MV-Adapter: Multi-View Consistent Image Generation Made Easy

Zehuan Huang[1] Yuan-Chen Guo[2†] Haoran Wang[3]
Ran Yi[3] Lizhuang Ma[3] Yan-Pei Cao[2✉] Lu Sheng[1✉]
[1]Beihang University [2]VAST [3]Shanghai Jiao Tong University

Project page: https://huanngzh.github.io/MV-Adapter-Page/

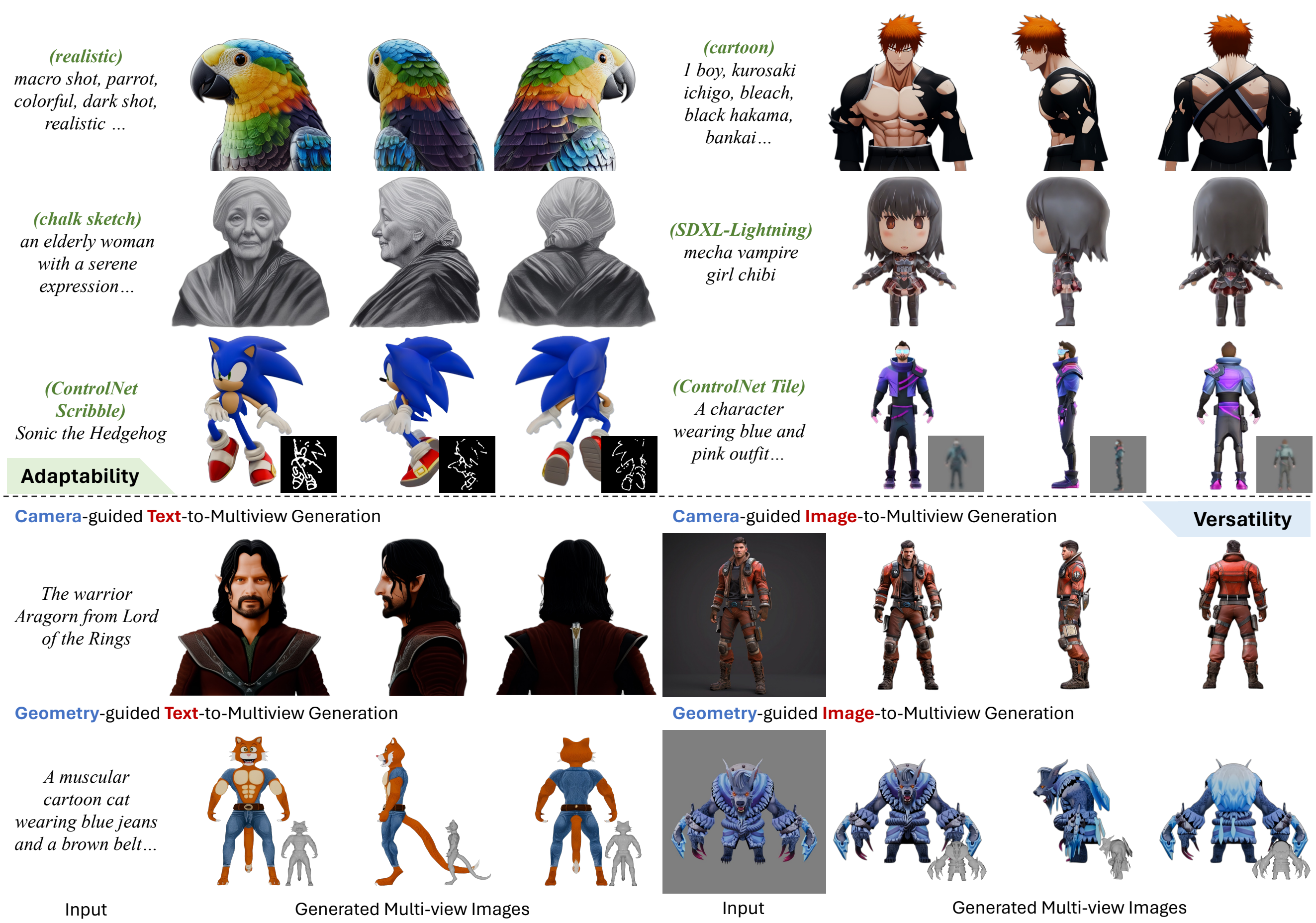

Figure 1. *MV-Adapter* is a versatile adapter that turns existing pre-trained text-to-image (T2I) diffusion models to multi-view image generators. ***Row 1,2,3***: results by integrating MV-Adapter with personalized T2I models, few-step T2I models, and ControlNets [68], demonstrating its **adaptability**. ***Row 4,5***: results under various control signals, including view-guided or geometry-guided generation with text or image inputs, showcasing its **versatility**.

Abstract

Existing multi-view image generation methods often make invasive modifications to pre-trained text-to-image (T2I) models and require full fine-tuning, leading to high computational costs and degradation in image quality due to scarce high-quality 3D data. This paper introduces MV-Adapter, an efficient and versatile adapter that enhances T2I models and their derivatives without altering the original network structure or feature space. To efficiently model

†Project lead; ✉ corresponding author

the 3D geometric knowledge within the adapter, we introduce innovative designs that include duplicated self-attention layers and parallel attention architecture, enabling the adapter to inherit the powerful priors of the pre-trained models to model the novel 3D knowledge. Moreover, we present a unified condition encoder that seamlessly integrates camera parameters and geometric information, facilitating applications such as text- and image-based 3D generation and texturing. MV-Adapter achieves multi-view generation at 768 resolution on Stable Diffusion XL (SDXL), and demonstrates adaptability and versatility. It can also be extended to arbitrary view generation, enabling broader applications. We demonstrate that MV-Adapter sets a new quality standard for multi-view image generation, and opens up new possibilities due to its efficiency, adaptability and versatility.

1. Introduction

Multi-view image generation is a fundamental task with significant applications in areas such as 2D/3D content creation, robotics perception, and simulation. With the advent of text-to-image (T2I) diffusion models [1, 33, 36, 37, 39, 40, 44], there has been considerable progress in generating high-quality single-view images. Extending these models to handle multi-view generation holds the promise of unifying text, image, and 3D data into a cohesive framework.

Recent attempts on multi-view image generation [10, 15, 17, 21, 24, 26, 48, 53, 54, 57, 69] involve fine-tuning T2I models on 3D datasets [5, 65] and propose modeling multi-view consistency by applying attention to relevant pixels in different views. However, it is computationally challenging when working with large T2I models and high-resolution images, as it requires at least n view images to be processed simultaneously during training. Advanced methods [19, 21] still struggle with 512 resolution, which is far from the 1024 or higher that modern T2I models can achieve. Moreover, the scarcity of high-quality 3D data exacerbates the optimization difficulty when performing full model fine-tuning, resulting in a degradation in the generation quality. These limitations primarily stem from the invasive changes to base models and full tuning.

A feasible way to address these challenges is to fine-tune a plug-and-play adapter, which helps preserve prior knowledge embedded in the pre-trained models. For example, NVS-Adapter [16] attaches an additional module to T2I models for novel view synthesis from a single image. For the adapter-based solution for multi-view image generation, a key issue is how to efficiently model multi-view consistency in the newly added networks while freezing the base model. It has been demonstrated that effectively achieving multi-view consistency demands the fundamental image prior [48], and this requirement applies to adapter-based approaches as well. NVS-Adapter [16] uses learnable tokens as a medium to transmit information among views in its view-consistency modules. Yet, because its adapter needs to be trained from scratch, it lacks fundamental image prior for learning multi-view consistency, leading to suboptimal performance (see Tab. 4 and Fig. 7). Moreover, it is restricted to the single image input and does not support native text input or geometry guidance, which significantly limits its applicability.

Therefore, we propose MV-Adapter, an efficient and versatile plug-and-play adapter that enhances T2I models and their derivatives for multi-view generation under various conditions. Unlike existing full-tuning methods [47, 48], which intrusively modify the base model's original self-attention layers to include multi-view or reference features, we duplicate the self-attention layers to create new multi-view attention and image cross-attention layers as an adapter. To efficiently model multi-view consistency in our adapter, we introduce a parallel structure for the newly added attention layers—those handle image-related features—ensuring they remain in the same domain as the base model's spatial self-attention. This design allows us to initialize the new attention weights with those already learned by the pre-trained self-attention, enabling the adapter to inherit the powerful image generation priors without having to relearn them from scratch (as is the case in NVS-Adapter [16]), and thus, the learning efficiency of multi-view consistency is greatly improved. Additionally, we introduce a unified condition embedding and encoder that seamlessly integrates camera parameters and geometric information into spatial map representations, enhancing both versatility and applicability of our approach.

By leveraging our adapter design, we successfully achieve the consistent multi-view generation at 768 resolution on SDXL [37]. As shown in Fig. 1, our MV-Adapter produces highly consistent images and demonstrates both adaptability and versatility. It seamlessly applies to derivatives of the base model [13, 43, 68] for customized or controllable generation, while simultaneously supporting camera and geometry guidance, which benefits applications in 3D generation and texture generation. Moreover, MV-Adapter can be extended to arbitrary view generation, enabling broader applications. In summary, our contributions are as follows:

- We design an innovative adapter, MV-Adapter, that inherits the pre-trained image prior and efficiently models multi-view consistency.
- MV-Adapter is a versatile plug-and-play adapter that enables T2I models and their derivatives to generate multi-view images under various conditions.
- Experiments demonstrate that MV-Adapter produces 768-resolution multi-view images from text, images and sketches, and supports generating arbitrary view images.

2. Related Work

Text-to-Image Diffusion Models. Text-to-image (T2I) generation [1, 14, 20, 28, 33, 36, 37, 39, 40, 44] has made remarkable progress, particularly with the advancement of diffusion models [7, 11, 12, 49]. Guided diffusion [7] and classifier-free guidance [11] improved text conditioning and generation fidelity. DALL-E2 [40] leverages CLIP [38] for better text-image alignment. The Latent Diffusion Model [42], also known as Stable Diffusion, enhances efficiency by performing diffusion in the latent space of an autoencoder. Stable Diffusion XL [37], a two-stage cascade diffusion model, has greatly improved the generation of high-frequency details.

Derivatives and Extensions of T2I Models. To facilitate creation with pre-trained T2Is, various derivative models and extensions have been developed, focusing on model distillation for efficiency [23, 27, 32, 51] and controllable generation [3, 30, 31, 66]. These derivatives encompass personalization [9, 13, 18, 29, 43, 46, 50, 56, 64], and spatial control [34, 68]. Typically, they employ adapters or fine-tuning methods to extend functionality while preserving the original feature space of the pre-trained models. Our work adhere to non-intrusive principle, ensuring compatibility with these derivatives or extensions for broader applications.

Multi-view Generation with T2I models. Multi-view generation methods [10, 15, 17, 21, 24, 26, 35, 48, 53, 54, 57, 61, 62, 69] extend T2I models by leveraging 3D datasets [5, 65]. For instance, MVDream [48] integrates camera embeddings and expands the self-attention mechanism from 2D to 3D for cross-view connections, while SPAD [17] enhances spatial relational modeling by applying epipolar constraints to cross-view attention. Era3D [21] introduces an efficient row-wise self-attention mechanism aligned with epipolar lines across views, facilitating high-resolution multi-view generation. However, these methods typically require extensive parameter updates, altering the feature space of pre-trained T2I models and limiting their compatibility with T2I derivatives. Our work addresses this by introducing a multi-view adapter that harmonizes with pre-trained T2Is, significantly expanding the potential for diverse applications.

3. Preliminary

We introduce the preliminary of multi-view diffusion models [17, 21, 48], which helps understand common strategies in modeling multi-view consistency in T2I models.

Multi-View Diffusion Models. Multi-view diffusion models enhance T2Is by introducing multi-view attention mechanism, enabling the generation of images that are consistent across different viewpoints. Several studies [48, 57] extend the self-attention of T2Is to include all pixels across multi-

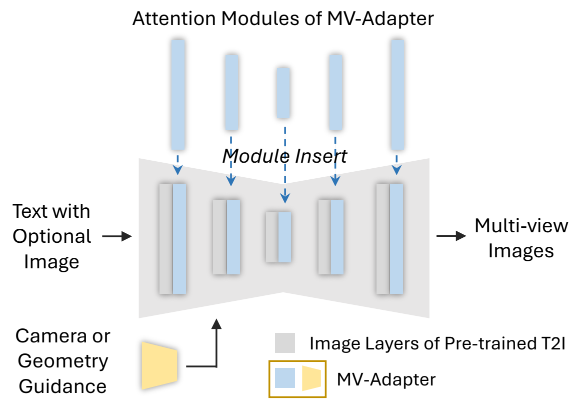

Figure 2. Inference pipeline.

view images. Let $\boldsymbol{f}^{in}$ denotes the input of the attention block, the dense multi-view self-attention extends $\boldsymbol{f}^{in}$ from the view itself to the concatenated feature sequence from n views. While this approach captures global dependencies, it is computationally intensive, as it processes all pixels of all views. To mitigate the computational cost, epipolar attention [15, 17] leverages geometric relationships between views. Specifically, methods like SPAD [17] extend the self-attention by restricting $\boldsymbol{f}^{in}$ to the view itself as well as patches along its epipolar lines.

Furthermore, when generating orthographic views at an elevation angle of 0°, the epipolar lines align with the image rows. Utilizing this property, row-wise self-attention [21] is introduced after the original self-attention layers in T2I models. The process is defined as:

$$\begin{aligned} \boldsymbol{f}^{self} &= \mathrm{SelfAttn}(\boldsymbol{f}^{in}) + \boldsymbol{f}^{in}; \\ \boldsymbol{f}^{mv} &= \mathrm{MultiViewAttn}(\boldsymbol{f}^{self}) + \boldsymbol{f}^{self} \end{aligned} \tag{1}$$

where MultiViewAttn performs attention across the same rows in different views, effectively enforcing multi-view consistency with reduced computational overhead.

4. Methodology

MV-Adapter is an efficient and versatile adapter that learns multi-view priors transferable to derivatives of T2Is without specific tuning, and enable them to generate multi-view consistent images under various conditions. As shown in Fig. 2, at inference, our MV-Adapter, which contains a condition guider and the decoupled attention layers, can be inserted into a personalized or distilled T2I to constitute the multi-view generator.

In detail, as shown in Fig. 3, the condition guider in Sec. 4.1 encodes the camera or geometry information, which supports both camera-guided and geometry-guided generation. Within the decoupled attention mechanism in Sec. 4.2, the additional multi-view attention layers learn multi-view consistency, while the optional image

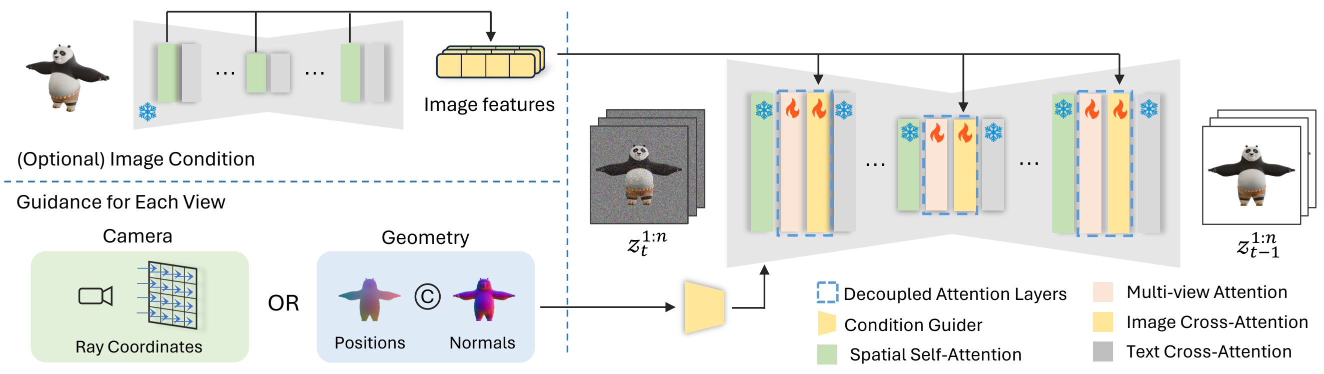

Figure 3. Overview of *MV-Adapter*. Our MV-Adapter consists of two components: 1) a condition guider that encodes camera or geometry condition; 2) decoupled attention layers that contain multi-view attention for learning multi-view consistency, and optional image cross-attention to support image-conditioned generation, where we use the pre-trained U-Net to encode the reference image.

cross-attention layers are for image-conditioned generation. These new layers are duplicated from pre-trained spatial self-attention and organized in a parallel architecture.

4.1. Condition Guider

We design a general condition guider that supports encoding both camera and geometric representations, enabling T2I models to perform multi-view generation under various guidance.

Camera Conditioning. To condition on the camera pose, we use a camera ray representation ("raymap") that shares the same height and width as the latent representations in the pre-trained T2I models and encodes the ray origin and direction at each spatial location [10, 45, 60].

Geometry Conditioning. Geometry-guided multi-view generation helps applications like texture generation. To condition on the geometry, we use a global, rather than view-dependent representation that contains position maps and normal maps [2, 22]. Each pixel in the position map represents the coordinates of the point on the shape, which provide point correspondences across different views. Normal maps provide orientation information and capture fine geometric details, helping produce detailed textures. We concatenate the position map and normal map along to form a composite geometric conditioning input for each view.

Encoder Design. To encode the camera or geometry representation, we design a simple and lightweight condition guider for the conditioning maps $\boldsymbol{c}_m$ ($\boldsymbol{c}_m \in \mathbb{R}^{n\times 6\times h\times w}$). The condition guider consists of a series of convolutional networks, which contain feature extraction blocks and downsampling layers to adapt the feature resolution to the features in the U-Net encoder. The extracted multi-scale features are then added to the corresponding scales in the U-Net, enabling the model to integrate the conditioning information seamlessly at multiple levels. In theory, the input to our encoder is not limited to specific types of conditions; it can also be extended to a wider variety of maps, such as depth maps and pose maps.

4.2. Decoupled Attention

We introduce a decoupled attention mechanism, where we retain the original spatial self-attention layers and duplicate them to create new multi-view attention layers as well as image cross-attention layers for image-conditioned generation. These three types of attention layers are organized in a parallel architecture, which ensures that the new attention layers can fully inherit the powerful priors of the pre-trained self-attention layers, thus enabling efficient learning of geometric knowledge.

Duplication of Spatial Self-Attention. Our design adheres to the principle of preserving the original network structure and feature space of the base T2I model. Existing methods like MVDream [48] and Zero123++ [47] modify the base model's self-attention layers to include multi-view or reference features, which disrupts the learned priors and requires full model fine-tuning. Here we duplicate the structure and weights of spatial self-attention layers to create new multi-view attention and image cross-attention layers, and initialize the output projections of these new attention layers to zero. This allows the new layers to learn geometric knowledge without interfering with the original model, ensuring excellent adaptability.

Parallel Attention Architecture. In the pre-trained T2I model, the spatial self-attention layer and text cross-attention layer are connected serially through residual connections. Suppose feature $\boldsymbol{f}^{in}$ is the input of the attention block, we can express the process as

$$\begin{aligned} \boldsymbol{f}^{self} &= \text{SelfAttn}(\boldsymbol{f}^{in}) + \boldsymbol{f}^{in}; \\ \boldsymbol{f}^{cross} &= \text{CrossAttn}(\boldsymbol{f}^{self}) + \boldsymbol{f}^{self} \end{aligned} \tag{2}$$

A straightforward method to incorporate new attention layers is to append them after the original layers, connecting

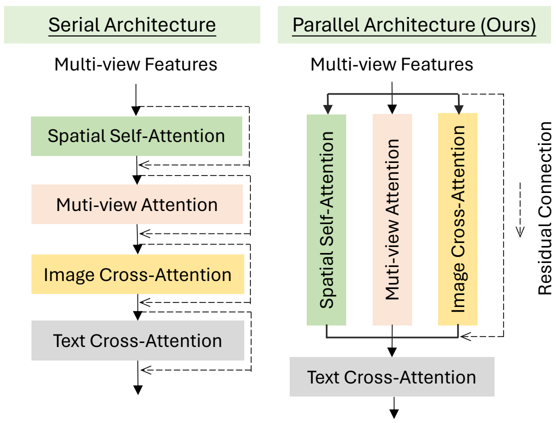

Figure 4. Serial vs parallel architecture.

them in a serial manner. However, the sequential arrangement may not effectively utilize the image priors modeled by the pre-trained self-attention layers, as it requires the new layers to learn from scratch. Even if we initialize the new layers with the pre-trained weights, the features input to these serially organized layers are in a different domain, causing the initialization to be ineffective. To fully exploit the effective priors of the spatial self-attention layers, we adopt a parallel architecture, as shown in Fig. 4. The process can be formulated as

$$\begin{aligned} \boldsymbol{f}^{self} = \text{SelfAttn}(\boldsymbol{f}^{in}) + \text{MultiViewAttn}(\boldsymbol{f}^{in}) \\ + \text{ImageCrossAttn}(\boldsymbol{f}^{in}, \boldsymbol{f}^{ref}) + \boldsymbol{f}^{in} \end{aligned} \tag{3}$$

where $\boldsymbol{f}^{ref}$ refers to features of the reference image. Since the features $\boldsymbol{f}^{in}$ fed into the new layers are the same as those to the self-attention layer, we can effectively initialize them with the pre-trained layers to transfer the image priors. We zero-initialize the output projection layer of the new layers to ensure that the initial output does not disrupt the original feature space. This architectural choice ***allows the model to build upon the established priors, facilitating efficient learning of multi-view consistency and image-conditioned generation***, while preserving the original space of the base text-to-image diffusion models.

Details of Multi-View Attention. We design different strategies for multi-view attention to meet the specific needs of different applications. For 3D object generation, we enable the model to generate multi-view images at an elevation of $0°$ and employ row-wise self-attention [21]. For 3D texture generation, considering the view coverage requirements, in addition to the four views evenly at elevation $0°$, we add two views from top and bottom. We then perform both row-wise and column-wise self-attention, enabling efficient information exchange among all views. For arbitrary view generation, we employ full self-attention [48] in our multi-view attention layers.

Details of Image Cross-Attention. To condition on reference images $\boldsymbol{c}_i$ and achieve, we propose a novel method for incorporating detailed information from the image without altering the original feature space of the T2I model. We employ the pre-trained and frozen T2I U-Net as our image encoder. We pass the clear reference image into this frozen U-Net, setting the timestep $t = 0$, and then extract multi-scale features from the spatial self-attention layers. These fine-grained features contain detailed information about the subject and are injected into the denoising U-Net through the decoupled image cross-attention layers. In this way, we leverage the rich representations learned by the pre-trained model, enabling precise control over the generated content.

5. Experiments

We implemented MV-Adapter on Stable Diffusion V2.1 (SD2.1) [42] and SDXL [37], training a 512×512 adapter for SD2.1 and a 768×768 adapter for SDXL using a subset of the Objaverse dataset [5]. Detailed configurations are provided in the supplementary materials.

5.1. Camera-Guided Multi-view Generation

Evaluation on Community Models and Extensions. We evaluated MV-Adapter using representative T2Is and extensions, including personalized models [13, 43], efficient distilled models [23, 27], and plugins such as ControlNet [68]. We present six qualitative results in Fig. 5. More results can be found in the supplementary materials.

Comparison with Baselines. For text-to-multiview generation, we compared our MV-Adapter with MVDream [48] and SPAD [17] on 1,000 prompts from the Objaverse dataset. The results are presented in Fig. 6 and Tab. 1. For image-to-multiview generation, we conduct comparison with full-tuning methods [21, 47, 55, 57, 59], and the adapter-based method NVS-Adapter [16] on the Google Scanned Objects (GSO) dataset [8], as results shown in Fig. 7 and Tab. 4. Compared to those full-tuning methods, it indicates that, by preserving the original feature space of T2I models, our MV-Adapter achieves higher visual fidelity and consistency with conditions. Compared to NVS-Adapter that needs to train new modules from scratch, our adapter inherits pre-trained prior and produces consistent multi-view images.

Quantitative comparisons on training efficiency with the baseline method Era3D [21] in Tab. 3 demonstrates that our MV-Adapter significantly reduces computational costs, facilitating high-resolution multi-view generation based on larger backbones.

5.2. Geometry-Guided Multi-view Generation

Evaluation on Community Models. We evaluated our geometry-guided model with T2I derivative models. The re-

Figure 5. Results of text-to-multi-view generation with community models and extensions.

Figure 6. Qualitative comparison on camera-guided text-to-multiview generation.

Table 1. Quantitative results on text-to-multiview generation.

Method	FID↓	IS↑	CLIP Score↑
MVDream [48]	32.15	14.38	31.76
SPAD [17]	48.79	12.04	30.87
Ours (SD2.1)	31.24	15.01	32.04
Ours (SDXL)	**29.71**	**16.38**	**33.17**

sults in Fig. 9 demonstrate the adaptability of MV-Adapter in seamlessly integrating with different base models.

Comparison with Baselines. We compare our text- and image-conditioned multi-view-based texture generation method (see details in Sec. 5.4) with four state-of-the-art methods, including TEXTure [41], Text2Tex [4], Paint3D [67], SyncMVD [25], and FlashTex [6]. For our image-to-texture model, we used ControlNet [68] to generate reference images conditioned on text and depth maps. As shown in Fig. 8 and Tab. 2, compared to these project-and-inpaint or synchronized multi-view texturing methods, our approach fine-tunes additional modules to model geometric associations and preserves the generative capabilities of the base T2I model, thereby producing multi-view consistent and high-quality textures. Additionally, testing on a single RTX 4090 GPU revealed that our method achieves faster generation speeds than the others.

5.3. Ablation Study

Parallel Attention Architecture. To assess the effectiveness of our proposed parallel attention architecture, we conducted ablation studies on image-to-multi-view generation setting. We report the quantitative and qualitative results of

Figure 7. Qualitative comparison on camera-guided image-to-multiview generation.

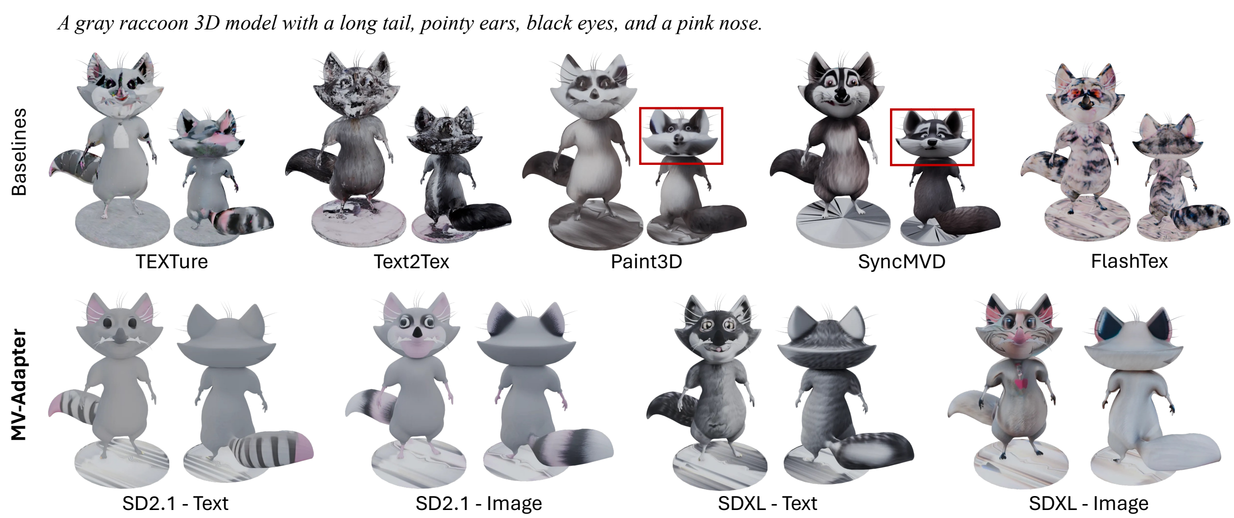

Figure 8. Qualitative comparison on texture generation. We compare our text- and image-conditioned models with baseline methods.

using serial or parallel architecture in Tab. 4 and Fig. 10. The results show that, the serial setting, which cannot leverage the pre-trained image prior, tends to produce artifacts and inconsistent details with the image input. In contrast, our parallel setting produces high-quality and highly consistent results with the reference image.

5.4. Applications

3D Generation. Following Era3D [21], we use StableNormal [63] to generate normal maps of multi-view images, which are then fed into NeuS [58] to reconstruct 3D meshes. We conducted comparison with Era3D [21]. Results in Tab. 3 show that our SD2.1-based MV-Adapter is comparable to Era3D, but our SDXL-based model shows significantly higher performance. These findings underline the scalability of MV-Adapter and its ability to leverage the strengths of state-of-the-art T2I models, providing additional benefits to 3D generation. Visualization results can be found in supplementary materials.

Table 2. Quantitative comparison on 3D texture generation. FID and KID ($\times 10^{-4}$) are evaluated on multi-view renderings. Our models achieves best texture quality with faster inference.

Method	FID↓	KID↓	Time↓
TEXTure [41]	56.44	61.16	90s
Text2Tex [4]	58.43	60.81	421s
Paint3D [67]	44.38	47.06	60s
SyncMVD [25]	36.13	42.28	50s
FlashTex [6]	50.48	56.36	186s
Ours (SD2.1 - Text)	38.19	42.83	**18s**
Ours (SD2.1 - Image)	33.93	38.73	19s
Ours (SDXL - Text)	32.75	35.18	32s
Ours (SDXL - Image)	**27.28**	**29.47**	33s

Texture Generation. We leverage back-projection and incidence-based weighted blending techniques [2] to map the generated multi-view images onto the UV texture map. We then perform view coverage analysis to identify uncov-

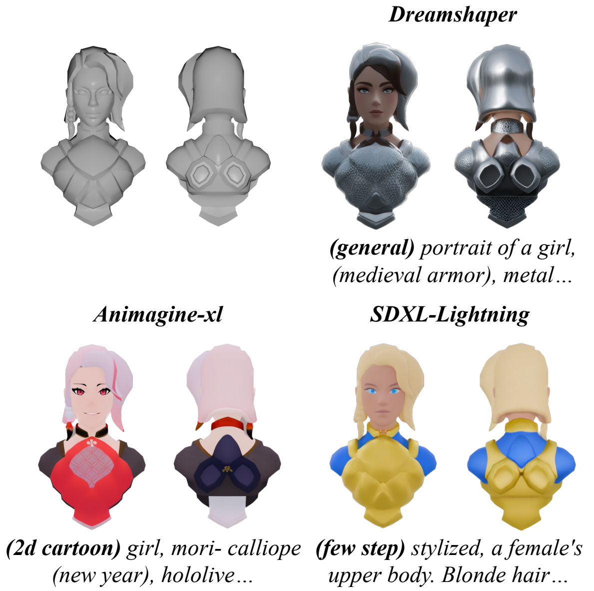

Figure 9. Results of geometry-guided text-to-multiview generation with community models.

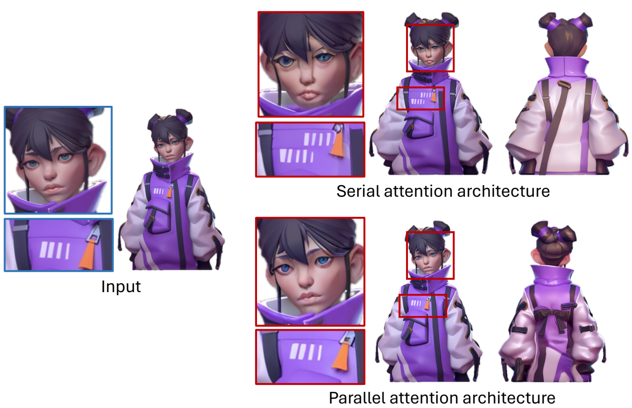

Figure 10. Qualitative ablation study on the attention architecture.

ered regions, render images from the current 3D texture for those views, and refine them using an efficient inpainting model [52]. More visualization results can be found in the supplementary materials.

Arbitrary View Generation. Following CAT3D [10], we perform multiple rounds of multi-view generation, with the number of views generated each time set to $n = 8$. Starting from text or an initial single image as input, we first generate eight anchor views that broadly cover the object. In practice, these anchor views are positioned at elevations of $0°$ and $30°$, with azimuth angles evenly distributed around the circle (*e.g.* every $45°$). For generating new target views, we cluster the viewpoints based on their spatial orientations, grouping them into clusters of 8. We then select the 4 nearest known views from the already generated anchor views to serve as conditions guiding the generation of each target view. When using four input views, we concatenate them into a long image and input this into the pre-trained T2I U-Net to extract features. Implementation details and visual results are provided in supplementary materials.

Table 3. Quantitative comparison on training efficiency (batch size set to 1) and 3D reconstruction. We compare with baseline method Era3D [21] on Training Params (TP), Memory Usage (MU), Training Speed (TS), as well as Chamfer Distance (CD) and Volume IoU (IoU) of reconstruction results.

Method	TP↓	MU↓	TS↑	CD↓	IoU↑
Era3D (SD2.1)	993M	36G	2.2iter/s	0.0329	0.5118
Ours (SD2.1)	**127M**	**17G**	**3.1iter/s**	0.0317	0.5173
Era3D (SDXL)	3.1B	>80G	-	-	-
Ours (SDXL)	**490M**	**60G**	**1.05iter/s**	**0.0206**	**0.5682**

Table 4. Quantitative results on image-to-multiview generation.

Method	PSNR↑	SSIM↑	LPIPS↓
ImageDream [57]	19.280	0.8472	0.1218
Zero123++ [47]	20.312	0.8417	0.1205
CRM [59]	20.185	0.8325	0.1247
SV3D [55]	20.042	0.8267	0.1396
Ouroboros3D [61]	20.810	0.8535	0.1193
Era3D [21]	20.890	0.8601	0.1199
NVS-Adapter [16]	17.236	0.8069	0.1476
Ours (SD2.1, Parallel Attention)	20.867	0.8695	0.1147
Ours (SDXL, Parallel Attention)	**22.131**	**0.8816**	**0.1002**
Ours (SDXL, Serial Attention)	20.687	0.8681	0.1149

6. Conclusion

In this paper, we present MV-Adapter, an efficient and versatile adapter that enhances text-to-image diffusion models and their derivatives without compromising quality or altering the original feature space. We introduce innovative adapter framework that includes duplicated self-attention layers and a parallel attention architecture, allowing the adapter to efficiently model 3D geometric knowledge. Additionally, we introduced a unified condition encoder that integrates camera parameters and geometric information into spatial map representations, enhancing the model's versatility and applicability in 3D generation and texture generation. Extensive evaluations highlight the efficiency, adaptability, and versatility of MV-Adapter across different models and conditions. MV-Adapter offers an efficient and flexible solution for multi-view image generation, presenting exciting possibilities for a wide range of applications.

Acknowledgment

This work was supported by National Natural Science Foundation of China (62132001), Capital's Funds for Health Improvement and Research (CFH 2024-2-40611), and the Fundamental Research Funds for the Central Universities.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2022. 2, 3

[2] Raphael Bensadoun, Yanir Kleiman, Idan Azuri, Omri Harosh, Andrea Vedaldi, Natalia Neverova, and Oran Gafni. Meta 3d texturegen: Fast and consistent texture generation for 3d objects, 2024. 4, 7

[3] Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey, 2024. 3

[4] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models, 2023. 6, 7

[5] 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, 2023. 2, 3, 5

[6] Kangle Deng, Timothy Omernick, Alexander Weiss, Deva Ramanan, Jun-Yan Zhu, Tinghui Zhou, and Maneesh Agrawala. Flashtex: Fast relightable mesh texturing with lightcontrolnet, 2024. 6, 7

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis, 2021. 3

[8] 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, 2022. 5

[9] 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, 2022. 3

[10] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models, 2024. 2, 3, 4, 8

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 3

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 3

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 2, 3, 5

[14] Zehuan Huang, Hongxing Fan, Lipeng Wang, and Lu Sheng. From parts to whole: A unified reference framework for controllable human image generation, 2024. 3

[15] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion, 2024. 2, 3

[16] Yoonwoo Jeong, Jinwoo Lee, Chiheon Kim, Minsu Cho, and Doyup Lee. Nvs-adapter: Plug-and-play novel view synthesis from a single image, 2025. 2, 5, 8

[17] Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers, 2024. 2, 3, 5, 6

[18] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023. 3

[19] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model, 2023. 2

[20] Jiaao Li, Yixiang Huang, Ming Wu, Bin Zhang, Xu Ji, and Chuang Zhang. Clip-sp: Vision-language model with adaptive prompting for scene parsing. *Computational Visual Media*, 10(4):741–752, 2024. 3

[21] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention, 2024. 2, 3, 5, 7, 8

[22] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d, 2023. 4

[23] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation, 2024. 3, 5

[24] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image, 2023. 2, 3

[25] Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion, 2023. 6, 7

[26] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion, 2024. 2, 3

[27] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023. 3, 5

[28] Hao Ma, Ming Li, Jingyuan Yang, Or Patashnik, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Clip-flow: Decoding images encoded in clip space. *Computational Visual Media*, 10(6):1157–1168, 2024. 3

[29] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning, 2024. 3

[30] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable

and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[31] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Leqi Shen, Chenyang Qi, Jixuan Ying, Chengfei Cai, Zhifeng Li, Heung-Yeung Shum, et al. Follow-your-click: Open-domain regional image animation via motion prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6018–6026, 2025. 3

[32] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models, 2023. 3

[33] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2023. 2, 3

[34] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2024. 3

[35] Tai-Jiang Mu, Hao-Xiang Chen, Jun-Xiong Cai, and Ning Guo. Neural 3d reconstruction from sparse views using geometric priors. *Computational Visual Media*, 9(4):687–697, 2023. 3

[36] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models, 2022. 2, 3

[37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis, 2024. 2, 3, 5

[38] 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, 2021. 3

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 2, 3

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2, 3

[41] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes, 2023. 6, 7

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 3, 5

[43] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. 2, 3, 5

[44] 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, 2022. 2, 3

[45] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations, 2022. 4

[46] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning, 2024. 3

[47] 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, 4, 5, 8

[48] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation, 2023. 2, 3, 4, 5, 6

[49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2020. 3

[50] Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. Moma: Multimodal llm adapter for fast personalized image generation, 2024. 3

[51] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023. 3

[52] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions, 2022. 8

[53] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion, 2023. 2, 3

[54] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction, 2024. 2, 3

[55] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion, 2024. 5, 8

[56] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation, 2024. 3

[57] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation, 2023. 2, 3, 5, 8

[58] 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, 2021. 7

[59] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model, 2024. 5, 8

[60] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models, 2022. 4

[61] Hao Wen, Zehuan Huang, Yaohui Wang, Xinyuan Chen, Yu Qiao, and Lu Sheng. Ouroboros3d: Image-to-3d generation via 3d-aware recursive diffusion, 2024. 3, 8

[62] Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan, and Lin Gao. Recent advances in 3d gaussian splatting. *Computational Visual Media*, 10(4):613–642, 2024. 3

[63] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal, 2024. 7

[64] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 3

[65] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images, 2023. 2, 3

[66] Bohan Zeng, Shanglin Li, Yutang Feng, Ling Yang, Hong Li, Sicheng Gao, Jiaming Liu, Conghui He, Wentao Zhang, Jianzhuang Liu, et al. Ipdreamer: Appearance-controllable 3d object generation with complex image prompts. *arXiv preprint arXiv:2310.05375*, 2023. 3

[67] Xianfang Zeng, Xin Chen, Zhongqi Qi, Wen Liu, Zibo Zhao, Zhibin Wang, Bin Fu, Yong Liu, and Gang Yu. Paint3d: Paint anything 3d with lighting-less texture diffusion models, 2024. 6, 7

[68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 2, 3, 5, 6

[69] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation, 2024. 2, 3