Coupled Diffusion Sampling for Training-Free Multi-View Image Editing

Hadi Alzayer[1,2] Yunzhi Zhang[1] Chen Geng[1] Jia-Bin Huang[2] Jiajun Wu[1]
[1]Stanford University [2]University of Maryland College Park
https://coupled-diffusion.github.io

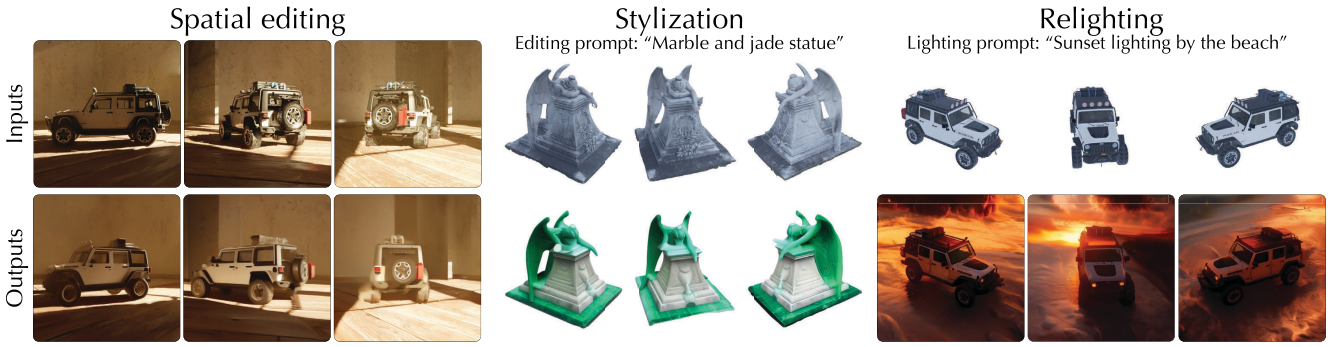


Figure 1. **Applications of coupled diffusion sampling.** Our approach enables lifting off-the-shelf 2D diffusion editing models into multi-view by combining the sampling process of 2D models with multi-view generation models to produce consistent edits. Here we showcase example view-consistent results using three 2D editing models: spatial editing, stylization, and text-based relighting.

Abstract

Given a collection of multi-view images, we perform consistent multi-view editing with a training-free framework using pre-trained 2D editing models and a generative multi-view model. While 2D editing models can independently edit each image in a set of multi-view images of a 3D scene, they do not maintain consistency across views. Existing approaches typically rely on explicit 3D representations to average out inconsistencies, but they suffer from lengthy optimization, instability under sparse views, and blurry results. We address the problem from a different lens, using the 2D editing model to guide a multi-view generative model during diffusion sampling. This is achieved through our novel coupled diffusion sampling process. We concurrently sample two trajectories from both a multi-view image distribution and a 2D edited image distribution, and connect the samples with a coupling term. Effectively, the two models guide each other during sampling, and the resulting sample from the multi-view model remains consistent while satisfying the desired edit. We validate the effectiveness and generality of this framework on three distinct multi-view image editing tasks, and demonstrate its applicability across various model architectures. We further illustrate the effects of coupling on SoTA image and video generation models, highlighting the potential of our method beyond multi-view editing.

1. Introduction

Diffusion-based image editing models have demonstrated unprecedented realism across diverse tasks via end-to-end training. Simply by curating paired editing datasets, we can perform relighting [20, 34, 61], spatial structure editing [2, 39, 49, 55], and stylization [60]. However, collecting and curating 3D data is significantly more challenging than 2D data. As a result, recent research has explored test-time optimization methods for multi-view editing that leverage pre-trained 2D image diffusion models [16, 41].

When editing a collection of photos of a 3D scene with a 2D model, we encounter the issue where each image is edited inconsistently, as shown in Fig. 2. To resolve those inconsistencies, most existing methods [16, 20] rely on explicit 3D representations, *i.e.,* NeRF [37] or 3D Gaussian Splatting [22]. The advantage of using explicit representations is to avoid the need to train new 3D editing models. However, these methods typically require time-consuming optimization and dense coverage of the input view, limiting their applicability to real-time, real-world scenarios. Furthermore, the lack of a data-driven prior can lead to inconsistencies that result in blurring or variants of the Janus problem.

We follow a different approach. Instead of using 2D models to supervise and train an explicit 3D representation, we use 2D models to *guide* implicit 3D generative models. Recent multi-view diffusion models [14, 63] were pro-

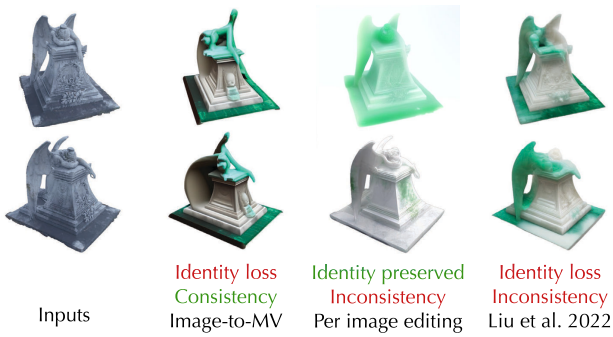


Figure 2. **Limitations of baselines.** Using image-to-multiview model conditioned on a single edited image cannot be faithful to the remaining input views, while editing each image individually with a 2D model produces inconsistent results. While prior work [31] proposes a method to compose diffusion models, their approach produces inconsistent and flickering results.

posed to perform novel view synthesis and generation using a data-driven prior. However, when naively attempting to edit a scene by editing a single image and synthesizing the remaining views, the outputs would lose the identity of the input, as shown in Fig. 2, because this process is underconstrained and many solutions exist. We address this by *steering* the multi-view model using the 2D editing model to sample a valid edit that is *faithful* to the input, and *consistent* across views. We call our novel diffusion sampling process *coupled diffusion sampling*. As shown in Fig. 1, our approach enables multi-view consistent image editing across diverse applications, including multi-view spatial editing, stylization, and relighting.

While prior work [13, 31] explored combining diffusion models within a modality, we observe that such approaches do not maintain multi-view consistency and can stray from the editing objective as shown in Fig. 2. Our approach is motivated by the observation that any sequence of images generated by a pre-trained multi-view image diffusion model inherently exhibits multi-view consistency. To this end, we embrace an implicit 3D regularization paradigm by leveraging scores estimated from multi-view diffusion models during the diffusion sampling process. Specifically, for any multi-view image editing task leveraging a pre-trained 2D model, we couple it with a foundation multi-view diffusion model and perform sampling under dual guidance from both models. This process ensures that the resulting samples satisfy both the editing objective and multi-view 3D consistency, yet without any additional explicit 3D regularization or training overhead.

We propose a practical sampling framework to achieve the above-mentioned goal by steering the standard diffusion sampling trajectory with an energy term coupling two sampling trajectories. This method ensures that each sample from a given diffusion model remains within its own distribution while being guided by the others. In particular, samples from the multi-view diffusion model maintain multi-view consistency while being steered by the content edits from the 2D model. Conversely, the 2D model is steered so that its edits remain faithful to the identity of the inputs while being consistent across independently edited frames.

Our solution is conceptually simple, broadly applicable, and adaptable to a variety of settings. We showcase its effectiveness across three distinct multi-view image editing tasks: multi-view spatial editing, stylization, and relighting. Through comprehensive experiments across each task, we demonstrate the advantages of our method over the state of the art. We further validate the generalizability of our approach to diverse backbones and latent spaces, and illustrate its effects on image and video generation, underscoring its promise as a general multi-view image editing engine with potential extensions to broader generative tasks.

2. Related work

Test-time diffusion guidance. Test-time scaling methods [24, 28, 33], such as rejection sampling or verifier-based search. In contrast, optimization-based guidance actively steers diffusion trajectories, offering a more efficient alternative. The diffusion sampling process can be steered with a classifier guidance [12], a differentiable objective function [4, 15], or by enforcing additional assumptions in the forward process, e.g., as in linear inverse restoration problems [11, 21, 51]. In cases where the edit can be characterized by a "prompt" edit, the model can be used to guide itself to edit the input [8, 25, 38]. In our method, we use diffusion models that capture complementary distributions to inform one another. Unlike prior work, this allows us to capture complex edits beyond what can be formulated by a differentiable function and to achieve multi-view editing without paired data.

3D and multiview editing. As diffusion models capable of producing high-quality 2D image edits [7, 8, 25, 38, 40, 45, 58], a natural question has been how to leverage those capabilities for 3D editing. One common approach is to optimize an explicit 3D representation, like NeRF, either by modifying the training dataset during the optimization loop [16, 54] or through score distillation sampling [35, 41, 46, 52, 57]. However, both approaches are prone to visual artifacts, which is fundamentally caused by the fact that 2D diffusion models lack 3D consistency awareness. To address this fundamental challenge, prior work has directly trained multi-view diffusion models [1, 6, 27, 30, 48] for consistent editing. However, training a multi-view diffusion model for each individual editing task is computationally expensive, and suitable training datasets are scarce. In our approach, we propose reusing existing multi-view *generation* models [14, 56, 63] for multi-view *editing* by combining them with a 2D editing model, thereby incurring

no additional training cost. In contrast to NeRF-based approaches, our method does not require a costly optimization process, as it relies solely on feed-forward sampling.

Compositional diffusion sampling. Compositional sampling methods for diffusion models have been proposed to combine the priors of multiple models. Examples include product-of-experts sampling [17, 62], which samples from the product distribution of individual models. However, this approach imposes a strict requirement that valid samples lie in the intersection of the supports of the models and fails when no such joint support exists. MultiDiffusion [5] and SyncTweedies [23] apply score composition for out-of-distribution scenarios like stitching panoramas or large images. On the other hand, our work emphasizes remaining within each model's prior distribution while steering generation toward satisfying cross-model constraints. Prior works [13, 31] address inference-time composition for diffusion models, but these works focus on the same data modality. In contrast, our work bridges 2D and 3D modalities to tackle the practical challenge of 3D data sparsity.

3. Method

3.1. Background

Diffusion Models. Let $x_0 \sim p_{\text{data}}(x_0)$ be a data sample and consider the forward noising process:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t \mid \sqrt{1-\sigma_t}x_{t-1}, \sigma_t I), \tag{1}$$

with a variance schedule $\{\sigma_t\}_{t=1}^T$. [19] proposes to train a neural network $\epsilon_\theta(x_t, t)$, where θ denotes network parameters, such that when starting with an initial noise $x_T \sim \mathcal{N}(0, I)$, it allows one to gradually denoise the sample to $x_0 \sim p_{\text{data}}(x_0)$ via

$$\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t)) \tag{2}$$

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_t) + \sigma_t z, \tag{3}$$

where $\alpha_t = 1 - \sigma_t, \bar{\alpha}_t := \prod_{s=1}^t \alpha_s$. The next-step prediction x_{t-1} is obtained by computing the *clean* image estimate $\hat{x}_0$ and re-injecting a decreasing amount of random noise $z \sim \mathcal{N}(0, I)$.

3.2. Coupled DDPM Sampling

Problem. Given two diffusion models ϵ_{θ^A} and ϵ_{θ^B} for a shared data domain $\mathbb{R}^d$ and with a shared DDPM schedule, our goal is to obtain two samples $x^A, x^B \in \mathbb{R}^d$ such that they follow the data distribution prescribed by the pretrained models $p^A_{\text{data}}(x)$ and $p^B_{\text{data}}(x)$, respectively, while staying close to each other. This objective can be interpreted as tilting the distribution $p^A_{\text{data}}(x)$ to be close to a sample $x^B(x) \sim p^B_{\text{data}}(x)$, and vice versa. We introduce a coupling

Algorithm 1 Coupled DDPM Sampling

θ_{2D}: Text2Image diffusion model
θ_{MV}: Text2MultiView diffusion model
$x_{T,2D}, x_{T,MV} \sim \mathcal{N}(0, I)$: initial latents
$x_{T,2D}, x_{T,MV}$ shapes: $N \times H \times W \times C$ where N is # of views
for $t \in T, \ldots, 0$ **do**
  x_0 prediction:
  $\hat{x}_{0,MV} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}}(x_{t,MV} - \sqrt{1-\bar{\alpha}_t}\epsilon_{\theta,MV}(x_{t,MV}))$
  $\hat{x}_{0,2D} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}}(x_{t,2D} - \sqrt{1-\bar{\alpha}_t}\epsilon_{\theta,2D}(x_{t,2D}))$
  DDPM step:
  $\hat{x}_{t-1,MV} \leftarrow \sqrt{\bar{\alpha}_{t-1,MV}}\hat{x}_{MV,0} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_{t,MV}) + \sigma_t z$
  $\hat{x}_{t-1,2D} \leftarrow \sqrt{\bar{\alpha}_{t-1,MV}}\hat{x}_{MV,0} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_{t,2D}) + \sigma_t z$
  Coupled guidance step:
  $x_{t-1,2D} \leftarrow \hat{x}_{t-1,2D} - \sqrt{1-\bar{\alpha}_{t-1}}\lambda(\hat{x}_{0,2D} - \hat{x}_{0,MV})$
  $x_{t-1,MV} \leftarrow \hat{x}_{t-1,MV} - \sqrt{1-\bar{\alpha}_{t-1}}\lambda(\hat{x}_{0,MV} - \hat{x}_{0,2D})$
end for

function $U : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that measures the closeness of two samples. A natural choice is the Euclidean Distance, and in this work, we use $U(x, x') = -\frac{\lambda}{2}\|x - x'\|_2^2$ with a constant coefficient $\lambda \in \mathbb{R}$. Formally, our objective is written as

$$\max_{x^A, x^B} \mathcal{J}^A(x^A, x^B) + \mathcal{J}^B(x^A, x^B), \quad \text{where} \tag{4}$$

$$\mathcal{J}^A(x; x') := p^A_{\text{data}}(x) \exp U(x, \text{sg}(x')), \tag{5}$$

$$\mathcal{J}^B(x; x') := p^B_{\text{data}}(x) \exp U(\text{sg}(x), x'), \tag{6}$$

where sg denotes stop grad. Taking the gradients:

$$\nabla_x \mathcal{J}^i(x, x') = \nabla_x \log p^i(x) + \nabla_x U(x, x'), \tag{7}$$

for $i \in \{A, B\}$. Here, the additional term $\nabla_x U(x, x')$ biases the sample trajectory $\{x^i_t\}_t$ from the standard diffusion trajectory following $p^i(x)$ to satisfy the goal. Tilting diffusion model sampling towards inference-time reward functions or constraints has been widely studied for preference alignment [53] and inverse problems [10, 11], with gradient likelihood of a form similar to Eq. (7), although typically under a fixed target. In contrast, in this work, the optimization target depends on another variable. The coefficient λ is a hyperparameter dictating the coupled guidance strength. A value of zero corresponds to standard DDPM sampling, and increasing it strengthens the coupling between the two models. Similar to prior guidance-based methods [4, 12, 18], applying too strong a guidance can cause the sampling process to collapse and produce out-of-distribution results.

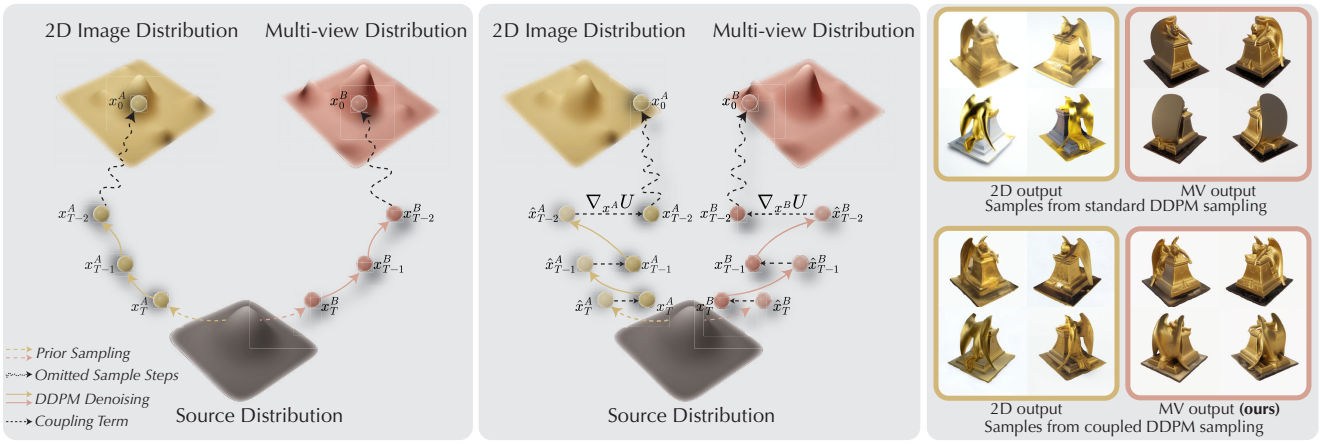
Algorithm. Let $x^A_t, x^B_t \in \mathbb{R}^d$ be two data samples to couple, and denote $\hat{x}^A_{t-1}, \hat{x}^B_{t-1}$ to be the denoised output in DDPM using Eq. (3). We compute the coupled samples by:

$$x^A_{t-1} = \hat{x}^A_{t-1} + \sqrt{1-\bar{\alpha}_{t-1}}\nabla_{\hat{x}^A_0} U(\hat{x}^A_0, \hat{x}^B_0) \tag{8}$$

$$x^B_{t-1} = \hat{x}^B_{t-1} + \sqrt{1-\bar{\alpha}_{t-1}}\nabla_{\hat{x}^B_0} U(\hat{x}^B_0, \hat{x}^A_0). \tag{9}$$

Notice that by fixing $\hat{x}^B_0$,

$$\exp U(\hat{x}^A_0, \hat{x}^B_0) \propto \exp -\frac{1}{2}\frac{\|\hat{x}^A_0 - \hat{x}^B_0\|_2^2}{1/\lambda} = \mathcal{N}(\hat{x}^B_0, 1/\sqrt{\lambda} I) \tag{10}$$



(a) Standard DDPM Sampling (b) Coupled DDPM Sampling (c) Samples from Two Methods

Figure 3. **Overview of the proposed coupled sampling method.** Given two target statistical distributions modeled with diffusion models: (a) standard DDPM sampling generates two instances independently, using scores from each distribution, which leads to samples without spatial alignment; (b) in contrast, the proposed coupled DDPM sampling introduces coupling terms ∇U that pull the two sample paths together, producing spatially and semantically aligned outputs; and (c) as illustrated, the 2D samples get increasingly consistent with coupling, and the coupled multi-view samples (which we set to be our final output) are faithful to the input views.

we find the interpretation that $\exp U(\hat{x}_0^A, \hat{x}_0^B)$ assigns low energy to $\hat{x}_0^A$ close to $\hat{x}_0^B$ in during the sampling process, and similarly for x_t^B. This term effectively serves as a soft regularization that encourages two samples to stay close. The gradient term $\nabla_x U(x, x') = -\lambda(x - x')$ is easy to compute with minimal computation overhead. The sampling algorithm is summarized in Algorithm 1. Our choice of scaling the guidance by $\sqrt{1-\bar{\alpha}_{t-1}}$ is to limit the KL divergence between the training and inference distribution across timesteps. More details can be found in the appendix.

Coupled sampling for multi-view editing. Given a collection of N posed multi-view images $\{I_k, P_k\}_{k=1}^N$, we condition the 2D editing model on each image independently I_k. On the other hand, we condition the multi-view model on a single edited image $I_{1,\text{edited}}$ and on the poses of the remaining views $\{P_k\}_{k=2}^N$ to synthesize as novel views. We perform the coupling between the 2D latent conditioned on image I_k, and the multi-view latent associated with its corresponding pose P_k. In Fig. 3, we illustrate our sampling process and show how the samples of each model change with coupling. We set the coupled multi-view sample as our final output, since we want it to lie within the multi-view distribution.

4. Experiments

To demonstrate the versatility of our method, we select tasks that highlight various editing aspects. 1) *Spatial editing*: We use Magic Fixup [2] to highlight the ability of making geometric changes in a scene. 2) *Stylization*: We perform stylization using Control-Net [60] with edge control, demonstrating how we can alter the general appearance of the input while preserving its overall shape. 3) *Relighting*: We perform relighting using Neural-Gaffer [20], which takes an explicit environment map as input. For each of these tasks, we begin with a collection of input images and additional task-specific conditioning. The 2D model edits each image individually, but cannot enforce consistency across multiple images. In contrast, the multi-view model [63] is a novel view synthesis model that takes a set of consistent images and generates novel views. Our pipeline is first edits a single image using the 2D model and then uses it as a reference for the multi-view model. Then, to enforce faithfulness to the remaining input views, we couple the two models, enabling the multi-view model to maintain the input identity while ensuring consistency across views. We perform the coupling in the latent space, and both the 2D models and the multi-view model in this section operate in the latent space of Stable Diffusion 2.1[44].

Baseline selection. For each task, we adopt Liu *et al*. [31] and Du *et al*. [13] as general-purpose baselines, as they allow composing two diffusion models, and can be directly compared against our framework. We further adopt strong task-specific baselines tailored to each scenario. We also compare against baselines of using the 2D editing model per image and of editing a single image, followed by image-to-MV generation.

User study. To provide a comprehensive evaluation, we conduct user studies using Prolific with 25 participants for each task, comparing our approach to all baselines using best-of-n preference questions. In each task, we conduct 9 comparisons with different scenes and edits.

4.1. Multi-view spatial editing

Spatial editing is challenging because it requires accurately harmonizing the scene, including object interactions and changes in shadows and reflections resulting from edits. No

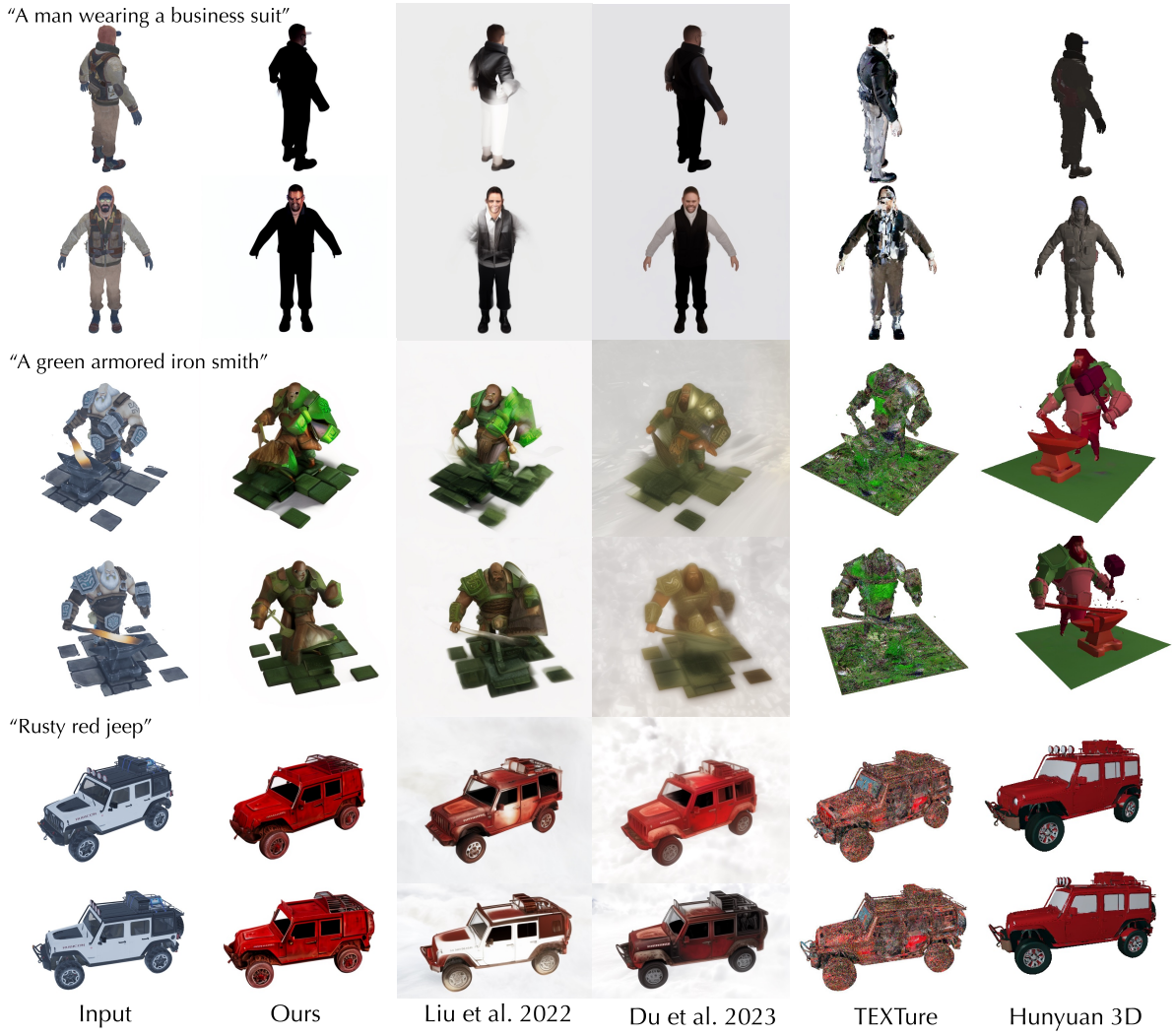


Figure 4. **Multi-view stylization.** We show three examples of multi-view stylization of our method against the baselines. Prior work on combining diffusion models [13, 31] suffer from inconsistencies across frames. When provided with a GT mesh, SDS based methods [43] suffer from severe artifacts while Hunyuan 3D's results follow the prompt loosely when doing retexturing.

Table 1. Quantitative comparison on spatial editing. We evaluate against GT renders of the target, and use MEt3r for consistency.

	Per-image metrics			MV metric	
Method	PSNR ↑	SSIM ↑	LPIPS ↓	MEt3r ↓	Users ↑
Per-image [2]	16.5	0.550	0.253	0.353	-
Image-to-MV [63]	12.84	0.400	0.556	0.417	-
Liu *et al.* [31]	16.5	0.530	**0.354**	0.368	9%
Du *et al.* [13]	16.7	0.548	0.411	0.344	1%
SDEdit [36]	15.4	0.458	0.468	0.393	11%
Coupled Sampling (Ours)	**17.0**	**0.550**	0.421	**0.335**	**80%**

Table 2. Quantitative comparison on stylization. We evaluate the temporal and subject consistency, and MEt3r score for geometric consistency. CLIP score is computed against the edit prompt. The best metrics across methods equivalent to ours are bolded.

	Per-img metric	MV metrics				
Method	CLIP ↑	Temporal ↑	Subject. ↑	MEt3r ↓	Users ↑	Mesh-free
Per-image [60]	30.0	0.922	0.740	0.546	-	✓
Image-to-MV [63]	29.5	0.927	0.787	0.382	-	✓
TEXTure [43]	28.4	0.967	0.748	0.426	14%	X
Hunyuan3D [47]	29.9	0.952	0.754	0.391	8%	X
Liu et al. [31]	30.1	0.934	0.759	0.461	19%	✓
Du et al. [13]	**30.2**	0.926	0.762	0.461	12%	✓
Coupled Sampling (Ours)	29.68	**0.946**	**0.807**	**0.392**	**47%**	✓

large-scale datasets are available for training spatial editing models. As a result, previous work on 2D spatial editing has relied on large-scale video datasets [2, 9, 55] to learn natural object motion. However, such data sources do not exist for multi-view datasets, as 4D datasets are extremely scarce and typically created only for evaluation. Our coupled sampling paradigm addresses this gap.

We use Magic Fixup [2] for the 2D editing model. This model takes the original image and a coarse edit that specifies the desired spatial changes. For multi-view editing, it is necessary to apply the edit consistently across all views. In our experiments, we unproject the target object in each image using a depth map. We then apply a 3D transformation to the object and reproject it into the image. As a baseline, we follow the proposed baseline in Magic Fixup,

Figure 5. **Qualitative comparison on multi-view spatial editing.** The baselines struggle in preserving the identity of the input, and produce flickering artifacts across edited frames, while our results achieve both editing targets and multi-view consistency.

using SDEdit [36], which allows coarse edit inputs. Figure 5 presents three different coarse edits, with two frames from each edit shown to illustrate consistency. In the first example, we find that our method correctly translates and rotates the car while preserving the identity of the input. By contrast, the baselines struggle to maintain the back view of the scene. In the final edit, our method produces smooth shadows that match the ground truth, whereas the baseline results in highly irregular shadows.

To quantitatively evaluate performance, we construct three synthetic scenes with a total of 10 spatial edits and render the ground-truth 3D transformations in Blender. Note that while many baselines exist for per-image spatial editing [9, 40, 45], we use Magic Fixup as the per-image baseline since it is the base model we use for coupling. We use standard reconstruction metrics, and MEt3r [3], which measures the 3D consistency of multi-view outputs. Table 1 demonstrates that our method achieves higher PSNR and SSIM scores, along with superior multi-view consistency.

4.2. Multi-view Stylization

Stylization is a common application of diffusion models: an input sequence provides the desired spatial structure, and a text prompt specifies the style. Control-Net [60] enables this type of stylization by incorporating geometry-related conditioning, such as the Canny edges of an image. A closely related task is 3D re-texturing, in which a 3D mesh is given, and a new texture is generated using a generative model. To assess our method, we rendered ten different objects and applied stylization to each

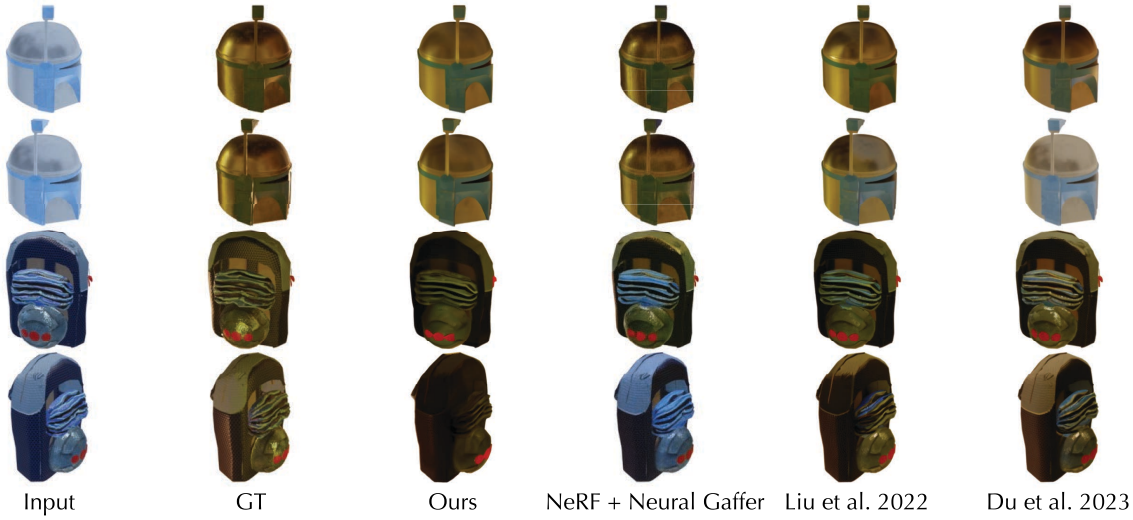


Figure 6. **Qualitative comparison on environment map based relighting.** Other methods tend to produce flickering artifacts (notice the change in color in the first two rows for [13, 31]), and using NeRF makes the lighting changes to be baked into the view dependent effects. Our method achieves the best overall result.

using user-defined prompts. While many stylization baselines exist [27, 42, 46, 64], we select two strong baselines that operate directly on ground truth 3D mesh, such as TEXTure [43], which synthesizes new textures using SDS [41], and Hunyuan3D [47], which employs a SoTA feed-forward multi-view model to generate textures. We omit Instruct-NeRF2NeRF as it fails on our inputs. Note that both baselines access the GT mesh, a highly privileged input that our method does not have access to. In Fig. 4, we present results from three representative examples. In the first example, score-averaging methods have difficulty preserving the identity of the edited subject, leading to color changes or changes in identity across frames. In contrast, TEXTure exhibits severe artifacts due to its SDS-based approach. Hunyuan3D produces very simple edits that often do not align with the text prompt. Prior work on composing diffusion models [13, 31] relies on the average scores of the two models, which pushes the samples off the multi-view manifold and causes flickering. On the other hand, our approach relies on softly steering the multi-view model to remain within the distribution and produce consistent results.

Although the quantitative evaluation of stylization remains challenging, we assess temporal and subject consistency in our generated videos using VBench [59] and measure geometric consistency with MEt3r [3], as shown in Table 2. Our results show that our method achieves superior temporal and subject consistency compared to previous approaches for combining diffusion models. For reference, we also report results from mesh-based methods on rendered videos, which are inherently temporally consistent due to the underlying mesh representation.

4.3. Multi-view Relighting

When the variance of the 2D diffusion results is low, meaning the sampling distribution is narrow, radiance fields can effectively regularize inconsistencies. However, this requires obtaining a consistent geometry beforehand. As an alternative, we demonstrate that a multi-view diffusion model can regularize inconsistencies in 2D relighting through coupled sampling. Figure 6 presents two relighting examples to illustrate this. We observe that prior methods for combining diffusion models [13, 31] can introduce flickering artifacts, as evidenced by abrupt color changes in the top two rows. In contrast, NeRF-based approaches (which we provide with privileged GT NeRF) may incorrectly attribute lighting variance to view-dependent effects, as illustrated in the bottom two rows of the backpack example. To quantitatively compare these methods, we use the 3D objects from Neural-Gaffer [20] and add both a diffuse and a glossy object, resulting in a total of 7 objects with 5 relightings each, for 35 comparisons in total. We compute per-image reconstruction metrics and geometric consistency using MEt3r, as shown in Table 3. Although these metrics do not capture subtle lighting flicker, our method achieves competitive results in both reconstruction and consistency. Importantly, we also report metrics for relighting each image individually, which serves as a coarse upper bound, and observe no degradation in performance.

Table 3. Quantitative comparison on relighting. We evaluate against GT relighting results in terms of per-image metrics, and evaluate multi-view consistency with MEt3r.

	Per-image metrics			MV metric	
Method	PSNR ↑	SSIM ↑	LPIPS ↓	MEt3r ↓	Users ↑
Per-image [20]	22.7	0.862	0.159	0.243	-
Image-to-MV [63]	19.3	0.815	0.193	0.229	-
Liu *et al.* [31]	**23.2**	**0.871**	**0.152**	0.220	10%
Du *et al.* [13]	22.1	0.863	0.158	**0.217**	19%
GT NeRF + Neural Gaffer [20]	22.4	0.865	0.162	**0.217**	25%
Coupled Sampling (Ours)	**23.2**	0.868	0.157	**0.217**	**46%**

5. Analysis Experiments

In this section, we demonstrate the effects of coupled sampling beyond multi-view models, and analyze the effects of varying the coupling strength. In the appendix, we in-

Figure 7. **Image space coupling.** Using Flux, we perform coupled sampling on different prompts. We show that the coupled samples are spatially aligned while being faithful to the prompt.

Figure 8. **Video space coupling.** Using Wan [50], we perform coupled sampling on videos conditioned on different prompts, and show that the coupled samples are highly aligned in each frame while being faithful to the prompt. *Refer to the project webpage for full videos.*

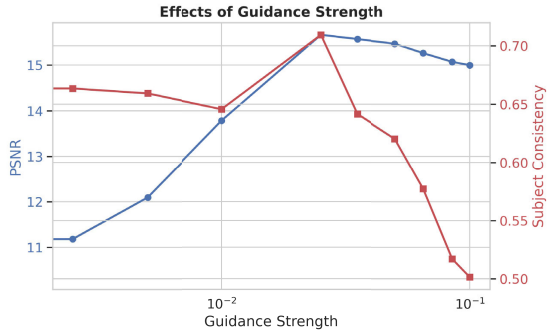


Figure 9. **Guidance strength analysis.** As the guidance strength increases, the reconstruction improves but the consistency drops.

clude experiments on additional backbone variations, and coupling across different latent spaces.

Coupling Text-to-image and video flow models. To illustrate the effects of coupled sampling beyond multi-view models, we implement our method using the text-to-image model Flux [26]. Although Flux is a flow-based model [29, 32], we show that our coupling approach remains effective. We test coupled sampling by generating two samples from the same model, each conditioned on a different prompt. As shown in Fig. 7, without coupling, the outputs are typically very distinct. With the coupled sampling, the outputs become spatially aligned while still reflecting their respective prompts. We further illustrate the effects of coupling with the video model Wan2.1 [50] in Fig. 8 and show the close alignment of coupled samples.

Guidance strength analysis. In Fig. 9 we quantitatively evaluate the effects of guidance strength λ on spatial editing performance. When λ is very small, the model output resembles image-to-MV sampling, resulting in low reconstruction performance, and as we increase λ, the reconstruction improves. However, when making the coupling strength too large, consistency across frames degrades and the performance eventually collapses. We highlight that there is a regime for the guidance strength, where both the reconstruction and consistency produce optimal results, emphasizing the value of coupled sampling.

6. Discussion and Conclusion

We introduce a simple and effective approach for coupling diffusion models, enabling 2D diffusion models to generate consistent multi-view edits when used with multi-view diffusion models. Our method is efficient, versatile, and achieves high-quality results. By guiding the diffusion sampling process, our approach produces outputs that retain the strengths of the underlying models, while also naturally inheriting their limitations.

We believe this coupling strategy has potential applications beyond multi-view editing. In the future, our paradigm could extend the capabilities of image-editing models to video editing by integrating with video diffusion models, without incurring additional computational overhead.

Acknowledgments. We would like to thank Gordon Wetzstein, Jon Barron, Ben Poole, Michael Gharbi, and Songwei Ge for the fruitful discussions. This work is in part supported by NSF RI #2211258, ONR MURI N00014-22-1-2740, the Stanford Institute for Human-Centered AI (HAI), the Magic Grant from the Brown Institute for Media Innovation, and Google Research Scholar Award.

References

[1] Hadi Alzayer, Philipp Henzler, Jonathan T. Barron, Jia-Bin Huang, Pratul P. Srinivasan, and Dor Verbin. Generative multiview relighting for 3d reconstruction under extreme illumination variation. In *CVPR*, 2025. 2

[2] Hadi Alzayer, Zhihao Xia, Xuaner (Cecilia) Zhang, Eli Shechtman, Jia-Bin Huang, and Michael Gharbi. Magic fixup: Streamlining photo editing by watching dynamic videos. *ACM Trans. Graph.*, 2025. 1, 4, 5

[3] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *CVPR*, 2025. 6, 7

[4] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *ICLR*, 2024. 2, 3

[5] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: fusing diffusion paths for controlled image generation. In *Int. Conf. Mach. Learn.*, 2023. 3

[6] Amir Barda, Matheus Gadelha, Vladimir G. Kim, Noam Aigerman, Amit H. Bermano, and Thibault Groueix. Instant3dit: Multiview inpainting for fast editing of 3d objects. In *CVPR*, 2025. 2

[7] Haoming Cai, Tsung-Wei Huang, Shiv Gehlot, Brandon Y Feng, Sachin Shah, Guan-Ming Su, and Christopher Metzler. Parametric shadow control for portrait generation in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18207–18217, 2025. 2

[8] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 2

[9] Yen-Chi Cheng, Krishna Kumar Singh, Jae Shin Yoon, Alexander Schwing, Liangyan Gui, Matheus Gadelha, Paul Guerrero, and Nanxuan Zhao. 3D-Fixup: Advancing Photo Editing with 3D Priors. In *Proceedings of the SIGGRAPH Conference Papers*, 2025. 5, 6

[10] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *NeurIPS*, 2022. 3

[11] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*, 2023. 2, 3

[12] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021. 2, 3

[13] Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Grathwohl. Reduce, reuse, recycle: compositional generation with energy-based diffusion models and mcmc. In *Int. Conf. Mach. Learn.*, 2023. 2, 3, 4, 5, 7

[14] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. In *NeurIPS*, 2024. 1, 2

[15] Daniel Geng and Andrew Owens. Motion guidance: Diffusion-based image editing with differentiable motion estimators. In *ICLR*, 2024. 2

[16] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *ICCV*, 2023. 1, 2

[17] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8): 1771–1800, 2002. 3

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

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 3

[20] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. In *NeurIPS*, 2024. 1, 4, 7

[21] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, 2022. 2

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[23] Jaihoon Kim, Juil Koo, Kyeongmin Yeo, and Minhyuk Sung. Synctweedies: A general generative framework based on synchronized diffusions. In *NeurIPS*, 2024. 3

[24] Jaihoon Kim, Taehoon Yoon, Jisung Hwang, and Minhyuk Sung. Inference-time scaling for flow models via stochastic generation and rollover budget forcing. In *NeurIPS*, 2025. 2

[25] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *ICCV*, 2025. 2

[26] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 8

[27] Peng Li, Suizhi Ma, Jialiang Chen, Yuan Liu, Congyi Zhang, Wei Xue, Wenhan Luo, Alla Sheffer, Wenping Wang, and Yike Guo. Cmd: Controllable multiview diffusion for 3d editing and progressive generation. In *ACM SIGGRAPH*, 2025. 2, 7

[28] Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, et al. Derivative-free guidance

in continuous and discrete diffusion models with soft value-based decoding. *arXiv preprint arXiv:2408.08252*, 2024. 2

[29] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023. 8

[30] Yehonathan Litman, Fernando De la Torre, and Shubham Tulsiani. Lightswitch: Multi-view relighting with material-guided diffusion. In *ICCV*, 2025. 2

[31] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 2, 3, 4, 5, 7

[32] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 8

[33] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. In *CVPR*, 2025. 2

[34] Nadav Magar, Amir Hertz, Eric Tabellion, Yael Pritch, Alex Rav-Acha, Ariel Shamir, and Yedid Hoshen. Lightlab: Controlling light sources in images with diffusion models. In *SIGGRAPH*, 2025. 1

[35] David McAllister, Songwei Ge, Jia-Bin Huang, David W. Jacobs, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Rethinking score distillation as a bridge between image distributions. In *NeurIPS*, 2024. 2

[36] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 5, 6

[37] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[38] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 2

[39] Jiteng Mu, Michaël Gharbi, Richard Zhang, Eli Shechtman, Nuno Vasconcelos, Xiaolong Wang, and Taesung Park. Editable image elements for controllable synthesis. In *ECCV*, 2024. 1

[40] Karran Pandey, Paul Guerrero, Metheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J. Mitra. Diffusion handles: Enabling 3d edits for diffusion models by lifting activations to 3d. In *CVPR*, 2024. 2, 6

[41] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 1, 2, 7

[42] Zhangyang Qi, Yunhan Yang, Mengchen Zhang, Long Xing, Xiaoyang Wu, Tong Wu, Dahua Lin, Xihui Liu, Jiaqi Wang, and Hengshuang Zhao. Tailor3d: Customized 3d assets editing and generation with dual-side images, 2024. 7

[43] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH Conference Proceedings*, 2023. 5, 7

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[45] Rahul Sajnani, Jeroen Vanbaar, Jie Min, Kapil D Katyal, and Srinath Sridhar. Geodiffuser: Geometry-based image editing with diffusion models. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2025. 2, 6

[46] Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. Vox-e: Text-guided voxel editing of 3d objects. In *ICCV*, 2023. 2, 7

[47] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025. 5, 7

[48] Alex Trevithick, Roni Paiss, Philipp Henzler, Dor Verbin, Rundi Wu, Hadi Alzayer, Ruiqi Gao, Ben Poole, Jonathan T. Barron, Aleksander Holynski, Ravi Ramamoorthi, and Pratul P. Srinivasan. Simvs: Simulating world inconsistencies for robust view synthesis. In *CVPR*, 2025. 2

[49] Vaibhav Vavilala, Seemandhar Jain, Rahul Vasanth, D. A. Forsyth, and Anand Bhattad. Generative blocks world: Moving things around in pictures, 2025. 1

[50] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 8

[51] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *ICLR*, 2023. 2

[52] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023. 2

[53] Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *NeurIPS*, 2023. 3

[54] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. Reconfusion: 3d reconstruction with diffusion priors. In *CVPR*, 2024. 2

[55] Ziyi Wu, Yulia Rubanova, Rishabh Kabra, Drew A. Hudson, Igor Gilitschenski, Yusuf Aytar, Sjoerd van Steenkiste, Kelsey R Allen, and Thomas Kipf. Neural assets: 3d-aware

multi-object scene synthesis with image diffusion models. In *NeurIPS*, 2024. 1, 5

[56] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *CVPR*, 2025. 2

[57] Runjie Yan, Yinbo Chen, and Xiaolong Wang. Consistent flow distillation for text-to-3d generation. In *ICLR*, 2025. 2

[58] Zixin Yin, Ling-Hao Chen, Lionel Ni, and Xili Dai. Consistedit: Highly consistent and precise training-free visual editing. In *SIGGRAPH Asia 2025 Conference Papers*, 2025. 2

[59] Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. *arXiv preprint arXiv:2412.09645*, 2024. 7

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 4, 5, 6

[61] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *ICLR*, 2025. 1

[62] Yunzhi Zhang, Carson Murtuza-Lanier, Zizhang Li, Yilun Du, and Jiajun Wu. Product of experts for visual generation. *arXiv preprint arXiv:2506.08894*, 2025. 3

[63] Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *ICCV*, 2025. 1, 2, 4, 5, 7

[64] Jingyu Zhuang, Di Kang, Yan-Pei Cao, Guanbin Li, Liang Lin, and Ying Shan. Tip-editor: An accurate 3d editor following both text-prompts and image-prompts. In *ACM SIGGRAPH*, 2024. 7