FlexGen: Flexible Multi-View Generation from Text and Image Inputs

Xinli Xu[1*] Wenhang Ge[1*] Jiantao Lin[1*] JiaweiFeng[1] Lie Xu[3]
Hanfeng Zhao[3] Shunsi Zhang[3] Ying-Cong Chen[1,2†]

HKUST(GZ)[1] HKUST[2] Quwan[3]

Abstract

In this work, we introduce FlexGen, a flexible framework designed to generate controllable and consistent multi-view images, conditioned on a single-view image, or a text prompt, or both. FlexGen tackles the challenges of controllable multi-view synthesis through additional conditioning on 3D-aware text annotations. We utilize the strong reasoning capabilities of GPT-4V to generate 3D-aware text annotations. By analyzing four orthogonal views of an object arranged as tiled multi-view images, GPT-4V can produce text annotations that include 3D-aware information with spatial relationship. By integrating the control signal with proposed adaptive dual-control module, our model can generate multi-view images that correspond to the specified text. FlexGen supports multiple controllable capabilities, allowing users to modify text prompts to generate reasonable and corresponding unseen parts. Additionally, users can influence attributes such as appearance and material properties, including metallic and roughness. Extensive experiments demonstrate that our approach offers enhanced multiple controllability, marking a significant advancement over existing multi-view diffusion models. This work has substantial implications for fields requiring rapid and flexible 3D content creation, including game development, animation, and virtual reality. Project page: https://xxu068.github.io/flexgen.github.io/.

Figure 1. FlexGen is a flexible framework designed to generate high-quality, consistent multi-view images conditioned on a single-view image, or a text prompt, or both. Our method allows editing of unseen regions and modification of material properties through user-defined text.

1. Introduction

Recent progress in generative models [15, 41] has significantly advanced 2D content creation, thanks to the rapid increase in 2D data volumes. However, 3D content creation remains challenging due to the limited accessibility of 3D assets, which are essential for diverse downstream applications, including game modeling [12, 18], computer animation [17, 30], and virtual reality [35]. Previous 3D generation methods primarily concentrate on optimization-based techniques using multi-view posed images [11, 43, 45, 50], or employ SDS-based distillation approaches derived from 2D generative models [23, 24, 32]. Although effective, these methods often demand significant optimization time, limiting their practicality in real-world applications.

Multi-view diffusion models [25, 26, 37, 39, 44] have demonstrated the potential of pre-trained 2D generative models for 3D content creation through the synthesis of consistent multi-view images. Despite their promising performance, the controllable generation of multi-view images

*Equal contribution.
†Corresponding author.

remains under-explored. Most existing multi-view diffusion models typically rely on a single-view image, which lacks 3D-aware controllable guidance and proves insufficient for robust multi-view image generation. For example, the generation of unseen regions continues to pose a significant challenge, often simply replicating information from the input view to unseen areas.

Several studies focus on conditional 3D generation. For instance, Coin3D [7] utilizes basic shapes as 3D-aware guidance, whereas Clay [54] leverages sparse point clouds and 3D bounding boxes. However, these guidance methods are not user-friendly. Other research [47, 48] employs sparse multi-view images as 3D-aware guidance for 3D generation. These methods effectively supplement most unseen regions due to the incorporation of additional views, thereby achieving promising results. Nevertheless, acquiring sparse multi-view images is not always feasible in practice. These approaches typically depend on a pre-trained multi-view diffusion model for this purpose, yet integrating 3D-aware controllable guidance into multi-view diffusion models remains a significant challenge. Inspired by controllable content creation in 2D domain with text prompts, similar design can be incorporated as extra condition for providing 3D-aware guidance. Text encompasses adequate relationship information, which has been proven in the 2D generative model [36]. Previous SDS-based distillation methods [23, 24, 32, 46] have successfully employed text prompts to guide 3D asset generation, yielding significant results and validating this approach.

However, designing 3D-aware text annotations for 3D assets is not trivial. Most prior works [1, 19] focus on 2D image captioning, which lacks sufficient 3D-aware information. To overcome this limitation, Cap3D [27] leverages the reasoning capabilities of BLIP-2 [19] to generate descriptions for each rendered view of a 3D asset. These descriptions are then aggregated by GPT-4 [1] to form a comprehensive caption. However, this approach often results in a high-level summary that misses detailed local captions. This limitation arises from two main factors: firstly, BLIP-2 primarily produces global descriptions, and secondly, individual view provides limited information about the object, often resulting in redundant or incomplete single-view annotations. Instant3D [20] is the most similar work to ours, utilizing text prompts from Cap3D to generate multi-view images. However, as discussed before, text annotations from Cap3D lack 3D-aware information and Instant3D only supports text-to-3D task, which is not flexible enough.

Alternatively, we propose leveraging the powerful recognition capabilities of GPT-4V(ision) to perform 3D-aware global-to-local captioning. By analyzing four orthogonal rendered views from a object, GPT-4V is capable of generating detailed descriptions that capture both global context and local features with 3D-aware information. This approach ensures the resulting text annotations are enriched with 3D-aware details with spatial relationship, providing a more comprehensive and perceptually accurate understanding. We show a comparison with Cap3D in Figure 2.

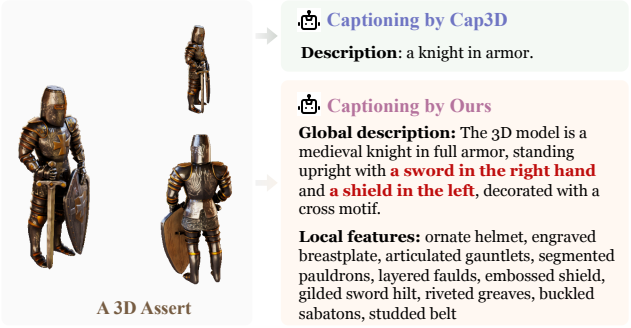

Figure 2. Comparison of the caption between Cap3D and Ours.

With 3D-aware text annotations, we can enhance the generative capabilities of existing multi-view diffusion models, enabling controllable generation guided by text prompts. By modifying the text prompt, our model is capable of generating diverse and consistent multi-view images that accurately correspond to the given textual descriptions. To this end, we propose an adaptive dual-control module that adaptively integrates image and text modalities through reference attention, enabling precise joint control over multi-view image generation by leveraging both visual input and detailed text prompts. In this work, we demonstrate three kinds of controllability of our model: firstly, our model can supplement unseen parts, achieving controllable unseen part generation; secondly, we introduce material controllability by adjusting the text prompt to modify the metallic and roughness properties when rendering multi-view images. For instance, appending the prompt with "high metallic" and "low roughness" can reflect these material characteristics accurately; finally, our model enables part-level control over the texture of the generated unseen parts, as we incorporate detailed texture information during the text annotation phase.

To summarize, our contributions are listed as follows.

- We propose FlexGen, a method for flexible generation of multi-view images, guided by both image and text inputs. This approach offers robust controllability and ensures that the generated images accurately correspond to the given textual descriptions.
- We annotate 3D-aware text guidance using GPT-4V and generate material prompts consistent with the materials used during rendering, achieving controllable generation of textures, materials, and unseen parts.
- Extensive experiments validate the effectiveness of our proposed methods, demonstrating the ability to controllable generation of multi-view images.

2. Related work

2.1. Diffusion Models for Multi-view Synthesis

Recent research has extensively explored the generation of multi-view images using diffusion models to achieve efficient and 3D-consistent results. These efforts include both text-based methods, such as MVDiffusion [6], MVDream [39] and ImageDream [44], and image-based methods like SyncDreamer [25], Wonder3D [26] and Zero123++ [37]. MVDiffusion, for instance, leverages text conditioning to simultaneously generate all images with a global transformer, facilitating cross-view interactions. Similarly, MVDream incorporates a self-attention layer to capture cross-view dependencies, ensuring consistency across different views. For image-based approaches, SyncDreamer [25] constructs a volume feature from the multi-view latent representation to produce consistent multi-view color images. Wonder3D enhances the quality of 3D results by explicitly encoding geometric information and employing cross-domain diffusion. Several methods [42, 49, 51] adopt these approaches to first generate multi-view images of an object and then reconstruct the 3D shape from these views using sparse reconstruction techniques. More recently, another line of work has emerged that treats multi-view synthesis as a video generation task, leveraging video diffusion models to produce view-consistent sequences, as demonstrated by CAT3D [10] and IM-3D [28]. Despite achieving reasonable results, these methods still face limitations in controllable generation.

2.2. Controllable Generative models

In recent years, adding conditional control to generative models has garnered increasing attention for enabling controllable generation. These efforts span both 2D and 3D domains. For instance, several 2D methods [3, 9, 13, 16, 31] focus on text-guided control by adjusting prompts or manipulating CLIP features. Additionally, ControlNet [53] allows for a series of 2D image hints for control tasks through a parallel model architecture. However, similar controllable capabilities in 3D generation [2, 4, 29] remain underexplored. Coin3D [7] introduces a framework for generating 3D assets guided by basic shapes. Basic 3D shapes are not user-friendly for general users, whereas text serves as a more intuitive and accessible conditional input. Along this line, TOSS [38] leverages text as a high-level semantic constraint for the task of novel view synthesis (NVS) from a single image, improving the plausibility of the generated views. While TOSS demonstrates the value of text guidance, our work, FlexGen, proposes a more flexible framework. We not only integrate text prompts for controllable generation from a single image, but also uniquely support generation from text or image inputs alone, offering greater versatility.

3. Method

FlexGen is a flexible multi-view generation framework that supports conditioning based on text, single-view images, or a combination of both. By incorporating 3D-aware text annotations derived from GPT-4V, our method effectively achieve controllable multi-view images generation, including reasonable unseen part generation, texture controllable generation and materials editing. We begin with a succinct problem formulation in Section 3.1. Then, we introduce how to annotate 3D-aware caption in Section 3.2. After that, adaptive dual-control module is introduced to add textual condition into the framework in Section 3.3. Finally, we introduce training and inference in Sectin 3.4. An overview framework of FlexGen is shown in Figure 3.

3.1. Problem Formulation

Given a single-view image I or a user-defined prompt T that describes an object or both, our goal is to develop a generative model G that produces a tiled image I_{out}. This image consists of a 2×2 layout, representing 4 views of an object - the front, left, back, and right - with each view at a resolution of 512×512. The multi-view images are aligned with both the single-view image and the prompt, ensuring consistency among them, as illustrated in Figure 1. Inspired by [21], our model is based on a large pre-trained text-to-image diffusion model. The design of the 2×2 image grid aligns better with the original data format used by the 2D diffusion model, facilitating the utilization of prior knowledge. Regardless of the focal length and pose of the input image I, we consistently generate orthographic images with a fixed focal length and an elevation angle of 5°. For example, when processing an input image captured at an elevation angle α and azimuth angle β, FlexGen generates multi-view images at azimuth angles $\{\beta, \beta + 90^\circ, \beta - 90^\circ, \beta + 180^\circ\}$, all with a fixed elevation of 5°.

3.2. 3D-Aware Caption Annotation

Cap3D [27] utilized BLIP [19] to annotate each single rendered view and then leverage GPT4 for holistic description. However, such method leads to text annotation lacking 3D-aware information. Alternatively, we construct a dataset consists of paired multi-view images and 3D-aware global-local text annotation, as shown in the Figure 4. Our dataset is based on Objaverse [5], which provides basic shape and textual for each object, lacking high-quality textual description for each 3D asset. Therefore, we utilized an advanced large-scale multimodal model, GPT-4V, to generate high-quality textual descriptions for each 3D asset. Given these orthogonal views, GPT-4V not only summarizes a comprehensive global description of the object but also captures the intricate 3D relationships between its components. Specifically, the dataset construction process consists of three steps: rendering, captioning, and merging. (1) In the ren-

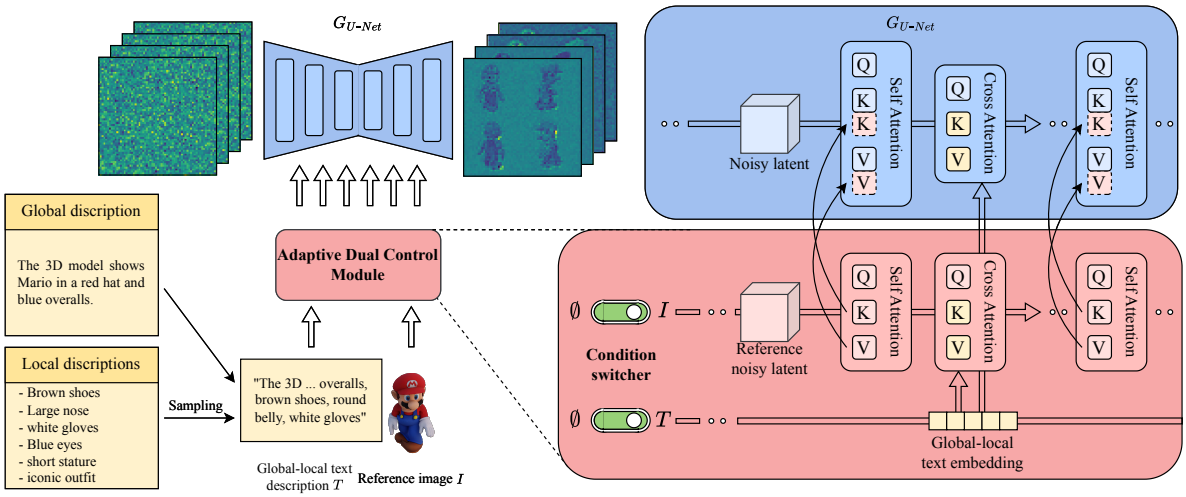

Figure 3. **Overview of the framework.** FlexGen is a flexible framework to generate controllable and consistent multi-view images, conditioned on a single-view image, or a text prompt, or both. The system incorporates a 3D-aware annotation method using GPT-4V and an adaptive dual-control module that integrates both a reference input image and text prompts for precise control. The condition switcher enhances flexibility, enabling the model to generate multi-view images based on image input, text input, or a combination of both.

dering stage, each 3D asset is rendered into four orthogonal multi-view images with a resolution of 512×512, which are tiled into a single image I_{out} in a 2×2 layout. (2) The captioning step is performed using GPT-4V with tiled image, which generates 3D-aware global and local captions. (3) In the final step, the global and local descriptions are merged to form the "global-local text description" of the 3D asset. During training, we randomly select a portion of the local descriptions to simulate user behavior.

Moreover, we incorporate material descriptions, such as metallic and roughness attributes, into the text annotations to enable material-controllable generation. We propose adding material descriptions that correspond to the materials used during rendering by blender [13]. For example, if multi-view images are rendered with high metallic and low roughness, we enhance the prompt with "high metallic" and "low roughness" to ensure the descriptions match the visual data. More details on the material rendering can be found in the Appendix.

3.3. Adaptive Dual-Control Module

Previous approaches, such as instant3d Li et al. [20] and Shi et al. [37], typically focus on single-modality inputs, conditioning on either a text prompt or a single-view image, without enabling joint control over both for generating multi-view images. To overcome this limitation, we propose an adaptive dual-control module that allows for simultaneous conditioning on both image and text inputs, enabling more precise and flexible multi-view image generation.

Our method builds upon the reference attention mechanism [52], which we extend to integrate both the reference image and text prompt effectively. This enhanced integration facilitates robust interaction between the two modalities, enabling our model to generate multi-view images that maintain (1) high fidelity to the input image and (2) coherence with both the global and fine-grained descriptions specified in the text.

Reference attention involves running the denoising UNet model on an additional reference image, appending the self-attention key and value matrices from the reference image to the corresponding attention layers during model denoising. We introduce a slight modification to this approach. In addition to the reference image, we inject the prompt information. Specifically, the user-defined prompt is processed through the CLIP encoder to obtain per-token CLIP text embeddings E with a shape of $L \times D$, where L represents the token length and D the embedding dimension. This prompt includes both global descriptions and local features. Using the cross-attention mechanism, we facilitate sufficient information interaction between the image and prompt, enabling more precise joint control. Once this interaction is complete, we append the self-attention key and value matrices from our adaptive dual-control module to the corresponding attention layers during model denoising.

3.4. Training and inference

Building on the adaptive dual-control module, our framework accommodates both prompt and image conditions to guide the generation of multi-view images. To increase flexibility, we introduced a condition switcher during training that supports both single-mode and dual-mode conditions, allowing for seamless transitions between different input scenarios. With a configurable probability, inputs can be left empty: when the text prompt is absent, it defaults to

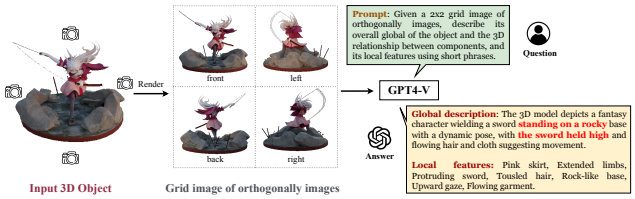

Figure 4. **3D-aware caption generation pipeline.** A 3D object is rendered into four orthogonal multi-view images (front, left, back, right) in a 2×2 grid layout. Using GPT-4V, the agent generates both global and local descriptions. The global description captures the overall attributes of the object and the 3D spatial relationships between its components, while local features detail specific aspects such as color, posture, and texture, thereby enriching the dataset with rich semantic annotations.

an empty string to ensure uninterrupted processing by the model. Similarly, in the absence of an image input, we substitute it with a black image.

During inference, this design facilitates flexible and controllable multi-view generation. When both modalities are available, the image and prompt collaborate to provide complementary information, enriching the generated output. If only the image is supplied, the model functions in an image to multi-view mode, generating multiple views based solely on the visual input. Conversely, when only the text is available, the model operates as a text to multi-view generator, producing views that align with the textual description.

4. Experiments

4.1. Evaluation Settings

Training Datasets. Given the inconsistent quality of the original Objaverse dataset [5], we initially excluded objects lacking texture maps and those with low polygon counts. We subsequently curated a collection of 147k high-quality objects to form the final training set. For rendering these objects, we employed Blender, setting the camera distance at 4.5 units and the field of view (FOV) to 30 degrees. We generated 24 ground-truth images for the target view set, maintaining a fixed elevation of 5 degrees while uniformly distributing the azimuth angles across the range [0, 360]. The input views were randomly sampled with elevation angles between -30 and 30 degrees and azimuth angles evenly distributed across [0, 360]. All images were rendered at a resolution of 512×512. For the purpose of 3D-aware caption annotation, we utilized four orthogonal images from the ground-truth set as inputs for GPT-4V.

Training Details. We utilized Stable Diffusion 2.1 as our base model, training it on eight NVIDIA A800 80GB GPUs over a period of 10 days, completing 180,000 iterations with a batch size of 32. The Adam optimizer was employed with a learning rate of 1e-5. During training, we configured the probabilities of using both the prompt and image, only the image, and only the text at 0.3 each, while the probability for both modalities being absent was set at 0.1. Sampling involved 75 steps using the DDIM methodology [40].

Evaluation Metrics. We evaluated our method across three distinct tasks: text to multi-view image generation, novel view synthesis (NVS), and sparse-view 3D reconstruction using the generated images as input to a reconstruction module. For text to multi-view generation, we employed the Frechet Inception Distance (FID) [14] and Inception Score (IS) [34] to assess image quality, and CLIP score [33] to evaluate alignment between the generated images and the textual descriptions. For NVS, we utilized LPIPS [55] to compare perceptual similarity between the generated novel-view images and the ground truth. The quality of 3D reconstruction was assessed using Chamfer Distance (CD) and volumetric Intersection over Union (IoU) between the reconstructed meshes and the ground truth ones.

Evaluation Datasets We employed the Google Scanned Object (GSO) dataset [8] for evaluation purposes. For the text to multi-view generation task, we randomly selected 300 samples from the GSO dataset. For each sample, we rendered a set of four orthogonal views and employed our captioning method to generate 300 text prompts that accurately describe the objects. These same 300 samples were also used for evaluation in the novel view synthesis (NVS) and sparse-view 3D reconstruction tasks.

4.2. Comparison with state-of-the-art methods

Novel view synthesis and sparse-view 3D reconstruction. First, we quantitatively compare our method with other single-image to multi-view approaches, including Zero123++ [37], Era3D [22], and SyncDreamer [25], as shown in Table 1. Our method outperforms the others across several key metrics, such as PSNR and LPIPS, demonstrating that the joint control of text prompts and images allows for more consistent and realistic multi-view generation, particularly in unseen areas. Qualitative results are presented in Figure 5. To further verify the consistency of multi-view generation in three dimensions, we reconstructed the multi-view images generated by all methods using the open-source reconstruction method InstantMesh [48] for a fair comparison. We report Chamfer Distance (CD) and FS metrics in Table 1, which show that our multi-view images lead to more accurate geometry reconstructions. Qualitative comparison of 3D reconstruction results are shown in Figure 7.

Text to multi-view. For text to multi-view evaluation, we conducted both qualitative and quantitative comparisons with the only existing open-source model, MVDream [39]. Table 2 presents quantitative results, highlighting differences in generation quality and text-image consistency. Our model achieved Inception Score (IS) and CLIP score metrics that were comparable to those of the validation set,

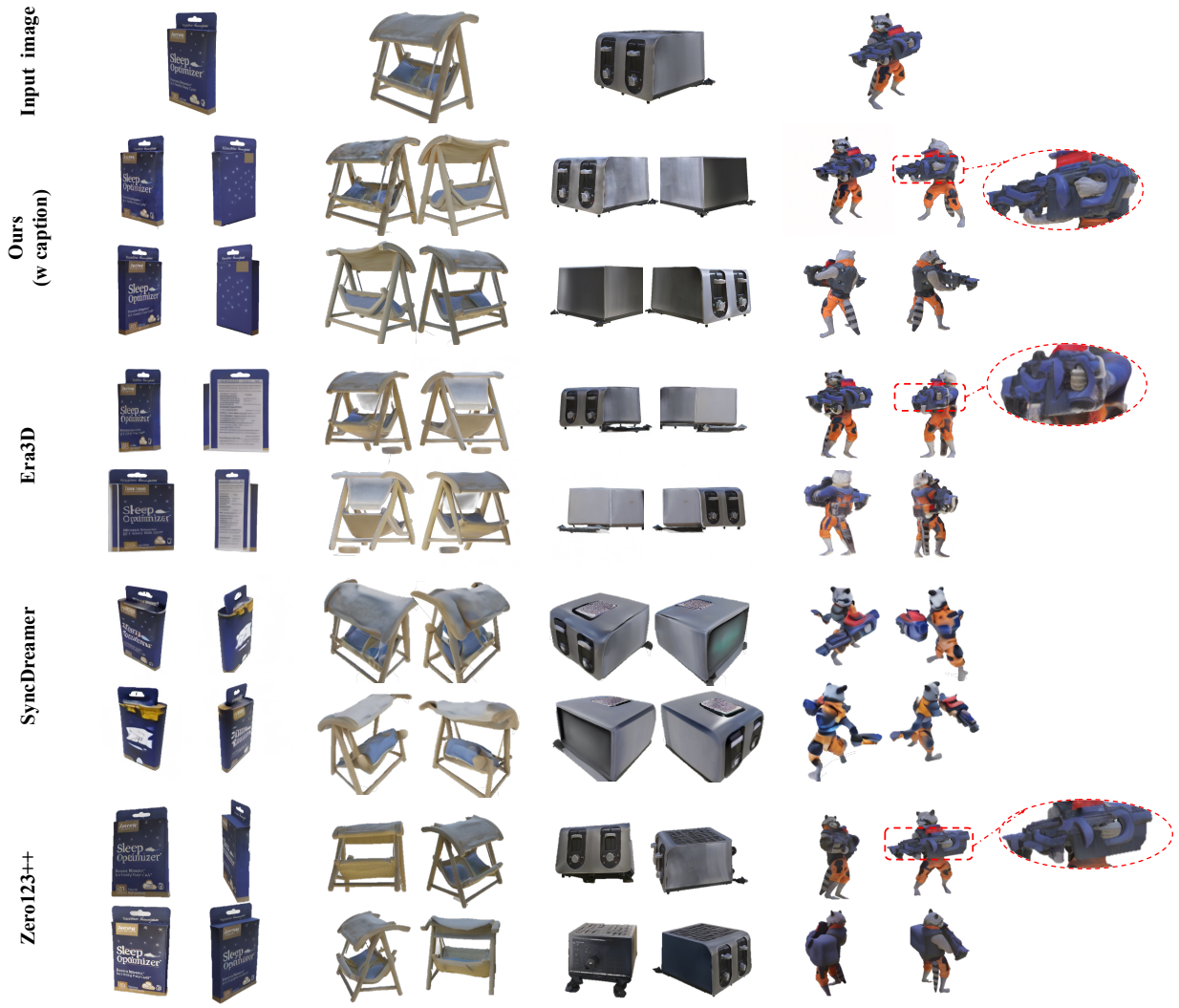

Figure 5. **Qualitative comparison of novel view synthesis.** Our method achieves superior generation quality. The text prompts used in our approach are generated by GPT-4V based on the input view. Some details are enlarged in the red circle

demonstrating strong image quality and text-image alignment. Furthermore, our model consistently outperformed MVDream across all evaluated metrics. Figure 6 showcases qualitative examples from the validation set, our model, and MVDream, illustrating that our model produces higher-quality, multi-view consistent images that more accurately match the text prompts. The images from our model exhibit superior fidelity and adherence to the input descriptions compared to those generated by MVDream, underscoring the effectiveness of our approach.

Table 1. The quantitative comparison in novel view synthesis and sparse-view reconstruction. We report PSNR, LPIPS, CD and FS on the GSO dataset.

Method	PSNR↑	LPIPS↓	CD↓	FS@0.1↑
Ours	22.31	0.12	0.076	0.928
Ours(w/o caption)	21.12	0.14	0.078	0.921
Zero123++	18.83	0.16	0.087	0.910
Era3D	18.52	0.19	0.245	0.713
SyncDreamer	17.66	0.21	0.126	0.833

Table 2. The quantitative comparison with MVDream in text to multi-view synthesis. "Ground truth" refers to the multiple views used to generate text by GPT-4V. Our method outperform MVDream by a large margin.

Method	FID ↓	IS↑	CLIP↑
Ours	35.56	13.41±0.87	0.83
Ground truth	N/A	13.81±1.40	0.89
MVDream	44.42	12.98±1.22	0.79

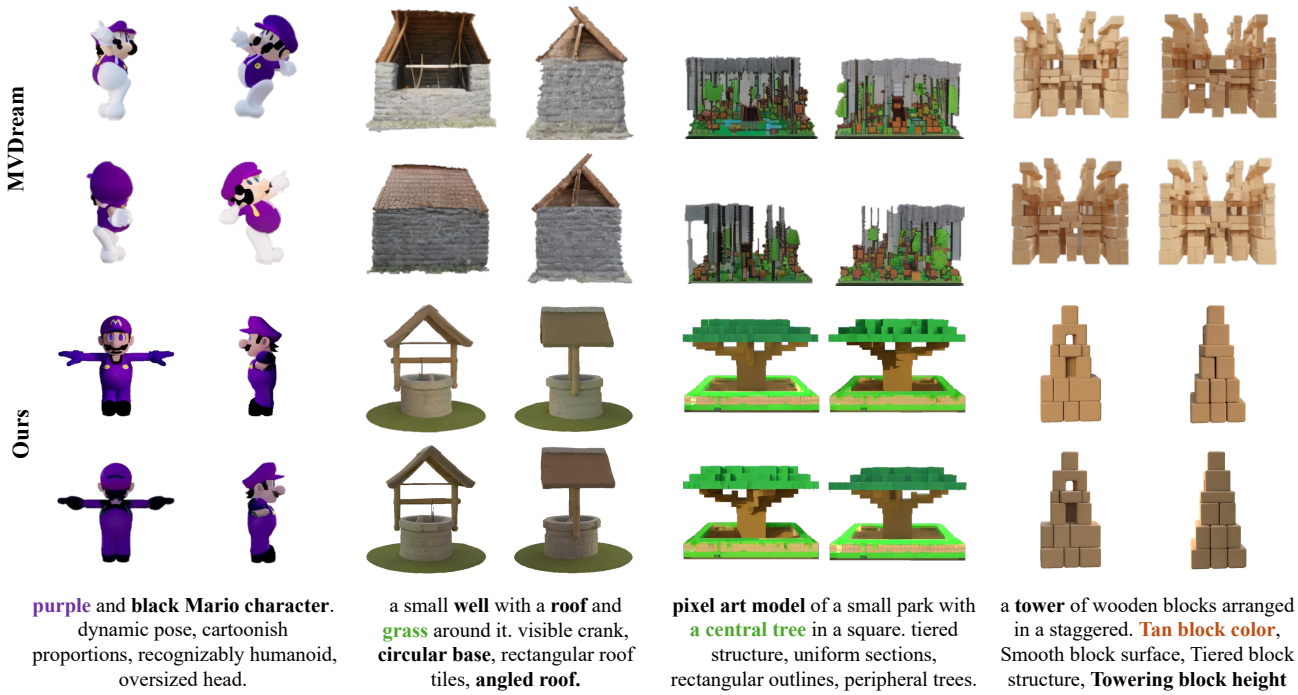

Figure 6. Qualitative comparison of text to multi-view. Our method significantly outperforms MVDream, achieving better generation quality that is consistent with the text.

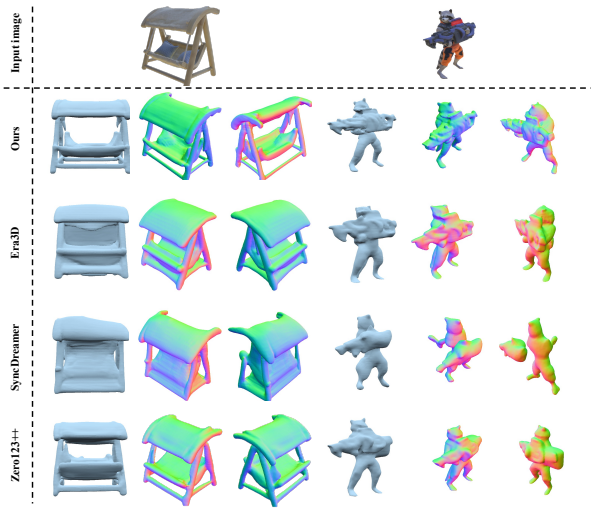

Figure 7. Qualitative comparison with Era3D, SyncDreamer and Zero123++ on sparse-view 3D reconstruction. Our method generates consistent multi-view images, achieving better results.

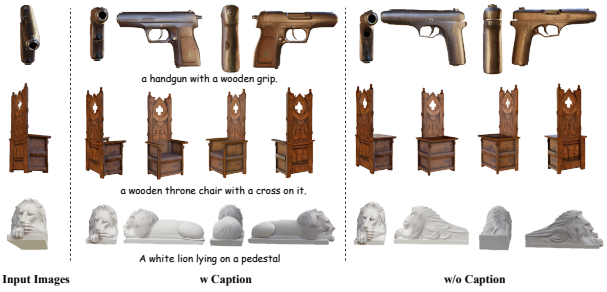

Figure 8. Ablation study to demonstrate that caption is capable of supplement unseen part, achieving significant better and reasonable results.

4.3. Abaltion Staudy

Adaptive Dual-Control Module. The adaptive dual-control module represents a key innovation within our framework, enabling simultaneous control over both image and text inputs. We extend reference attention to integrate CLIP-encoded caption features via cross-attention, enabling spatially aligned fusion of text and image cues. While this modification may appear simple, it is crucial for grounding fine-grained semantics in the generated views. As shown in Table 3 ("No Cross-Attention") and Fig. 10, removing this module significantly reduces visual fidelity and 3D consistency, validating its importance. This dual-input capability empowers users to guide the generation process using both image and textual prompts, significantly enhancing the model's ability to produce coherent and contextually appropriate multi-view images, as demonstrated in Figure 8. Additionally, we illustrate the module's capability to edit mate-

Figure 9. Material editing in generated multi-view images is facilitated by providing prompts such as "high-metallic" and "low-roughness".

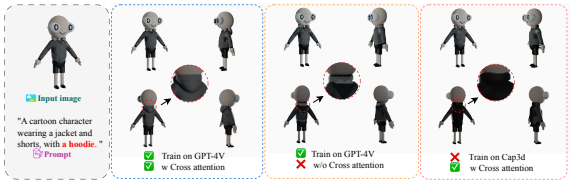

Figure 10. *Controllability analysis under varying model variants.*

rial properties in the generated multi-view images, as shown in Figure 9.

Table 3. **Ablation study on caption quality, architectural design, and inference inputs.** Replacing GPT-4V captions with Cap3D weakens supervision quality. Removing cross-attention affects multi-modal grounding. Omitting captions at inference reduces 3D consistency, confirming each component's necessity.

Variant	PSNR↑	LPIPS↓	CD↓	FS@0.1↑
Cap3D Captions	21.68	0.13	0.078	0.923
No Cross-Attention	21.89	0.13	0.077	0.926
Image-only Inference	21.12	0.14	0.078	0.921
Ours (Full)	**22.31**	**0.12**	**0.076**	**0.928**

The impact of 3D-Aware Captioning. To assess the effectiveness of 3D-aware captioning, we trained our model using annotations from the Cap3D dataset and our own dataset separately, comparing the control capabilities afforded by image and text prompts. As illustrated in Table 3 and Figure 11 , training with Cap3D data limits the ability for fine control, whereas our method, through its integration of 3D-aware information, enables more effective editing of text prompts.

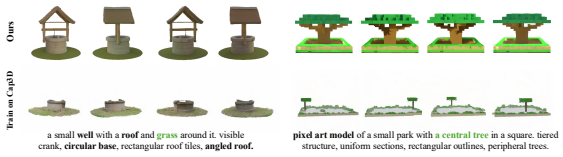

Figure 11. *Text to Multi-view using GPT-4V vs. Cap3D.*

5. Conclusion

In this work, we present FlexGen, a multi-view diffusion model designed to generate consistent and controllable multi-view images guided by a single image, text, or a combination of both. To achieve this, we harness the powerful recognition capabilities of GPT-4V to perform 3D-aware text annotations by reasoning over orthogonal views of an object, arranged as tiled multi-view images. Additionally, we introduce an adaptive dual-control module that enables text-based conditioning to be incorporated directly into the generation phase. By embedding spatial relationships, texture, and material descriptions into the text annotations, we achieve enhanced control over the output. Extensive experiments demonstrate the effectiveness of our proposed approach.

Limitations and future works. Although FlexGen introduces the innovative capability to jointly control both image and text inputs, our method occasionally encounters difficulties with complex user-defined instructions. This limitation likely stems from the constraints imposed by the availability of high-quality datasets. In the future, we plan to expand the dataset size and enhance the control capabilities to more effectively accommodate intricate instructions.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Shariq Farooq Bhat, Niloy Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 3

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[4] Dana Cohen-Bar, Elad Richardson, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Set-the-scene: Global-local training for generating controllable nerf scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 3, 5

[6] Zijun Deng, Xiangteng He, Yuxin Peng, Xiongwei Zhu, and Lele Cheng. Mv-diffusion: Motion-aware video diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 3

[7] Wenqi Dong, Bangbang Yang, Lin Ma, Xiao Liu, Liyuan Cui, Hujun Bao, Yuewen Ma, and Zhaopeng Cui. Coin3d: Controllable and interactive 3d assets generation with proxy-guided conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 2, 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. In *2022 International Conference on Robotics and Automation (ICRA)*, 2022. 5

[9] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision (ECCV)*, 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. *arXiv preprint arXiv:2405.10314*, 2024. 3

[11] Wenhang Ge, Tao Hu, Haoyu Zhao, Shu Liu, and Ying-Cong Chen. Ref-neus: Ambiguity-reduced neural implicit surface learning for multi-view reconstruction with reflection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[12] Jason Gregory. *Game engine architecture*. AK Peters/CRC Press, 2018. 1

[13] Roland Hess. *Blender foundations: The essential guide to learning blender 2.5*. Routledge, 2013. 3, 4

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeuIPS)*, 2017. 5

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeuIPS)*, 2020. 1

[16] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022. 3

[17] John Lasseter. Principles of traditional animation applied to 3d computer animation. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, 1987. 1

[18] Michael Lewis and Jeffrey Jacobson. Game engines. *Communications of the ACM*, 2002. 1

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 2023. 2, 3

[20] 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. *https://arxiv.org/abs/2311.06214*, 2023. 2, 4

[21] 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. *arXiv preprint arXiv:2311.06214*, 2023. 3

[22] 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. *arXiv preprint arXiv:2405.11616*, 2024. 5

[23] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023. 1, 2

[24] 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 (CVPR)*, 2023. 1, 2

[25] 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. *arXiv preprint arXiv:2309.03453*, 2023. 1, 3, 5

[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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 3

[27] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *arXiv preprint arXiv:2306.07279*, 2023. 2, 3

[28] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Natalia Neverova, Andrea Vedaldi, Oran Gafni, and Filippos Kokkinos. Im-3d: Iterative multiview diffusion and reconstruction for high-quality 3d generation. *arXiv preprint arXiv:2402.08682*, 2024. 3

[29] Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J Mitra. Diffusion handles enabling 3d edits for diffusion models by lifting activations to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[30] Rick Parent. *Computer animation: algorithms and techniques*. Newnes, 2012. 1

[31] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 3

[32] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 2021. 5

[34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems (NeuIPS)*, 2016. 5

[35] Martijn J Schuemie, Peter Van Der Straaten, Merel Krijn, and Charles APG Van Der Mast. Research on presence in virtual reality: A survey. *Cyberpsychology & behavior*, 2001. 1

[36] Guibao Shen, Luozhou Wang, Jiantao Lin, Wenhang Ge, Chaozhe Zhang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, Guangyong Chen, et al. Sg-adapter: Enhancing text-to-image generation with scene graph guidance. *arXiv preprint arXiv:2405.15321*, 2024. 2

[37] 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. *arXiv preprint arXiv:2310.15110*, 2023. 1, 3, 4, 5

[38] Yukai Shi, Jianan Wang, He Cao, Boshi Tang, Xianbiao Qi, Tianyu Yang, Yukun Huang, Shilong Liu, Lei Zhang, and Heung-Yeung Shum. Toss: High-quality text-guided novel view synthesis from a single image. *arXiv preprint arXiv:2310.10644*, 2023. 3

[39] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1, 3, 5

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[42] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 3

[43] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[44] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 1, 3

[45] 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. *arXiv preprint arXiv:2106.10689*, 2021. 1

[46] 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. *Advances in Neural Information Processing Systems (NeuIPS)*, 2024. 2

[47] Chao Xu, Ang Li, Linghao Chen, Yulin Liu, Ruoxi Shi, Hao Su, and Minghua Liu. Sparp: Fast 3d object reconstruction and pose estimation from sparse views. *arXiv preprint arXiv:2408.10195*, 2024. 2

[48] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2, 5

[49] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pages 1–20. Springer, 2024. 3

[50] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[51] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 3

[52] Lyumin Zhang. Reference-only control. In *Reference-only control*, pages https://github.com/ Mikubill/sd–webui–controlnet/discussions/1236. github, 2023. 4

[53] 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 (ICCV)*, 2023. 3

[54] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creat-

ing high-quality 3d assets. *arXiv preprint arXiv:2406.13897*, 2024. 2

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 5