# ViewFusion: Towards Multi-View Consistency via Interpolated Denoising

Xianghui Yang[1,2*], Yan Zuo[1], Sameera Ramasinghe[1], Loris Bazzani[1], Gil Avraham[1], Anton van den Hengel[1,3]
[1]Amazon, [2]The University of Sydney, [3]The University of Adelaide

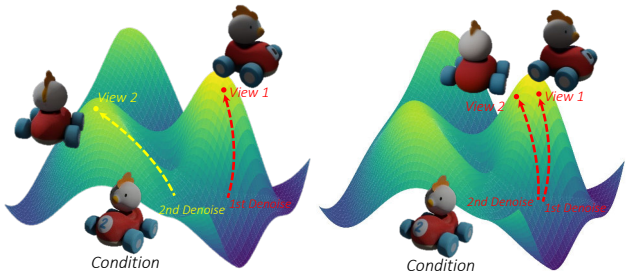
## Abstract

*Novel-view synthesis through diffusion models has demonstrated remarkable potential for generating diverse and high-quality images. Yet, the independent process of image generation in these prevailing methods leads to challenges in maintaining multiple view consistency. To address this, we introduce ViewFusion, a novel, training-free algorithm that can be seamlessly integrated into existing pre-trained diffusion models. Our approach adopts an auto-regressive method that implicitly leverages previously generated views as context for next view generation, ensuring robust multi-view consistency during the novel-view generation process. Through a diffusion process that fuses known-view information via interpolated denoising, our framework successfully extends single-view conditioned models to work in multiple-view conditional settings without any additional fine-tuning. Extensive experimental results demonstrate the effectiveness of ViewFusion in generating consistent and detailed novel views.*

## 1. Introduction

Humans have a remarkable capacity for visualizing unseen perspectives from just a single image view – an intuitive process that remains complex to model. Such an ability is known as Novel View Synthesis (NVS) and necessitates robust geometric priors to accurately infer three-dimensional details from flat imagery; lifting from a two-dimensional projection to a three-dimensional form involves assumptions and knowledge about the nature of the object and space. Recently, significant advancements in NVS have been brought forward by neural networks [15, 16, 31, 42, 61, 65, 66, 73, 76, 77], where novel view generation for downstream reconstruction shows promising potential [35, 70].

Specifically, diffusion models [20, 52] and their ability to generate high-quality 2D images have garnered significant attention in the 3D domain, where pre-trained, text-conditioned 2D diffusion models have been re-purposed for 3D applications via distillation [4, 32, 41, 48, 50, 58, 64, 69, 74]. Follow-up approaches [35, 70] remove the requirement of text conditioning and instead take an image and target pose as conditions for NVS. However, distillation [64] is still required as the diffusion model cannot produce the multi-view consistent outputs that are appropriate for certain downstream tasks (*e.g.*, optimizing Neural Radiance Fields (NeRFs) [42]).




(a) Inconsistent generation. (b) Consistent generation (ours).

Figure 1. The cause of multi-view inconsistency in diffusion-based novel-view synthesis models. (a) Diffusion models incorporate randomness for diversity and better distribution modeling; this independent generation process produces realistic views under specific instances but may produce different plausible views for various instances, lacking alignment across adjacent views. (b) In contrast, ViewFusion incorporates an auto-regressive process to reduce uncertainty and achieve multi-view consistency, by ensuring a correlated denoising process that ends at the same high-density area, fostering consistency across views.

Under the single-view setting, maintaining multi-view consistency remains particularly challenging since there may exist several plausible outputs for a novel view that are aligned with the given input image. For diffusion-based approaches which generate novel views in an independent manner [35, 70], this results in synthesized views containing artifacts of multi-view inconsistency (Fig. 1a). Previous work [34, 37, 39, 54, 75, 78] focuses on improving the robustness of the downstream reconstruction to address the inconsistency issue, including feature projection layers in the NeRF [34] or utilising three-dimensional priors to constrain NeRF optimization [37, 78], yet these techniques require training or fine-tuning to align additional modules to the original diffusion models.

In this work, we address the multi-view inconsistency that arises during the process of view synthesis. Rather than

*Work done during internship at Amazon

independently synthesizing views conditioned only on the initial reference image, we develop a novel approach where each subsequently generated view is also conditioned on the *entire set* of previously generated views. Specifically, our method incorporates an auto-regressive process into the diffusion process to model the joint distribution of views, guiding our novel-view synthesis by maintaining the denoising direction towards the same high density area of already generated views (Fig. 1b).

Our framework, named ViewFusion, relaxes the single-view conditioning requirement of typical diffusion models through an interpolated denoising process. ViewFusion offers several additional advantages: 1) it can utilize all available views as guidance, thereby enhancing the quality of generated images by incorporating more information; 2) it does not require any additional fine-tuning, effortlessly converting pre-trained single-view conditioned diffusion models into multi-view conditioned diffusion models; 3) it provides greater flexibility in setting adaptive weights for condition images based on their relative view distance to the target view.

The contributions of this paper are the following:

- We propose a *training-free* algorithm which can be directly applied to pre-trained diffusion models to improve multi-view consistency of synthesized views and supports multiple conditional inputs.
- Our method utilizes a novel, auto-regressive approach which we call *Interpolated Denoising*, that implicitly addresses key limitations of previous auto-regressive approaches for view synthesis.
- Extensive empirical analysis on ABO [6] and GSO [10] show that our method is able to achieve better 3D consistency in image generation, leading to significant improvements in novel view synthesis and 3D reconstruction of shapes under single-view and multi-view image settings over other baseline methods.

## 2. Related Work

### 2.1. 3D-adapted Diffusion Models

Diffusion models have excelled in image generation using conditional inputs [21, 47, 51, 85] and given this success in the 2D domain, recent works have tried to extend diffusion models to 3D content generation [1, 3, 5, 11, 17, 19, 23, 25, 26, 28, 38, 44–46, 67, 83, 84] – although the scarcity of 3D data presents a significant challenge to directly train these diffusion models. Nonetheless, pioneer works such as DreamFusion [48] and Score Jacobian Chaining [64] leverage pre-trained text-conditioned diffusion models to craft 3D models via distillation. Follow-up approaches [4, 32, 58, 69] improve this distillation in terms of speed, resolution and shape quality. Approaches such as [41, 50, 58, 74] extend upon this to support image conditions through the use of captions with limited success due to the non-trivial nature of textual inversion [14].

### 2.2. Novel View Synthesis Diffusion Models

Another line of research [2, 9, 18, 29, 36, 57, 59, 62, 63, 70, 72, 79, 81, 87] directly applies 2D diffusion models to generate multi-view images for shape reconstruction. To circumvent the weakness of text-conditioned diffusion models, novel-view synthesis diffusion models [35, 70] have also been explored, which take an image and target pose as conditions to generate novel views. However, for these approaches, recovering a 3D consistent shape is still a key challenge. To mitigate 3D inconsistency, Liu et al. [34] suggests training a Neural Radiance Field (NeRF) with feature projection layers. Concurrently, other works [37, 39, 71, 75, 78] add modules to original diffusion models for multi-view consistency, including epipolar attention [78], synchronized multi-view noise predictor [37] and cross-view attention [39, 71]; although these methods require fine-tuning an already pre-trained model. We adopt a different paradigm, instead of extending a single-view diffusion model with additional trainable models that incorporate multi-view conditions, our training-free method enables pre-trained diffusion models to incorporate previously generated views via the denoising step and holistically extends these models into multi-view settings.

### 2.3. Other Single-view Reconstruction Methods

Before the prosperity of generative models used in 3D reconstruction, many works [12, 13, 15, 16, 27, 30, 31, 60, 65, 76] reconstructed 3D shapes from single-view images using regression [15, 16, 30, 65, 76] or retrieval [60], both of which face difficulties in generalizing to real data or new categories. Methods based on Neural Radiance Fields (NeRFs) [42] have found success in novel-view synthesis, but these approaches typically depend on densely captured images with accurately calibrated camera positions. Currently, several studies are investigating the adaptation of NeRF to single-view settings [22, 33, 53, 80]; although, reconstructing arbitrary objects from single-view images is still a challenging problem.

## 3. Method

### 3.1. Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models (DDPM) [20, 55] are a class of generative models that model the real data distribution $q(x_0)$ with a tractable model distribution $p_\theta(x_0)$ by learning to iteratively denoise samples. It learns a probability model $p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T}$ to convert unstructured noise $\mathbf{x}_T$ to real samples $\mathbf{x}_0$ in the form of a Markov chain, with Gaussian transitions. The Gaussian transition is

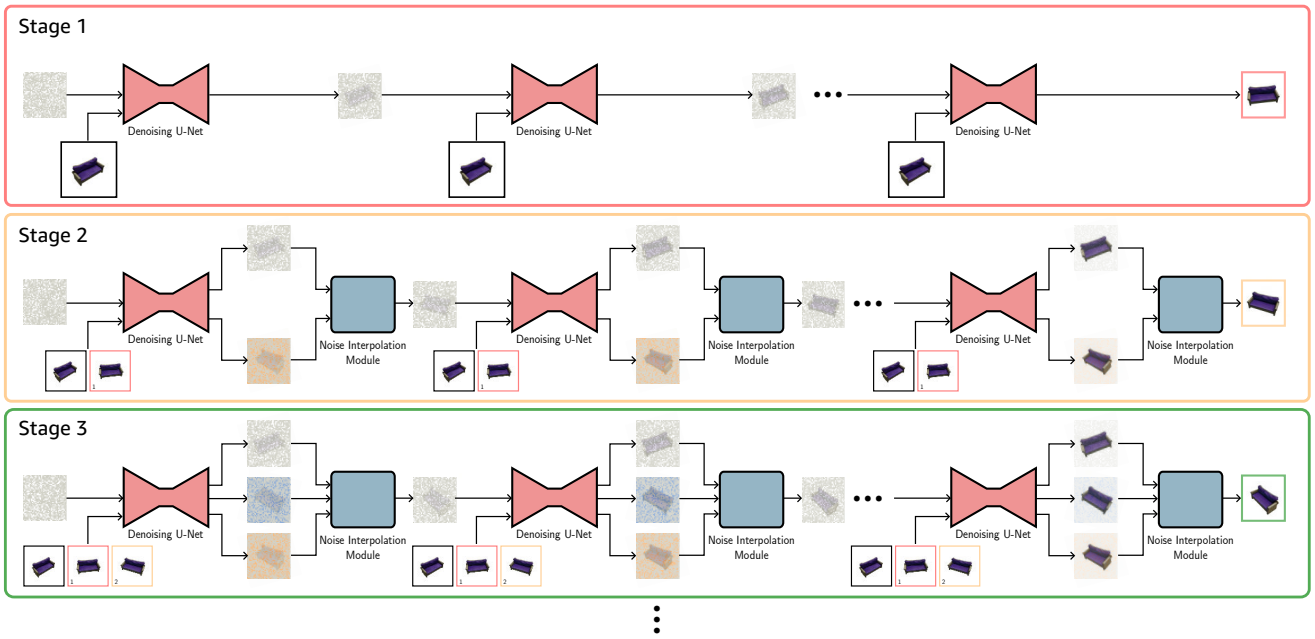



Figure 2. Illustration of the Auto-Regressive Generation Process. In our approach, we extend a pre-trained diffusion model from single-stage to multi-stage generation and we maintain a view set that contains all generated views. For each stage, we construct $N$ reverse diffusion processes and sharing a common starting noise. At each time step within this generation stage, the diffusion model predicts $N$ noises individually. These $N$ noises are then subjected to weighted interpolation through the *Noise Interpolation Module*, concluding the denoising step with the a shared interpolated noise for subsequent denoising steps.

defined as:

$$q(\mathbf{x}_T|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where $\beta_t, t \in \{1, ..., T\}$ are the variance schedule parameter and timestep in the denoising process respectively. The reverse denoising process starts from a noise sampled from a Gaussian distribution $q(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and is constructed as:

$$p_\theta(\mathbf{x}_0|\mathbf{x}_T) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}), \tag{2}$$

where the variance $\sigma_t^2$ is a time-dependent constant [20], and $\mu_\theta(\mathbf{x}_t, t)$ is the mean from the learned *noise predictor* $\epsilon_\theta$:

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

Here, $\alpha_t$ and $\bar{\alpha}_t$ are constants derived from $\beta_t$. The objective of noise predictor $\epsilon_\theta$ is simplified to:

$$\ell = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|_2\right], \tag{4}$$

where $\epsilon$ is a random variable sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ [20].

### 3.2. Pose-Conditional Diffusion Models

Similar to other generative models [43, 56], diffusion models inherently possess the capability to model conditional distributions of the form $p_\theta(x_{t-1}|x_t, y)$ where $y$ is the condition. We employ a conditional denoising autoencoder, denoted as $\epsilon_\theta(\mathbf{x}_t, t, y)$ which enables controlling the synthesis process through a variety of input modalities, including textual descriptions [51], semantic maps [21, 47], or other image-to-image translation tasks [21]. In the following, we present a range of approaches to novel-view synthesis, exploring how various works, including our own, approach the concept of a single reverse diffusion step. Through this comparison, we clarify and establish the underlying relationships between these different methodologies. The notation will follow that bottom subscript $(\cdot)_t$ indicates the diffusion step and upper subscript $(\cdot)^i$ relates to the view index. Subsequently, the $i$-th condition image and its relative pose to the target view are defined as $\mathbf{y}^i$ and $\pi^i$, respectively, and the noisy image to be denoised at timestep $t$ is defined as $\mathbf{x}_t$.

**Direct condition** was applied by Zero 1-to-3 [35] to the reverse process when given a single input image and target pose $\mathbf{y}^1, \pi^1$:

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}^1, \pi^1). \tag{5}$$

**Stochastic conditioning** was formulated by [70] which can leverage multiple views sampled from a collection of views $p_{\mathbf{y},\pi}(\mathcal{Y}, \pi)$:

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}^i, \pi^i), \{\mathbf{y}^i, \pi^i\} \sim p_{\mathbf{y},\pi}(\mathcal{Y}, \pi), \tag{6}$$

where the sampling of image and pose happens at each diffusion step $t$.

**Joint output distribution** was shown in SyncDreamer[37] which learns a joint distribution of many views given an image condition $y^1$:

$$p(\mathbf{x}_{t-1}^{1:N}|\mathbf{x}_t^{1:N}, \mathbf{y}^1, e^1), \tag{7}$$

where $N$ is the number of generated novel views and $e$ is the elevation condition (partial pose information). We note that in this formulation the target poses are not fully specified as part of the condition allowing for diverse pose generation of outputs.

**Auto-regressive distribution** is an auto-regressive distribution setting which can generate an arbitrary number of views given a single or multiple condition images and poses contained in the set of $\mathbf{y}^{1:N-1}, \pi^{1:N-1}$:

$$p(\mathbf{x}_{t-1}^{N}|\mathbf{x}_t^{N}, \mathbf{y}^{1:N-1}, \pi^{1:N-1}). \tag{8}$$

Our approach falls in the auto-regressive category and for the remainder of this section we detail the implementation to achieve this sampling strategy.

### 3.3. Interpolated Denoising

The standard DDPM model has been adapted for novel-view image synthesis by using an image and target pose (*i.e.*, rotation and translation offsets) as conditional inputs [70]. Following training on a large-scale dataset, this approach has demonstrated the capability for zero-shot reconstruction [35]. To address the challenge of maintaining multi-view consistency, we employ an auto-regressive approach for generating sequential frames (See Fig. 2). Instead of independently producing each frame from just the input images – a process prone to significant variations between adjacent images – we integrate an auto-regressive algorithm into the diffusion process. This integration enables us to model a conditional joint distribution, ensuring smoother and more consistent transitions between frames.

To guide the synthesis of novel views using images under different views, we design an interpolated denoising process. For the purpose of this derivation, we assume access to an image set containing $N-1$ images denoted as $\{\mathbf{y}^1, ..., \mathbf{y}^{N-1}\}$. We want to model the distribution of the $N$-th view image conditioned on these $N-1$ views $q(\mathbf{x}_{1:T}^N|\mathbf{y}^{1:N-1})$, where the relative pose offsets $\pi^i, i \in \{1, N-1\}$ between the condition images $\{\mathbf{y}^1, ..., \mathbf{y}^{N-1}\}$ and target image $\mathbf{x}_0^N$ are omitted for simplicity. The forward process of the multi-view conditioned diffusion model is a direct extension of the vanilla DDPM in Eq. 1, where noises are added to every view independently by

$$q(\mathbf{x}_{1:T}^N|\mathbf{y}^{1:N}) = \prod_{t=1}^{T} q(\mathbf{x}_t^N|\mathbf{x}_{t-1}^N, \mathbf{y}^{1:N}) \tag{9}$$

where $q(\mathbf{x}_t^N|\mathbf{x}_{t-1}^N, \mathbf{y}^{1:N}) = \mathcal{N}(\mathbf{x}_t^N; \sqrt{1-\beta_t}\mathbf{x}_{t-1}^N, \beta_t\mathbf{I})$. The initial is defined as $\mathbf{x}_0^N := \mathbf{y}^N$. Similarly, following Eq. 2, the log reverse process is constructed as

$$\begin{aligned}
\log p_\theta(\mathbf{x}_0^N|\mathbf{x}_T^N, \mathbf{y}^{1:N-1}) &= \sum_{t=1}^{T} \log p_\theta(\mathbf{x}_{t-1}^N|\mathbf{x}_t^N, \mathbf{y}^{1:N-1}) \\
&\underset{(1)}{\approx} \sum_{t=1}^{T} \log \prod_{n=1}^{N-1} p_\theta(\mathbf{x}_{t-1}^N|\mathbf{x}_t^N, \mathbf{y}^n) \\
&= \sum_{t=1}^{T} \sum_{n=1}^{N-1} \log \mathcal{N}(\mathbf{x}_{t-1}^N; \mu_\theta^n(\mathbf{x}_t^N, \mathbf{y}^n, t), \sigma_t^2\mathbf{I}) \\
&= \sum_{t=1}^{T} \mathcal{N}\left(\mathbf{x}_{t-1}^N; \bar{\mu}_\theta(\mathbf{x}_t^N, \mathbf{y}^{1:N-1}, t), \bar{\sigma_t}^2\mathbf{I}\right).
\end{aligned} \tag{10}$$

Where $\bar{\mu_\theta}, \bar{\sigma_t}^2$ are taken as the mean and variance of the summation of $N-1$ *log*-normal distributions. A note on subscript (1) in Eq.10; to avoid cluttering the derivation, we assume $N-1$ independent inferences of the same random variable $\mathbf{x}_{t-1}^N$ using *a different* $\mathbf{y}^n$ that results in $N-1$ independent normal distributions, which would require an additional subscript that we omitted for clarity.

### 3.4. Single and Multi-view Denoising

*In practice*, however, we may not have all $N-1$ views but a single view or a handful of views. For the reminder of this section, we treat an estimated view as $\mathbf{x}_0^n$, to be the $n$-th view $\mathbf{y}^n$ after a full reverse diffusion process. We use $\bar{\mu}_\theta(\mathbf{x}_t, \mathbf{y}^{1:N-1}, t)$ as the weighted average of $\mu_\theta^n(\mathbf{x}_t, \mathbf{y}^n, t)$. For computing $\bar{\mu}_\theta$ using both given views and estimated views we adopt an approach where different views contribute differently to the target view, and we assign the weight $\omega_n$ for the $n$-th view in practice while satisfying the constraint $\sum_{n=1}^{N-1} w_n = 1$. The *Noise Interpolation Module* in Fig. 2 is modeled as:

$$\begin{aligned}
\bar{\mu}_\theta(\mathbf{x}_t, \mathbf{y}^{1:N-1}, t) &= \sum_{n=1}^{N-1} \omega_n \mu_\theta^n(\mathbf{x}_t, \mathbf{y}^n, t) \\
&= \sum_{n=1}^{N-1} \omega_n \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, \mathbf{y}^n, t)\right) \\
&= \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \sum_{n=1}^{N} \omega_n \epsilon_\theta(\mathbf{x}_t, \mathbf{y}^n, t)\right).
\end{aligned} \tag{11}$$

In our approach, as the full view set is not given to us, we approximate this process by an auto-regressive way and

grow the condition set during the generation. We define the weight parameter $\omega_n$ based on the angle offset, *i.e.*, azimuth ($\Delta_a^n$), elevation ($\Delta_e^n$), and distance ($\Delta_d^n$), between the target view and the $n-th$ condition view. The core idea is to assign higher importance to near-view images during the denoising process while ensuring that the weight for the initial condition image does not diminish too rapidly, even when the target view is positioned at a nearly opposite angle. We use an exponential decay weight function for the initial condition image, defined as $\omega_n = e^{-\frac{\Delta^n}{\tau_c}}$. Here, $\tau_c$ is the temperature parameter that regulates the decay speed, and $\Delta^n$ is the sum of the absolute relative azimuth ($\Delta_a^n$), elevation ($\Delta_e^n$), and distance ($\Delta_d^n$) between the target and condition poses. We calculate $\Delta^n$ as $\Delta^n = |\Delta_a^n|/\pi + |\Delta_e^n|/\pi + |\Delta_d^n|$.

For the weights of the remaining images denoted as $\{x_0^2, ..., x_0^N\}$, all generated from the initial condition image $y^1 := x_0^1$, we use a softmax function to define the weights $\omega_n$:

$$\omega_n = \text{Softmax}(\frac{e^{-\frac{\Delta^n}{\tau_g}}}{\sum_{n=2}^{N} e^{-\frac{\Delta^n}{\tau_g}}}), n = 2, ..., N \tag{12}$$

Similarly, $\Delta_n$ represents the relative pose offset between target view and the $n$-th generated view, and $\tau_g$ represents the temperature parameter for generated views. As an example, in the single-view case, the weights are expressed as follows,

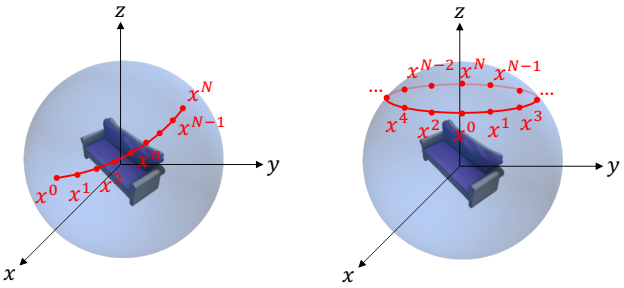
$$\omega_n = \begin{cases} exp(-\frac{\Delta^n}{\tau_c}), & n = 1 \\ (1-\omega_1)\text{Softmax}(\frac{e^{-\frac{\Delta^n}{\tau_g}}}{\sum_{n=2}^{N} e^{-\frac{\Delta^n}{\tau_g}}}), & n \neq 1 \end{cases} \tag{13}$$

we apply the term $1-\omega_1$ on the generated image weights to ensure the requirement of $\sum_{n=1}^{N-1} w_n = 1$ will be met. In practice, Eq.16 is generalised to allow the condition set can be larger than 1, *i.e.*, multi-view generation (see supplementary).

### 3.5. Step-by-step Generation

**Single image generation.** When applying the auto-regressive approach to image generation, we have devised a generation trajectory, as illustrated in Fig. 3a. We uniformly sample views along this trajectory in sequence. Each previously generated view image on this trajectory is incorporated into the condition set, providing guidance for the subsequent denoising process via our interpolated denoising method. To determine the number of steps, denoted as $S$, needed for this trajectory, we use the following formula:

$$S = max\left(\lceil\frac{\Delta_a^N}{\delta}\rceil, \lceil\frac{\Delta_e^N}{\delta}\rceil\right). \tag{14}$$




(a) Single image generation. (b) Spin video generation.

Figure 3. Illustration of Step-by-step Generation. (a) we uniformly sample views along this trajectory in sequence to generate a novel-view image; (b) we sample views from nearest to furthest views according to to view distance to generate a 360° spin video.

Here, we set the maximum offset per step $\delta$ to determine the step count $S$, also based on the target view offsets $\Delta_a^N$ and $\Delta_e^N$. We then proceed to sample the $n$-th view using the following equation:

$$(\Delta_a^n, \Delta_e^n, \Delta_d^n) = (\frac{\Delta_a^N}{S} * n, \frac{\Delta_e^N}{S} * n, \frac{\Delta_d^N}{S} * n) \tag{15}$$

**Spin videos generation.** In contrast to generating a single target image, the process of spin video generation begins from an initial image and concludes at the same position. To achieve this, we need to modify the generation order to leverage the broad range of rotation images, rather than simply following the rotation degree range of $[0°, 360°]$ in sequence. This is because, at $\Delta_a = \pi$, the view is opposite to the conditioning view, marking the end of the generation process. To establish the generation order for spin video generation, we introduce the minimum azimuth offset, denoted as $\delta$, and employ a skip trajectory with the following order: $\{\delta, -\delta, 2\delta, -2\delta..., N\delta\}$, shown in Fig. 3b. For simplicity, we only consider rotation along the azimuth dimension in this context.

## 4. Experiments

**Datasets.** We evaluate our method and compare to baselines on the ABO [6] and GSO [10] datasets. These datasets are out-of-the-distribution as all baselines are trained on the Objaverse [8]. We also provide qualitative results on real images to showcase performance of our method on in-the-wild images in the supplementary. For additional results, please refer to the videos contained in the supplementary.

**Metrics.** We assess our novel-view synthesis on three main criteria:

1. *Image Quality*: LPIPS [86], PSNR, and SSIM [68] metrics to help gauge the similarity between synthesized and ground-truth views.

| Dataset | Method | Free Renderings | | | SyncDreamer Renderings | | |
|---|---|---|---|---|---|---|---|
| | | SSIM↑ | PSNR↑ | LPIPS↓ | SSIM↑ | PSNR↑ | LPIPS↓ |
| ABO | Zero123 [35] | 0.8796 | 21.33 | 0.0961 | 0.7822 | 18.27 | 0.1999 |
| | SyncDre. [37] | 0.7712 | 13.43 | 0.2182 | **0.8031** | **19.07** | **0.1816** |
| | Ours | **0.8848** | **21.43** | **0.0923** | 0.7983 | 18.75 | 0.1985 |
| GSO | Zero123 [35] | 0.8710 | 20.33 | 0.1029 | 0.7925 | 18.06 | 0.1714 |
| | SyncDre. [37] | 0.8023 | 14.42 | 0.1833 | 0.8024 | 18.20 | **0.1647** |
| | Ours | **0.8820** | **20.73** | **0.0958** | **0.8076** | **18.40** | 0.1703 |

Table 1. Quantitative results on ABO and GSO datasets with **arbitrary** (left) and **discrete** (right) rotation and translation. Free renderings are a set of arbitrary rotation and translation as target generation view, while SyncDreamer renderings are a fixed set of 16 views with discrete azimuth, fixed elevation and distance, *i.e.*, azimuth$\in \{0°, 22.5°, 45°, ..., 315°, 337.5°\}$, elevation=30°. Note that SyncDreamer [37] *cannot* generate images under arbitrary views apart from the predefined 16 camera positions.

2. *Multi-View Consistency*: Using SIFT [40], LPIPS [86] and CLIP [49], we measure the uniformity of images across various perspectives.
3. *3D Reconstruction*: Chamfer distances and F-score between ground-truth and reconstructed shapes determine geometrical consistency.

### 4.1. Novel-view synthesis

In Tab. 1, we show quantitative results for novel-view synthesis under arbitrary and fixed-view settings. The fixed-view setting uses the rendering set of [37] and ensures a fair comparison to [37] which is limited to this fixed-view generation setting. As shown in Tab. 1, our method can produce comparable results to [37], without any fine-tuning on the rendering set of [37]. Under the arbitrary-view setting, we sample the nearest view from the rendering set of [37] to the designated target view for generation; [37] underperforms both [35] and our approach. Overall, regardless of the evaluation setting, our method consistently outperforms [35] and even outperforms [37] on the GSO dataset under the favourable fixed-view setting for [37].

The task of novel view synthesis serves as a precursor for enabling more significant downstream applications, such as 3D reconstruction, by generating requisite input views. Thus, in addition to image quality, thorough evaluation needs to encompass the multi-view consistency of the synthesized perspectives. This additional criterion ensures that the generated imagery not only appears visually compelling but also aligns geometrically across different viewpoints. In the following sections, we evaluate against baselines on 3D consistency and show that the significant improvements our approach offers.

**Multi-view Consistency.** For measuring multi-view consistency of synthesized views, quantitative results are shown in Tab. 2, while qualitative comparisons are shown in Figs. 4 and 5. Here, we can see when compared to the baselines of [35] and [37], our approach excels in generating images that are both semantically consistent with the input image and maintains multi-view consistency in terms of colors and geometry under arbitrary-view settings.

| Dataset | Method | View | SIFT↑ | LPIPS↓ | CLIP↑ |
|---|---|---|---|---|---|
| ABO | Zero123 [35] | 16 | 13.38 | 0.1782 | 0.9604 |
| | SyncDreamer [37] | | 12.36 | 0.1895 | 0.9584 |
| | Ours | | **13.51** | **0.1602** | **0.9664** |
| | Zero123 [35] | 36 | 17.03 | 0.1231 | 0.9725 |
| | Ours | | **18.01** | **0.0966** | **0.9812** |
| GSO | Zero123 [35] | 16 | 12.58 | 0.1411 | 0.9482 |
| | SyncDreamer [37] | | 13.24 | 0.1315 | 0.9532 |
| | Ours | | **13.83** | **0.1187** | **0.9601** |
| | Zero123 [35] | 36 | 15.20 | 0.1056 | 0.9592 |
| | Ours | | **17.95** | **0.0676** | **0.9773** |

Table 2. Quantitative results for multi-view consistency. We report the SIFT matching point number, LPIPS and CLIP similarity between adjacent frames to evaluate the multi-view consistency. Note that SyncDreamer can only generate 16 view images for a spin video due to the constraints imposed by its training.

| Dataset | Method | SSIM↑ | PSNR↑ | LPIPS↓ |
|---|---|---|---|---|
| ABO | Zero123 [35] (1 view) | 0.8820 | 21.51 | 0.0945 |
| | Ours (1 view) | 0.8870 | 21.61 | 0.0904 |
| | Ours (2 views) | 0.8913 | 21.92 | 0.0887 |
| | Ours (3 views) | **0.8995** | **22.74** | **0.0815** |
| GSO | Zero123 [35] (1 view) | 0.8721 | 20.42 | 0.1017 |
| | Ours (1 view) | 0.8830 | 20.87 | 0.0948 |
| | Ours (2 views) | 0.8901 | 21.25 | 0.0893 |
| | Ours (3 views) | **0.8979** | **21.95** | **0.0792** |

Table 3. Quantitative results for multi-conditioned generation. Our approach outperforms the Zero-1-to-3 baseline [35] by extending its single-view reconstruction framework to a multi-view reconstruction framework. Note that 3 condition views are removed from test set, and thus the results are slightly different to Tab. 1.

### 4.2. Multi-view conditional setting for NVS.

Additionally, our approach allows for the extension of single-view conditioned models into multi-view conditioned models which can take multiple conditional images as input. The results presented in Tab. 3 showcase the advantages our method offers, as novel-view synthesis quality improves with an increasing number of conditional input views. This is a significant improvement over the Zero-1-to-3 baseline and demonstrates the efficacy of our proposed method in the multi-view setting.

### 4.3. 3D Reconstruction

For shape reconstruction, we present our results both quantitatively (Tab. 4) and qualitatively (Fig. 6). We compare against baselines which synthesize novel views as well as direct image-to-shape approaches [24, 45]. For the former approach, instead of using distillation [64], we train NeuS [66] on synthesized images to recover a shape.

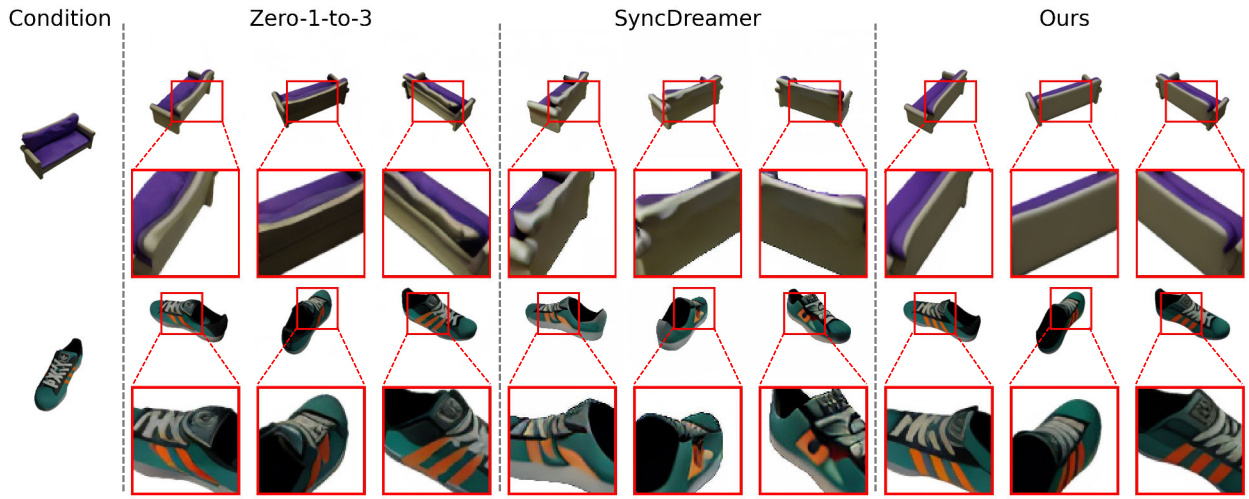



Figure 4. Qualitative results for 360° Spin Video Generation. Note the additional consistency in generated views our approach offers over the competing baselines shown in the bounding boxes.

Figure 5. Qualitative comparison for Motion Smoothness. We visualize the output videos using space-time Y-t slices through frames of the generated spin video (along the scanline shown in the condition).

**Single-view Reconstruction.** We first compare our proposed method with several baselines under the single-view reconstruction setting, including Point-E [45] and Shap-E [24], Zero-1-to-3 [35], One-2-3-45 [34] and SyncDreamer [37]. Qualitative results are shown in Fig. 6; Point-E and Shap-E tend to generate incomplete shapes due to the single-view setting. Multi-view alignment methods such as Syncdreamer [37] and One-2-3-45 [34] capture the general geometry but tend to lose fine details. In comparison, our proposed method achieves the highest reconstruction quality amongst all approaches, where we can generate smooth surfaces and capture detailed geometry with precision.

**Multi-view Reconstruction** Given our approach can also synthesis novel views under the multi-view setting, we also show results where we use 3 conditional input views to synthesize 36 views for shape reconstruction. Compared with existing multi-view reconstruction frameworks such as

| Cond. views | Method | Gen. views | ABO | | GSO | |
|---|---|---|---|---|---|---|
| | | | CD↓ | F-score↑ | CD↓ | F-score↑ |
| 1 | Point-E [45] | N/A | 0.0428 | 0.7144 | 0.0672 | 0.6340 |
| | Shap-E [24] | N/A | 0.0466 | 0.7364 | 0.0384 | 0.7313 |
| | One-2-3-45 [34] | 32 | 0.0419 | 0.6665 | 0.0408 | 0.6490 |
| | SyncDreamer [37] | 16 | 0.0160 | 0.8187 | 0.0229 | 0.7767 |
| | Zero123 [35] | 16 | 0.0147 | 0.8226 | 0.0206 | 0.8045 |
| | Zero123 [35] | 36 | 0.0139 | 0.8247 | 0.0207 | 0.8078 |
| | Ours | 16 | 0.0133 | 0.8423 | 0.0177 | 0.8274 |
| | Ours | 36 | **0.0126** | **0.8472** | **0.0164** | **0.8436** |
| 3 | MonoSDF [82] | N/A | 0.1020 | 0.3963 | 0.0830 | 0.4581 |
| | Ours | 36 | **0.0115** | **0.8587** | **0.0124** | **0.8628** |

Table 4. Quantitative results on reconstructing 3D Shapes using the generated images.

| Dataset | Method | SSIM↑ | PSNR↑ | LPIPS↓ |
|---|---|---|---|---|
| ABO | Zero123 (no interp.) | 0.8796 | 21.33 | 0.0961 |
| | Standard auto-regression | 0.8010 | 16.77 | 0.1854 |
| | Interpolated conditions | 0.7243 | 13.26 | 0.3770 |
| | Interpolated outputs | **0.8925*** | **21.95*** | 0.1246 |
| | Stochastic conditioning [70] | 0.8699 | 20.64 | 0.1106 |
| | Interpolated denoising | 0.8848 | 21.43 | **0.0923** |
| GSO | Zero123 (no interp.) | 0.8710 | 20.33 | 0.1029 |
| | Standard auto-regression | 0.8094 | 16.30 | 0.1801 |
| | Interpolated conditions | 0.7661 | 14.38 | 0.3427 |
| | Interpolated outputs | 0.8799 | 20.44 | 0.1659 |
| | Stochastic conditioning [70] | 0.8658 | 19.97 | 0.1119 |
| | Interpolated denoising | **0.8820** | **20.73** | **0.0958** |

Table 5. Quantitative comparison between our method to different auto-regressive variants. Note that the *Interpolated outputs* variant achieves the highest SSIM and PSNR values on the ABO dataset, but generates blurred images.

MonoSDF [82], our approach capture more details and generates smoother surfaces (note that [82] also relies on additional depth and normal estimations for its reconstruction whereas ours does not).

### 4.4. Ablations

Here, we study the effectiveness of our proposed interpolated denoising process by exploring various auto-regressive generation variants (in Tab. 5 and Fig. 7). Specifically, we investigate four variants of denoising: stan-

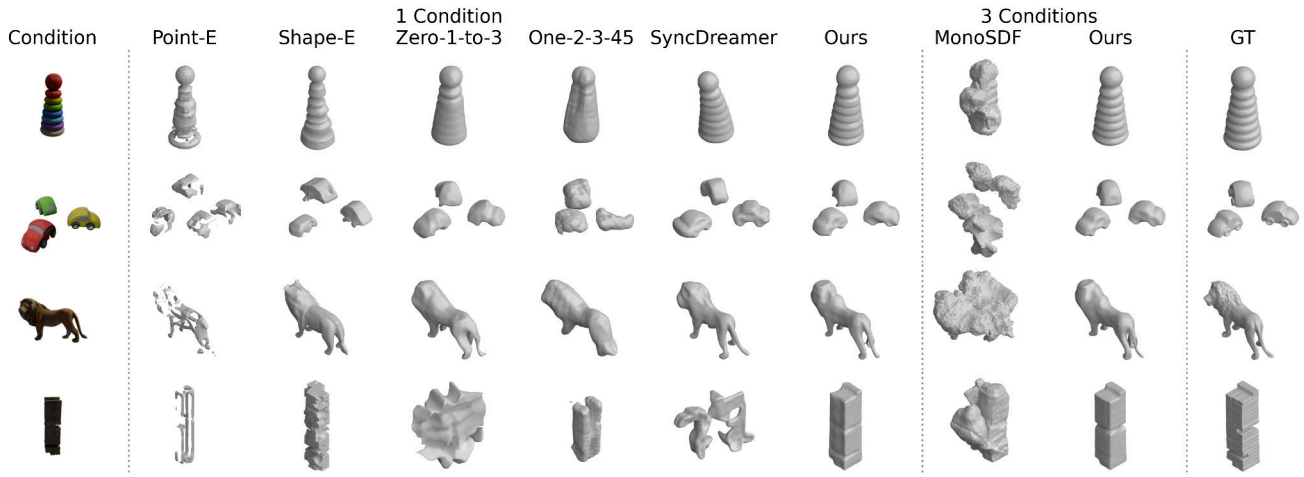



Figure 6. Qualitative results for single-view (1 condition image) and multi-view (3 condition images) 3D Reconstruction. Our method offers the most consistent results across a variety of different reconstructed shapes.

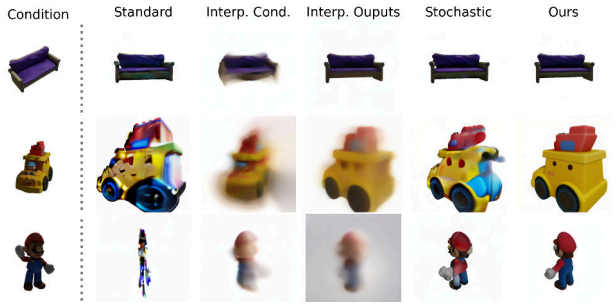



Figure 7. Qualitative comparison for different auto-regressive generations.

dard auto-regression, interpolated conditions, interpolated outputs, and stochastic conditioning which was proposed in [70].

**Standard Auto-regression.** One initial approach to auto-regression involves using the last generated view as the subsequent conditioning. However, this method exhibits bad generation quality, due to the accumulation of errors during the sequential generation process. As each subsequent view relies on the accuracy of the previous one, any inaccuracies or imperfections in earlier stages can compound, leading to a degradation in overall image quality. This limitation highlights the need for more sophisticated auto-regressive strategies to address the issue of error propagation and enhance the quality of generated views.

**Interpolated Conditions and Interpolated Outputs.** Interpolated Conditions and Interpolated Outputs are two straightforward approaches to introduce auto-regressive generation into an existing diffusion model. The former method involves interpolating feature embeddings from condition images and poses, while the latter interpolates the final image feature maps produced by the model. Despite SSIM and PSNR metrics showing favorable results for Interpolated Outputs over others in Tab. 5, as illustrated in visual comparisons in Fig. 7 shows that it leads to blurring of the output views and this is corroborated by larger LPIPS distance.

**Stochastic Conditioning.** We also explore the application of the Stochastic Conditioning Sampler proposed by [70] to the Zero-1-to-3 model. We observe that Stochastic Conditioning fails to deliver the desired auto-regressive generation results; this may be attributed to the specific category on which the diffusion model used by Watson et al. [70] was trained, allowing it to handle plausible condition images more effectively. By contrast, Zero-1-to-3 [35] was trained on a cross-category dataset and designed for zero-shot reconstruction. Additionally, our evaluation data contains out-of-the-distribution data (*i.e.*, ABO [6] and GSO [10]).

## 5. Conclusion

In this work, we have developed ViewFusion, a novel algorithm that addresses the challenge of multi-view consistency in novel-view synthesis with diffusion models. Our approach circumvents the need for fine-tuning or additional modules by integrating an auto-regressive mechanism that incrementally refines view synthesis, utilizing the entire history of previously generated views. Our proposed diffusion interpolation technique extends the denoising process in pre-trained diffusion models from a single-view setting to a multi-view setting without training requirements. Empirical evidence underscores ViewFusion's capability to produce consistently high-quality views, and we achieve significant steps forward in novel view synthesis and 3D reconstruction applications.

## References


[1] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *CVPR*, 2023. 2

[2] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *ICCV*, 2023. 2

[3] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *ICCV*, 2023. 2

[4] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 1, 2

[5] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *CVPR*, 2023. 2

[6] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022. 2, 5, 8

[7] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 2

[8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 5, 2

[9] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *CVPR*, 2023. 2

[10] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, 2022. 2, 5, 8

[11] Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. *arXiv preprint arXiv:2303.17015*, 2023. 2

[12] George Fahim, Khalid Amin, and Sameh Zarif. Single-view 3d reconstruction: A survey of deep learning methods. *Computers & Graphics*, 94:164–190, 2021. 2

[13] Kui Fu, Jiansheng Peng, Qiwen He, and Hanxiao Zhang. Single image 3d object reconstruction based on deep learning: A review. *Multimedia Tools and Applications*, 80:463–498, 2021. 2

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[15] Georgia Gkioxari, Justin Johnson, and Jitendra Malik. Mesh r-cnn. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[16] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. A papier-mache approach to learning 3d surface generation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[17] Jiatao Gu, Qingzhe Gao, Shuangfei Zhai, Baoquan Chen, Lingjie Liu, and Josh Susskind. Learning controllable 3d diffusion models from single-view images. *arXiv preprint arXiv:2304.06700*, 2023. 2

[18] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *ICML*, 2023. 2

[19] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023. 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2, 3

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2, 3

[22] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021. 2

[23] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 2

[24] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023. 6, 7

[25] Animesh Karnewar, Niloy J Mitra, Andrea Vedaldi, and David Novotny. Holofusion: Towards photo-realistic 3d generative modeling. In *ICCV*, 2023. 2

[26] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *CVPR*, 2023. 2

[27] Hiroharu Kato and Tatsuya Harada. Learning view priors for single-view 3d reconstruction. In *CVPR*, 2019. 2

[28] Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *CVPR*, 2023. 2

[29] Jiabao Lei, Jiapeng Tang, and Kui Jia. Generative scene synthesis via incremental view inpainting using rgbd diffusion models. In *CVPR*, 2022. 2

[30] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *ECCV*, 2020. 2

[31] Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. Sdf-srn: Learning signed distance 3d object reconstruction from static images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[32] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 1, 2

[33] Kai-En Lin, Lin Yen-Chen, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023. 2

[34] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization, 2023. 1, 2, 7

[35] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 1, 2, 3, 4, 6, 7, 8

[36] Xinhang Liu, Shiu-hong Kao, Jiaben Chen, Yu-Wing Tai, and Chi-Keung Tang. Deceptive-nerf: Enhancing nerf reconstruction using pseudo-observations from diffusion models. *arXiv preprint arXiv:2305.15171*, 2023. 2

[37] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 1, 2, 4, 6, 7

[38] Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. In *ICLR*, 2023. 2

[39] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion, 2023. 1, 2

[40] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157 vol.2, 1999. 6

[41] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *CVPR*, 2023. 1, 2

[42] 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, 2

[43] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 3

[44] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *CVPR*, 2023. 2

[45] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 6, 7

[46] Evangelos Ntavelis, Aliaksandr Siarohin, Kyle Olszewski, Chaoyang Wang, Luc Van Gool, and Sergey Tulyakov. Autodecoding latent 3d diffusion models. *arXiv preprint arXiv:2307.05445*, 2023. 2

[47] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

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

[49] 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. 6

[50] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. 1, 2

[51] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 2, 3

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[53] Kyle Sargent, Jing Yu Koh, Han Zhang, Huiwen Chang, Charles Herrmann, Pratul Srinivasan, Jiajun Wu, and Deqing Sun. Vq3d: Learning a 3d-aware generative model on imagenet. *arXiv preprint arXiv:2302.06833*, 2023. 2

[54] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation, 2023. 1

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[56] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Neural Information Processing Systems*, 2015. 3

[57] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. *arXiv preprint arXiv:2306.07881*, 2023. 2

[58] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *ICCV*, 2023. 1, 2

[59] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint arXiv:2307.01097*, 2023. 2

[60] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019. 2

[61] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B Goldman, and M. Zollhöfer. State of the art on neural rendering. *Computer Graphics Forum*, 39(2):701–727, 2020. 1

[62] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B Tenenbaum, Frédo Durand, William T Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *arXiv preprint arXiv:2306.11719*, 2023. 2

[63] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, 2023. 2

[64] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023. 1, 2, 6

[65] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 1, 2

[66] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 1, 6

[67] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, 2023. 2

[68] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 5

[69] 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. *arXiv preprint arXiv:2305.16213*, 2023. 1, 2

[70] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 1, 2, 3, 4, 7, 8

[71] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, C. L. Philip Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3d object synthesis, 2023. 2

[72] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3d-aware image generation using 2d diffusion models. *arXiv preprint arXiv:2303.17905*, 2023. 2

[73] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, 2022. 1

[74] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360 views. *arXiv e-prints*, pages arXiv–2211, 2022. 1, 2

[75] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion, 2023. 1, 2

[76] Xianghui Yang, Guosheng Lin, and Luping Zhou. Single-view 3d mesh reconstruction for seen and unseen categories. *IEEE Transactions on Image Processing*, pages 1–1, 2023. 1, 2

[77] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 1

[78] Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models, 2023. 1, 2

[79] Paul Yoo, Jiaxian Guo, Yutaka Matsuo, and Shixiang Shane Gu. Dreamsparse: Escaping from plato's cave with 2d frozen diffusion model given sparse views. *CoRR*, 2023. 2

[80] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[81] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *ICCV*, 2023. 2

[82] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 7

[83] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. In *NeurIPS*, 2022. 2

[84] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. In *SIGGRAPH*, 2023. 2

[85] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[86] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 6

[87] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 2

# ViewFusion: Towards Multi-View Consistency via Interpolated Denoising

## Supplementary Material

## 6. Implementation

We implement our auto-regressive techniques on the pre-trained Zero-1-to-3 [35]. To facilitate single-view generation and spin video generation, we set a maximum offset per step, denoted as $\delta = 10°$ for most cases except 16 view spin video generation. For a fair comparison with SyncDreamer, we modify our setup to match their conditions, where $\delta = 22.5°$ to generate 16 view images, aligning with SyncDreamer's configuration. We have conducted an investigation into various values for the temperature parameters, $\tau_c$ and $\tau_g$, in Eqs.(12) and (13). Our experiments reveal that setting $\tau_c$ to 0.5 and $\tau_g$ to 1.0 leads to superior results, as evidenced by the data presented in Tab. 6. The *Interpolated Denoising* process is illustrated in Algorithm 1. For reconstruction, we optimize the NeuS [66] using the generated multi-view images with their corresponding masks from Zero-1-to-3 [35], SyncDreamer [37] and ours. For One-2-3-45 [34], we directly follow the their pipeline, which requires elevation estimation. To apply ViewFusion on in-the-wild images, we apply an off-the-shelf background removal tool CarveKit to remove the background and adjust the object ratio on the image.

**Algorithm 1** Interpolated denoising with classifier-free guidance

Input: conidition $y$, unconditional scale $u$, $\alpha_t$, $\sigma_t$, $\tau_c$, $\tau_g$
Determine generated trajectory $x_0^1, x_0^2, ..., x_0^N$
Add $x_0^1 \leftarrow y$ to condition set
$\{w_1, w_2..., w_N\} \leftarrow Eq.13$
**for** $n$ from 2 to $N$ **do**
  $x_T \leftarrow$ Sample from $\mathcal{N}(\mathbf{0}, \mathbf{I})$
  **for** $t$ from $T$ to 1 **do**
    $y^i \leftarrow$ Sample $x_0^i$ from condition set
    $\epsilon_t^i \leftarrow \epsilon_t^i(x_t, \emptyset) + u\left(\epsilon_t^i(x_t, y^i) - \epsilon_t^i(x_t, \emptyset)\right)$
    $\epsilon' \leftarrow$ Sample from $\mathcal{N}(\mathbf{0}, \mathbf{I})$
    $x_{t-1}^i \leftarrow \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_t}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\cdot\epsilon_t^i + \sigma_t\epsilon'$
    $x_{t-1}^n \leftarrow \sum_{i=1}^n \omega_i x_{t-1}^i$
  **end for**
  Add $x_0^n$ to condition set
**end for**

## 7. Multi-view generation.

We formulate the weights to single-view generation in Eq 13. In the general case, when given k views, the weights are expressed as follows,

$$\omega_n = \begin{cases} exp(-\frac{\Delta^n}{\tau_c})\text{Softmax}(\frac{e^{-\frac{\Delta^n}{\tau_c}}}{\sum_{n=1}^k e^{-\frac{\Delta^n}{\tau_c}}}) & n = 1, \ldots, k \\ (1 - \sum_{i=1}^k \omega_i)\text{Softmax}(\frac{e^{-\frac{\Delta^n}{\tau_g}}}{\sum_{n=k+1}^N e^{-\frac{\Delta^n}{\tau_g}}}), & n > k \end{cases} \tag{16}$$

where we apply the term $1 - \sum_{i=1}^k \omega_i$ on the generated image weights to ensure sum of all weights equals 1 as a requirement for the objective $\sum_{n=1}^N w_n = 1$.

## 8. Image Rendering

We organize the testing data by using the rendering scripts provided by both Zero-1-to-3 and SyncDreamer respectively. It's important to note that there are slight variations in the camera and lighting settings between the two approaches.

**Camera.** Zero-1-to-3 employs random sampling for the camera distance within a range of $[1.5, 2.2]$. The azimuth and elevation angles for both condition and target images are randomly selected. SyncDreamer maintains a fixed camera distance of 1.5 and samples azimuth angles from a discrete angle set $\{0°, 22.5°, 45°, ..., 337.5°\}$ for both condition and target images. The condition elevation is randomly sampled within the range of $[0°, 30°]$, while the target elevation is fixed at $30°$.

**Lighting.** Zero-1-to-3 uses point light as its lighting model. SyncDreamer, on the other hand, employs a uniform environment light setup. This choice of lighting leads to differences in the rendering results. Specifically, renderings from Zero-1-to-3 exhibit shadows on the backside of the objects, whereas those from SyncDreamer do not.

These discrepancies in rendering impact the evaluation of 3D reconstructions. As we take Zero-1-to-3 as our baseline, we adopt the consistent rendering settings with Zero-1-to-3 to organize test data for fair comparison.

## 9. SSIM and PSNR

In the main manuscript, we mentioned the limitations of SSIM and PSNR in effectively capturing blur, as detailed in Tab. 5. We further underscore these limitations with illustrative examples, as depicted in Fig. 8, where images with higher SSIM and PSNR scores exhibit pronounced blurriness. Our findings highlight the comparative shortcomings of output interpolation when compared with diffusion interpolation. Importantly, we stress that LPIPS provides a more precise assessment of image quality.

| Dataset | Method | Image Quality | | | Multi-view Consistency | | |
|---|---|---|---|---|---|---|---|
| | | SSIM↑ | PSNR↑ | LPIPS↓ | SIFT↑ | LPIPS↓ | CLIP↑ |
| ABO | Zero123 | 0.8796 | 21.33 | 0.0961 | 16.69 | 0.1234 | 0.9725 |
| | $\tau_c = 0.33 + \tau_g = 0.1$ | 0.8633 | 19.78 | 0.1168 | 18.32 | 0.0965 | 0.9804 |
| | $\tau_c = 0.33 + \tau_g = 0.5$ | 0.8788 | 20.86 | 0.0984 | 17.95 | **0.0945** | **0.9815** |
| | $\tau_c = 0.33 + \tau_g = 1.0$ | 0.8804 | 21.06 | 0.0961 | 17.94 | 0.0948 | 0.9812 |
| | $\tau_c = 0.50 + \tau_g = 0.1$ | 0.8753 | 20.56 | 0.1045 | **18.46** | 0.0968 | 0.9813 |
| | $\tau_c = 0.50 + \tau_g = 0.5$ | **0.8848** | 21.35 | 0.0933 | 18.12 | 0.0964 | 0.9813 |
| | $\tau_c = 0.50 + \tau_g = 1.0$ | 0.8848 | **21.43** | **0.0923** | 18.01 | 0.0966 | 0.9812 |
| GSO | Zero123 | 0.8710 | 20.33 | 0.1029 | 15.15 | 0.1054 | 0.9592 |
| | $\tau_c = 0.33 + \tau_g = 0.1$ | 0.8632 | 19.15 | 0.1193 | **19.43** | 0.0675 | 0.9760 |
| | $\tau_c = 0.33 + \tau_g = 0.5$ | 0.8770 | 20.18 | 0.1020 | 18.64 | **0.0664** | **0.9779** |
| | $\tau_c = 0.33 + \tau_g = 1.0$ | 0.8789 | 20.38 | 0.0994 | 18.54 | 0.0671 | 0.9778 |
| | $\tau_c = 0.50 + \tau_g = 0.1$ | 0.8725 | 19.89 | 0.1081 | 19.13 | 0.0675 | 0.9764 |
| | $\tau_c = 0.50 + \tau_g = 0.5$ | 0.8812 | 20.62 | 0.0969 | 18.30 | 0.0689 | 0.9773 |
| | $\tau_c = 0.50 + \tau_g = 1.0$ | **0.8820** | **20.73** | **0.0958** | 17.95 | 0.0676 | 0.9773 |

Table 6. Experiments about condition image weights.

Figure 8. Visual comparison for SSIM and PSNR limitation in capturing blur.

## 10. Limitation

While our model, ViewFusion, demonstrates promising performance in significantly enhancing the multi-view consistency of the original Zero-1-to-3 framework, there are certain limitations that remain unaddressed by the current framework.

First, ViewFusion relies on using all generated images to guide the generation process. This requirement necessitates additional memory to store these images and imposes a sequential nature to the generation process. In contrast, the original Zero-1-to-3 can proceed spin video generation as a batch process and generate views in parallel, resulting in a more time-efficient approach. The sequential generation nature of ViewFusion can lead to additional time consumption. Considering to generate a single image, Zero-1-to-3 takes roughly 4s, while our methods takes $4s \sim 45s$ (from 1 condition to 24 conditions) depending on the size of the condition set.

Second, ViewFusion heavily relies on the pre-trained Zero-1-to-3 model. While it is generally effective, there are still instances where it fails, particularly under certain specific views. Even with the integration of auto-regressive generation, it cannot entirely mitigate this limitation, as demonstrated in the Fig. 9 1st and 2nd examples.

Third, although current pose-conditional diffsuion models [35, 37, 59, 63, 71, 75] have been trained on large-scale 3D dataset, *i.e.*, Objaverse [7, 8], they are still struggling to deal with scenes that comprise intricate details (*e.g.*, human faces, detailed textures) as shown by the 3rd and 4th examples in the Fig. 9, complex scenes, as shown by 5th examples in Fig. 9, and the models may struggle with elevation angle ambiguity, as demonstrated by the 6th example in Fig. 9. In these cases, the model's performance may be limited in capturing all the fine-grained information and nuances.

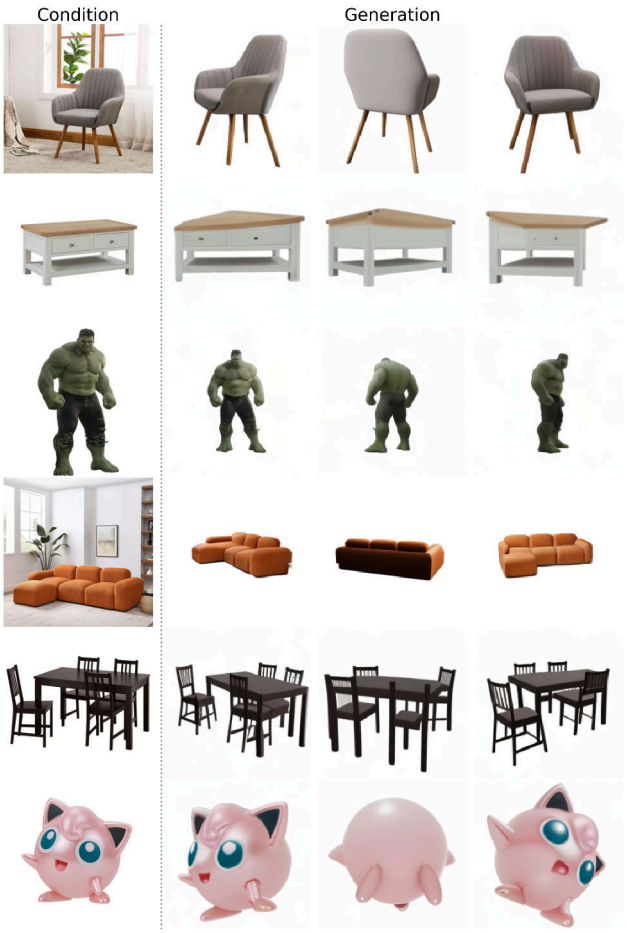



Figure 9. Visual examples for failure cases. The failure cases mainly includes failure under specific views (1st and 2nd rows), face (3rd row), detailed textures (4th row), complex scenes (5th row), and elevation angle ambiguity (6th row)

## 11. Application

**Multi-view generation.** As mentioned in the main manuscript, thanks to the multi-view conditioned ability by the introduced interpolated denoising process, we could extend the single-view conditioned model into multi-view conditioned model easily, thus enabling support for multi-view reconstruction. The quantities results presented in Tab. 3 and we provide qualitative comparison in Fig. 11 here to further demonstrate the advantages of our method, as it consistently yields improved reconstructions with an increasing number of views. This clear improvement demonstrate the effectiveness of our proposed techniques in handling multi-view condition images.

**Consistent BRDF decomposition.** In our experimental observations, we identified a particular challenge encountered by the pre-trained decoder, which often struggles to effectively distinguish between shadows and surface textures in images. To overcome this limitation, we introduced a dedicated decomposition decoder, specifically designed to meticulously separate these visual elements. When this decomposition decoder is integrated with our interpolated denoising approach, it not only upholds multi-view consistency but also exhibits the potential to excel in novel-view decomposition and rendering tasks.

This novel combination of techniques offers promising possibilities. By leveraging decomposed BRDF (Bidirectional Reflectance Distribution Function) maps, we gain greater control over the lighting and shape geometry of the scenes. The availability of normal maps enhances our ability to manipulate the lighting conditions, promising more flexibility in rendering as shown in Fig. 13. With this level of control, we can explore various exciting applications, such as dynamic relighting, creative scene composition, and the generation of captivating visual effects. This opens up new avenues for artistic and practical image and video manipulation, granting artists and professionals the tools to craft engaging and visually stunning content.

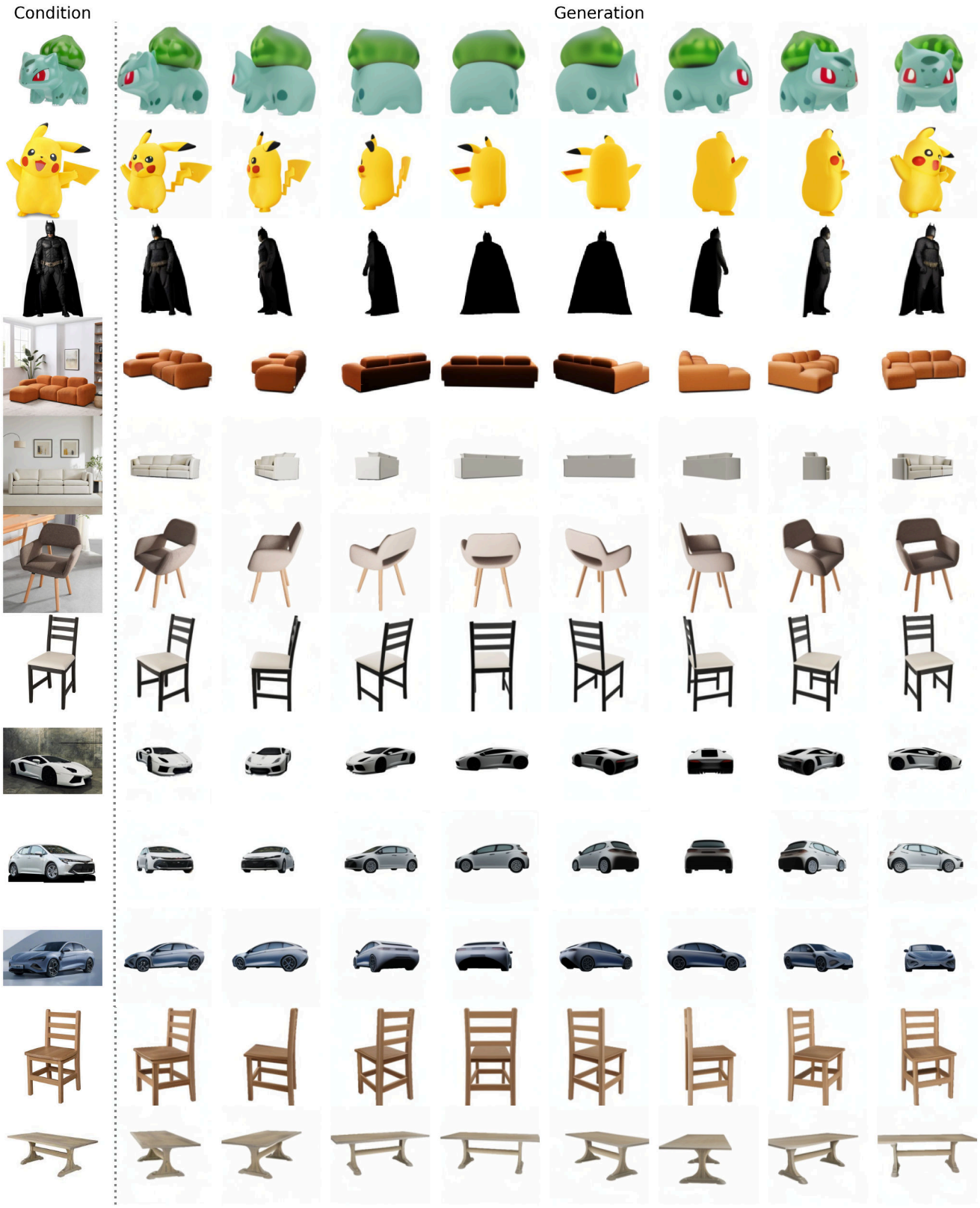



Figure 10. Visualization on real images. Images were downloaded online, where foreground objects were segmented and the image was resized to be aligned with pre-training images.

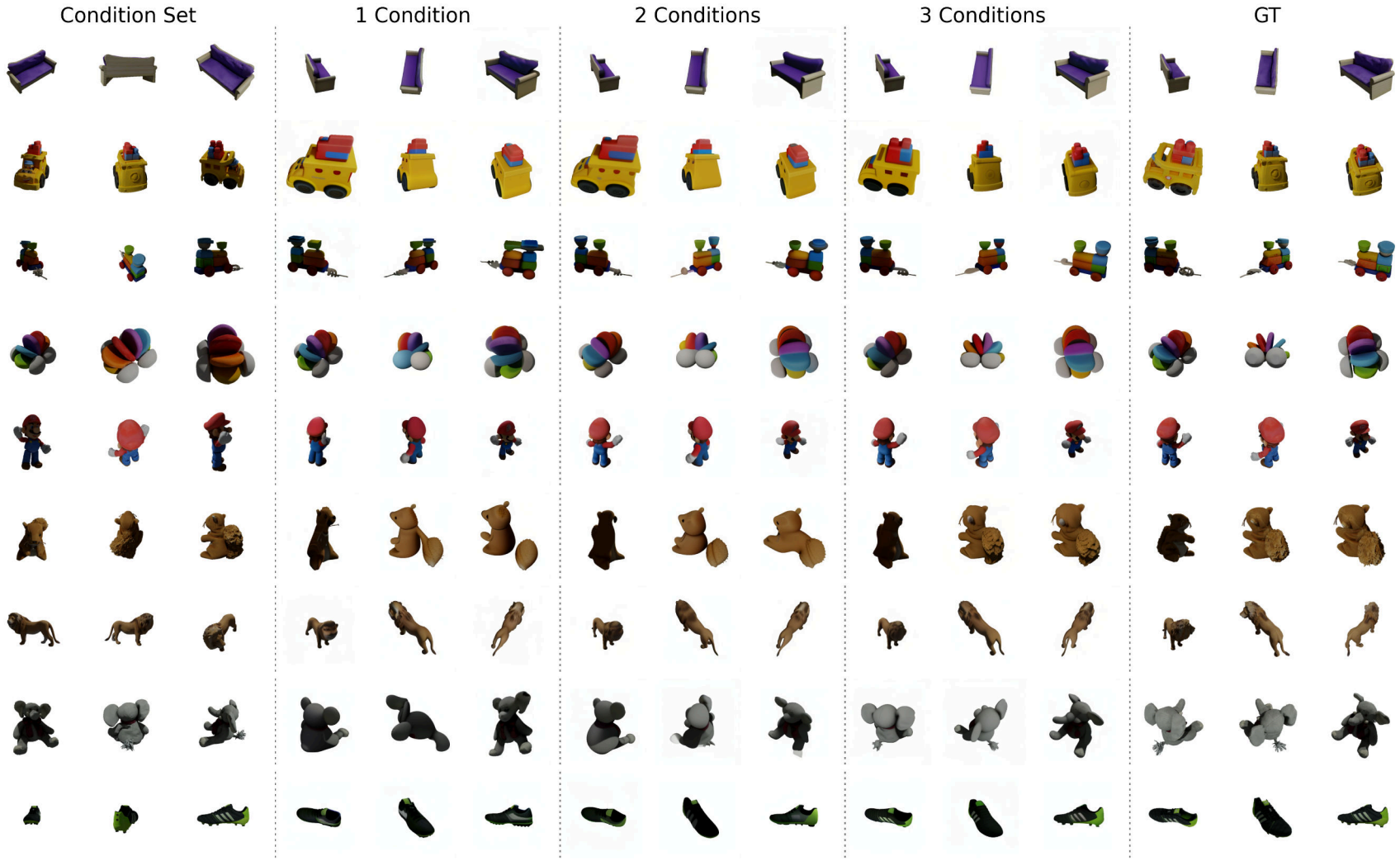



Figure 11. Qualitative comparison for Multi-view reconstruction.

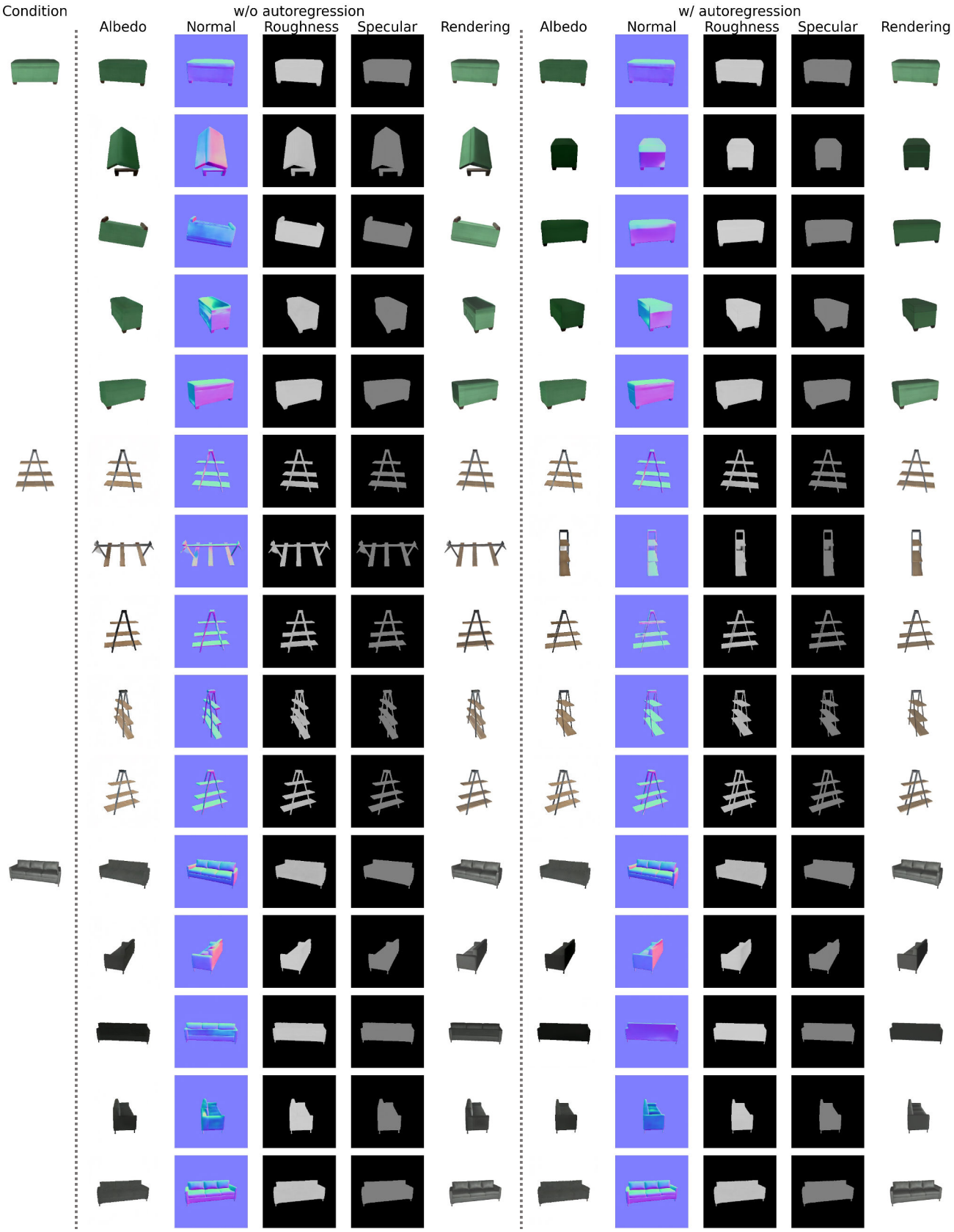



Figure 12. Qualitative comparison for BRDF decomposition w/o vs w/ autoregression.

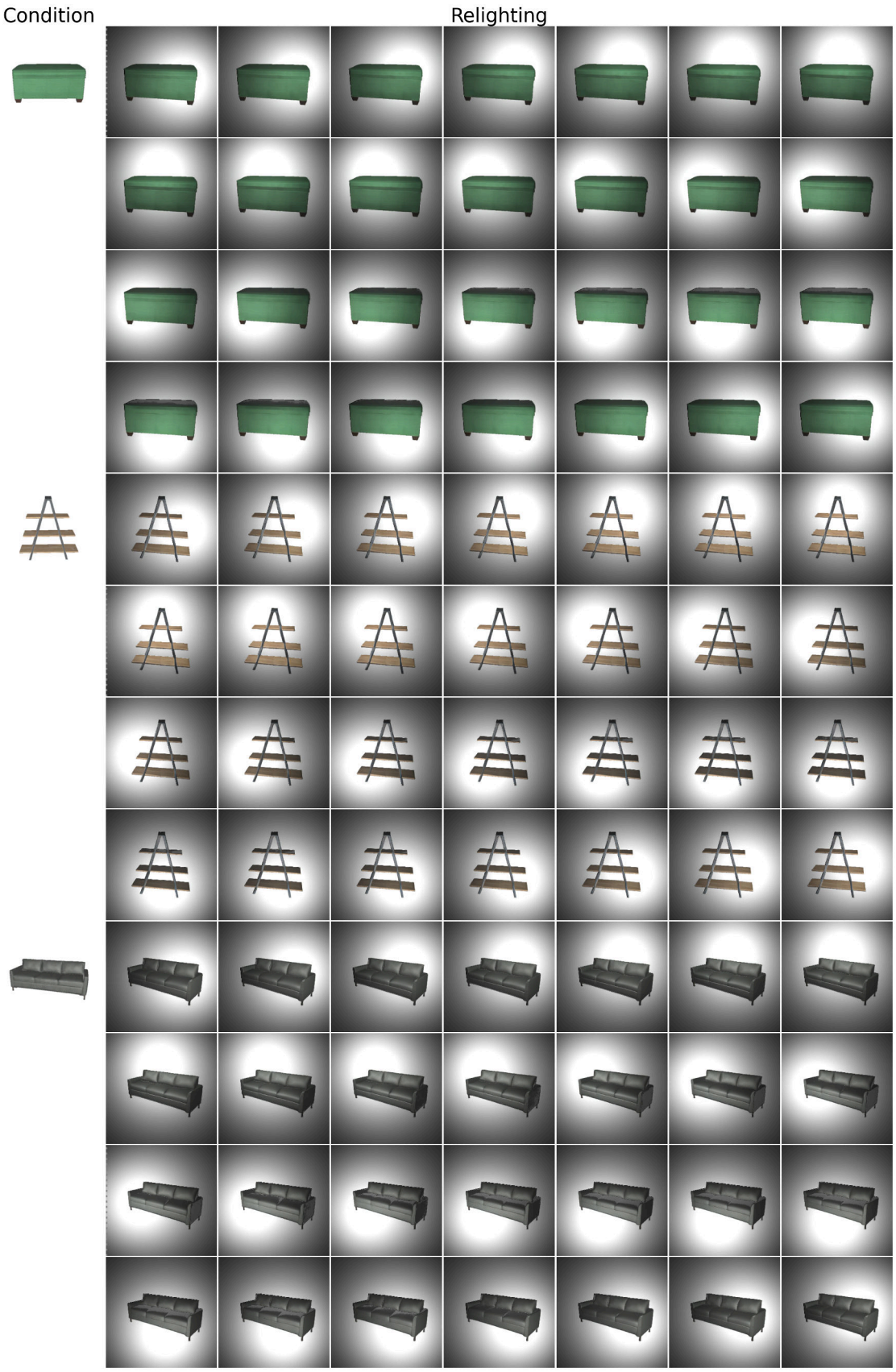



Figure 13. Qualitative comparison for relighting.