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, τ_c and τ_g, in Eqs.(12) and (13). Our experiments reveal that setting τ_c to 0.5 and τ_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, α_t, σ_t, τ_c, τ_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.

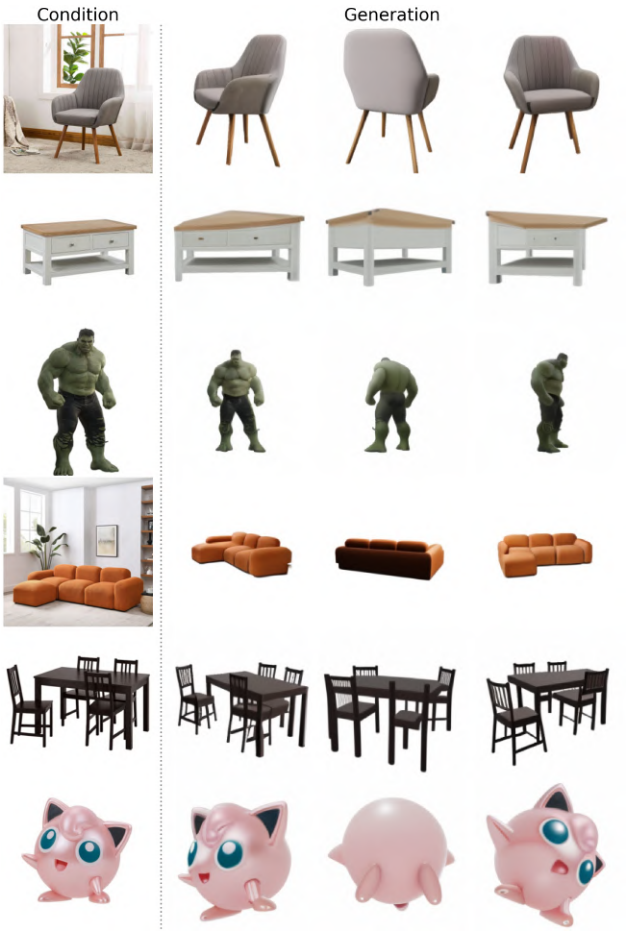

Figure 9. Visual examples for failure cases. The failure cases mainly includes failure under specific views (1st and 2nd rows), face (3rd row), detailed textures (4th row), complex scenes (5th row), and elevation angle ambiguity (6th row)

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.

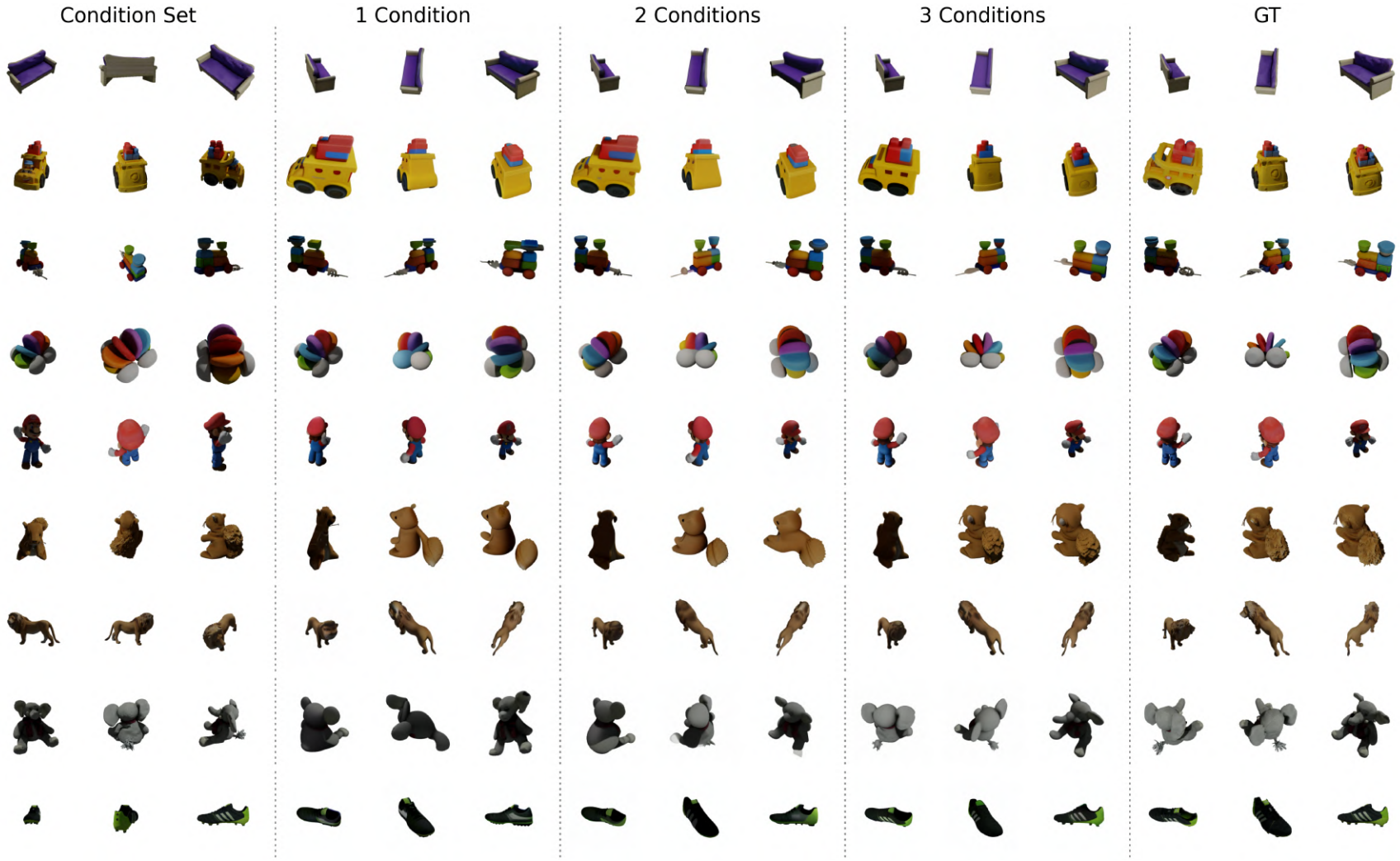

Figure 11. Qualitative comparison for Multi-view reconstruction.

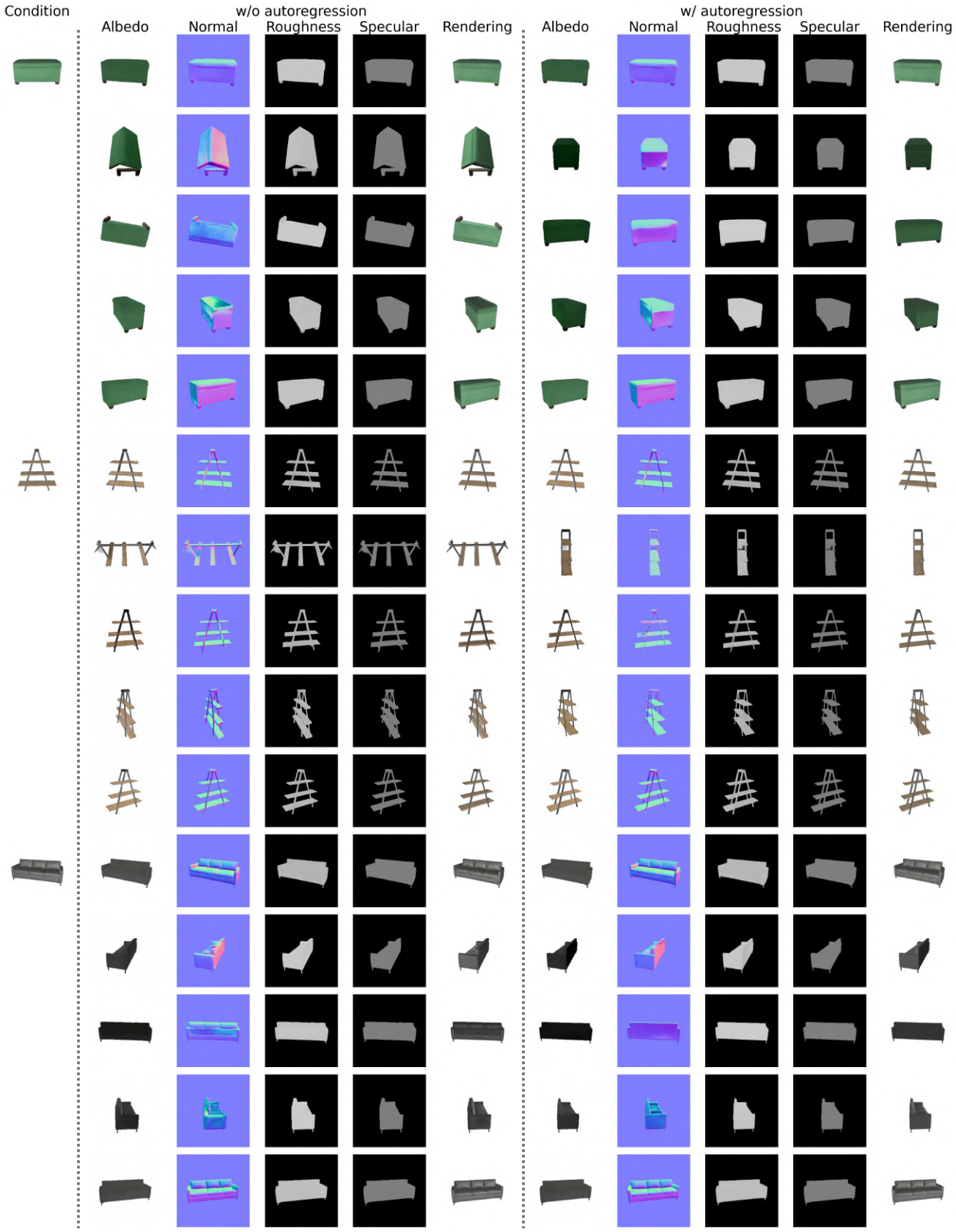

Figure 12. Qualitative comparison for BRDF decomposition w/o vs w/ autoregression.

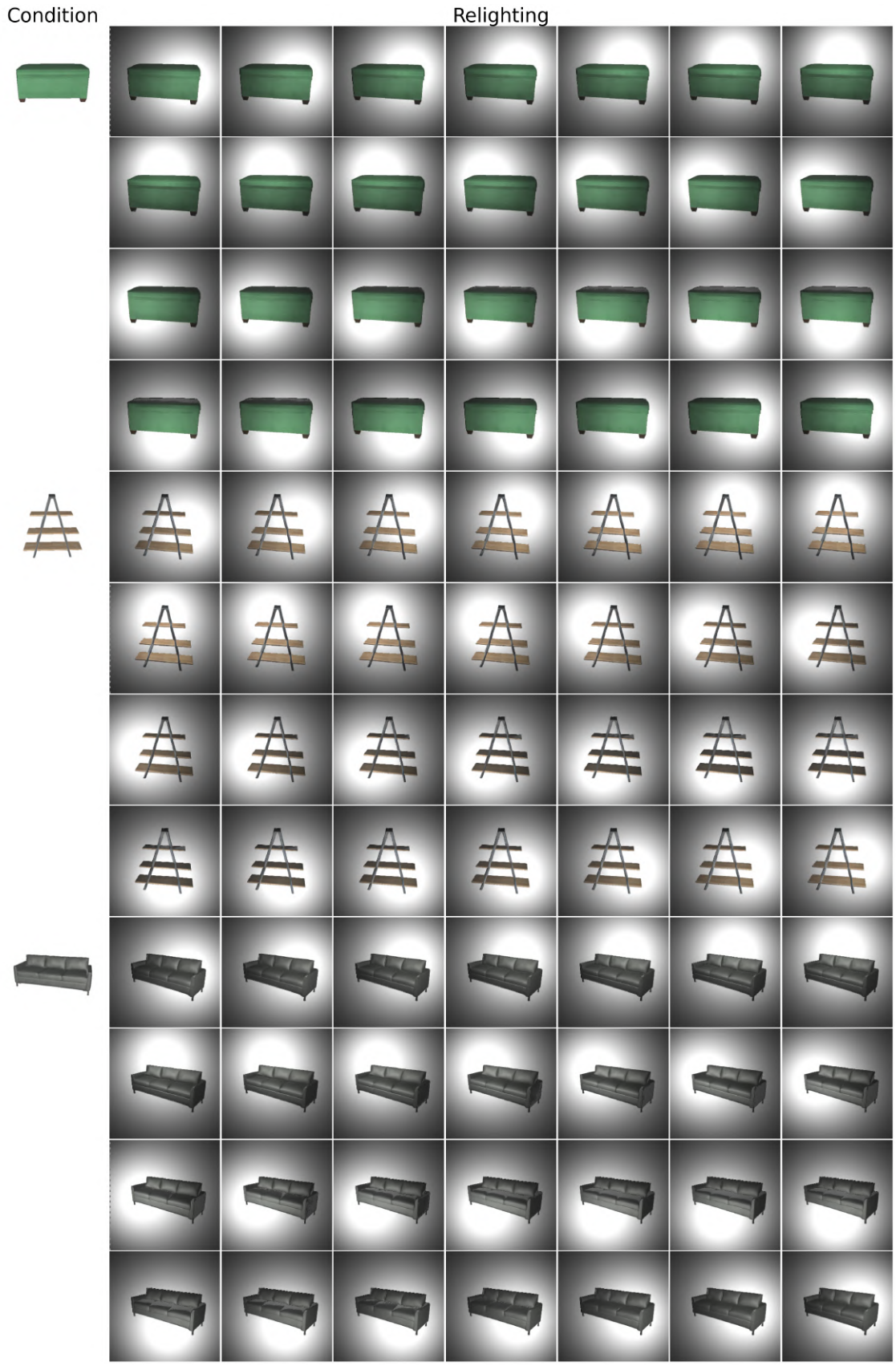

Figure 13. Qualitative comparison for relighting.