

SyncSDE: A Probabilistic Framework for Diffusion Synchronization

Hyunjun Lee^{1*}Hyunsoo Lee^{1*}Sookwan Han^{1,2†}¹ECE, Seoul National University²Republic of Korea Air Force

{hj11013, philip21, jellyheadandrew}@snu.ac.kr

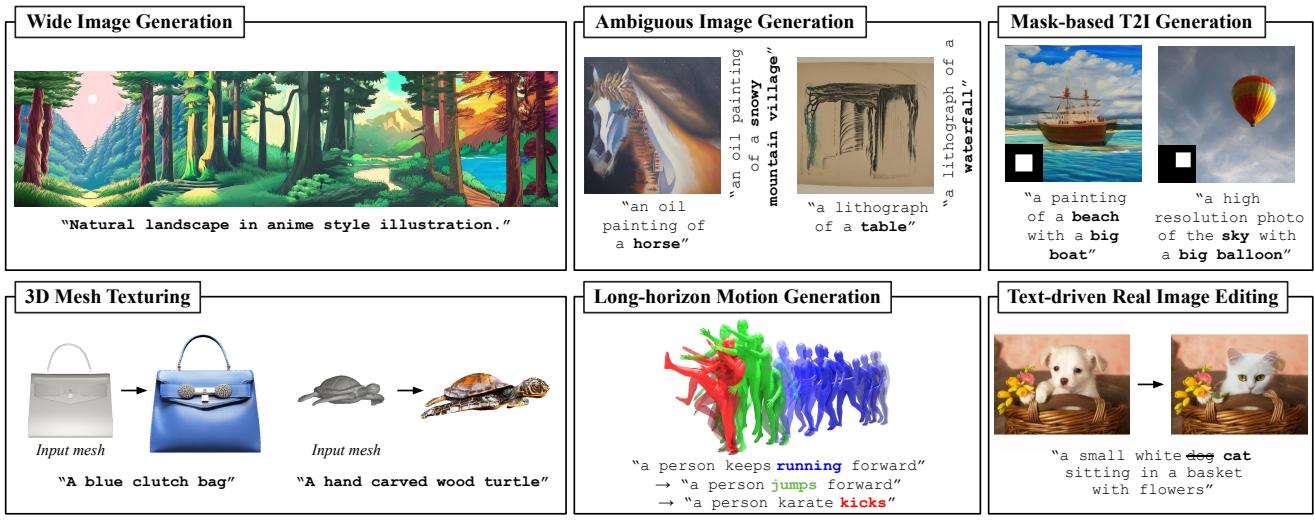


Figure 1. **Applications of SyncSDE.** SyncSDE analyzes diffusion synchronization to identify where the correlation strategies should be focused, enabling coherent and high-quality results across diverse collaborative generation tasks.

Abstract

There have been many attempts to leverage multiple diffusion models for collaborative generation, extending beyond the original domain. A prominent approach involves synchronizing multiple diffusion trajectories by mixing the estimated scores to artificially correlate the generation processes. However, existing methods rely on naive heuristics, such as averaging, without considering task specificity. These approaches do not clarify why such methods work and often fail when a heuristic suitable for one task is blindly applied to others. In this paper, we present a probabilistic framework for analyzing why diffusion synchronization works and reveal where heuristics should be focused—modeling correlations between multiple trajectories and adapting them to each specific task. We further identify optimal correlation models per task, achieving better results than previous approaches that apply a single heuristic across all tasks without justification.

1. Introduction

Diffusion models [15, 16, 40, 48] have achieved remarkable success in generating high-quality images [33, 36, 38], 3D scenes [26, 43, 52], human and motion [12, 19, 20, 44, 45, 51], and videos [17, 24, 49]. Despite their success, these models are typically trained on fixed-domain data, limiting their ability to generate diverse data formats (e.g., varying shapes or dimensions). This constraint reduces the flexibility of diffusion models and narrows the range of generative tasks they can effectively handle.

To harness the diverse generative capabilities of multiple diffusion models with varying characteristics, existing approaches [9, 11, 18, 30] use heuristics to synchronize diffusion trajectories, ensuring consistency across the generations managed by each trajectory. For instance, Visual Anagrams [9] generates images with optical illusions from different perspectives, while SyncTweedies [18] explores multiple heuristics to align generation paths, enabling the creation of panoramic images and even 3D textures.

Although previous approaches with diverse synchronization heuristics demonstrate promising results in collaborative

*Equal Contributions.

†Project Lead.

generation tasks, they do not explain why synchronization works, relying solely on empirical evidence. This lack of theoretical grounding limits both inter-task and intra-task generalizability, leading to inconsistent performance across tasks. As a result, users must experiment extensively to find optimal synchronization strategies for each new task, hindering the scalability of using multiple diffusion models beyond familiar scenarios. For instance, SyncTweedies [18] tests 60 different synchronization strategies to approximate optimal results for given tasks. Repeating this process for each new compositional generation task would severely limit the practical use of multiple diffusion models—especially without theoretical support to validate whether the results are truly optimal.

Our paper addresses the *why* behind synchronization by introducing a probabilistic framework that formulates it as the optimization of two distinct terms. In particular, one term models the correlation between diffusion trajectories, providing a foundation for applying human heuristics as strategic choices. Supported by theoretical analysis, we investigate which synchronization strategies yield the best results across both existing and novel tasks, showing that naive application of heuristics often leads to suboptimal outcomes. This work is the first to analyze *why* synchronization works and to leverage this understanding to guide *where* strategy selection should be focused for future tasks. We demonstrate scalability by applying our method to a wide range of tasks and show that, while naive strategies frequently fall short, our approach consistently achieves superior results. We refer to this method as **SyncSDE**: Synchronization of Stochastic Differential Equations. The main contributions of our work are summarized as follows:

- We introduce a probabilistic framework for diffusion synchronization, providing a theoretical foundation to understand *why* synchronization works.
- Our approach reduces redundant empirical testing by identifying *where* heuristics should be applied, mitigating the suboptimal outcomes of existing naive strategies.
- Extensive experiments across diverse diffusion synchronization tasks demonstrate the superior performance of our method over state-of-the-art baselines, reinforcing its generalizability to novel tasks.

2. Related work

Recent advances in diffusion models have unlocked a wide range of applications, powered by foundation models such as Stable Diffusion [36], DeepFloyd [1], ControlNet [53]. Numerous studies build on these pretrained models to tackle specific tasks, including compositional generation.

Diffusion Models. Diffusion models generate realistic images by progressively denoising Gaussian noise. DDPM [15] and DDIM [40] implement this process via discrete sampling, which is known to approximate stochastic differential equa-

tions (SDEs) [48], forming the theoretical basis for diffusion models. Latent diffusion models [36] improve efficiency by operating in latent space, with Stable Diffusion being the most widely used. Pixel-based models like DeepFloyd [1] are also gaining attention. Beyond image synthesis, diffusion models are applied to tasks such as image-to-image translation [3, 7, 10, 28, 29, 42], human motion generation [44, 45, 51] where models edit target details while preserving source structure. For instance, Imagic [3] fine-tunes pretrained models for this purpose.

Diffusion Synchronization. Diffusion synchronization enables collaborative generation by synchronously sampling from multiple diffusion trajectories while maintaining consistency across them. It extends the capabilities of a single diffusion model to support tasks such as generating images of arbitrary sizes [18, 30, 47, 54], creating seamless textures [5, 8, 18, 27, 34, 50], producing optical illusions [9, 11, 18], and synthesizing complex motions. By leveraging the prior knowledge encoded in pretrained diffusion models, these methods require no additional training, expanding applicability across diverse domains. For example, SyncTweedies [18] addresses a range of synchronization tasks by empirically testing 60 strategies, ultimately adopting an averaging method based on Tweedie’s formula [41]. Other works target specific tasks: MultiDiffusion [30] focuses on wide image and mask-based T2I generation using bootstrapping for improved localization; Visual Anagram [9] produces ambiguous images that shift with view transformations; and SyncMVD [27] generates UV texture maps from 3D meshes and text prompts. These methods synchronize trajectories by averaging intermediate signals—such as predicted noise or latents—but offer no theoretical justification for why this works. In contrast, we propose a probabilistic framework that explicitly models correlations between diffusion trajectories, providing the first theoretical foundation for diffusion synchronization.

3. Method

3.1. Preliminaries

3.1.1. Diffusion sampling

Starting from Gaussian noise $p_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the DDIM [40] reverse process samples $\mathbf{x}_T \sim p_T$ and then sequentially samples \mathbf{x}_{t-1} from \mathbf{x}_t using the distribution:

$$\begin{aligned} q_\sigma(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) &= \mathcal{N}(\sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1 - \alpha_t}}, \sigma_t^2 \mathbf{I}). \end{aligned} \quad (1)$$

where $\{\alpha_t\}_{t=0}^T$ is a predefined increasing sequence, and σ_t controls the stochasticity of the diffusion trajectory. Since the distribution of \mathbf{x}_t given \mathbf{x}_0 is modeled as $q_\sigma(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1 - \alpha_t)\mathbf{I})$, the ground truth \mathbf{x}_0 can be approxi-

mated using Tweedie's formula [41] as:

$$\hat{\mathbf{x}}_0(\mathbf{x}_t, t) := \frac{\mathbf{x}_t - \sqrt{1 - \alpha_t} \cdot \epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{\alpha_t}}, \quad (2)$$

where $\epsilon_\theta(\cdot, \cdot)$ is a noise prediction network, typically implemented using a U-Net [37] or Transformer architecture [31]. Generally, we set $\sigma_t = 0$ which makes the deterministic DDIM reverse process as

$$\begin{aligned} \mathbf{x}_{t-1} &= \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \mathbf{x}_t + (1 - \alpha_t) \gamma_t \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) \\ &\approx \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \mathbf{x}_t - \sqrt{1 - \alpha_t} \gamma_t \epsilon_\theta(\mathbf{x}_t, t) \end{aligned} \quad (3)$$

where $\gamma_t := \sqrt{\alpha_{t-1}/\alpha_t} - \sqrt{(1 - \alpha_{t-1})/(1 - \alpha_t)}$.

3.1.2. Stochastic differential equation

As shown in Song *et al.* [48], the noise perturbation in DDPM [15] and DDIM [40] can be modeled as a stochastic process governed by a discretized forward SDE [22]:

$$d\mathbf{x} = f(\mathbf{x}, t) dt + g(t) d\bar{w}, \quad (4)$$

which has a corresponding reverse SDE that denotes the reverse process of the diffusion model as follows:

$$d\mathbf{x} = [f(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})] dt + g(t) d\bar{w}. \quad (5)$$

3.1.3. Notations

Before introducing our method, we define the notations used throughout the paper. Let $\{\mathbf{X}_t\}_{t=0}^T$ denote the original objective we aim to generate. We define $\{f_i\}_{i=1}^N$ as a set of mapping functions that project \mathbf{X}_t into N different patches $\{\mathbf{y}_t^i\}_{i=1}^N$, where $\mathbf{y}_t^i := f_i(\mathbf{X}_t)$. Since each \mathbf{y}_t^i has a resolution compatible with the diffusion model, we apply the diffusion process to each patches, resulting in diffusion trajectories $\{\mathbf{y}_t^i\}_{t=0}^T$ for each i .

For example, in the case of wide image generation (Sec. 3.3.3), \mathbf{X}_t corresponds to the wide image itself, and f_i is a cropping function that extracts patch \mathbf{y}_t^i from \mathbf{X}_t . In contrast, for 3D mesh texturing (Sec. 3.3.5), \mathbf{X}_t represents the texture map of an input object, and f_i transforms \mathbf{X}_t into a rendered image of the mesh from a specific viewpoint. Additionally, we define $\tilde{\mathbf{X}}^i := \cup_{j=1}^{i-1} \{\mathbf{y}_t^j\}_{t=1}^T$ and $\tilde{\mathbf{X}}_t^i := \cup_{j=1}^{i-1} \mathbf{y}_t^j$. We elaborate on how the union is defined for each task in the following sections.

3.2. Proposed framework: SyncSDE

We propose a synchronous generation process that sequentially generates the trajectories of $\{\mathbf{y}_t^i\}_{t=0}^T$. First, we generate the trajectory $\{\mathbf{y}_t^1\}_{t=0}^T$. Then, $\{\mathbf{y}_t^2\}_{t=0}^T$ is generated, conditioned on the previously generated trajectory $\{\mathbf{y}_t^1\}_{t=0}^T$. This process continues iteratively, where each trajectory

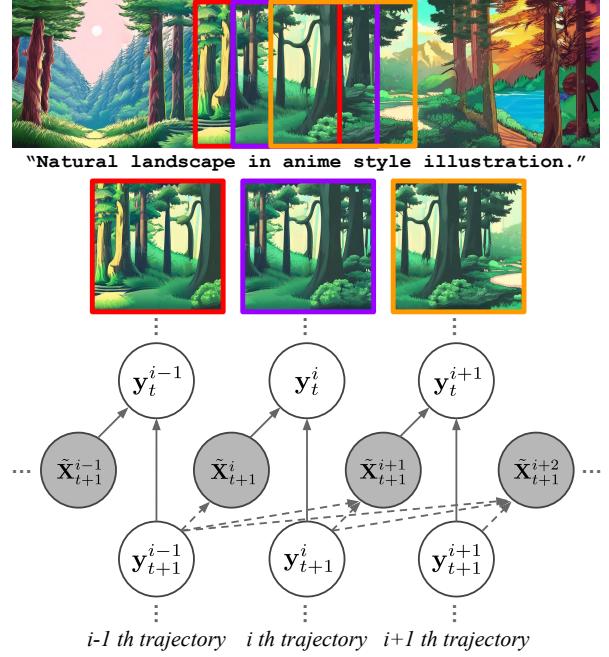


Figure 2. **Graphical diagram of our method.** We sequentially generate views conditioned on previous generations.

$\{\mathbf{y}_t^i\}_{t=0}^T$ is conditioned on the previously generated trajectories $\{\mathbf{y}_t^1\}_{t=0}^T, \{\mathbf{y}_t^2\}_{t=0}^T, \dots, \{\mathbf{y}_t^{i-1}\}_{t=0}^T$. A graphical illustration of the proposed generation process is shown in Fig. 2. We model the relationship between trajectories only at the same timestep, *i.e.*, for all $t_1 \neq t_2$ and $i \neq j$, we assume $\mathbf{y}_{t_1}^i \perp \!\!\! \perp \mathbf{y}_{t_2}^j$. Inspired by the conditional score estimation proposed in [23], we derive the score function for the conditional generation process as follows:

$$\begin{aligned} \nabla_{\mathbf{y}_t^i} \log p(\mathbf{y}_t^i | \tilde{\mathbf{X}}^i) &= \nabla_{\mathbf{y}_t^i} \log p(\mathbf{y}_t^i | \tilde{\mathbf{X}}_t^i) \\ &= \nabla_{\mathbf{y}_t^i} \log p(\mathbf{y}_t^i) + \nabla_{\mathbf{y}_t^i} \log p(\tilde{\mathbf{X}}_t^i | \mathbf{y}_t^i). \end{aligned} \quad (6)$$

The first term of Eq. 6 corresponds to the original score function estimated by the pretrained diffusion model. The second term captures the relationship between generations, which we explicitly model. By substituting Eq. 6 into Eq. 3, the reverse SDE sampling update for the i^{th} view at timestep t is given by:

$$\begin{aligned} \mathbf{y}_{t-1}^i &= \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \mathbf{y}_t^i + (1 - \alpha_t) \gamma_t \nabla_{\mathbf{y}_t^i} \log p(\mathbf{y}_t^i | \tilde{\mathbf{X}}_t^i) \\ &\approx \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \mathbf{y}_t^i - \sqrt{1 - \alpha_t} \gamma_t \epsilon_\theta(\mathbf{y}_t^i, t) \\ &\quad + (1 - \alpha_t) \gamma_t \nabla_{\mathbf{y}_t^i} \log p(\tilde{\mathbf{X}}_t^i | \mathbf{y}_t^i). \end{aligned} \quad (7)$$

Note that we focus on modeling the heuristic to estimate $\nabla_{\mathbf{y}_t^i} \log p(\tilde{\mathbf{X}}_t^i | \mathbf{y}_t^i)$. Our approach significantly reduces the search space for empirical testing when identifying optimal diffusion synchronization strategies. Following the setup

in [23], we model this conditional probability term for each task using a single tunable hyperparameter, λ .

3.3. Applications of SyncSDE

We explore a range of collaborative generation tasks to demonstrate the applicability of SyncSDE, including mask-based text-to-image generation, text-driven real image editing, wide image generation, ambiguous image generation, 3D mesh texturing, and long-horizon motion generation.

3.3.1. Mask-based Text-to-Image generation

For mask-based text-to-image generation, we use three prompts: p^{bg} for the background, p^{fg} for the masked region, and p^{img} for the overall image that semantically integrates p^{bg} and p^{fg} . We generate the final image using three variables: \mathbf{y}_t^1 for the background conditioned on p^{bg} , \mathbf{y}_t^2 for the masked region using p^{fg} , and \mathbf{y}_t^3 to refine the entire image based on p^{img} . Our goal is to generate a high-quality image \mathbf{X} , defined as \mathbf{y}_0^3 . We first generate \mathbf{y}_t^1 , then synthesize \mathbf{y}_t^2 —representing the foreground region—using the conditional score function defined in Eq. 6. The conditional distribution is defined as:

$$p(\tilde{\mathbf{X}}_t^2 \mid \mathbf{y}_t^2) \sim \mathcal{N}(\mathbf{y}_t^2, \lambda(1 - \alpha_t)\mathbf{M}^{-1}), \quad (8)$$

where $\tilde{\mathbf{X}}_t^2$ is equal to \mathbf{y}_t^1 , \mathbf{M} is a diagonal precision matrix indicating the background region of the image, and λ is a tunable hyperparameter. The diagonal elements of \mathbf{M} are constructed as:

$$\text{Diag}(\mathbf{M}) = \text{Reshape}(\mathbf{B}), \quad (9)$$

where $\mathbf{B} \in \mathbb{R}^{H \times W}$ is a binary mask such that $\mathbf{B}[h, w] \in \{0, 1\}$ indicates background (1) and foreground (0) regions. The $\text{Reshape}(\cdot)$ operation flattens the matrix from $\mathbb{R}^{H \times W}$ to a vector in \mathbb{R}^{HW} . The intuition behind Eq. 8 is that foreground pixels require higher variance to allow object formation, while background pixels should maintain lower variance to ensure consistency. Accordingly, we assign smaller variance to the background and larger variance to the foreground. We finally generate \mathbf{y}_t^3 using the following conditional probability:

$$p(\tilde{\mathbf{X}}_t^3 \mid \mathbf{y}_t^3) \sim \mathcal{N}(\mathbf{y}_t^3, \lambda(1 - \alpha_t)\mathbf{M}^{-1}) \cdot \mathcal{N}(\mathbf{y}_t^3, \lambda(1 - \alpha_t)(\mathbf{1} - \mathbf{M})^{-1}), \quad (10)$$

where $\mathbf{1}$ is a diagonal precision matrix with all diagonal elements set to 1. Note that $\tilde{\mathbf{X}}_t^3$ is computed as:

$$\tilde{\mathbf{X}}_t^3 = \mathbf{M} \odot \mathbf{y}_t^1 + (\mathbf{1} - \mathbf{M}) \odot \mathbf{y}_t^2, \quad (11)$$

where \odot is the Hadamard product. We choose $\mathbf{X} = \mathbf{y}_0^3$.

3.3.2. Text-driven real image editing

The text-driven real image editing task requires precise modifications, as it aims to manipulate only the foreground region while preserving the background. Given a source image \mathbf{x}^{src} , a source prompt p^{src} , and a target prompt p^{tgt} , we first invert the source image using the forward SDE [22] to obtain the latent sequence $\{\mathbf{x}_t^{\text{src}}\}_{t=0}^T$. Following CSG [23], we generate a soft mask $\tilde{\mathbf{B}}$ that identifies the background region of the source image using attention maps [46] from the pretrained diffusion model. Details of the soft mask generation process are provided in the Appendix. We then obtain the binary mask \mathbf{B} as follows:

$$\mathbf{B}[h, w] = \chi(\tilde{\mathbf{B}}[h, w] \geq \tau), \quad (12)$$

where χ outputs 1 if the given condition is true, and 0 otherwise, and $\tau \in [0, 1]$ is a threshold for attention values. Following the logic of mask-based T2I generation, we generate the target image \mathbf{x}^{tgt} . We first apply the binary mask obtained from Eq. 12 to Eq. 9 to construct \mathbf{M} . Next, we replace each \mathbf{y}_t^1 with $\mathbf{x}_t^{\text{src}}$, and set $\mathbf{y}_T^3 = \mathbf{x}_T^{\text{src}}$. Finally, using p^{tgt} , we sample $\{\mathbf{y}_t^2\}_{t=0}^T$ and $\{\mathbf{y}_t^3\}_{t=0}^{T-1}$, and obtain the edited image as $\mathbf{x}^{\text{tgt}} = \mathbf{y}_0^3$.

3.3.3. Wide image generation

To generate a wide image, we define the operation f_i as a cropping function that extracts the image patch \mathbf{y}_t^i from the wide image \mathbf{X}_t . The patches $\{\mathbf{y}_0^i\}_{i=1}^N$ are defined to be *partially overlapped*. We then design the conditional probability term as follows:

$$p(\tilde{\mathbf{X}}_t^i \mid \mathbf{y}_t^i) \sim \mathcal{N}(\mathbf{y}_t^i, \lambda(1 - \alpha_t)\mathbf{M}_i^{-1}), \quad (13)$$

where

$$\tilde{\mathbf{X}}_t^i = (\mathbf{1} - \mathbf{M}_i) \odot f_i(f_{i-1}^{-1}(\mathbf{y}_t^{i-1})), \quad (14)$$

and \mathbf{M}_i is a binary mask that indicates the non-overlapping pixels between the i^{th} and $(i-1)^{\text{th}}$ patches. After generating all N patches, we apply an overlapping operation φ to combine them into the initial reconstruction \mathbf{X}_0 :

$$\mathbf{X}_0 = \varphi(\{f_i^{-1}(\mathbf{y}_0^i)\}_{i=1}^N). \quad (15)$$

The operation φ ensures that patches with larger i values are placed on top in overlapping regions. Finally, we decode \mathbf{X}_0 using the VAE decoder [21] of LDM [36] to obtain the final wide image \mathbf{X} .

3.3.4. Ambiguous image generation

An ambiguous image is designed to support multiple interpretations through visual transformations f_i . These transformations include operations such as identity mapping, (counter) clockwise rotation, skewing, and flipping. The specific types of f_i used in our experiments are described in Sec. 4.2.4. We define the conditional probability as:

$$p(\tilde{\mathbf{X}}_t^i \mid \mathbf{y}_t^i) \sim \mathcal{N}(\mathbf{y}_t^i, \lambda(1 - \alpha_t)\mathbf{1}), \quad (16)$$

where $\tilde{\mathbf{X}}_t^i$ is defined as

$$\tilde{\mathbf{X}}_t^i = f_i(f_{i-1}^{-1}(\mathbf{y}_t^{i-1})), \quad (17)$$

and λ is a tunable hyperparameter.

3.3.5. 3D mesh texturing

For 3D mesh texturing, we define the variable \mathbf{y}^i as the projected image of a 3D mesh observed from the i^{th} viewpoint. We then design the conditional probability term as:

$$p(\tilde{\mathbf{X}}_t^i | \mathbf{y}_t^i) \sim \mathcal{N}(\mathbf{y}_t^i, \lambda(1 - \alpha_t)\mathbf{M}_i^{-1}), \quad (18)$$

where λ is a tunable hyperparameter, and $\tilde{\mathbf{X}}_t^i$ is defined as:

$$\tilde{\mathbf{X}}_t^i = f_i(f_{i-1}^{-1}(\{\mathbf{y}_t^j\}_{j=1}^{i-1}))[i]. \quad (19)$$

Here, \mathbf{M}_i is a binary mask indicating the background region of the i^{th} view, generated during the rendering process. The function f_{i-1}^{-1} is an inverse-projection function that composes a texture map from images captured at the first $i-1$ viewpoints, while f_i is a projection function that renders the texture map into i viewpoint-specific images. We take the i^{th} image from the output of f_i to obtain $\tilde{\mathbf{X}}_t^i$.

3.3.6. Long-horizon motion generation

For long-horizon motion generation, we generate short-duration motion segments with MDM [44] with overlapping timestamps to smoothly form an extended, coherent motion sequence. We define the operation f_i as a query function for extracting motion segment \mathbf{y}_t^i from the total motion sequence \mathbf{X}_t . These segments, $\{\mathbf{y}_0^i\}_{i=1}^N$, are constructed to have *partial temporal overlaps*. To achieve coherent transitions, we define the conditional probability as:

$$p(\tilde{\mathbf{X}}_t^i | \mathbf{y}_t^i) \sim \mathcal{N}(\mathbf{y}_t^i, \lambda(1 - \alpha_t)\mathbf{M}_i^{-1}), \quad (20)$$

where

$$\tilde{\mathbf{X}}_t^i = (\mathbf{1} - \mathbf{M}_i) \odot f_i(f_{i-1}^{-1}(\mathbf{y}_t^{i-1})), \quad (21)$$

and \mathbf{M}_i is a binary mask indicating non-overlapping timestamps between the i^{th} and $(i-1)^{\text{th}}$ motion segments. After generating N motion segments, we combine them using an overlapping operation φ , ensuring smooth continuity at overlapping timestamps, as follows:

$$\mathbf{X}_0 = \varphi(\{f_i^{-1}(\mathbf{y}_0^i)\}_{i=1}^N). \quad (22)$$

The operation φ prioritizes later segments (larger i) in regions where timestamps overlap, ensuring temporal consistency in the complete long-horizon motion sequence \mathbf{X} .

4. Experiments

In this section, we qualitatively and quantitatively evaluate the performance of SyncSDE. We compare it against SyncTweedies [18] across tasks presented in

Table 1. **Quantitative results of mask-based T2I generation.** We generate images using the pretrained Stable Diffusion [36]. KID score is scaled by 10^3 .

Method	KID [6] \downarrow	FID [13] \downarrow	CLIP-S [32] \uparrow
MultiDiffusion [30]	47.694	84.225	0.330
SyncTweedies [18]	117.360	149.470	0.307
SyncSDE ($1/\lambda = 5$)	43.774	82.878	0.332
SyncSDE (best)	34.859	72.118	0.331

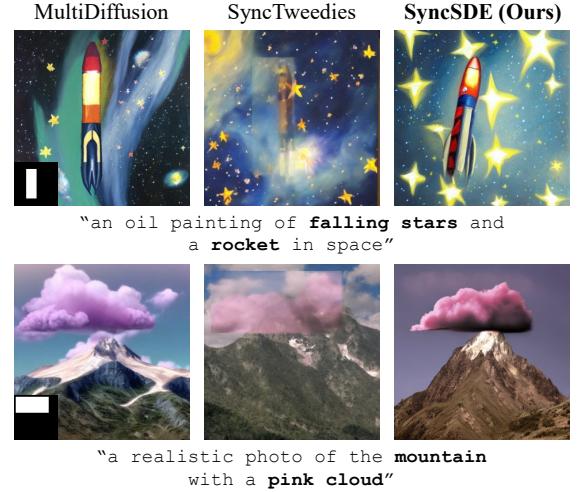


Figure 3. **Qualitative results of mask-based T2I generation.** We find that SyncSDE effectively models the correlation between background and foreground giving comparable result with task specific model, MultiDiffusion [30], while SyncTweedies [18] fails and tends to generate blurry image in the masked region.

Sec. 3.3.1~3.3.5, and against task-specific methods [7, 9, 10, 27, 28, 30, 42] for text-driven image editing in Sec. 3.3.2. We then present results for long-horizon motion generation (Sec. 3.3.6), highlighting the scalability of our approach. Finally, we analyze the impact of the hyperparameter λ .

4.1. Implementation details

We implement our method based on the official codebases of CSG* [23] and SyncTweedies† [18]. All experiments are conducted using the DDIM [40] sampler as the numerical solver for the reverse SDE. For fair comparison, we use the same number of DDIM sampling steps across all methods and tasks. We also apply classifier-free guidance [14] with a consistent guidance scale across all baselines. For ease of implementation, we use $1/\lambda$ in place of λ , and employ a scheduler that linearly decreases $1/\lambda$ as the timestep t decreases. Task-specific details are provided in the following subsections and in the Appendix.

*<https://github.com/Hleephilip/CSG>

†<https://github.com/KAIST-Visual-AI-Group/SyncTweedies>

Table 2. **Quantitative results of text-driven real image editing.** We use real images from LAION-5B dataset [39] and the pretrained Stable Diffusion [36]. SyncSDE shows better performance compared to task-specific methods [7, 10, 28, 42].

Method	CLIP-S [32] \uparrow	LPIPS [55] \downarrow	BG-LPIPS [55] \downarrow
DDIB [42]	0.294	0.379	0.350
SDEdit [28]	0.298	0.407	0.369
PTI [10]	0.322	0.409	0.379
MasaCtrl [7]	0.285	0.290	0.341
SyncSDE ($1/\lambda = 5$)	0.311	0.281	0.266
SyncSDE (best)	0.313	0.254	0.222

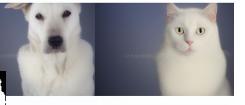
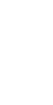
Reconstruction	Edit	Reconstruction	Edit
			
Cat \rightarrow Dog	Dog \rightarrow Cat		
			
Horse \rightarrow Zebra	Zebra \rightarrow Horse		

Figure 4. **Qualitative results of text-driven real image editing.** We visualize the results of various real image editing tasks using the real images sampled from LAION-5B dataset [39]. SyncSDE shows a good performance on text-driven image manipulation.

4.2. Results

We present quantitative and/or qualitative results of the proposed method and baseline algorithms [7, 9, 10, 18, 27, 28, 30, 42] across the collaborative generation tasks discussed in Section 3.3. For SyncSDE, we report results using two different values of λ : one with a fixed setting of $1/\lambda = 5$, which performs well across tasks, and another which is tuned for each task to achieve the best results. In all tables, we highlight the best (■) and second-best (□) results in each column using colored cells. Additional results and implementation details are provided in the Appendix.

4.2.1. Mask-based Text-to-Image generation

We compare the proposed method with SyncTweedies [18] and MultiDiffusion [30] using the pretrained Stable Diffusion [36]. Since the official implementation of SyncTweedies does not include mask-based T2I generation, we modified the code to produce the results. We generate two diffusion trajectories (w_0, w_1), representing the background and masked region, respectively. Final output image z is computed by using w_0 as the background while averaging the masked region with w_1 as follows:

$$z = M \odot w_0 + (1 - M) \odot (w_0 + w_1) \quad (23)$$

where M is the background mask explained in Section 3.3.1, \odot is Hadamard product operator. We follow the notation of

Table 3. **Quantitative results of wide image generation.** We generate wide images using the pretrained Stable Diffusion [36]. Note that KID score is scaled by 10^3 .

Method	KID [6] \downarrow	FID [13] \downarrow	CLIP-S [32] \uparrow
SyncTweedies [18]	51.024	78.333	0.328
SyncSDE ($1/\lambda = 5$)	17.311	44.969	0.324
SyncSDE (best)	16.872	44.707	0.324

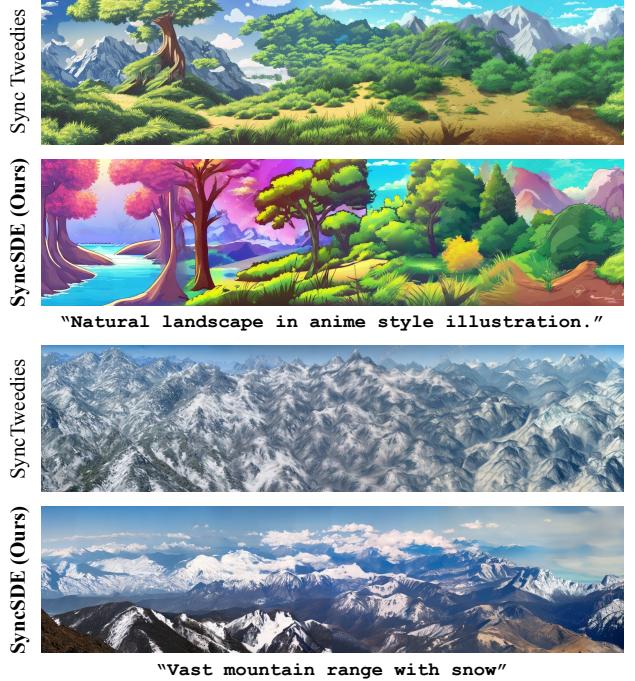


Figure 5. **Qualitative results of wide image generation.** We visualize 2048×512 sized wide image generated by our method and SyncTweedies [18]. SyncSDE generates better quality of wide images in a continuous manner compared to SyncTweedies.

SyncTweedies for variable z and w which are denoted as canonical and instance variable respectively.

For quantitative evaluations, we use KID [6] and FID [13] to quantify the fidelity of the generated images, with CLIP-S [32] to measure the similarity between the generated images and the given text prompts. As shown in Table 1, SyncSDE significantly outperforms SyncTweedies and MultiDiffusion across all three metrics. Also, SyncSDE generates plausible images that seamlessly composites foreground with the background, as illustrated in Figure 3. In contrast, SyncTweedies struggles with localization and synchronization, failing to blend the object into the overall image. While MultiDiffusion relies on an additional bootstrapping strategy specifically for object localization, SyncSDE achieves superior performance without any extra components, highlighting the effectiveness of our method.

Table 4. Quantitative results of ambiguous image generation. We generate ambiguous images using the pretrained Deepfloyd [1]. Note that KID score is scaled by 10^3 .

Method	KID [6] ↓	FID [13] ↓	CLIP-S [32] ↑
Visual Anagrams [9]	195.286	215.082	0.290
SyncTweedies [18]	215.119	226.922	0.262
SyncSDE ($1/\lambda = 5$)	173.590	212.196	0.273
SyncSDE (best)	174.902	208.788	0.272



Figure 6. Qualitative results of ambiguous image generation. We visualize the ambiguous images generated by SyncSDE, SyncTweedies [18], and Visual Anagrams [9]. The first row applies identity and skew transformations, while the second row applies identity and flip transformations. SyncSDE generates realistic images that blends in both prompts, while SyncTweedies fails to integrate two prompts.

4.2.2. Text-driven real image editing

We compare our method with prior works [7, 10, 28, 42] with the pre-training stable diffusion [36]. Since the official implementation of PTI is unavailable, we reproduce it. For comparisons, we sample 1,000 real images from the LAION-5B dataset [39]. We generate the source prompt using the pretrained image captioning model BLIP [25]. Then, the target prompt corresponding to the edited image is generated by swapping the words of the source prompt. We evaluate each methods with CLIP-S score [32] to measure the similarity between the edited image and the target prompt. In addition, we measure LPIPS [55] to quantify the perceptual similarity between the source and the edited image. To further evaluate background preservation, we calculate LPIPS using the background region (BG-LPIPS), which is segmented using the pretrained image segmentation model Detic [56]. As shown in Table 2 and Figure 4, SyncSDE shows superior performance in text-driven real image editing. Note that ‘Reconstruction’ denotes the reconstructed source image y_0^1 , while ‘Edit’ means the edited image y_0^3 .

4.2.3. Wide image generation

We generate 2048×512 resolution wide image using the pretrained Stable Diffusion [36] as backbone. Quantitative results of wide image generation is presented in Table 3. Our method outperforms SyncTweedies [18] in terms of KID [6], FID [13], and CLIP-S score [32]. Additionally, Figure 5

Table 5. Quantitative results of 3D mesh texturing. We generate textures using the pretrained ControlNet [53]. Note that KID score is scaled by 10^3 .

Method	KID [6] ↓	FID [13] ↓	CLIP-S [32] ↑
SyncMVD [27]	196.341	189.268	0.314
SyncTweedies [18]	186.648	183.387	0.311
SyncSDE ($1/\lambda = 5$, best)	184.704	183.180	0.311



Figure 7. Qualitative result of 3D mesh texturing. We qualitatively compare the performance of the proposed method on 3D mesh texturing with SyncMVD [27] and SyncTweedies [18]. SyncSDE gives high-quality textured images corresponding to the given text prompt.

shows that our method produces more plausible and high-fidelity results compared to SyncTweedies. Notably, in the wide image generated with the prompt “Vast mountain range with snow”, SyncTweedies fails to synthesize realistic views, highlighting the limitations of patch averaging.

4.2.4. Ambiguous image generation

We compare our method with SyncTweedies [18] and Visual Anagrams [9]. To generate ambiguous images, we use the pretrained DeepFloyd-IF [1]. We generate a single image using two prompts, modeling two semantics within the image by choosing f_1 and f_2 . We fix f_1 as identity mapping, and choose f_2 from 4 types of transformation explained in Section 3.3.4; (1) $\pm 90^\circ$ rotation, (2) 180° rotation, (3) vertical flip, and (4) skew transformation. Note that skew transformation shifts columns of image pixels with offset.

Table 4 shows that SyncSDE outperforms all baselines across KID [6] and FID [13] scores, and has comparable CLIP-S score [32]. Figure 6 further illustrates that our method generates significantly better results compared to prior works. Especially, SyncTweedies tend to generate blurry images that appear as simple averages of the two different images from each view, rather than plausibly blended images. In some cases, the objects given in two different text prompts become unidentifiable. We claim that this issues is due to the lack of theoretical foundation behind patch averaging. In contrast, SyncSDE generates more coherent images, where views are seamlessly blended and properly correlated. Additionally, we emphasize that a significant

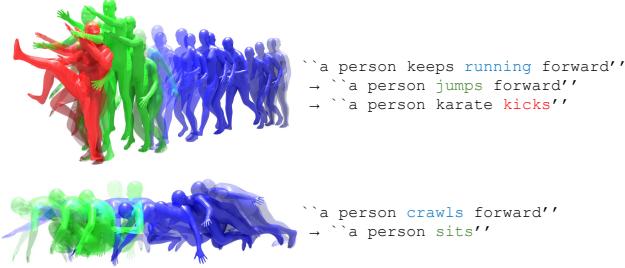


Figure 8. Qualitative results of long-horizon motion generation. Our method generates coherent long-horizon motion sequences by synchronizing multiple output trajectories from a motion diffusion model [44], where each trajectory produces a short motion segment.

benefit of our method is its ability to produce high-quality results compared to Visual Anagrams, despite Visual Anagrams being specifically designed only for ambiguous image generation.

4.2.5. 3D mesh texturing

We compare the proposed method with SyncTweedies [18] and SyncMVD [27]. Note that we utilize the pretrained depth-conditioned ControlNet [53] as backbone architecture. We synchronize ten different diffusion processes, to generate a texture for a given 3D mesh. Eight of the processes correspond to views with azimuth values evenly spaced by 45° within the range $[0^\circ, 360^\circ]$, and two auxiliary trajectories encoding views with azimuth 0° and 180° , both at an elevation of 30° . Each diffusion process is modeled with ControlNet [53] conditioned on depth information extracted from input mesh. The results are then compared based on the projected images of the generated texture map.

Table 5 shows that SyncSDE ($1/\lambda = 5$, best) outperforms prior works in terms of KID [6] and FID [13], and has comparable values of CLIP-S score [36]. As shown in Figure 7, our method qualitatively surpasses the performance of SyncTweedies and SyncMVD, while SyncTweedies tend to blur the details of the texture.

4.2.6. Long-horizon motion generation

We demonstrate the broad applicability of SyncSDE through long-horizon motion generation, as shown in Fig. 8. Specifically, we use the motion diffusion model [44], where each trajectory generates a short human motion sequence of 120 frames. To compose a continuous motion, we set $1/\lambda = 3$ and apply a 0.25 overlap ratio across timesteps (*i.e.*, 30 frames overlap). SyncSDE successfully synchronizes the motion segments, producing a coherent long-horizon sequence with smooth transitions between segments.

4.3. Effects of λ

We analyze the effect of the hyperparameter λ by generating ambiguous images using different values of λ . Theoretically, λ controls the degree of collaboration between mul-

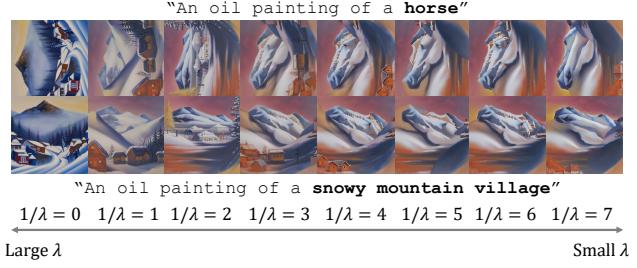


Figure 9. Effects of λ . We show the balance of collaboration between two prompts. When λ is large, the effect of 2nd prompt dominates, while the effect of 1st prompt becomes significant for smaller λ value.

tiple diffusion trajectories. In the probabilistic formulation of $p(\tilde{\mathbf{X}}_t^i \mid \mathbf{y}_t^i)$ in Eq. 16, a smaller value of λ reduces the variance of the N^{th} trajectory relative to the 1st $\sim (N-1)^{\text{th}}$ trajectories, thereby preserving semantic features encoded in 1st $\sim (N-1)^{\text{th}}$ trajectories. In contrast, larger λ values increase variance, allowing the N^{th} trajectory to deviate more and depend less on earlier trajectories.

This theoretical analysis is visually supported by Figure 9. When $1/\lambda = 0$ (*i.e.*, $\lambda = \infty$), the 2nd trajectory becomes independent of the 1st, generating an image that only aligns with the 2nd prompt. Mathematically, this eliminates the conditional probability term in Eq. 16, resulting in a failure of integrating both prompts, which is consistent with our theoretical explanation. As $1/\lambda$ increases (*i.e.*, λ decreases), trajectory correlation strengthens, generating a well-blended image incorporating both prompts. If $1/\lambda$ further increases, the balance between two trajectories collapses, causing the 1st prompt to dominate. This effect is clearly visualized with $1/\lambda = 6$ and $1/\lambda = 7$, where the features of the horse dominate the features of the mountain village.

5. Conclusion

We propose a probabilistic framework for diffusion synchronization, providing a theoretical analysis of why it works. By designing conditional probabilities between diffusion trajectories, we establish synchronization across multiple trajectories. Based on the proposed method, we focus on efficient heuristic modeling by identifying which probability term to model, significantly reducing the empirical testing to find optimal solutions. We evaluate our method on various collaborative generation tasks, comparing its performance with prior works. Experimental results demonstrate that our method is widely applicable and consistently outperforms baseline algorithms. We hope this work inspires future research on more robust and principled models of inter-trajectory correlations to further advance diffusion synchronization.

Acknowledgment

This paper was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (Ministry of Education) (P0025681-G02P22450002201-10054408, “Semiconductor”-Specialized University)

References

- [1] DeepFloyd Lab at StabilityAI. Deepfloyd if. <https://www.deepfloyd.ai/deepfloyd-if>, 2023. 2, 7, 11, 12, 13
- [2] Franz Aurenhammer. Voronoi diagrams—a survey of a fundamental geometric data structure. *CSUR*, 1991. 11
- [3] Oran Lang Omer Tov Huiwen Chang Tali Dekel Inbar Mosseri Bahjat Kawar, Shiran Zada and Michal Irani. Imagic: Text-based real image editing with diffusion models. *CVPR*, 2023. 2
- [4] Romain Beaumont. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. <https://github.com/rom1504/clip-retrieval>, 2022. 11
- [5] Raphael Bensadoun, Yanir Kleiman, Idan Azuri, Omri Harosh, Andrea Vedaldi, Natalia Neverova, and Oran Gafni. Meta 3d texturegen: Fast and consistent texture generation for 3d objects. *arXiv preprint arXiv:2407.02430*, 2024. 2
- [6] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *ICLR*, 2018. 5, 6, 7, 8, 11
- [7] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactr: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 2, 5, 6, 7
- [8] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023. 2
- [9] Inbum Park Daniel Geng and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. *CVPR*, 2024. 1, 2, 5, 6, 7
- [10] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *ICCV*, 2023. 2, 5, 6, 7
- [11] Daniel Geng, Inbum Park, and Andrew Owens. Factorized diffusion: Perceptual illusions by noise decomposition. In *ECCV*, 2024. 1, 2
- [12] Sookwan Han and Hanbyul Joo. Chorus: Learning canonized 3d human-object spatial relations from unbounded synthesized images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15835–15846, 2023. 1
- [13] 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. *NeurIPS*, 2017. 5, 6, 7, 8, 11
- [14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021. 5
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1, 2, 3
- [16] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022. 1
- [17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022. 1
- [18] Kyeongmin Yeo Jaihoon Kim, Juil Koo and Minhyuk Sung. Sync tweedies: A general generative framework based on synchronized diffusions. *NeurIPS*, 2024. 1, 2, 5, 6, 7, 8, 11
- [19] Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, and Hanbyul Joo. Chupa: Carving 3d clothed humans from skinned shape priors using 2d diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15965–15976, 2023. 1
- [20] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models. In *European Conference on Computer Vision*, pages 400–419. Springer, 2024. 1
- [21] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 4
- [22] Peter E Kloeden, Eckhard Platen, Peter E Kloeden, and Eckhard Platen. *Stochastic differential equations*. Springer, 1992. 3, 4
- [23] Hyunsoo Lee, Minsoo Kang, and Bohyung Han. Conditional score guidance for text-driven image-to-image translation. *NeurIPS*, 2023. 3, 4, 5, 11
- [24] Vahram Tadevosyan Roberto Henschel Zhangyang Wang Shant Navasardyan Levon Khachatryan, Andranik Moysian and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *ICCV*, 2023. 1
- [25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 7, 11
- [26] Yuheng Liu, Xinke Li, Xuetong Li, Lu Qi, Chongshou Li, and Ming-Hsuan Yang. Pyramid diffusion for fine 3d large scene generation. In *ECCV*, 2024. 1
- [27] Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion. In *SIGGRAPH Asia*, 2024. 2, 5, 6, 7, 8, 11
- [28] 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. *ICLR*, 2022. 2, 5, 6, 7
- [29] 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
- [30] Yaron Lipman Omer Bar-Tal, Lior Yariv and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *ICML*, 2023. 1, 2, 5, 6

- [31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 3
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5, 6, 7, 11
- [33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022. 1
- [34] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *SIGGRAPH*, 2023. 2
- [35] Dominik Lorenz, Patrick Esser, Robin Rombach, Andreas Blattmann, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022.
- [36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 4, 5, 6, 7, 8, 11, 12, 13
- [37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention*, 2015. 3
- [38] 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. *NeurIPS*, 2022. 1
- [39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 6, 7, 11, 12
- [40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021. 1, 2, 3, 5, 11
- [41] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, 1981. 2, 3
- [42] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *ICLR*, 2023. 2, 5, 6, 7
- [43] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *CVPR*, 2024. 1
- [44] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 1, 2, 5, 8
- [45] Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit H Bermano, and Michiel van de Panne. Closd: Closing the loop between simulation and diffusion for multi-task character control. *arXiv preprint arXiv:2410.03441*, 2024. 1, 2
- [46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 4
- [47] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 360-degree panorama generation from few unregistered nfov images. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 2
- [48] Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, Yang Song, Jascha Sohl-Dickstein, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021. 1, 2, 3
- [49] Niv Haim, Yaniv Nikankin, and Michal Irani. Sinfusion: Training diffusion models on a single image or video. *ICML*, 2023. 1
- [50] Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In *CVPR*, 2024. 2
- [51] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023. 1, 2
- [52] Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonsenes: Generating commonsense 3d indoor scenes with scene graph diffusion. *NeurIPS*, 2023. 1
- [53] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 7, 8, 11, 12, 14
- [54] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *CVPR*, 2023. 2
- [55] 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. 6, 7
- [56] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 7

A. Task-specific experimental details

In this section, we provide the experimental details of each diffusion synchronization task.

A.1. Mask-based Text-to-Image generation

We use the pretrained Stable Diffusion v2 checkpoint [36] for image generation, resulting in 512×512 resolution image. Using 10 prompts, we generate 250 images per prompt with a fixed background mask for each. KID [6] and FID [13] we use 2,000 images per prompt using the same pretrained model. We use 50 steps for DDIM [40] sampling.

A.2. Text-driven real image editing

Firstly, we explain the details of soft mask generation. Note that we follow CSG [23] to generate the soft mask $\tilde{\mathbf{B}}$ which indicates the background region of the source image. Following paragraph summarizes the procedure introduced in [23].

We extract the self-attention and cross-attention map of the source image using the pretrained Stable Diffusion [36], each denoted as $\mathbf{M}_{\text{self}} \in \mathbb{R}^{H \times W \times H \times W}$ and $\mathbf{M}_{\text{cross}} \in \mathbb{R}^{N \times H \times W}$, where N denotes the number of word tokens defined in the pretrained Stable Diffusion model. Then we generate the background mask $\tilde{\mathbf{B}}$ as follows:

$$\tilde{\mathbf{B}} = \mathbf{1} - \mathbf{M}_{\text{fg}}, \quad (24)$$

where each element of $\mathbf{M}_{\text{fg}} \in \mathbb{R}^{H \times W}$ is defined as

$$\mathbf{M}_{\text{fg}}[h, w] = \text{tr}(\mathbf{M}_{\text{self}}[h, w]\mathbf{M}_{\text{cross}}^\top[u]). \quad (25)$$

Note that u denotes the index of the word token corresponds to the object that we want to manipulate.

We use the pretrained Stable Diffusion v1-4 model for experiments, generate images in 512×512 resolution. Also, we use four image editing tasks for evaluation: cat \rightarrow dog, dog \rightarrow cat, horse \rightarrow zebra, and zebra \rightarrow horse. For each task, we sample 250 real images from the LAION-5B dataset [39]. To find the most relevant images for the source word (*e.g.* ‘cat’ in cat-to-dog task) within the dataset, we leverage CLIP retrieval [4]. The source prompt is generated using the pretrained BLIP model [25], while the target prompt is made by replacing the source word with the target word. For instance, in the ‘horse \rightarrow zebra’ task, we swap the word ‘horse’ in the source prompt with ‘zebra’ to generate the target prompt. We use DDIM [40] sampling with 50 steps.

A.3. Wide image generation

We use the pretrained Stable Diffusion v2 checkpoint [36] for wide image generation. With four different text prompts, we generate 250 images per prompt at a resolution of 2048×512 . To measure KID [6], FID [13], and CLIP-S score [32], we randomly crop the generated wide images to a resolution of 512×512 . We generate 2,000 images per prompt to construct the reference image set using the same pretrained model. We use 50 steps for DDIM [40] sampling.

A.4. Ambiguous image generation

We use the pretrained DeepFloyd v1.0 checkpoint [1] for experiments, synthesizing images at 256×256 resolution. The DeepFloyd-IF model employs a two-stage sampling process for image generation. Note that we apply the proposed synchronization strategy only to the 1st stage, while the 2nd stage’s sampling is performed without synchronization. We use 5 prompt pairs, where each pair consists of two prompts describing the semantics to be modeled in resulting ambiguous image. For each prompt pair, we set f_1 as identity mapping and choose f_2 from one of 4 visual transforms: $\pm 90^\circ$ rotation, 180° rotation, vertical flip, and skew transformation. We then generate 250 images per prompt pair. In case of reference images for measuring KID [6] and FID [13], we generate 2,000 images per prompt in each prompt pair with the same pretrained model. Total 50 timesteps are used for DDIM [40] sampling.

A.5. 3D mesh texturing

We use the pretrained depth-conditioned ControlNet v1-1 [53] for mesh texturing. Using 6 meshes and a single prompt for each mesh, we generate 100 textures per mesh. Each generated texture is projected onto a fixed single view, resulting in a 768×768 resolution RGB image. To generate reference images, we use the same pretrained model and sample 2,000 images per prompt using the equivalent mesh map as depth condition. In addition, following SyncMVD [27] and SyncTweedies [18], we use the self-attention modification technique proposed in [27] along with Voronoi Diagram-guided filling [2]. We use 30 steps for DDIM [40] sampling. Like SyncTweedies, we do not use diffusion synchronization during the last 20% of the sampling steps.

Table 6. Comparison of computational overhead between SyncSDE and SyncTweedies [18].

Method	Time (s/image)	GPU memory (GB)
SyncTweedies [18]	7.721	2.44
SyncSDE (Ours)	5.664	2.78

B. Computational overhead

We analyze the computational overhead in terms of both time and GPU memory required to generate a single image. The measured computational overhead of the proposed method and SyncTweedies [18] is reported in Table 6. We use a single NVIDIA RTX 3090 GPU for measurement. Notably, SyncSDE exhibits a comparable computational overhead to SyncTweedies.

C. Additional qualitative results

We visualize additional qualitative results of SyncSDE in Figure 10, 11, 12, 13, and 14. As shown in the figures, SyncSDE

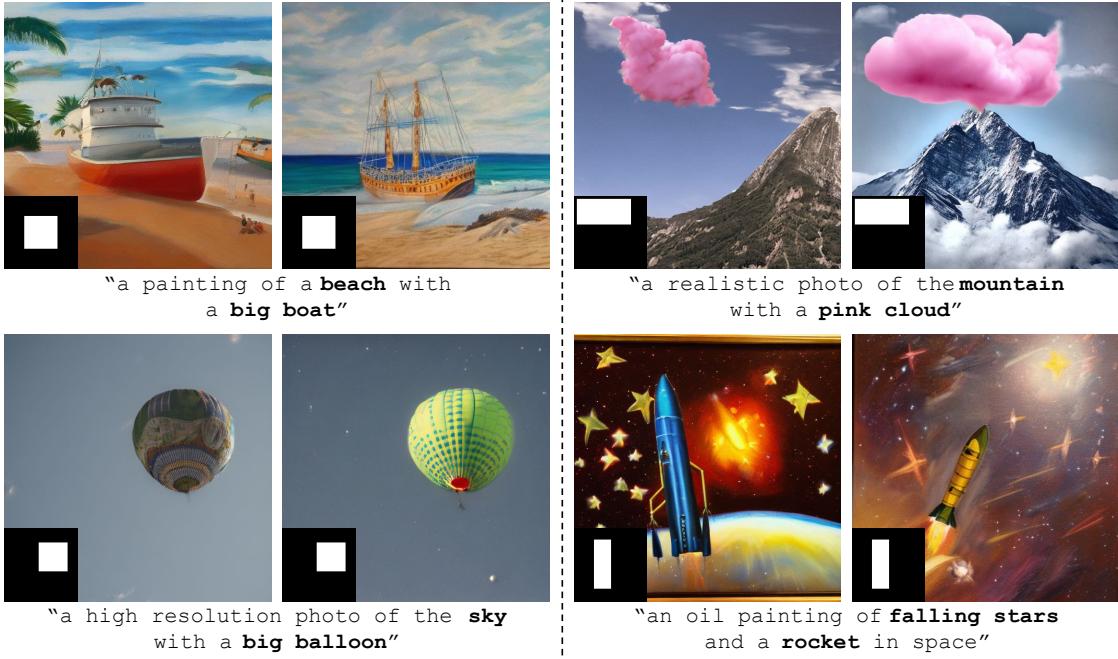


Figure 10. **Additional qualitative results of mask-based T2I generation.** SyncSDE shows strong performance on mask-based T2I generation task. We use the pretrained Stable Diffusion [36] for image generation.

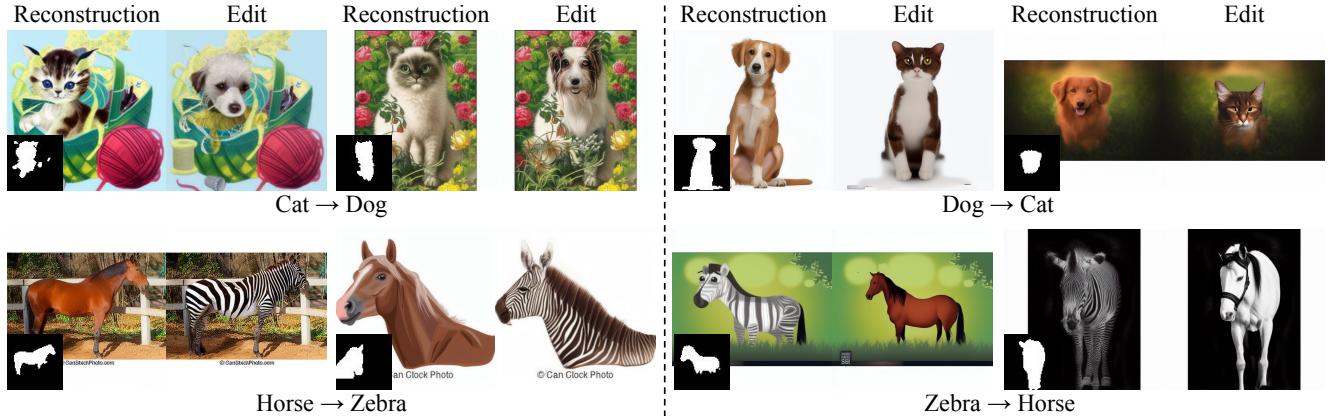


Figure 11. **Additional qualitative results of text-driven real image editing.** We edit the real images sampled from the LAION-5B dataset [39] by leveraging SyncSDE combined with the pretrained Stable Diffusion [36]. We also visualize the foreground region defined by the generated mask.

shows outstanding performance in multiple image generation tasks, including mask-based T2I generation, text-driven real image editing, wide image generation, ambiguous image generation, and 3D mesh texturing. The experimental results demonstrate that the proposed method successfully models the correlation between multiple diffusion trajectories, thus smoothly blending the generated patches.

D. Limitations and social impacts

Since our method uses a pretrained text-to-image diffusion model [1, 36, 53], the proposed method may result in sub-optimal outcomes depending on the pretrained backbone model. For instance, due to the limitations of the pretrained diffusion model, it may struggle to synthesize images with complex structures or multiple fine details. Furthermore, the proposed method may generate harmful images due to the shortcomings of the pretrained diffusion model.

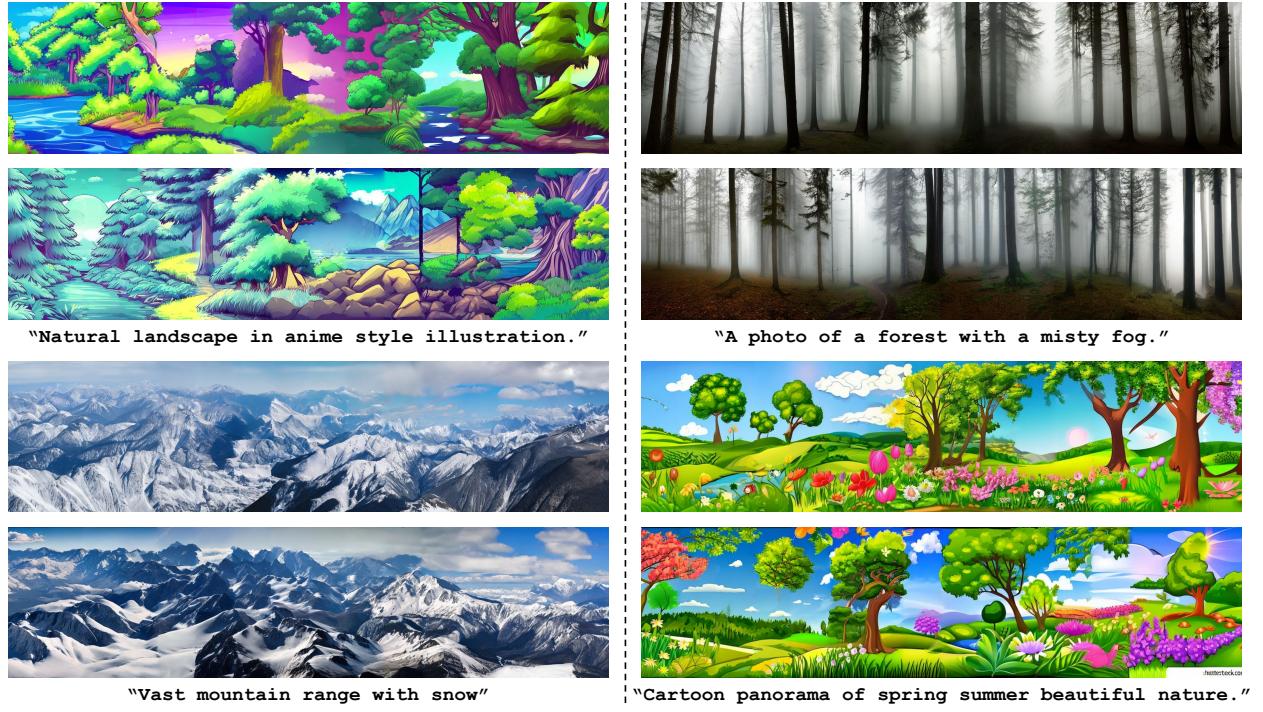


Figure 12. **Additional qualitative results of wide image generation.** We visualize wide images generated by SyncSDE using the pretrained Stable Diffusion [36] for image generation.

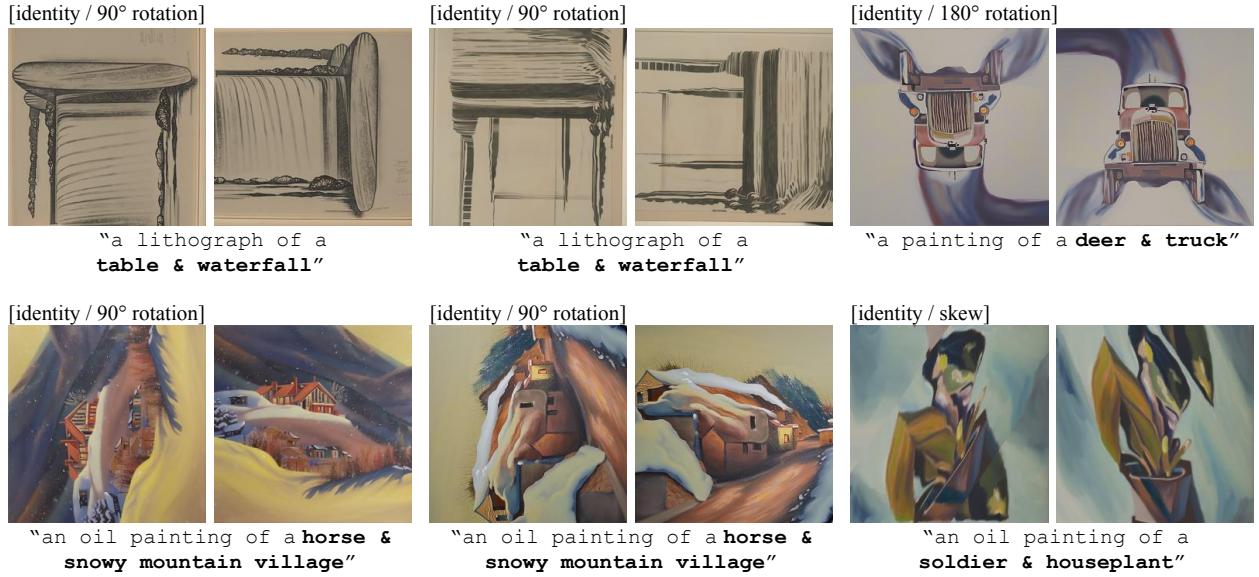


Figure 13. **Additional qualitative results of ambiguous image generation.** Using the pretrained Deepfloyd-IF model [1], we generate ambiguous image with various prompt pairs and visual transformations. SyncSDE generates high-quality ambiguous images.

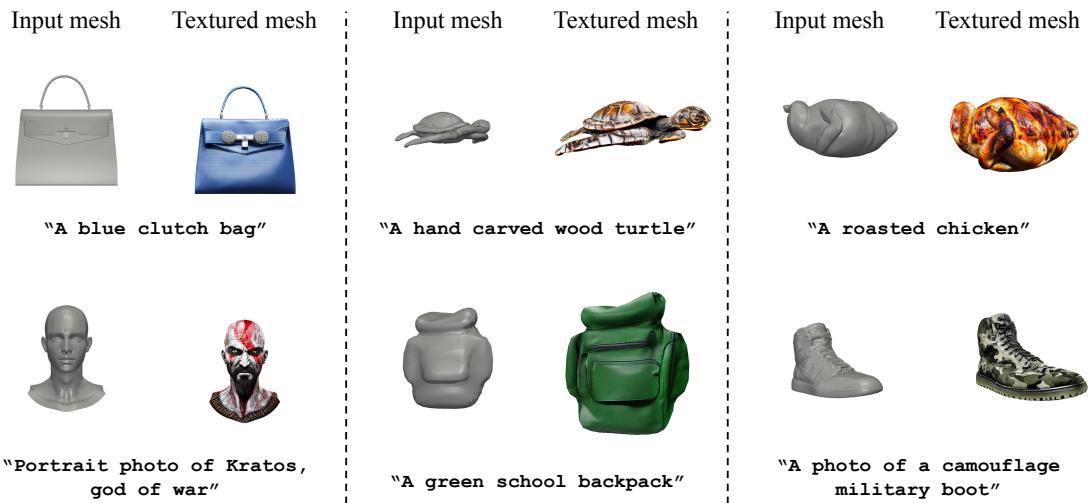


Figure 14. **Additional qualitative results of 3D mesh texturing.** We use the pretrained depth-conditioned ControlNet [53] for mesh texturing. Given an input mesh and the text prompt, SyncSDE generates remarkable texture images.