this work, we extend this approach, demonstrating how merging noise estimates can enhance generation quality when applying CFG to fine-tuned models.

3. Background

3.1. Diffusion Models

Diffusion models [28, 53, 54] generate data by sampling from a given distribution (*e.g.*, Gaussian) and applying iterative denoising. In the forward process, random noise is applied to clean data $\mathbf{x}_0$ following:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\bar{\alpha}_t \in [0,1]$[1] [28]. In the reverse process, the noisy data $\mathbf{x}_t$ is denoised by modeling the transition as a Gaussian distribution:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}) \tag{2}$$

where the variance σ_t^2 is predefined, and predicting the posterior mean $\mu_\theta(\mathbf{x}_t, t)$ can be reparameterized as a noise prediction task using Tweedie's formula [18] as below:

$$\mu_\theta(\mathbf{x}_t, t) = \tilde{\mu}_t\left(\mathbf{x}_t, \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t))\right) \tag{3}$$

$$=: \tilde{\mu}\left(\mathbf{x}_t, g(\mathbf{x}_t, \epsilon_\theta(\mathbf{x}_t))\right) \tag{4}$$

where $\epsilon_\theta(\mathbf{x}_t)$[2] is the noise prediction from a diffusion model, and $\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0)$ is the forward process posterior mean. Eq. 3 can be further interpreted as updating the posterior mean towards the *prediction of a clean observation* from $\mathbf{x}_t$ using Tweedie's formula. We define this clean observation predicted from Tweedie's formula as the function $g(\cdot,\cdot)$ and denote the predicted clean observation by $\mathbf{x}_{0|t}$.

DDIM Sampling. DDIM [54] enables efficient sampling for diffusion models by modeling the *non-Markovian* transition $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$, conditioned on $\mathbf{x}_0$. One denoising step of DDIM is presented as the following deterministic transition:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}g(\mathbf{x}_t, \epsilon_\theta(\mathbf{x}_t)) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\mathbf{x}_t). \tag{5}$$

3.2. Classifier-Free Guidance (CFG) [27]

For a diffusion model to perform *conditional* generation given a condition c, it needs to sample from the conditional distribution $p(\mathbf{x}|c)$. One approach is to use a classifier to guide the sampling process toward the conditional distribution [14]; however, it comes at the cost of training a separate classifier. Alternatively, Ho and Salimans [27] eliminated the need for a separate classifier by introducing Classifier-Free Guidance (CFG), a straightforward modification to the training and sampling process of diffusion models. In CFG training, the model learns to predict the noise in $\mathbf{x}_t$ at timestep t not only when a condition c is given, but also when a *null condition* $\emptyset$ is given. That is, the diffusion model performs both *conditional* (*i.e.*, $\epsilon_\theta(\mathbf{x}_t, c)$) and *unconditional* (*i.e.*, $\epsilon_\theta(\mathbf{x}_t, \emptyset)$) noise prediction. This is achieved by setting the condition c to the null condition $\emptyset$ with a certain probability during training. With a model trained using this technique, CFG can be applied in the sampling process by replacing $\epsilon_\theta(\mathbf{x}_t)$ in Eq. 5 with:

$$\epsilon_\theta^{(\gamma)}(\mathbf{x}_t, c) = \epsilon_\theta(\mathbf{x}_t, \emptyset) + \gamma(\epsilon_\theta(\mathbf{x}_t, c) - \epsilon_\theta(\mathbf{x}_t, \emptyset)), \tag{6}$$

where γ is a guidance scale. A detailed algorithm of CFG with DDIM sampling [54] is shown in Alg. 1.

Analysis of CFG. Based on the connection between diffusion models and score-based models [55, 56], $\epsilon_\theta(\mathbf{x}_t, c)$ and $\epsilon_\theta(\mathbf{x}_t, \emptyset)$ model the conditional score $\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|c)$

[1] With variance schedule $\{\beta_t\}_{t=1}^T$, $\alpha_t := 1-\beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$.

[2] We omit timestep t in $\epsilon_\theta(\mathbf{x}_t)$ and $g(\mathbf{x}_t, \epsilon_\theta(\mathbf{x}_t))$ for brevity.

Algorithm 1 DDIM Sampling with CFG

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\quad \epsilon_\theta^{(\gamma)}(\mathbf{x}_t, c) = \epsilon_\theta(\mathbf{x}_t, \emptyset) + \gamma(\epsilon_\theta(\mathbf{x}_t, c) - \epsilon_\theta(\mathbf{x}_t, \emptyset))$
4: $\quad \mathbf{x}_{0|t} = g(\mathbf{x}_t, \epsilon_\theta^{(\gamma)}(\mathbf{x}_t, c))$
5: $\quad \mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta^{(\gamma)}(\mathbf{x}_t, c)$
6: **end for**
7: **return** $\mathbf{x}_0$

and the unconditional score $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$ (up to a scaling factor), respectively. We can interpret the CFG noise $\epsilon_\theta^{(\gamma)}(\mathbf{x}_t, c)$ in Eq. 6 as an approximation of the true score $\nabla_{\mathbf{x}_t} \log p_\gamma(\mathbf{x}_t|c)$ where $p_\gamma(\mathbf{x}_t|c) := p(\mathbf{x}_t)\left(\frac{p(\mathbf{x}_t|c)}{p(\mathbf{x}_t)}\right)^\gamma$ is the *gamma-powered* distribution. We can rewrite $p_\gamma(\mathbf{x}_t|c)$ as follows:

$$p_\gamma(\mathbf{x}_t|c) = p(\mathbf{x}_t)\left(\frac{p(\mathbf{x}_t, c)}{p(c)p(\mathbf{x}_t)}\right)^\gamma \propto p(\mathbf{x}_t)p(c|\mathbf{x}_t)^\gamma.$$

Thus, CFG guides the samples via the implicit classifier $p(c|\mathbf{x}_t)^\gamma$. Using $\gamma > 1$ results in sharpening the mode corresponding to c which leads to better condition-alignment [27] and image quality [36].

4. Unconditional Priors Matter

In this section, we discuss the negative impact of degraded unconditional noise estimates in diffusion models *fine-tuned* on a narrower task-specific data distribution (Sec. 4.1). We then present a simple yet effective approach to enhance the generation quality of these models by leveraging richer unconditional noise estimates from other pre-trained diffusion models (Sec. 4.2).

4.1. Poor Unconditional Priors Affect Conditional Generation

The CFG training technique, introduced in Sec. 3, is also commonly used for *fine-tuning* an existing diffusion model to incorporate new types of input conditions [7, 35, 44, 52, 61]. Consider a pretrained diffusion model, such as Stable Diffusion [50], referred to as the *base model*, parameterized by ψ. This base model can be fine-tuned on task-specific datasets to incorporate certain types of input conditions, such as camera poses [44] or a reference image [61]. We refer to the resulting model as the *fine-tuned* model, parameterized by θ.

After fine-tuning the base model with CFG training, conditional generation with the fine-tuned model is performed by combining conditional and unconditional noise predictions following Eq. 6, yielding the CFG noise $\epsilon_\theta^{(\gamma)}(\mathbf{x}_t, c)$. However, we observe a significant quality drop in *unconditional* generation with the fine-tuned model compared to the base model. As shown in Fig. 2, the unconditional outputs from fine-tuned models [7, 44, 61] clearly lack detailed semantics and exhibit lower image quality. These results are expected, as (1) the unconditional distribution is inherently more complex than the conditional distribution, and (2) only a small fraction of the training data is utilized in each training iteration due to the low CFG dropping probability (typically 5-20%).

Importantly, we observe that the quality of the *conditional* generation in these fine-tuned models is negatively impacted by poor unconditional priors. This degradation can be understood through the CFG [27] formulation: CFG is designed to sample from the gamma-powered distribution $p_\gamma(\mathbf{x}_t|c) \propto p(\mathbf{x}_t)p(c|\mathbf{x}_t)^\gamma$. Poor unconditional priors introduce approximation errors to $p(\mathbf{x}_t)$, which in turn affects both $p(\boldsymbol{x}_t)$ and $p(c|\boldsymbol{x}_t) \propto \frac{p(\boldsymbol{x}_t|c)}{p(\boldsymbol{x}_t)}$. Although a key advantage of CFG is the joint modeling of the unconditional and conditional distributions, we find that under limited data or multiple fine-tuning, the fine-tuned model loses the rich unconditional prior of the base model, leading to quality degradation.

4.2. Finding Richer Unconditional Priors

Can we improve the quality of conditional generation by incorporating better unconditional priors during sampling? Note that for fine-tuned diffusion models, we have access to a diffusion model with reliable unconditional priors: its base model. Therefore, we propose a simple yet effective fix, combining the *unconditional* noise prediction from the base model with the *conditional* noise prediction from the fine-tuned model. For this, the CFG noise in line 3 of Alg. 1 is modified as follows:

$$\epsilon_{\theta,\psi}^{(\gamma)}(\mathbf{x}_t, c) = \epsilon_\psi(\mathbf{x}_t, \emptyset) + \gamma(\epsilon_\theta(\mathbf{x}_t, c) - \epsilon_\psi(\mathbf{x}_t, \emptyset)), \quad (7)$$

where ϵ_ψ and ϵ_θ denote the base model and fine-tuned model, respectively. Then DDIM sampling step becomes:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}g(\mathbf{x}_t, \epsilon_{\theta,\psi}^{(\gamma)}(\mathbf{x}_t, c)) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_{\theta,\psi}^{(\gamma)}(\mathbf{x}_t, c)$$

Surprisingly, this simple modification results in significant improvements in the output quality of conditional generation. We demonstrate this through both qualitative and quantitative evaluations in Sec. 5.

A natural next question that arises is: should the unconditional noise come from the base model or from another unconditional model? We find that the base model does not necessarily need to be the true base model from which the new conditional model was fine-tuned from, but can instead be another diffusion model with good unconditional priors. In our experiments, we show that even though some models have been fine-tuned on `SD1.x`, using unconditional predictions of `SD2.1` or `PixArt-`α results in further improvements as shown in Sec. 5.

Combining Diffusion Models. Although the base model and the fine-tuned model have different model weights,

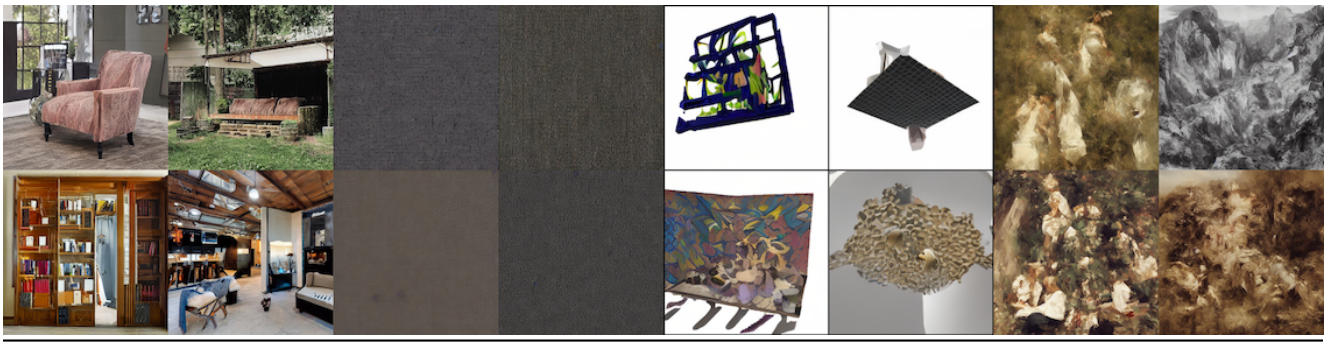


Figure 2. **Unconditional samples from different diffusion models.** Stable Diffusion [4], which often serves as the base model for fine-tuning conditional diffusion models, generates plausible images, whereas other fine-tuned diffusion models fail to sample realistic images.

their noise predictions can be combined as done in previous works [11, 19, 45]. Based on the connection between energy-based models (EBMs) and diffusion models [43], at each timestep, t, our method is equivalent to sampling from the time-annealed distribution $p_\psi(\mathbf{x}_t)^{1-\gamma} p_\theta(\mathbf{x}_t|c)^\gamma$ where $p_\psi(\mathbf{x}_t)$ and $p_\theta(\mathbf{x}_t)$ are the distributions modeled by the base and fine-tuned models, respectively. Notably, $p_\psi(\mathbf{x}_t)$ can be modeled by any pretrained diffusion model which may have different weights or even different architecture from the fine-tuned model so long as $p_\psi(\mathbf{x}_t)$ is a better approximation of the *true* unconditional distribution than $p_\theta(\mathbf{x}_t)$.

5. Experiments

We validate our method on five conditional diffusion models, each trained for distinct conditional generation tasks: Zero-1-to-3 [44], Versatile Diffusion (VD) [61], DiT [48], DynamiCrafter [59], and InstructPix2Pix [7]. For experiments on models fine-tuned from Stable Diffusion, we present the results using unconditional noise predictions from both the true base model of the fine-tuned networks and other diffusion models, Stable Diffusion 2.1 (`SD2.1`) [50] and `PixArt-`α [10]. Notably, even when the fine-tuned model is a *UNet*, our method yields improvements when using `PixArt-`α, which is a *DiT*, in place of the base model. We refer readers to **supplementary** for details on the experimental setup for each application.

5.1. Single-Condition CFG Formulation

In this section, we provide experimental results of our method when applied to diffusion models that use single-condition CFG formulation. Models of this category samples using noise of the form provided in Eq. 6.

Zero-1-to-3 [44] Zero-1-to-3 is a conditional diffusion model for novel view synthesis, taking a reference image and relative camera poses as input. It is fine-tuned from Stable Diffusion Image Variations (`SD-IV`) [35], which itself is originally fine-tuned from `SD1.4`. Due to the multiple fine-tuning stages, we opted to use `SD1.4` as the base model. We evaluate the samples on the Google Scanned Objects (GSO) dataset [15] using LPIPS [66], PSNR, and SSIM. As shown in Tab. 1, our method, which incorporates unconditional noise predictions from base models, achieves significant improvements across all three metrics, with the best performances observed when using `SD2.1` as the unconditional prior. Fig. 3 shows that using improved unconditional noise from the base model enhances lighting quality (row 1), reduces color saturation (row 2) and shape distortions (rows 3 and 4).

Method	LPIPS↓	PSNR↑	SSIM↑
Zero-1-to-3 [44]	0.182	16.647	0.824
Ours w/ `SD1.4`	<u>0.163</u>	<u>17.514</u>	<u>0.842</u>
Ours w/ `SD2.1`	**0.158**	**17.801**	**0.848**
Ours w/ `PixArt-`α	0.169	17.069	0.825

Table 1. **Novel View Synthesis with Zero-1-to-3 [44].** Our method improves quality of novel view images (**bold** represents the best, and <u>underline</u>, the second best method).

Method	FID ↓	$\text{FD}_{\text{DINOv2}}$ ↓	CLIP-I ↑	DINOv2 ↑
VD [61]	8.38	167.65	0.93	0.91
Ours w/ `SD1.4`	<u>6.68</u>	156.77	**0.94**	**0.92**
Ours w/ `SD2.1`	7.80	<u>151.48</u>	**0.94**	**0.92**
Ours w/ `PixArt-`α	**6.29**	**148.48**	**0.94**	**0.92**

Table 2. **Image Variations with Versatile Diffusion [61].** Our sampling method achieves best performances across all metrics (**bold** represents the best, and <u>underline</u>, the second best method).

Versatile Diffusion [61] Versatile Diffusion (VD) is a multi-task diffusion model designed to handle text-to-image, image variations, and image-to-text tasks within a unified architecture. VD is progressively fine-tuned from `SD1.4` in three stages to handle additional image conditions on top of text condition. Due to the cascaded fine-tuning scheme, VD displays the worst image-unconditional generation quality as shown in Fig. 2. We focus on using VD for image variations to generate semantically similar images from a reference image. We report FID [26] and $\text{FD}_{\text{DINOv2}}$ [57] on COCO-Captions [41] for image quality

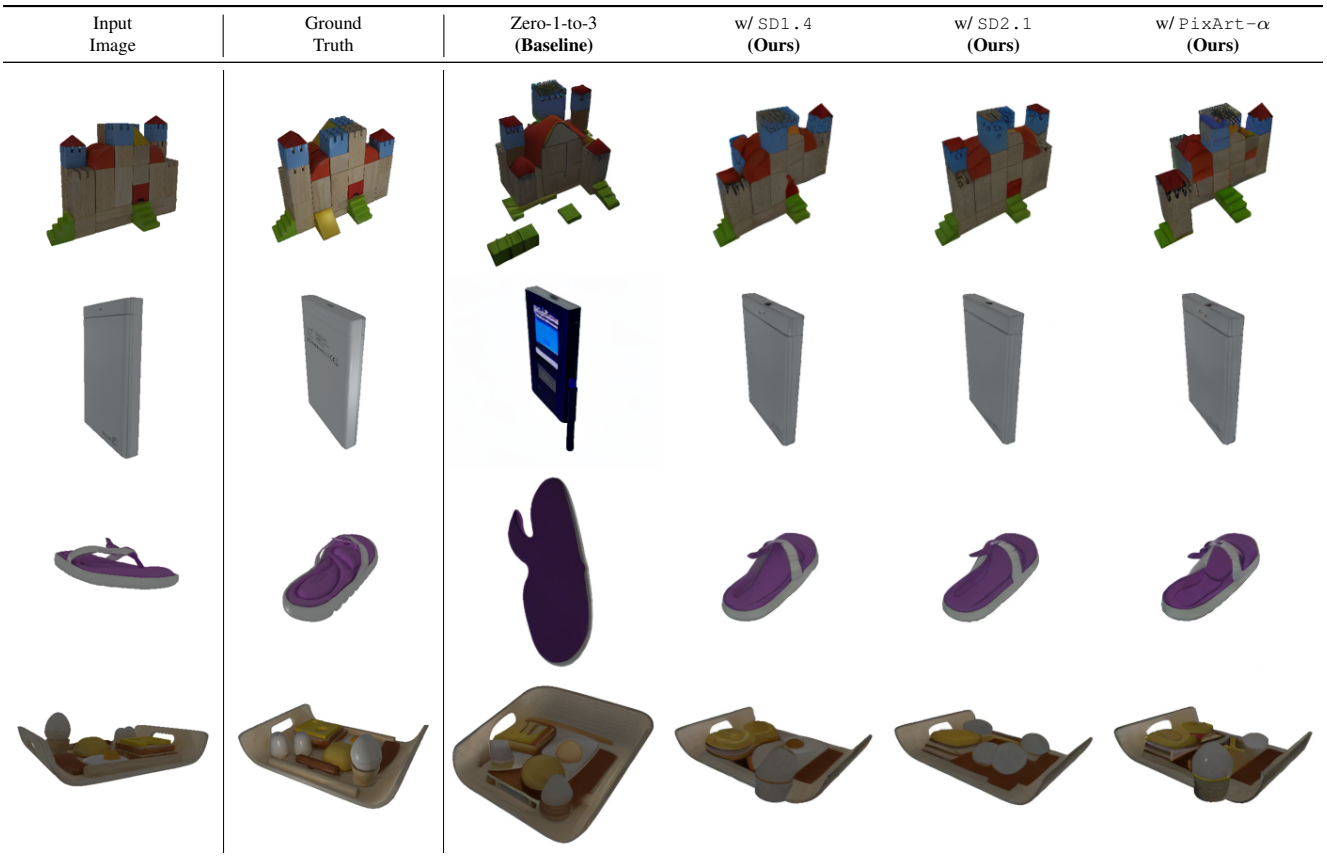


Figure 3. **Novel View Synthesis with Zero-1-to-3 [44].** Outputs from Zero-1-to-3 often show inaccuracies in lighting or shape distortions during novel view synthesis. By incorporating unconditional noise predictions from Stable Diffusion [50] or PixArt-α [10], our method achieves clear improvements in output quality.

assessment and CLIP-I [25] and DINOv2 [47] image similarity metrics to evaluate condition alignment. As shown in Tab. 2, using unconditional noise prediction from the base models yields better FID and $\text{FD}_{\text{DINOv2}}$ while retaining similar CLIP-I and DINOv2 image similarity, showing a performance improvement while maintaining condition alignment. As shown in Fig. 4, VD often generates images with highly saturated colors (rows 1 and 2) and distorted objects (row 3) while our method corrects both.

DiT [48] Although the experiments so far have been conducted using diffusion models with the UNet architecture, we show that our method holds for fine-tuned DiTs [48] as well. Since there are no publicly available fine-tuned DiT models, we fine-tune DiT-XL/2 on the standard downstream tasks SUN397 [58], Food101 [6], and Caltech101 [23]. The FID for the different fine-tuning tasks are shown in Tab. 3. Incorporating the unconditional noise from the base DiT-XL/2 results in improved FID. We also observe larger benefits when the fine-tuning dataset is large (Food101 and SUN397) corroborating our observation that the degradation in unconditional priors is amplified by the limited computational budget. While we observe no improvement for Caltech101, the dataset is over 10 times smaller than both SUN397 and Food101, thus the model is given sufficient time to fit the dataset despite the low CFG condition drop rate (10%). In practice, large-scale fine-tuning like in Zero-1-to-3 and Versatile Diffusion often suffer from limited data and computation. In these practical settings, the degradation of the unconditional prior becomes highly detrimental as we have shown.

Method	SUN397	Food101	Caltech101
Fine-tuned DiT-XL/2	17.12	18.31	**24.05**
Ours	**14.51**	**17.67**	24.15

Table 3. **Class-conditional generation with DiT[48].** FID-5k evaluated on three fine-tuning tasks (SUN397 [58], Food101 [6], and Caltech101 [23]). Our method improves FID of the fine-tuned models (**bold** represents the best method).

5.2. Dual-Condition CFG Formulation

In this section, we provide experimental results on diffusion models which use the dual-condition CFG formulation. Diffusion models in this category are conditioned on two conditions and sample using the modified CFG noise

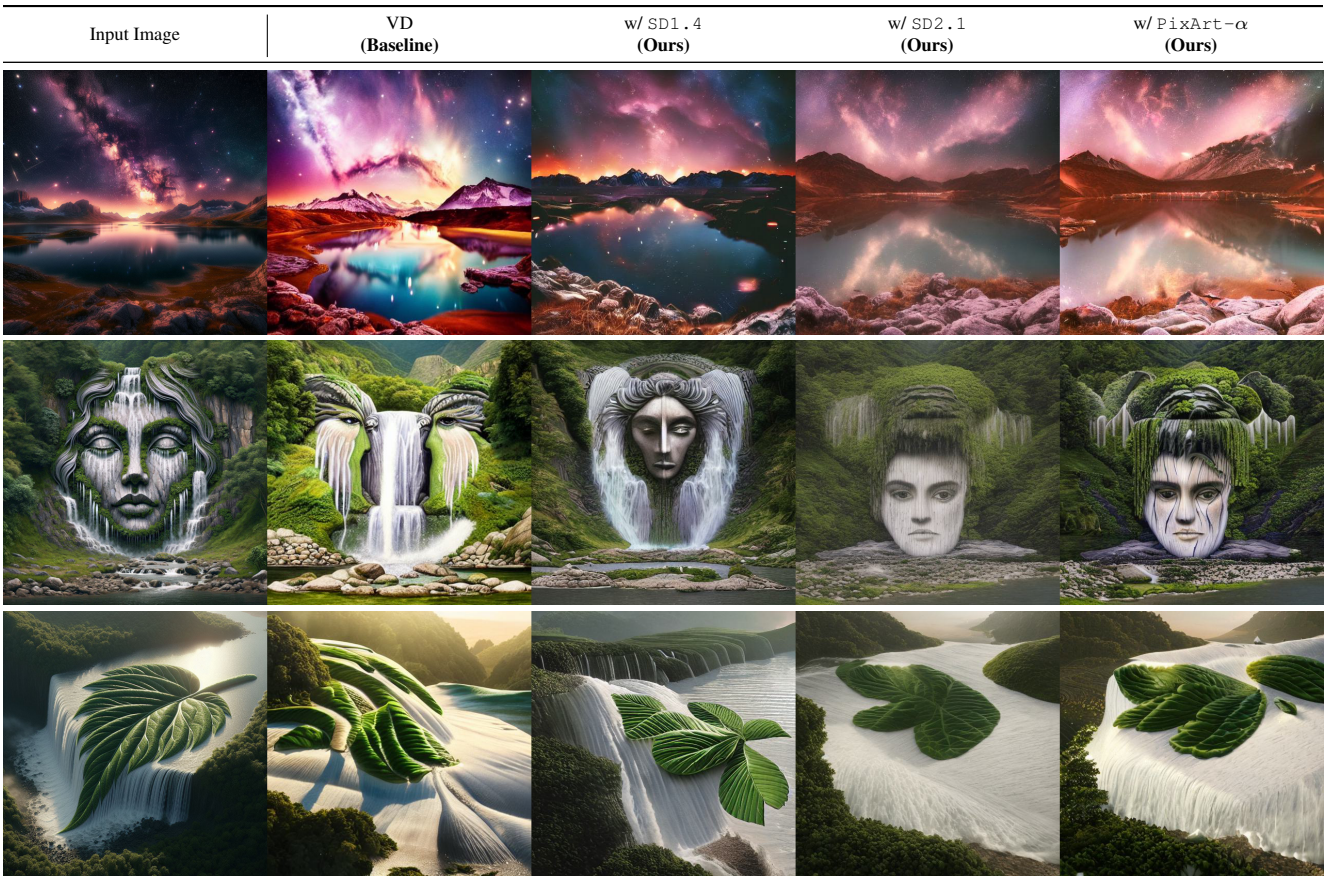


Figure 4. **Image Variations with Versatile Diffusion [61].** Versatile Diffusion often suffers from style and detail degradation—excessive saturation (rows 1 and 3) or loss of key content (row 2). In contrast, our method, leveraging `SD1.4`, `SD2.1`, or `PixArt-`α as unconditional priors, achieves noticeable improvements in performance.

$$\begin{aligned}\epsilon_\theta(\mathbf{x}_t, c_1, c_2) = \epsilon_\theta(\mathbf{x}_t, \emptyset, \emptyset) \\ + \gamma_1(\epsilon_\theta(\mathbf{x}_t, c_1, \emptyset) - \epsilon_\theta(\mathbf{x}_t, \emptyset, \emptyset)) \\ + \gamma_2(\epsilon_\theta(\mathbf{x}_t, c_1, c_2) - \epsilon_\theta(\mathbf{x}_t, c_1, \emptyset)). \quad (8)\end{aligned}$$

Notably, the dual-condition CFG formulation has two unconditional terms trained using CFG condition dropout: $\epsilon_\theta(\mathbf{x}_t, \emptyset, \emptyset)$ and $\epsilon_\theta(\mathbf{x}_t, c_1, \emptyset)$. We replace $\epsilon_\theta(\mathbf{x}_t, \emptyset, \emptyset)$ with the base model unconditional noise prediction $\epsilon_\psi(\mathbf{x}_t, \emptyset)$. However, since the other unconditional term $\epsilon_\theta(\mathbf{x}_t, c_1, \emptyset)$ is not replaced, we observe less improvements in this case than in the single-condition CFG formulation. This is to be expected as the quality degradation stems from training using a low dropout rate for the condition which is applied to both $\epsilon_\theta(\mathbf{x}_t, \emptyset, \emptyset)$ and $\epsilon_\theta(\mathbf{x}_t, c_1, \emptyset)$, only one of which is replaced by the better base model unconditional prior.

DynamiCrafter [59] We apply our method to DynamiCrafter [59], a text-and-image-to-video diffusion model fine-tuned from the text-to-video diffusion model VideoCrafterT2V [9]. DynamiCrafter incorporates an image condition c_I as c_1 and a text condition c_T as c_2. For our method, we replace the DynamiCrafter unconditional noise with the VideoCrafterT2V unconditional noise. We report quantitative results using VBenchI2V [12, 34] which measures video quality and temporal consistency across multiple dimensions. For more details on the metrics, please refer to VBench [34]. Quantitative results are reported in Tab. 4. Our method outperforms the baseline in 7 out of 9 metrics, yielding more consistent videos with higher aesthetic quality. As shown in Fig. 5, our method is more temporally consistent (first video) and less distorted (second video). Video results are also included in the **supplementary**.

InstructPix2Pix [7] InstructPix2Pix (IP2P) tackles instruction-based image editing by fine-tuning `SD1.5` to condition on both text (editing instruction, as c_2) and image (as c_1) to generate the edited image. We evaluate the performance on the EditEvalv2 benchmark [33]. To assess identity preservation, we compute the CLIP image similarity (C-I) [25] between the edited and input images. To evaluate the faithfulness of the edited images, we measure CLIP text alignment (C-T), CLIP Directional Similarity (C-D), Image Reward (IR) [60], and PickScore (PS) [38] based on the edited image prompt. The reported PS are compared against IP2P. As shown in Tab. 5, our method shows better alignment with the prompt while

Method	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	I2V Subject	I2V Background
DynamiCrafter [59]	90.80	96.73	95.21	96.67	**59.59**	57.06	**64.10**	92.77	94.56
Ours	**91.49**	**97.03**	**95.34**	**96.86**	57.32	**57.51**	63.15	**93.49**	**94.72**

Table 4. **Video Generation with DynamiCrafter [59].** All metrics are scored out of 100, higher indicates better performance.



Figure 5. **Image-to-Video Generation with DynamiCrafter [59].** Our method is more temporally consistent (lighting on the biker) and less distorted (the hand and face in the second video).

preserving the identity of the source image. We observe improvements in both IR and PS which have been observed to better align with human preference [38, 60] with slight underperformance in CLIP-T (for `SD1.5` and `SD2.1`). Qualitative results are shown in Fig. 6. Our method generates faithful, high-fidelity edited images (rows 1 and 2) whereas IP2P creates distorted images (row 3).

Method	C-I ↑	C-T ↑	C-D ↑	IR ↑	PS ↑
IP2P [7]	0.909	0.294	0.174	-0.510	–
Ours w/ `SD1.5`	0.911	0.291	**0.186**	-0.460	0.514
Ours w/ `SD2.1`	0.913	0.290	0.184	-0.464	0.518
Ours w/ `PixArt-`α	**0.915**	**0.297**	0.185	**-0.363**	**0.532**

Table 5. **Image Editing with InstructPix2Pix (IP2P) [7].** We normalize text and image similarity scores: C-I, C-T, and C-D. (**bold** represents the best, and underline represents the second best method.)

6. Conclusion

We presented a novel, training-free approach to improving the generation quality of a CFG-based fine-tuned conditional diffusion model by replacing the low-quality unconditional noise with richer unconditional noise from a separate pretrained base diffusion model. We validated our approach across a range of diffusion models trained for distinct conditional generation, including image variation [61],

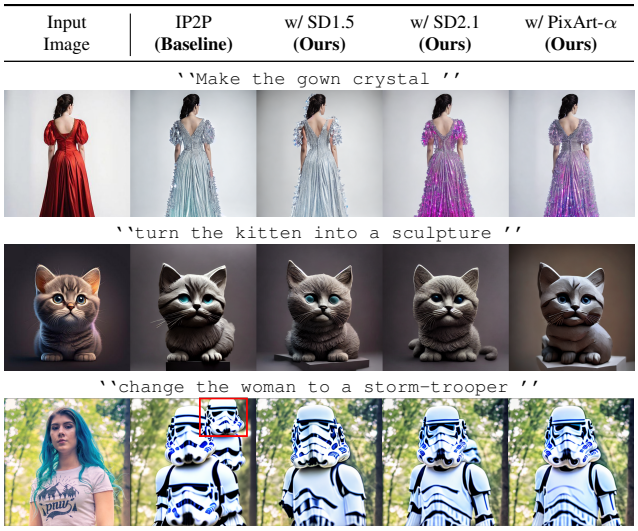


Figure 6. **Image Editing with InstructPix2Pix (IP2P) [7].** Applying our method improves alignment with the editing prompt while preserving the identity of the source image.

image editing [7], novel view synthesis [44], and video generation [59]. Notably, we find that the separate pretrained diffusion model can have different weights and architecture from the original base model.

Limitations and Discussions. Although our method is training-free, it involves loading a second model into memory which increases memory cost. Furthermore, we can no longer parallelize computation as done in CFG, resulting in slight inference time overhead. However, the inference speed is only slightly affected as shown in the **supplementary**. Our method significantly improves diverse fine-tuned diffusion models, but is less effective for adapter-based fine-tuning methods such as ControlNet [64] and GLIGEN [40], which exhibit less degradation in unconditional priors. Identifying unconditional priors for these advanced fine-tuning techniques would be a valuable future direction.

Acknowledgements.

This work was supported by the IITP grants (RS-2024-00399817, RS-2025-25441313, RS-2025-25443318, RS-2025-02653113); and the Industrial Technology Innovation Program (RS-2025-02317326), all funded by the Korean government (MSIT and MOTIE), as well as by the DRB-KAIST SketchTheFuture Research Center.

References

[1] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *ECCV*, 2024. 2

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 3

[3] Benjamin Biggs, Arjun Seshadri, Yang Zou, Achin Jain, Aditya Golatkar, Yusheng Xie, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Diffusion soup: Model merging for text-to-image diffusion models. In *ECCV*, 2024. 3

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 5

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2

[6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 6

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1, 2, 4, 5, 7, 8

[8] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204*, 2024. 3

[9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 7

[10] Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*. 5, 6

[11] Ziyang Chen, Daniel Geng, and Andrew Owens. Images that sound: Composing images and sounds on a single canvas. In *NeurIPS*, 2024. 3, 5

[12] VBench Contributors. Vbench. *GitHub repository*, 2023. 7

[13] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *NeurIPS*, 2023. 2

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 3

[15] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 5

[16] Yilun Du and Leslie Kaelbling. Compositional generative modeling: A single model is not all you need. 2024. 3

[17] Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pages 8489–8510. PMLR, 2023. 3

[18] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 3

[19] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *ICCV*, 2023. 3, 5

[20] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *CVPR*, 2024. 3

[21] Daniel Geng, Inbum Park, and Andrew Owens. Factorized diffusion: Perceptual illusions by noise decomposition. In *ECCV*, 2024. 3

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[23] Gregory Griffin, Alex Holub, Pietro Perona, et al. Caltech-256 object category dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena, 2007. 6

[24] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS*, 2023. 3

[25] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6, 7

[26] 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. In *NeurIPS*, 2017. 5

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1, 2, 3, 4

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022. 2

[30] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *ICCV*, 2023. 2

[31] 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. In *arXiv preprint arXiv:2106.09685*, 2021. 3

[32] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. 2023. 2

[33] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. *arXiv preprint arXiv:2402.17525*, 2024. 7

[34] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 7

[35] Justin. Experiments with stable diffusion. https://github.com/justinpinkney/stable-diffusion. 4, 5

[36] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *arXiv preprint arXiv:2406.02507*, 2024. 2, 4

[37] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[38] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023. 7, 8

[39] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *arXiv preprint arXiv:2404.07724*, 2024. 2

[40] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 8

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[42] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. 2

[43] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 5

[44] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 1, 2, 4, 5, 6, 8

[45] Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel. Unite and conquer: Plug & play multi-modal synthesis using diffusion models. In *CVPR*, 2023. 3, 5

[46] Nithin Gopalakrishnan Nair, Jeya Maria Jose Valanarasu, and Vishal M Patel. Maxfusion: Plug&play multi-modal generation in text-to-image diffusion models. In *ECCV*, 2024. 3

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[48] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2, 5, 6

[49] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

[50] 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. 2, 4, 5, 6

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[52] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 4

[53] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. 2015. 2, 3

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3

[55] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 3

[56] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2, 3

[57] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[58] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6

[59] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 1, 5, 7, 8

[60] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023. 7, 8

[61] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 1, 2, 4, 5, 7, 8

[62] Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*, 2024. 3

[63] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *IJCV*, 2024. 2

[64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 8

[65] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *CVPR*, 2023. 3

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[67] Jincheng Zhong, XiangCheng Zhang, Jianmin Wang, and Mingsheng Long. Domain guidance: A simple transfer approach for a pre-trained diffusion model. In *The Thirteenth International Conference on Learning Representations*. 2