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

DreamComposer: Controllable 3D Object Generation via Multi-View Conditions

Yunhan Yang[1*] Yukun Huang[1*] Xiaoyang Wu[1] Yuan-Chen Guo[3,4]
Song-Hai Zhang[4] Hengshuang Zhao[1] Tong He[2] Xihui Liu[1†]

[1] The University of Hong Kong [2] Shanghai Artificial Intelligence Lab [3] VAST [4] Tsinghua University
[*] Equal Contribution Project Page: https://yhyang-myron.github.io/DreamComposer/

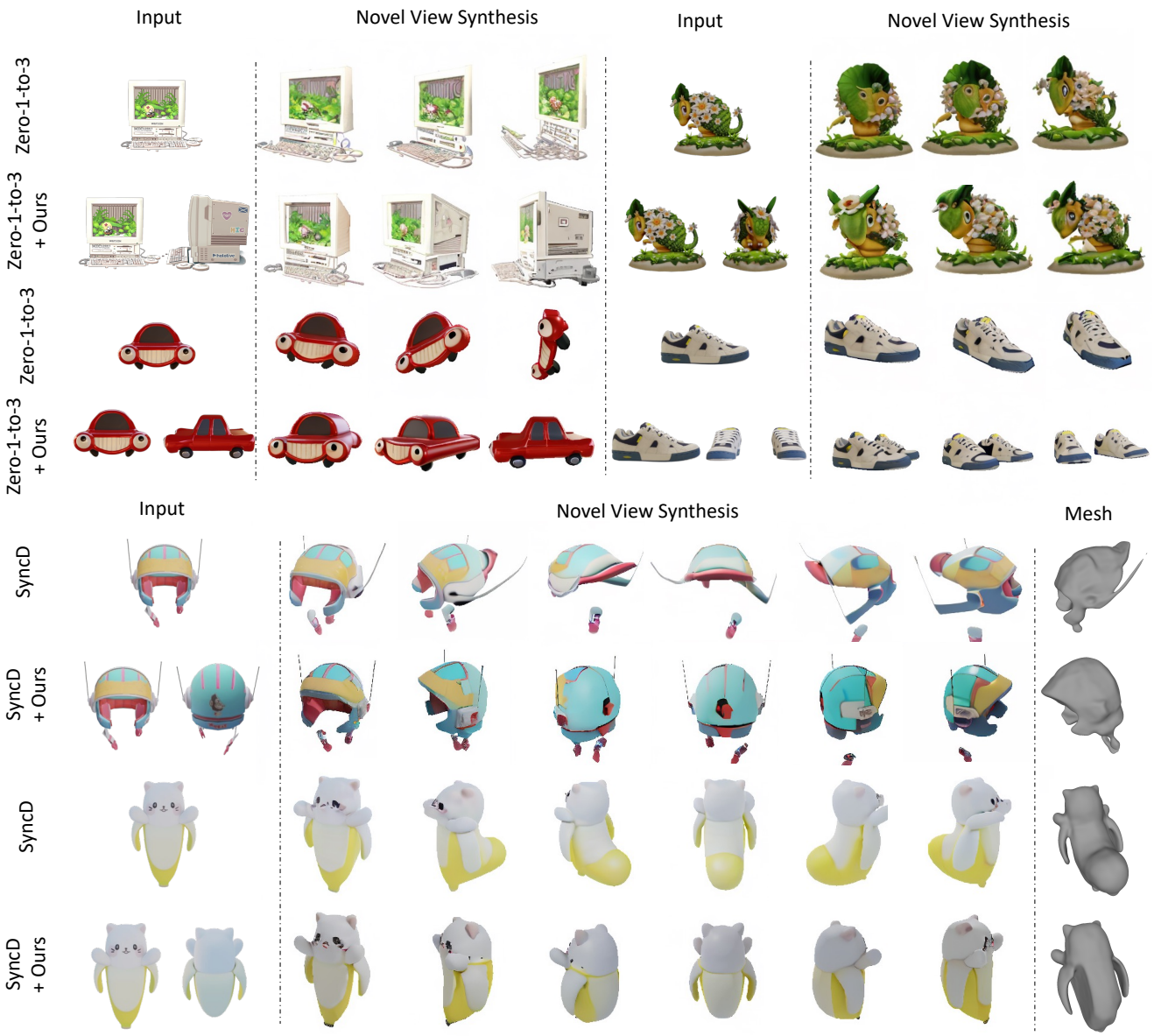

Figure 1. **DreamComposer** is able to generate controllable novel views and 3D objects via injecting multi-view conditions. We incorporate the method into the pipelines of Zero-1-to-3 [17] and SyncDreamer (SyncD) [18] to enhance the control ability of those models.

† Corresponding author.

Abstract

Utilizing pre-trained 2D large-scale generative models, recent works are capable of generating high-quality novel views from a single in-the-wild image. However, due to the lack of information from multiple views, these works encounter difficulties in generating controllable novel views. In this paper, we present DreamComposer, a flexible and scalable framework that can enhance existing view-aware diffusion models by injecting multi-view conditions. Specifically, DreamComposer first uses a view-aware 3D lifting module to obtain 3D representations of an object from multiple views. Then, it renders the latent features of the target view from 3D representations with the multi-view feature fusion module. Finally the target view features extracted from multi-view inputs are injected into a pre-trained diffusion model. Experiments show that DreamComposer is compatible with state-of-the-art diffusion models for zero-shot novel view synthesis, further enhancing them to generate high-fidelity novel view images with multi-view conditions, ready for controllable 3D object reconstruction and various other applications.

1. Introduction

3D object generation is an emerging research topic in computer vision and graphics, serving a wide range of applications such as augmented reality (AR), virtual reality (VR), film production, and game industry. With 3D object generation models, users and designers can easily create the desired 3D assets with text or image prompts, without requiring considerable human endeavors by human experts.

Recently, diffusion models [9, 29] achieve remarkable success in generating 2D images from texts, which inspires the exploration of 3D object generation using 2D diffusion priors [5, 12, 15, 26, 35, 39]. Although great 3D generation results have been achieved [39], 2D diffusion models lack view control and struggle to provide view-consistent supervision, resulting in various quality issues of 3D generation such as multiple faces and blurry details. To alleviate this problem, Zero-1-to-3 [17] empowers 2D diffusion models with viewpoint conditioning, enabling zero-shot novel view synthesis (NVS) conditioned on a single-view image and image-to-3D object generation. Considering the inconsistent output of Zero-1-to-3, a series of subsequent works [18, 19, 31, 41, 44] are proposed to improve the 3D consistency of the generated multi-view images. However, limited by the incomplete information of single-view input, these methods inevitably encounter unpredictable and implausible shapes and textures when predicting novel views. For example, as shown on the right side of the third row of Figure 1, the actual number of shoes cannot be determined if only given a side view of the shoes. In other words, novel view synthesis and 3D object generation are not fully controllable with only single-view image conditions.

To address this problem, our core idea is to introduce flexible multi-view image conditioning to diffusion models, enabling more controllable novel view synthesis and 3D object reconstruction. For example, based on the front view, back view, and side view of an object drawn by designers, the model will generate images of other viewpoints that are consistent with the multiple input images. It also allows interactive 3D generation where users can provide conditioning images from new viewpoints if the generated 3D objects do not follow the user intention. However, such an attempt is challenging for two reasons. Firstly, it is non-trivial to integrate arbitrary numbers of input views into consistent 3D representations that can guide the generation of the target view image. Secondly, it is challenging to design a flexible framework that is compatible with and can be plugged into existing models such as Zero-1-to-3 [17] and SyncDreamer [18] to empower multi-view conditioning for various models.

To this end, we propose DreamComposer, a scalable and flexible framework that can extend existing view-conditioned models to adapt to an arbitrary number of multi-view input images. DreamComposer comprises three stages: target-aware 3D lifting, multi-view feature fusion, and target-view feature injection. (i) Target-Aware 3D Lifting encodes multi-view images into latent space and then lifts the latent features to 3D tri-planes [2]. The tri-plance representation with latent features is compact and efficient, and the target-view-aware 3D lifting design allows the network to focus more on building 3D features related to the target view. (ii) Multi-View feature fusion renders and fuses the 3D features from different views to target-view 2D features with a novel composited volume rendering approach. (iii) Target-View Feature Injection injects the latent features from the previous stage into the diffusion models with a ControlNet-like structure. The injection module takes the relative angle as condition, allowing for adaptive gating of multi-view conditions. DreamComposer can be flexibly plugged into existing models, such as Zero-1-to-3 [17] and SyncDreamer [18], and endow them with the ability to handle multi-view input images, as shown in Figure 1.

In summary, we propose DreamComposer, a scalable and flexible framework to empower diffusion models for zero-shot novel view synthesis with multi-view conditioning. The scalability and flexibility of DreamComposer are empowered by our novel design of the target-aware 3D lifting, multi-view feature fusion, and target-view feature injection modules. Extensive experiments show that DreamComposer is compatible with recent state-of-the-art methods, endowing high-fidelity novel view synthesis, controllable 3D object reconstruction, and various other applications such as controllable 3D object editing and 3D character modeling with the ability to take multi-view inputs.

2. Related Work

Zero-shot Novel View Synthesis. Previous works [8, 14, 22] on novel view synthesis are generally trained on datasets with limited scenes or categories and cannot generalize to in-the-wild image inputs. Recently, diffusion models [29, 30] trained on large-scale Internet data have demonstrated powerful open-domain text-to-image generation capabilities. This success inspired the community to implement zero-shot novel view synthesis by fine-tuning these pre-trained diffusion models. Zero-1-to-3 [17] fine-tuned the Stable Diffusion model [29] on the large 3D dataset Objaverse [6], achieving viewpoint-conditioned image synthesis of an object from a single in-the-wild image. Based on Zero-1-to-3, several subsequent works [16, 18, 19, 31, 41, 44] aim to produce multi-view consistent images from a single input image to create high-quality 3D objects. However, limited by the ambiguous information of single input image, these models might produce uncontrollable results when rendering novel views.

Diffusion Models for Novel View Synthesis. In addition to fine-tuning directly on the pre-trained text-image diffusion models, Some recent works [3, 8, 14, 40, 47, 49] also attempt to combine diffusion models with 3D priors for novel view synthesis. GeNVS [3] integrates geometry priors in the form of a 3D feature volume into the 2D diffusion backbone, producing high-quality, multi-view-consistent renderings on varied datasets. NerfDiff [8] distills the knowledge of a 3D-aware conditional diffusion model into NeRF at test-time, avoiding blurry renderings caused by severe occlusion. While remarkable outcomes have been obtained for particular object categories from ShapeNet [4] or Co3D [28], the challenge of designing a generalizable 3D-aware diffusion model for novel view synthesis from any in-the-wild inputs remains unresolved.

3D Object Generation. Due to the limited size of existing 3D datasets, it remains challenging to train generative 3D diffusion models [13, 23, 24, 37] using 3D data. With pre-trained text-to-image diffusion models and score distillation sampling [26], DreamFusion-like methods [5, 10, 12, 15, 21, 26, 35, 39, 48] have achieved remarkable text-to-3D object generation by distilling 2D image priors into 3D representations. Some methods [20, 27, 33, 34, 42, 43] utilize similar distillation approaches to execute image-to-3D tasks. Since these works, which rely on an optimization strategy, have not previously encountered real 3D datasets, they face the Janus (multi-face) problem, making it challenging to generate high-quality 3D object shapes.

3. Method

DreamComposer aims to empower existing diffusion models for zero-shot novel view synthesis [17, 18, 44] with multi-view conditions. It consists of three components: (i) *Target-Aware 3D Lifting* extracts 2D features from multi-view inputs and transforms them into 3D representations (Sec. 3.1); (ii) *Multi-View Feature Fusion* renders and fuses the 3D features from different views to target-view 2D features with a novel composited volume rendering approach (Sec. 3.2); (iii) *Target-View Feature Injection* injects the target-view features extracted from multi-view inputs into the diffusion models for multi-view controllable novel view synthesis (Sec. 3.3). All components are optimized in an Adapter [11, 45] fashion (Sec. 3.4). An overview pipeline of DreamComposer is demonstrated in Figure 2.

Formulation. Given a main view $\mathbf{x}_1 \in \mathbb{R}^{H\times W\times 3}$ and several additional views $\mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n$ of an object, our target is to synthesize the novel view $\hat{\mathbf{x}}$ with the relative angle $\Delta\gamma$ to the main view. With the relative angle $\Delta\gamma$, we can calculate the relative camera rotation $R \in \mathbb{R}^{3\times 3}$ and translation $T \in \mathbb{R}^3$. In general, we aim to learn a model $\mathcal{M}$ that can synthesize a novel view $\hat{\mathbf{x}}_{R,T}$ from a main view $\mathbf{x}_1$ and multiple conditional views $\mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n$:

$$\hat{\mathbf{x}}_{R,T} = \mathcal{M}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n, R, T). \tag{1}$$

3.1. Target-Aware 3D Lifting

Existing diffusion models [17, 18] for zero-shot novel view synthesis are specialized for single-view input and therefore cannot handle an undefined number of multi-view inputs. For a scalable solution, we propose to lift 2D features from different views into 3D representations, ready for view-conditional control.

2D-to-3D Feature Lifting. Given an input image $\mathbf{x}_i \in \mathbb{R}^{H\times W\times 3}$ from the camera view i, we first utilize the image encoder of Stable Diffusion [29] to encode it into latent feature $f_i \in \mathbb{R}^{H'\times W'\times 4}$, where $H' \times W'$ is down-sampled image size. Then, we introduce a 3D lifting module with a convolutional encoder structure with self-attention and cross-attention layers. The 3D lifting module lifts the 2D latent feature f_i into a 3D representation $F_i \in \mathbb{R}^{H'\times W'\times 32\times 3}$ conditioned on the relative angle $\Delta\gamma$. We adopt the triplane [2] feature $F_i = \{F_i^{xy}, F_i^{xz}, F_i^{yz}\}$ as the 3D representation as it is *compact and efficient enough to alleviate the high training cost caused by multi-view inputs*. Note that the 2D-to-3D feature lifting is performed in latent space, which *significantly reduces the computational cost*.

The network structure of the 3D lifting module includes self-attention layers, cross-attention layers, and convolutional layers. Here we design a view conditioning mechanism based on cross-attention, enabling adaptive 3D lifting. Specifically, we take the angle difference between the input view and the target view as a condition and inject it into the 3D lifting module through the cross-attention layers. This mechanism allows 3D lifting to *focus more on building 3D features related to the target view, rather than trying to construct a complete 3D representation.*

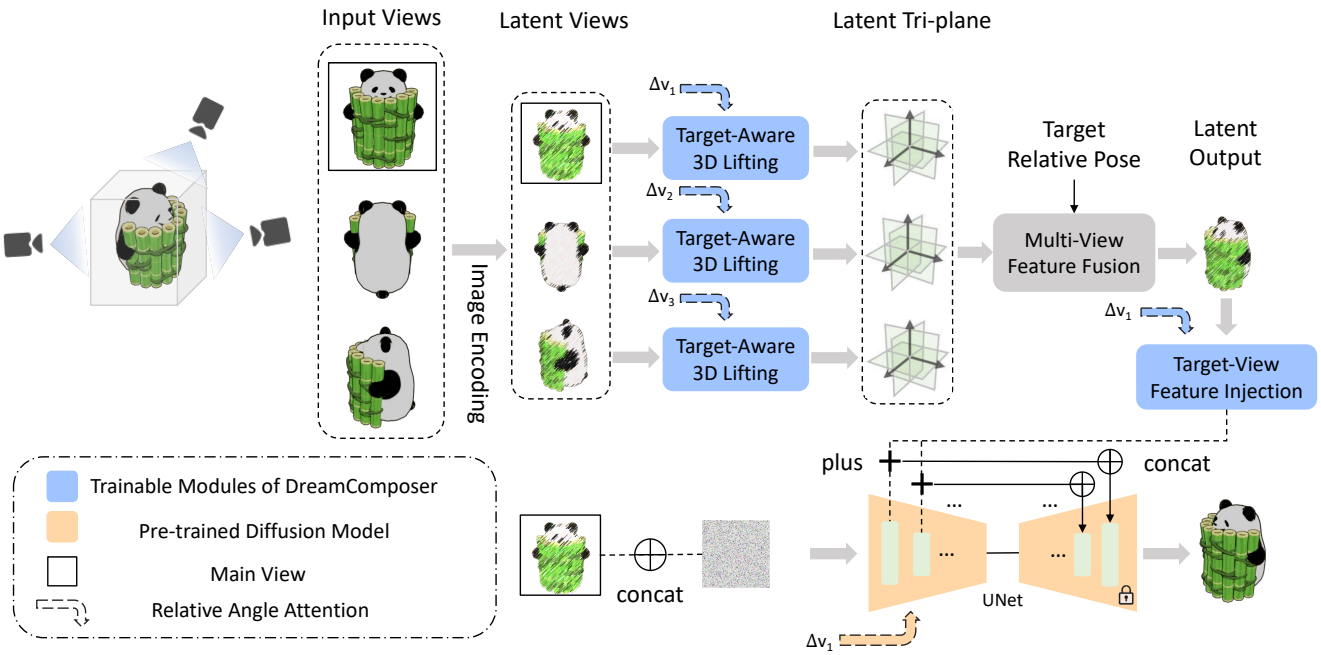

Figure 2. An overview pipeline of **DreamComposer**. Given multiple input images from different views, DreamComposer extracts their 2D latent features and uses a 3D lifting module to produce tri-plane 3D representations. Then, the multi-view condition rendered from 3D representations is injected into the pre-trained diffusion model to provide target-view auxiliary information.

Multi-View Cases. Given multiple input images from n different views, i.e. $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$, we can obtain their tri-plane features $\{F_1, F_2, ..., F_n\}$ via 2D image encoding and 2D-to-3D feature lifting. These tri-plane features are ready for providing target-view auxiliary information in subsequent multi-view conditioning.

3.2. Multi-View Feature Fusion

After obtaining the 3D features $\{F_1, F_2, ..., F_n\}$ of input images from n different views, target-view latent feature f_t can be extracted from these 3D features as the condition for the diffusion model.

To render the target-view latent feature f_t, 3D features from different views need to be fused. However, this is tricky because these 3D features are lifted in different camera spaces and are not aligned. To deal with it, we use a composited volume rendering approach: (1) sampling ray points from the target view; (2) projecting these points onto different input-view camera spaces; (3) indexing and aggregating 3D point features from different views; (4) integrating point features along the target-view rays to render the desired latent feature f_t.

In particular, we adopt a weighting strategy to adaptively aggregate 3D point features from different inputs, considering that different input views contribute differently to the target view. Given n input views, the azimuth differences between them and the target view are denoted as $\Delta\theta_1, \Delta\theta_2, ..., \Delta\theta_n$. Then, the weight of input view i can be formulated as:

$$\lambda_i = \frac{\cos \Delta\theta_i + 1}{2}, \tag{2}$$

and the weighted 3D point feature aggregation across different views is formulated as:

$$f_p^t = \sum_{i=1}^{n} \bar{\lambda}_i \cdot f_p^i, \tag{3}$$

where f_p^t and f_p^i denote feature embeddings of 3D point p from target view and input view i, respectively; while $\bar{\lambda}_i$ is the nomalized weight of input view i calculated by $\lambda_i / \sum_{i=1}^{n} \lambda_i$. Finally, all sampled 3D point's features f_p^t are integrated along the target-view rays using the volume rendering equation [22] and yield f_t.

3.3. Target-View Feature Injection

Latent feature f_t contains rich target-view information extracted from multi-view inputs. We inject f_t into the diffusion model's UNet to provide multi-view conditions. To achieve this, we follow ControlNet [45] structure for target-view feature injection. Specifically, we clone the network blocks of the diffusion model's UNet to trainable copies. These copies, serving as target-view feature injection modules, take the latent feature f_t as conditional input and predict residuals added to the intermediate outputs of UNet.

Most details are consistent with ControlNet [45], except that the input layer needs to be modified to match the size

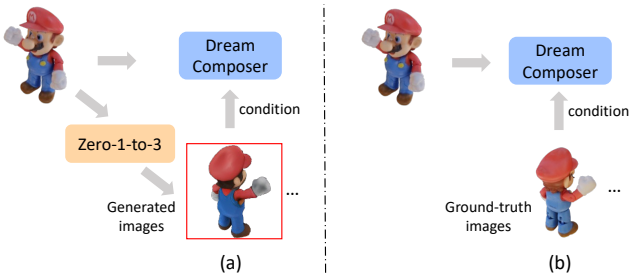

Figure 3. **Different numbers of ground-truth inputs.** Our model is capable of handling a variety of ground-truth input quantities.

of latent input f_t. Besides, we also take the angle difference between the main view and the target view as a condition, and inject it into the multi-view injection module through the cross-attention layers. This design enables adaptive gating of multi-view conditions: auxiliary information from multiple views is less important when the main view and the target view are close.

3.4. Training and Inference

In Sec. 3.1, Sec. 3.2, and Sec. 3.3, we respectively introduce the target-aware 3D lifting, multi-view feature fusion, and target-view feature injection modules, empowering the pre-trained diffusion model with multi-view inputs. Among those modules the target-aware 3D lifting and target-view injection modules are trainable. To train these additional modules, we always sample three views of objects in each iteration, including a front view, a back view, and a random view. This sampling strategy improves training efficiency while encouraging generalization to arbitrary view inputs. Given multi-view input images, we further propose a two-stage training paradigm.

In the first stage, we pre-train the target-aware 3D lifting module on the proxy task of sparse view reconstruction. Given several input views of an object, the 3D lifting module is encouraged to predict novel views correctly, with a mean square error (MSE) loss in latent space as objective.

In the second stage, a pre-trained diffusion model such as Zero-1-to-3 [17] is introduced as the frozen backbone. To enhance it with multi-view conditioning, our target-aware 3D lifting, multi-view feature fusion, and target-view feature injection are integrated and optimized jointly. We use diffusion loss and MSE loss as in the first stage for training.

In the inference stage, the trained model is flexible and can take one or more images from different views as inputs, enabling zero-shot novel view synthesis under multi-view conditions. It also benefits downstream 3D reconstruction and generation tasks with scalability and controllability.

4. Experiments

We evaluate the effectiveness of DreamComposer on zero-shot novel view synthesis and 3D object reconstruction.

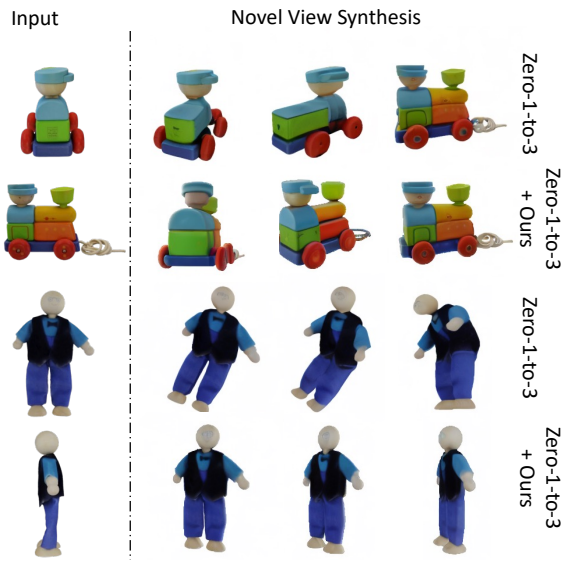

Figure 4. Qualitative comparisons with Zero-1-to-3 [17] in controllable novel view synthesis. DC-Zero-1-to-3 effectively generates more controllable images from novel viewpoints by utilizing conditions from multi-view images.

(a) Elevation Degree - 0

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-1-to-3 [17]	20.82	0.840	0.139
Zero-1-to-3+Ours	**25.25**	**0.888**	**0.088**

(b) Elevation Degree - 15

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-1-to-3	21.38	0.837	0.131
Zero-1-to-3+Ours	**25.85**	**0.891**	**0.083**

(c) Elevation Degree - 30

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-1-to-3	21.66	0.837	0.128
Zero-1-to-3+Ours	**25.63**	**0.885**	**0.086**

Table 1. Quantitative analysis of novel view synthesis using the GSO dataset is presented, employing four distinct angles as inputs. The closest image to the desired viewpoint serves as the input for Zero-1-to-3 and the primary view for DC-Zero-1-to-3, with the remaining three images acting as auxiliary views for DC-Zero-1-to-3. Additionally, we compute results for input elevation angles set at 0, 15, and 30 degrees, respectively.

Method	PSNR↑	SSIM↑	LPIPS↓
Realfusion [20]	15.26	0.722	0.283
Zero-1-to-3 [17]	18.93	0.779	0.166
SyncDreamer [18]	20.05	0.798	0.146
SyncDreamer+Ours	**20.52**	**0.828**	**0.141**

Table 2. Quantitative comparisons of novel view synthesis on GSO dataset. We employ images generated from diffusion models as our additional condition-view.

Datasets, evaluation metrics, and implementation details are provided in Section 4.1 and Section 4.2. Our model is able

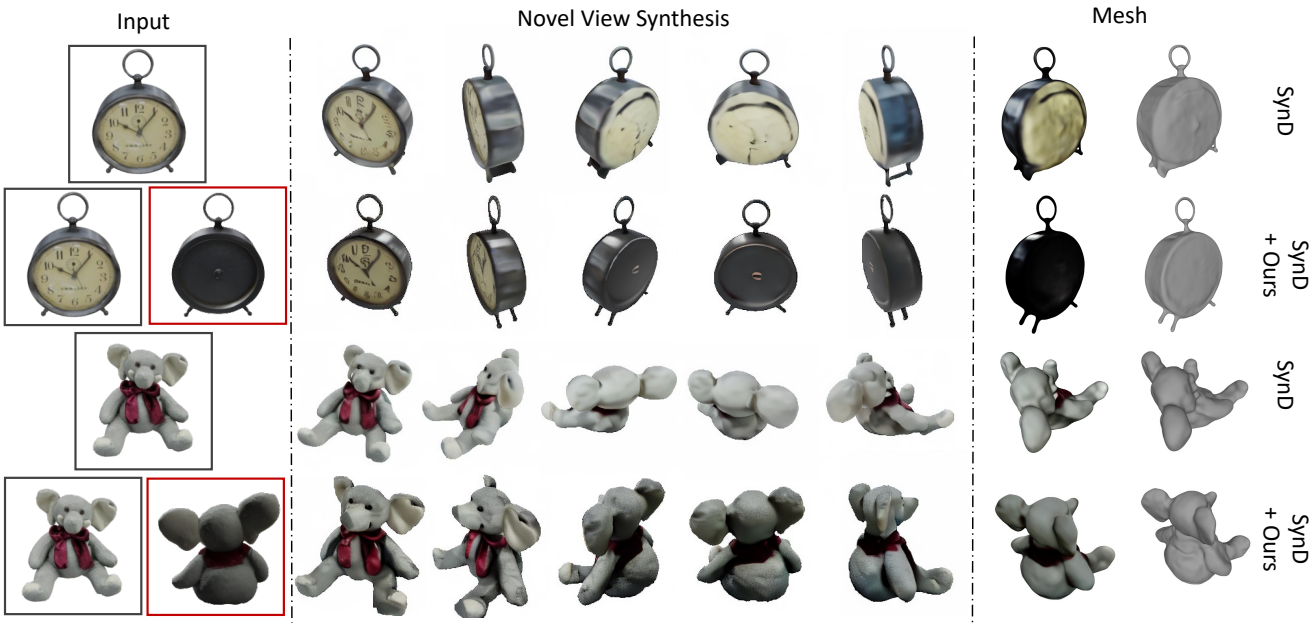

Figure 5. Qualitative comparison with SyncDreamer (SyncD) [18] in controllable novel view synthesis and 3D reconstruction. The image in □ is the main input, and the other image in □ is the conditional input generated from Zero-1-to-3 [17]. With more information in multi-view images, DC-SyncDreamer is able to generate more accurate back textures and more controllable 3D shapes.

Method	PSNR↑	SSIM↑	LPIPS↓
trainable UNet	15.96	0.762	0.209
w/o reconstruction loss	16.18	0.766	0.206
w/o view conditioning	19.04	0.805	0.166
full model	**20.52**	**0.828**	**0.141**

Table 3. Ablation study on GSO dataset. Eliminating the reconstruction loss and training the UNet are both factors that negatively impact the final outcome. With view conditioning in the 3D lifting module, our model not only ensures more stable training but also yields the most optimal results.

to accept any number of ground-truth images as input. To show the flexibility of our framework, we integrate DreamComposer into Zero-1-to-3 [17] with multi-view ground-truth inputs and SyncDreamer [18] with the single-view ground-truth input, as described in Section 4.3 and Section 4.4 respectively. We further demonstrate the applications of DreamComposer in Section 4.5, including controllable editing and 3D character modeling. We further conduct ablation study in Section 4.6.

4.1. Datasets and Evaluation Metrics

Training Dataset. We train DreamComposer (DC) on the large-scale Objaverse [6] dataset containing around 800k 3D objects. We randomly pick two elevation angles for every object and render N images with the azimuth evenly distributed in $[0°, 360°]$. We set N to 36 for DC-Zero-1-to-3 and 16 for DC-SyncDreamer. For training and inference, image sizes are 256×256 and background is set to white.

Evaluation Dataset. To evaluate the generalization of our model to out-of-distribution data, we extend our evaluation dataset from Objaverse to Google Scanned Objects (GSO) [7], which contains high-quality scans of everyday household items. This evaluation setting is consistent with that for SyncDreamer [18], comprising 30 objects that include both commonplace items and various animal species.
Evaluation Metrics. Following previous works [17, 18], We utilize Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [38], and Learned Perceptual Image Patch Similarity (LPIPS) [46] as metrics.

4.2. Implementation Details

During the entire training process, we randomly pick a target image as ground truth and utilize a set of three images as inputs: two images captured from opposing angles and one image from a random angle. Benefited from the image triplet training scheme, our model can adapt to two or more inputs. This data sampling strategy not only improves the efficiency of the model's optimization but also preserves its scalability and adaptability to various input configurations.

4.3. Multi-view Input

In this section, we evaluate the performance of DreamComposer plugged into the Zero-1-to-3 with multi-view ground-truth inputs, as depicted in Figure 3 (a).
Evaluation Protocols. When provided with an input image of an object, Zero-1-to-3 [17] has the ability to generate new perspectives of the same object. We take the four or-

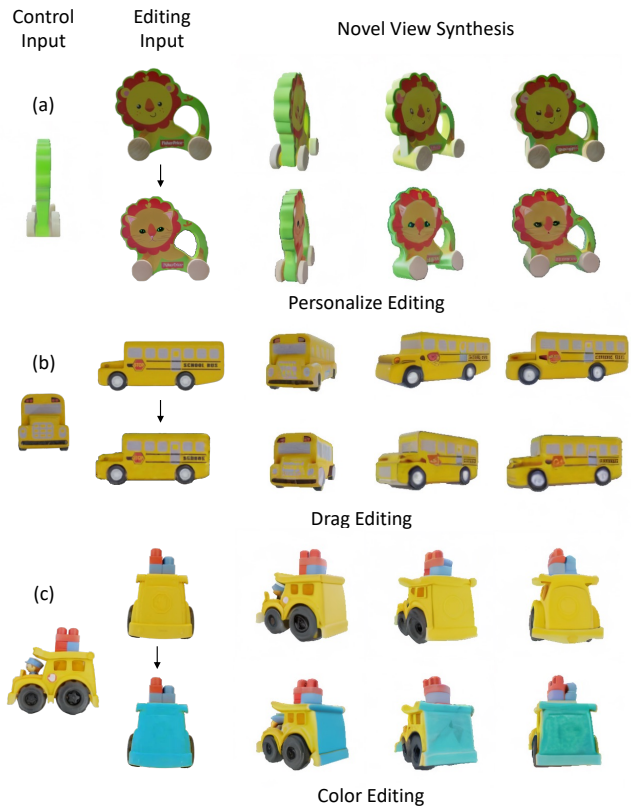

Figure 6. **Controllable Editing.** We present personalize editing with InstructPix2Pix [1] in (a), drag editing with DragGAN [25], DragDiffusion [32] in (b), and color editing in (c).

thogonal angles as input and set the image closest to the target perspective as the input for Zero-1-to-3 as well as the main view for DC-Zero-1-to-3. The remaining three images serve as the additional condition-views for DC-Zero-1-to-3. We calculate the results for input elevation angles of 0, 15, and 30 degrees respectively.

Evaluation on NVS. The comparison of quantitative results is shown in Table 1, and the comparison of qualitative results is shown in Figure 4. While Zero-1-to-3 possesses the ability to produce visually plausible images from novel views, the absence of multi-view inputs compromises the accuracy of these unseen viewpoints. Our DC-Zero-1-to-3, by conditioning on multi-view images, ensures the controlled generation of new viewpoints while maintaining the integrity of its diffusion model's generative capabilities. DC-Zero-1-to-3 significantly surpasses other methods in terms of the quality and consistency of generated images across various angles.

4.4. Single-view Input

In this section, we evaluate the performance of DreamComposer plugged into the SyncDreamer [18] with single-view ground-truth inputs, as depicted in Figure 3 (b).

Evaluation Protocols. We compare our method with SyncDreamer [18], Zero-1-to-3 [17], and RealFusion [20]. Given an input image of an object, Zero-1-to-3 can synthesize novel views of the object, and SyncDreamer is able to generate consistent novel views from 16 fixed views. RealFusion [20] is a single-view reconstruction method based on Stable Diffusion [29] and SDS [26]. The inverse perspective of the input, generated using Zero-1-to-3 [17], serves as an additional condition-view for DC-SyncDreamer. We adhere to the identical input configurations as established in SyncDreamer. The mesh is directly reconstructed from multi-view images by NeuS [36].

Evaluation on NVS and 3D Reconstruction. The comparison of quantitative results is shown in Table 2, and the comparison of qualitative results is shown in Figure 5. While SyncDreamer is able to generate consistent novel views, the shape of the object and the texture on the back may still appear unreasonable. DC-SyncDreamer not only maintains multi-view consistency in colors and geometry but also enhances the control over the shape and texture of the newly generated perspectives.

4.5. Applications

We explore the various applications of DreamComposer, including controllable 3D object editing with DC-Zero-1-to-3 and 3D character modeling with DC-SyncDreamer.

Controllable 3D object Editing. DreamComposer is able to perform controllable editing by modifying or designing images from certain perspectives, as shown in Figure 6. We designate an image from a specific viewpoint as the "control input", which remains unaltered. Concurrently, we manipulate an "editing input", which represents an image from an alternate viewpoint. We utilize InstructPix2Pix [1], DragGAN [25] and DragDiffusion [32] to manipulate the image, thereby achieving our desired style, corresponding to (a), (b) in Figure 6 respectively. And we modify the color of the editing input ourselves in (c). Subsequently, we employ the modified images in conjunction with the control input to synthesize novel views.

3D Character Modeling. With DC-SyncDreamer, 3D characters can be modeled from only a few 2D paintings, as shown in Figure 7. This can significantly improve the efficiency of existing 3D pipelines, and is expected to be connected with ControlNet for text-to-3D character creation.

4.6. Ablation Analysis

We conduct ablation studies on DC-SyncDreamer. For the ablation study, quantitative results are included in Table 3 and qualitative samples are included in Figure 8.

Necessity of reconstruction loss. First, we remove the reconstruction MSE loss in the second step of training as discussed in Section 3.4. As shown in Figure 8 and Table 3, without reconstruction MSE loss, the multi-view 3D lifting

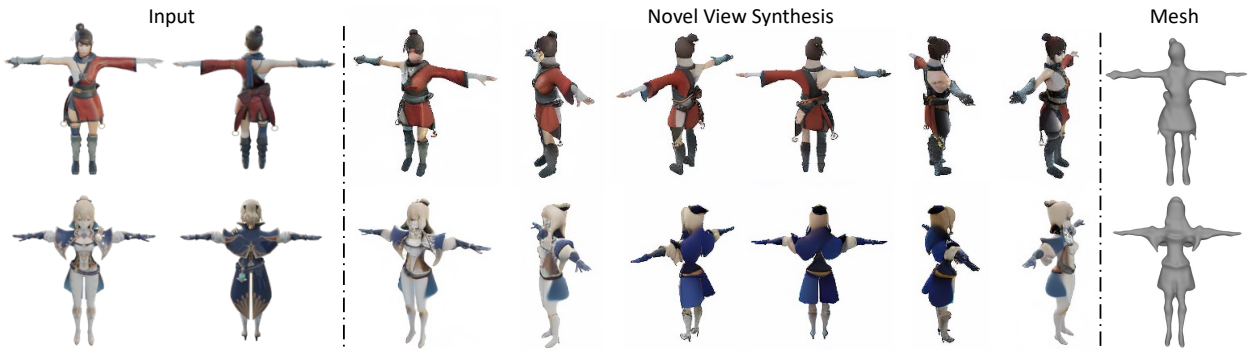

Figure 7. **3D Character Modeling.** DC-SyncDreamer is able to reconstruct arbitrary objects with rarely sparse inputs. We present the results of 3D character modeling from multi-view 2D paintings.

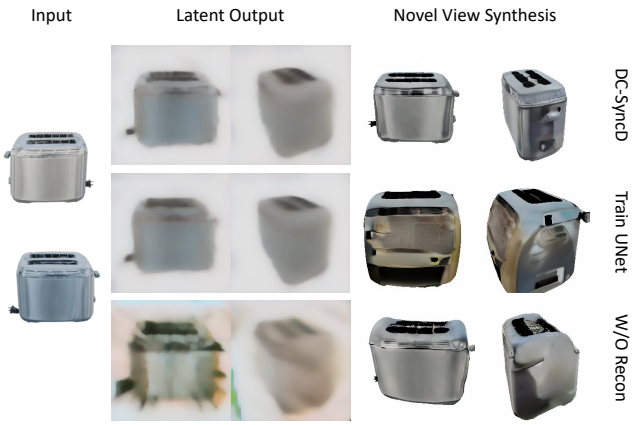

Figure 8. Ablation studies to verify the designs of our approach. "DC-SyncD" means our full model incorporating with SyncDreamer [18] pipeline. "Train UNet" indicates finetuning the UNet with our modules without freezing it. "W/O Recon" means removing the reconstruction MSE loss in the second step of training. The Latent Output is derived by rendering and pooling features in the tri-planes, as shown in Figure 2. Without reconstruction loss, DC-SyncDreamer fails to produce precise latent outputs.

module is unable to produce effective latent outputs, resulting in the inability to synthesize satisfactory novel views.

Finetuning v.s. freezing the diffusion U-Net. In our design, the pretrained diffusion U-Net is frozen during training DreamController. We attempt to finetune the diffusion U-Net with DreamComposer's modules in the second stage of training. As shown in Figure 8 and Table 3, the model performance decreases when we finetune the U-Net together with our modules.

Necessity of view-conditioning for 3D lifting. We remove the view conditioning cross-attention of the 2D-to-3D lifting module. As shown in Table 3, removing the view conditioning leads to worse performance. We also empirically observe that the training is unstable without view conditioning. More results are shown in supplementary material.

Scalability for arbitrary numbers of input views. We validate our model's flexibility and scalability in managing arbitrary numbers of inputs. As shown in Figure 9, our model can handle arbitrary numbers of input views, and its controllability is strengthened proportionally with the increasing number of input views.

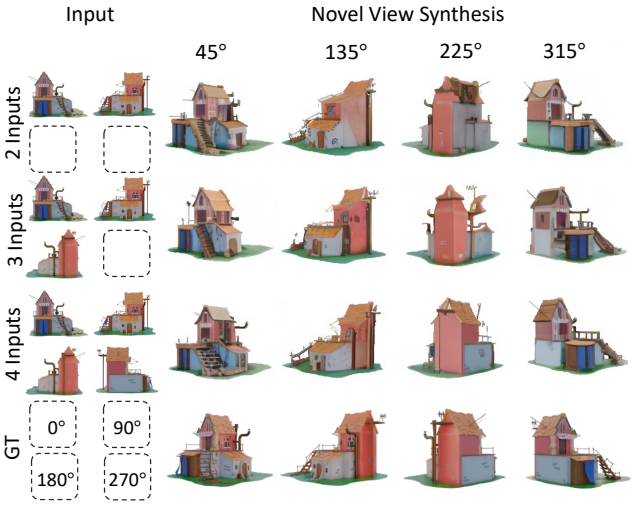

Figure 9. Ablation study to demonstrate the scalability of our model. Our model has the capacity to process arbitrary inputs, and its ability to control outcomes enhances correspondingly with the increasing information of input data.

5. Conclusion and Discussions

We propose DreamComposer, a flexible and scalable framework to empower existing diffusion models for zero-shot novel view synthesis with multi-view conditioning. DreamComposer is scalable to the number of input views. It can be flexibly plugged into a range of existing state-of-the-art models to empower them to generate high-fidelity novel view images with multi-view conditions, ready for controllable 3D object reconstruction. More discussions and limitations are presented in the supplementary materials.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 7

[2] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2, 3

[3] 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. *arXiv preprint arXiv:2304.02602*, 2023. 3

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[5] 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. 2, 3

[6] 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, 2022. 3, 6

[7] 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. 6

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

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

[10] Susung Hong, Donghoon Ahn, and Seungryong Kim. Debiasing scores and prompts of 2d diffusion for view-consistent text-to-3d generation, 2023. 3

[11] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3

[12] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 2, 3

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

[14] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision*, pages 198–216. Springer, 2022. 3

[15] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 2, 3

[16] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion, 2023. 3

[17] 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, 2023. 1, 2, 3, 5, 6, 7

[18] 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, 3, 5, 6, 7, 8

[19] 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. 2, 3

[20] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023. 3, 5, 7

[21] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures, 2022. 3

[22] 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. *Communications of the ACM*, 65(1):99–106, 2021. 3, 4

[23] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion, 2023. 3

[24] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts, 2022. 3

[25] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold, 2023. 7

[26] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022. 2, 3, 7

[27] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both

2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 3

[28] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 3

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 3, 7

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[31] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 2, 3

[32] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing, 2023. 7

[33] 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. *arXiv preprint arXiv:2303.14184*, 2023. 3

[34] 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, 2023. 3

[35] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2, 3

[36] 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, 2023. 7

[37] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion, 2022. 3

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 6

[39] 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. 2, 3

[40] 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. 3

[41] 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, 3

[42] 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 360deg views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4479–4489, 2023. 3

[43] 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, 2023. 3

[44] 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. 2, 3

[45] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 4

[46] 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

[47] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12588–12597, 2023. 3

[48] Junzhe Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance, 2023. 3

[49] Zi-Xin Zou, Weihao Cheng, Yan-Pei Cao, Shi-Sheng Huang, Ying Shan, and Song-Hai Zhang. Sparse3d: Distilling multiview-consistent diffusion for object reconstruction from sparse views. *arXiv preprint arXiv:2308.14078*, 2023. 3