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

Supplementary Material

1. Background

Stable Diffusion (SD) and Stable Diffusion XL (SDXL). We adopt Stable Diffusion [17] and Stable Diffusion XL [15] as our base T2I models, since they have a well-developed community with many powerful derivatives for evaluation. SD and SDXL perform the diffusion process within the latent space of a pre-trained autoencoder $\mathcal{E}(\cdot)$ and $\mathcal{D}(\cdot)$. In training, an encoded image $z_0 = \mathcal{E}(x_0)$ is perturbed to z_t at step t by the forward diffusion. The denoising network ϵ_θ learns to reverse this process by predicting the added noise, encouraged by an MSE loss:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(\boldsymbol{x}_0),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}),\boldsymbol{c},t}[\|\boldsymbol{\epsilon} - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{c}, t)\|_2^2] \tag{1}$$

where $\boldsymbol{c}$ denotes the conditioning texts. In SD, ϵ_θ is implemented as a UNet [18] consisting of pairs of down/up sample blocks and a middle block. Each block contains pairs of spatial self-attention layers and cross-attention layers, which are serially connected using the residual structure. SDXL leverages a three times larger UNet backbone than SD for high-resolution image synthesis and introduces a refinement denoiser to improve the visual fidelity.

2. Implementation Details

2.1. Dataset

We trained MV-Adapter on a filtered high-quality subset of the Objaverse dataset [3], comprising approximately 70,000 samples, with captions from Cap3D [13]. To accommodate the efficient multi-view self-attention mechanism, we rendered orthographic views to train the the model to generate $n = 6$ views per sample. For the camera-guided generation, we rendered views of 3D models with the elevation angle set to $0°$ and azimuth angles at $\{0°, 45°, 90°, 180°, 270°, 315°\}$. This distribution aligns with the setting used in Era3D [11], facilitating the application of a similar image-to-3D pipeline for 3D generation tasks. For the geometry-guided generation, we included four views at an elevation of $0°$ with azimuth angles of $\{0°, 90°, 180°, 270°\}$, added two additional views from the top and bottom. In addition to the target views, we rendered five random views within a certain frontal range of the models to serve as reference images during training.

2.2. Training

During training, we only optimize the MV-Adapter, while freezing weights of the pre-trained T2I models. We train MV-Adapter on the dataset with pairs of a reference image, text and n views, using the same training objective as T2I models:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(\boldsymbol{x}_0^{1:n}),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}),\boldsymbol{c}_t,\boldsymbol{c}_i,\boldsymbol{c}_m,t}[\|\boldsymbol{\epsilon}-\epsilon_\theta(\boldsymbol{z}_t^{1:n}, \boldsymbol{c}_t, \boldsymbol{c}_i, \boldsymbol{c}_m, t)\|_2^2] \tag{2}$$

where $\boldsymbol{c}_t$, $\boldsymbol{c}_i$ and $\boldsymbol{c}_m$ represent texts, reference images and conditioning maps (*i.e.* camera or geometry conditions) respectively. We randomly zero out the features of the reference image to drop image conditions, enabling classifier-free guidance at inference. Similar to prior work [1, 7], we shift the noise schedule towards high noise levels as we move from the T2Is to the multi-view diffusion model that captures data of higher dimensionality. We shift the log signal-to-noise ratio by $\log(n)$, where n is the number of generated views.

We utilized two versions of Stable Diffusion [17] as the base models for training. Specifically, we trained a 512-resolution model based on Stable Diffusion 2.1 (SD2.1) and a 768-resolution model based on Stable Diffusion XL (SDXL). During training, we randomly dropped the text condition with a probability of 0.1, the image condition with a probability of 0.1, and both text and image conditions simultaneously with a probability of 0.1. Following prior work [1, 7], we shifted the noise schedule to higher noise levels by adjusting the log signal-to-noise ratio (SNR) by $\log(n)$, where $n = 6$ is the number of the generated views. For the specific training configurations, we used a learning rate of 5×10^{-5} and trained the MV-Adapter on 8 NVIDIA A100 GPUs for 10 epochs.

2.3. Inference

At inference, for generation conditioned solely on text, we set the guidance scale to 7.0. For image-conditioned generation, we set the guidance scale of image condition α and text condition β to 3.0. Following TOSS [21], the calculation can be expressed as:

$$\begin{aligned}\hat{\epsilon}_\theta(\boldsymbol{z}_t^{1:n}, \boldsymbol{c}_t, \boldsymbol{c}_i, \boldsymbol{c}_m, t) &= \epsilon_\theta(\boldsymbol{z}_t^{1:n}, \emptyset, \emptyset, \boldsymbol{c}_m, t) \\ &+ \alpha \left[\epsilon_\theta(\boldsymbol{z}_t^{1:n}, \emptyset, \boldsymbol{c}_i, \boldsymbol{c}_m, t) - \epsilon_\theta(\boldsymbol{z}_t^{1:n}, \emptyset, \emptyset, \boldsymbol{c}_m, t)\right] \\ &+ \beta \left[\epsilon_\theta(\boldsymbol{z}_t^{1:n}, \boldsymbol{c}_t, \boldsymbol{c}_i, \boldsymbol{c}_m, t) - \epsilon_\theta(\boldsymbol{z}_t^{1:n}, \emptyset, \boldsymbol{c}_i, \boldsymbol{c}_m, t)\right]\end{aligned} \tag{3}$$

where $\boldsymbol{c}_t$, $\boldsymbol{c}_i$ and $\boldsymbol{c}_m$ represent texts, reference images and conditioning maps (*i.e.* camera or geometry conditions) respectively. Since we did not drop $\boldsymbol{c}_m$ during training, we do not use the classifier-free guidance method for it.

2.4. Comparison with Baselines

We conducted comprehensive comparisons with baseline methods across three settings: text-to-multiview genera-

tion, image-to-multiview generation, and texture generation. In these experiments, we evaluated both versions of MV-Adapter based on Stable Diffusion 2.1 (SD2.1) [17] and Stable Diffusion XL (SDXL) [15], demonstrating the performance gains brought by MV-Adapter due to its efficient training and scalability.

For text-to-multiview generation, we selected MVDream [22] and SPAD [9] as baseline methods. MVDream extends the original self-attention mechanism of T2I models to the multi-view domain. SPAD introduces epipolar constraints into the multi-view attention mechanism. We tested on 1,000 prompts selected from the Objaverse dataset [3]. We computed Fréchet Inception Distance (FID), Inception Score (IS), and CLIP Score on all generated views to assess the quality of the generated images and their alignment with the textual prompts.

For image-to-multiview generation, we compared our method with ImageDream [25], Zero123++[20], CRM[26], SV3D [24], O3D [27], and Era3D [11]. ImageDream, Zero123++, CRM, and Era3D generally fall into the category of modifying the original network architecture of T2I models to extend them for multi-view generation. SV3D and Ouroboros3D fine-tune text-to-video (T2V) models to achieve multi-view generation. We selected 100 assets covering multiple object categories from the Google Scanned Objects (GSO) dataset [5] as our test set. For each asset, we rendered input images from front-facing views, with input views randomly distributed in azimuth angles between -45° and 45° and elevation angles between -10° and 30°. We evaluated the generated multi-view images by computing Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) between the generated images and the ground truth, assessing both the consistency and quality of the outputs.

For 3D texture generation, we compared our text-based and image-based models with project-and-paint methods such as TEXTure [16], Text2Tex [2], and Paint3D [30], the synchronized multi-view texturing method SyncMVD [12], and the optimization-based method FlashTex [4]. We randomly selected 200 models along with their captions from the Objaverse [3] dataset for testing. Multiple views were rendered from the generated 3D textures, and we computed FID and Kernel Inception Distance (KID) of them to evaluate the quality of the generated textures. Additionally, we recorded the texture generation time to assess the inference efficiency of each method.

2.5. Community Models and Extensions

To ensure a comprehensive benchmark, we selected a diverse set of representative T2I derivative models and extensions from the community for evaluation. As illustrated in Table 1, these models include personalized models that encompass various domains such as anime, stylistic paintings, and realistic photographic images, as well as efficient distilled models and plugins for controllable generation. They cover a wide range of subjects, including portraits, animals, landscapes, and more. This selection enables a thorough evaluation of our approach across different styles and content, demonstrating the adaptability and generality of MV-Adapter in working with various T2I derivatives and extensions.

3. Additional Discussions

3.1. MV-Adapter vs. Multi-view LoRA

LoRA (Low-Rank Adaptation) [8] offers an alternative approach to achieving plug-and-play multi-view generation. Specifically, using a condition encoder to inject camera representations, we extend the original self-attention mechanism to operate across all pixels of multiple views. During training, we introduce trainable LoRA layers into the network, allowing these layers to learn multi-view consistency or, optionally, generate images conditioned on a reference view. This approach requires the spatial self-attention mechanism to simultaneously capture spatial image knowledge, ensure multi-view consistency, and align generated images with reference views.

However, the multi-view LoRA approach has a notable limitation. The "incremental changes" it introduces to the network are **not orthogonal or decoupled** from those induced by T2I derivatives, such as personalized T2I models or LoRAs. Specifically, layers fine-tuned by multi-view LoRA and those tuned by personalized LoRA often overlap. Note that each weight matrix learned by both represents a linear transformation defined by its columns, so it is intuitive that the merger would retain the information available in these columns only when the columns that are being added are orthogonal to each other [19]. Clearly, the multi-view LoRA and personalized models are not orthogonal, which often leads to challenges in retaining both sets of learned knowledge. This can result in a trade-off where either multi-view consistency or the fidelity of concepts (such as style or subject identity) is compromised.

In contrast, our proposed **decoupled** attention mechanism encourages different attention layers to specialize in their respective tasks without needing to fine-tune the original spatial self-attention layers. In this design, the layers we train do not overlap with those in the original T2I model, thereby better preserving the original feature space and enhancing compatibility with other models.

We conducted a series of experiments to test these approaches. We trained two versions of multi-view LoRA, targeting different modules: (1) inserting LoRA layers only into the attention layers, and (2) inserting LoRA layers into multiple layers, including the convolutional layers, down-

Table 1. Community models and extensions for evaluation.

Category	Model Name	Domain	Model Type
Personalized T2I	Dreamshaper[1]	General	T2I Base Model
	RealVisXL[2]	Realistic	T2I Base Model
	Animagine-xl[3]	2D Cartoon	T2I Base Model
	3D Render Style XL[4]	3D Cartoon	LoRA
	Pokemon Trainer Sprite PixelArt[5]	Pixel Art	LoRA
	Chalk Sketch SDXL[6]	Chalk Sketch	LoRA
	Chinese Ink LoRA[7]	Color Ink	LoRA
	Zen Ink Wash Sumi-e[8]	Wash Ink	LoRA
	Watercolor Style SDXL[9]	Watercolor	LoRA
	Papercut SDXL[10]	Papercut	LoRA
	Furry Enhancer[11]	Enhancer	LoRA
	White Pitbull Dog SDXL[12]	Concept	LoRA
	Spider spirit fourth sister[13]	Concept	LoRA
Distilled T2I	SDXL-Lightning[14]	Few Step	T2I Base Model
	LCM-SDXL[15]	Few Step	T2I Base Model
Extension	ControlNet Openpose[16]	Spatial Control	Plugin
	ControlNet Scribble[17]	Spatial Control	Plugin
	ControlNet Tile[18]	Image Deblur	Plugin
	T2I-Adapter Sketch[19]	Spatial Control	Plugin
	IP-Adapter[20]	Image Prompt	Plugin

sampling, up-sampling layers, etc. For both settings, we set the LoRA rank to 64 and alpha to 32. As shown in Fig. 1 and Fig. 2, while the multi-view LoRA approach can generate multi-view consistent images when the base model is not changed, it often struggles to maintain multi-view consistency when switching to a different base model or when integrating a new LoRA. In contrast, as demonstrated in Fig. 3, our MV-Adapter, equipped with the decoupled attention mechanism, maintains consistent multi-view generation even when used with personalized models.

Compared to the LoRA mechanism, our decoupled attention-based approach proves more robust and adaptable for extending T2I models to multi-view generation, offering greater flexibility and compatibility with various pre-trained models.

3.2. Image Restoration Capabilities

During the training of MV-Adapter, we probabilistically compress the resolution of reference images in the training data pairs to enhance the robustness of multi-view generation from images. We observed that the model trained with this approach is capable of generating high-resolution, detailed multi-view images even when the input is low-resolution, as depicted in Fig. 4. Through such training strategy, MV-Adapter has inherent image restoration capabilities and automatically enhances and refines input images during the generation process.

3.3. Details of Arbitrary View Synthesis

In the main text, we introduced a novel adapter architecture—comprising parallel attention layers and a unified condition encoder—to achieve multi-view generation. We implemented efficient row-wise and column-wise attention mechanisms tailored for two specific applications: 3D object generation and 3D texture mapping, generating six views accordingly. However, our adapter framework is not limited to these configurations and can be extended to perform arbitrary view synthesis. To explore this capability, we designed a corresponding approach and conducted experiments, training a new version of MV-Adapter to handle arbitrary viewpoints.

Following CAT3D [6], 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 clus-

(Base model: SDXL) *Daenerys Targaryen from game of throne, full body, blender 3d, art station*

(Base model: Animagine-xl) *1 girl, pink hair, pink shirts, smile, shy, masterpiece*

(LoRA: Watercolor Style) *painting, Burmese Cat, wearing ral-wtrclr, Comic book art*

Figure 1. Results of multi-view LoRA (set target modules to attention layers). The azimuth angles of the images from left to right are $0°, 45°, 90°, 180°, 270°, 315°$, corresponding to the front, front-left, left, back, right, and front-right of the object.

(Base model: SDXL) *Daenerys Targaryen from game of throne, full body, blender 3d, art station*

(Base model: Animagine-xl) *1 girl, pink hair, pink shirts, smile, shy, masterpiece*

(LoRA: Watercolor Style) *painting, Burmese Cat, wearing ral-wtrclr, Comic book art*

Figure 2. Results of multi-view LoRA (set target modules to attention layers, convolutional layers, etc.). The azimuth angles of the images from left to right are $0°, 45°, 90°, 180°, 270°, 315°$, corresponding to the front, front-left, left, back, right, and front-right of the object.

(Base model: SDXL) *Daenerys Targaryen from game of throne, full body, blender 3d, art station*

(Base model: Animagine-xl) *1 girl, pink hair, pink shirts, smile, shy, masterpiece*

(LoRA: Watercolor Style) *painting, Burmese Cat, wearing ral-wtrclr, Comic book art*

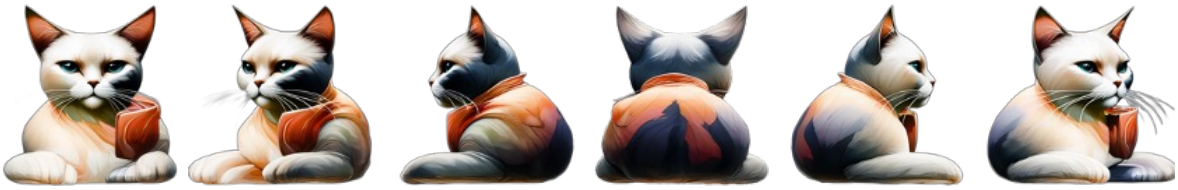

Figure 3. Results of MV-Adapter, which introduces decoupled attention mechanism rather than LoRA. The azimuth angles of the images from left to right are $0°, 45°, 90°, 180°, 270°, 315°$, corresponding to the front, front-left, left, back, right, and front-right of the object.

ters 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.

In terms of implementation, the overall framework of our MV-Adapter remains unchanged. We adjust its inputs and specific attention components to accommodate arbitrary view synthesis. First, we set the number of input images to either 1 or 4. 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. This simple yet effective method allows the images from the four views to interact within the pre-trained U-Net without requiring additional camera embeddings to represent these views. Second, we utilize full self-attention in the multi-view attention component, expanding the attention scope to enable the generation of target views with more flexible distributions.

To train an MV-Adapter capable of generating arbitrary viewpoints, we rendered data from 40 different views, with elevations of $-10°, 0°, 10°, 20°, 30°$, and azimuth angles evenly distributed around 360 degrees at each elevation layer. We trained the model for 16 epochs. During the first 8 epochs, the model was trained using a setting of one conditional view and eight target anchor views. In the subsequent 8 epochs, we trained with an equal mixture of one condition plus eight target views and four conditions plus eight target views.

As shown in Fig. 5, the visualization results demonstrate that MV-Adapter can generate consistent, high-quality multi-view images beyond the six views designed for specific applications. This extension further verifies the scalability and practicality of our adapter framework, showcasing its potential for arbitrary view synthesis in diverse applications. More results can be found in the supplementary materials.

3.4. Limitations and Future Works

Limitation: Dependency on image backbone. Within our MV-Adapter, we only fine-tune the additional multi-view attention and image cross-attention layers, and do not disturb the original structure or feature space. Consequently, the overall performance of MV-Adapter is heavily dependent on the base T2I model. If the foundational model struggles to generate content that aligns with the provided prompt or produces images of low quality, MV-Adapter is unlikely to compensate for these deficiencies. On the other hand, employing superior image backbones can enhance the synthetic results. We present a comparison of outputs generated using SDXL [15] and SD2.1 [17] models in Fig. 6, which confirms this observation, particularly in text-conditioned multi-view generation. We believe that MV-Adapter can be further developed by utilizing advanced T2I models [10, 23] based on the DiT architecture [14], to

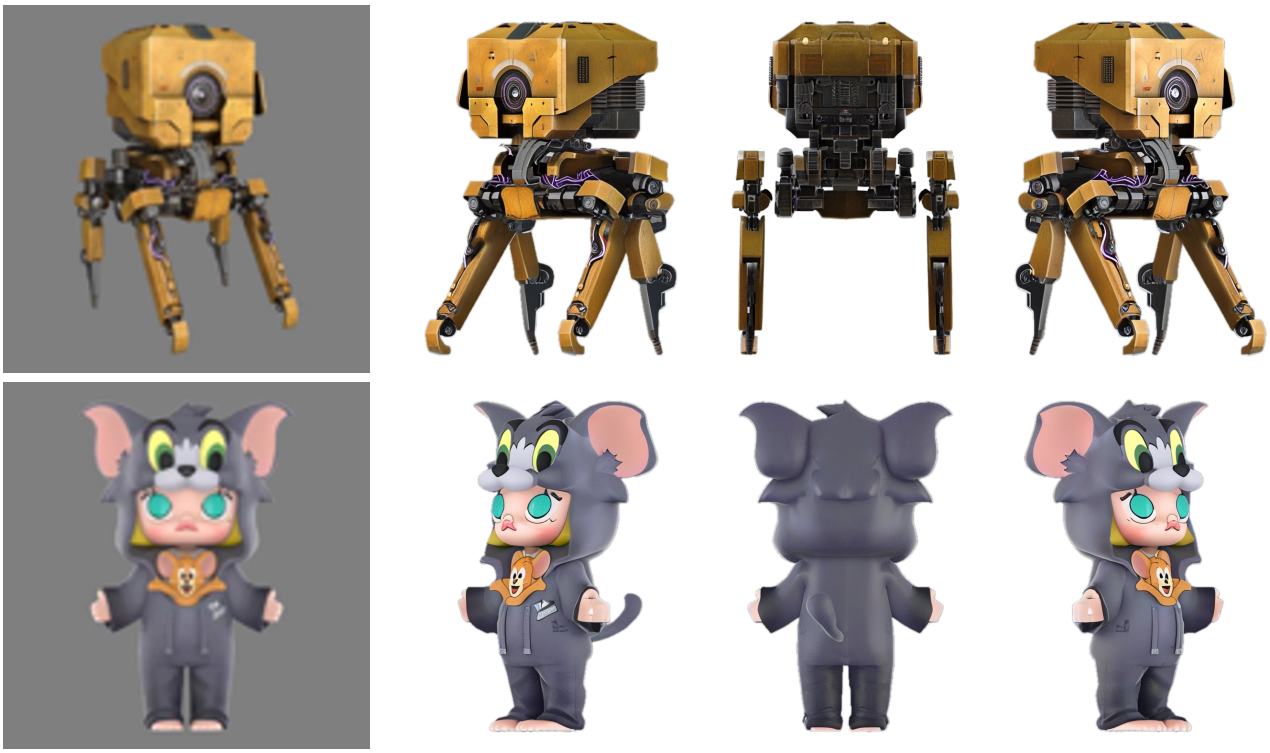

Figure 4. Results on camera-guided image-to-multiview generation with low-resolution images as input.

Figure 5. Visualization results using MV-Adapter to generate arbitrary viewpoints.

Figure 6. Qualitative comparison of our MV-Adapter based on SD2.1 and SDXL.

achieve higher visual quality in the generated images.

Future works: 3D scene generation, dynamic multi-view video generation, inspiration for modeling new knowledge. This paper provides extensive analyses and enhancements for our novel multi-view adapter, MV-Adapter. While our model has significantly improved efficiency, adaptability, versatility, and performance compared to previous models, we identify several promising areas for future work:

- 3D scene generation. Our method can be extended to scene-level multi-view generation, accommodating both camera- and geometry-guided approaches with text or image conditions.
- Dynamic multi-view video generation. Exploring dynamic multi-view video generation using a similar approach as MV-Adapter within text-to-video generation models [28, 32] presents a valuable opportunity for further advancements.
- Inspiration for modeling new knowledge. Our approach of decoupling the learning of geometric knowledge from the image prior can be applied to learning zoom in/out effects, consistent lighting, and other viewpoint-dependent properties. It also provides valuable insights for modeling physical or temporal knowledge based on image priors.

4. More Comparison Results

4.1. Image-to-Multi-view Generation

To provide a more in-depth analysis of our quantitative results on image-to-multi-view generation, we conducted a user study comparing MV-Adapter (based on SD2.1 [17]) with baseline methods [11, 20, 24–27]. The study aimed to evaluate both multi-view consistency and image quality preferences. We selected 30 samples covering a diverse range of categories, such as toy cars, medicine bottles, stationery, dolls, and sculptures. A total of 50 participants were recruited to provide their preferences between the outputs of different methods.

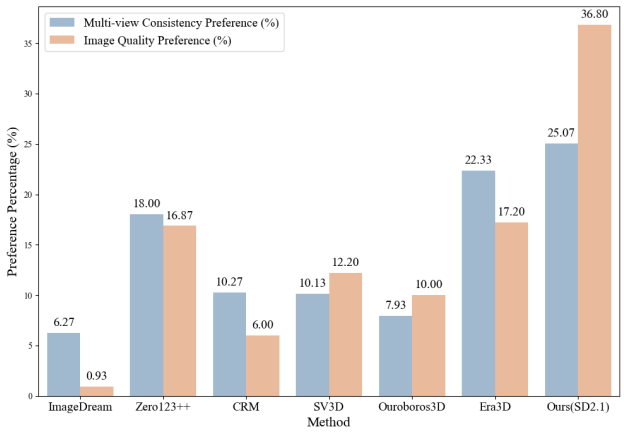

Figure 7. Results of user study on image-to-multi-view generation.

Participants were presented with pairs of multi-view images generated by MV-Adapter and the baseline methods. For each pair, they were asked to choose the one they preferred in terms of multi-view consistency and image quality. The results of the user study are summarized in Fig. 7. The findings indicate that, in terms of multi-view consistency, MV-Adapter performs comparably to Era3D, with preference rates of 25.07% and 22.33%, respectively. However, regarding image quality, MV-Adapter demonstrates a significant advantage, receiving a higher preference rate of 36.80% compared to the baseline methods. The improved image quality can be attributed to MV-Adapter's ability to leverage the strengths of the underlying T2I models without full fine-tuning, preserving the original feature space and benefiting from the high-quality priors of the base models.

Figure 8. Additional results on camera-guided text-to-multiview generation with community models.

5. More Visual Results

In Fig. 8 and Fig. 9, we show more visual results of MV-Adapter on camera-guided text-to-multiview generation with community models and extensions, such as ControlNet [31] and IP-Adapter [29]. In Fig. 10, we show more visual results on camera-guided image-to-multiview generation. In Fig. 11, we show more visual results on text-to-3D generation. In Fig. 12, we show more visual results on

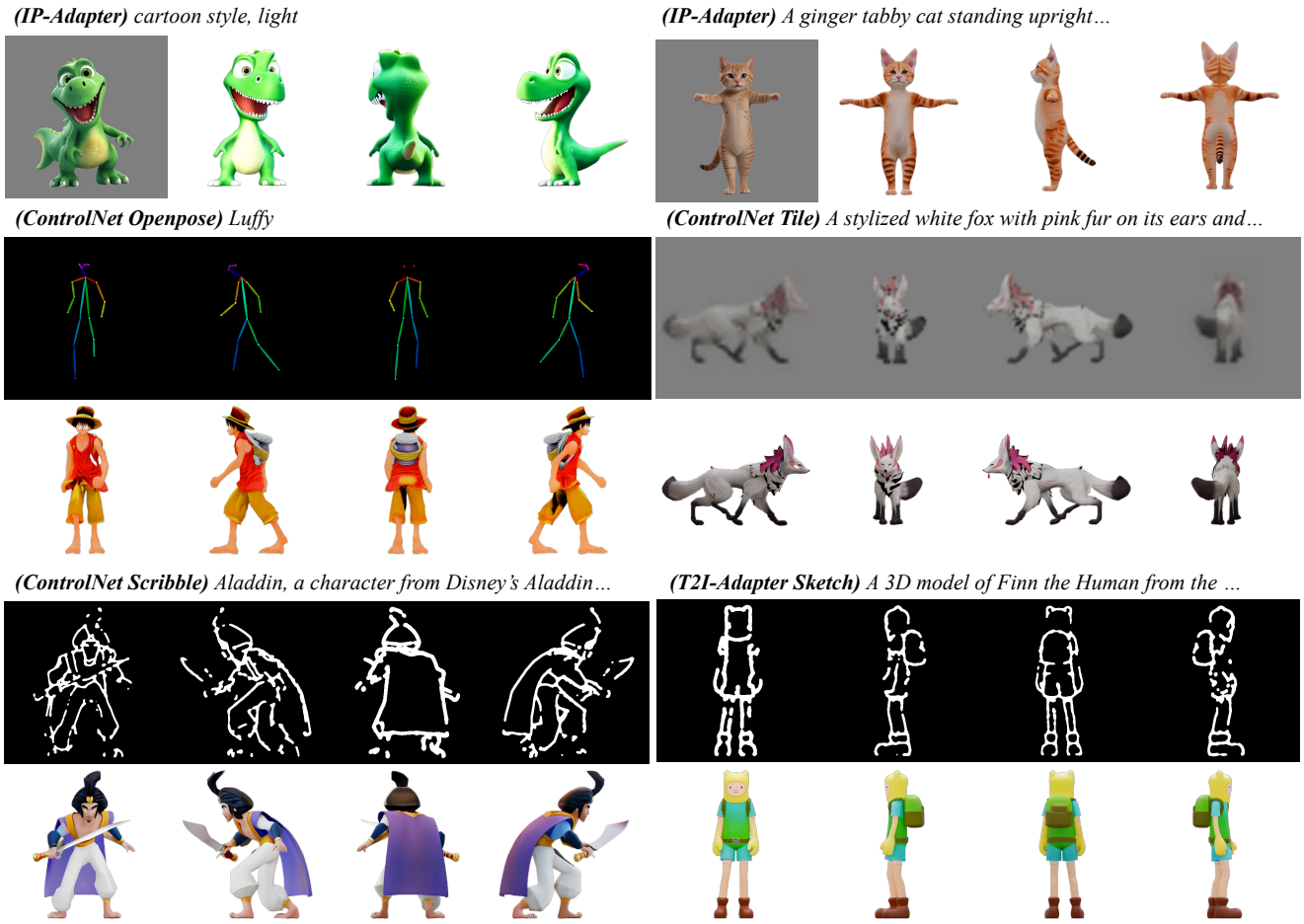

Figure 9. Additional results on camera-guided text-to-multiview generation with extensions.

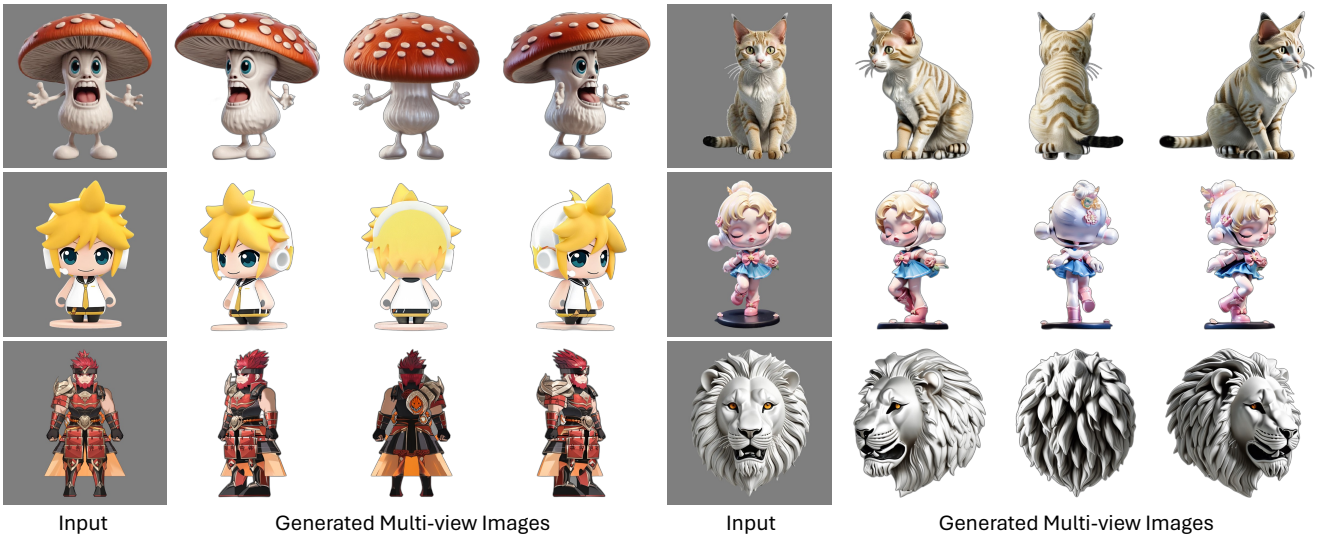

Figure 10. Additional results on camera-guided image-to-multiview generation.

Figure 11. Visual results on text-to-3D generation.

image-to-3D generation. In Fig. 13, we show more visual results on geometry-guided text-to-texture generation. In Fig. 14, we show more visual results on geometry-guided image-to-texture generation. Note that we have removed the background of the generated images in the visual results.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 1

[2] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models, 2023. 2

[3] 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. 1, 2

[4] Kangle Deng, Timothy Omernick, Alexander Weiss, Deva Ramanan, Jun-Yan Zhu, Tinghui Zhou, and Maneesh Agrawala. Flashtex: Fast relightable mesh texturing with lightcontrolnet, 2024. 2

[5] 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. 2

[6] 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. 3

Figure 12. Visual results on image-to-3D generation.

[7] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images, 2023. 1

[8] 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

[9] 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

[10] Black Forest Labs. Flux. [Online], 2024. https://github.com/black-forest-labs/flux. 5

[11] 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. 1, 2, 7

[12] Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion, 2023. 2

[13] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models, 2024. 1

[14] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. 5

[15] 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. 1, 2, 5

[16] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes, 2023. 2

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 2, 5, 7

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 1

[19] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras, 2023. 2

[20] 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, 7

[21] Yukai Shi, Jianan Wang, He Cao, Boshi Tang, Xianbiao Qi, Tianyu Yang, Yukun Huang, Shilong Liu, Lei Zhang, and

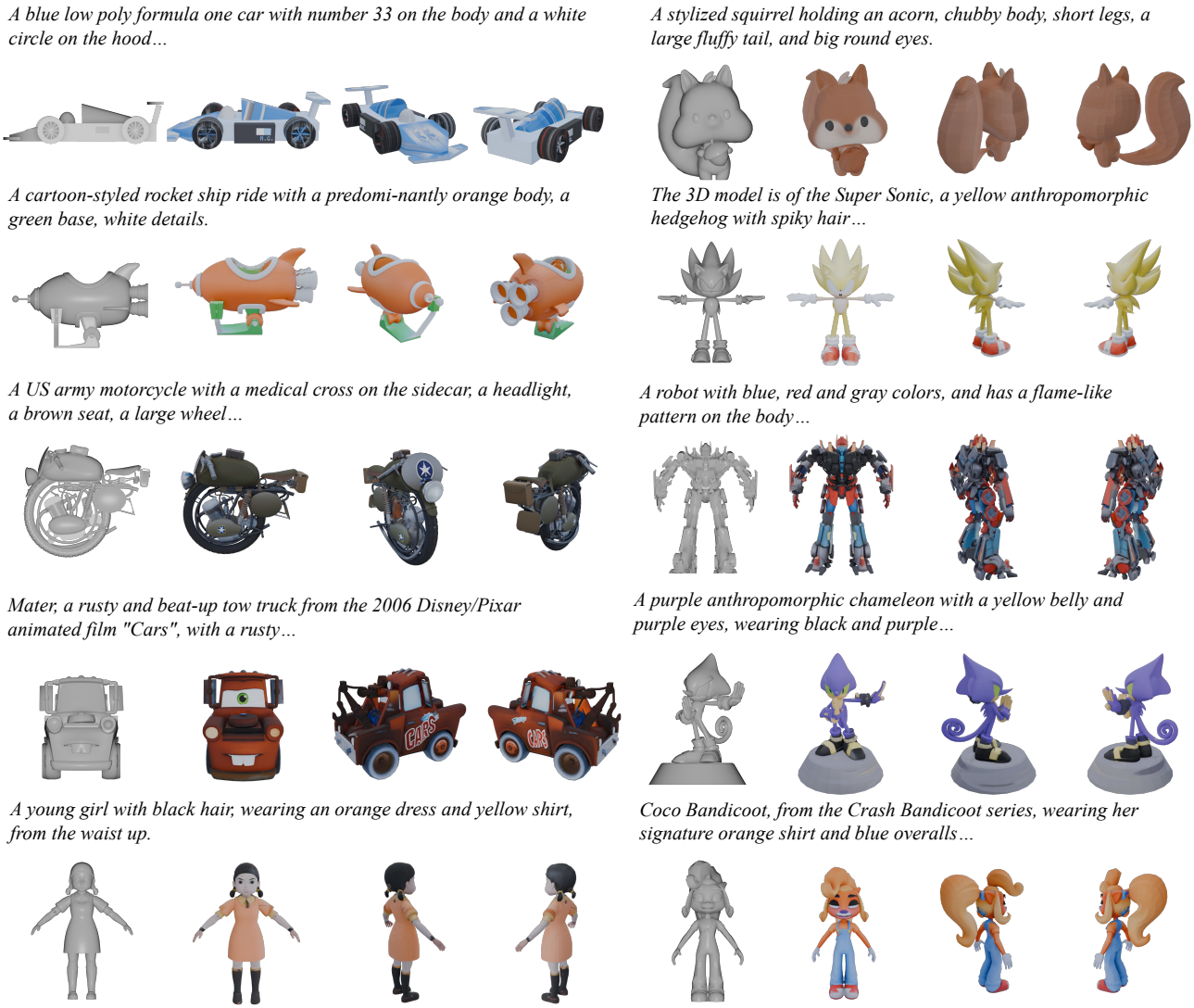

Figure 13. Additional results on geometry-guided text-to-texture generation.

Heung-Yeung Shum. Toss: High-quality text-guided novel view synthesis from a single image, 2023. 1

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

[23] Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis, 2024. 5

[24] 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. 2, 7

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

[26] 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. 2

[27] Hao Wen, Zehuan Huang, Yaohui Wang, Xinyuan Chen, Yu Qiao, and Lu Sheng. Ouroboros3d: Image-to-3d generation via 3d-aware recursive diffusion, 2024. 2, 7

[28] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer, 2024. 7

[29] 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. 8

[30] 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. 2

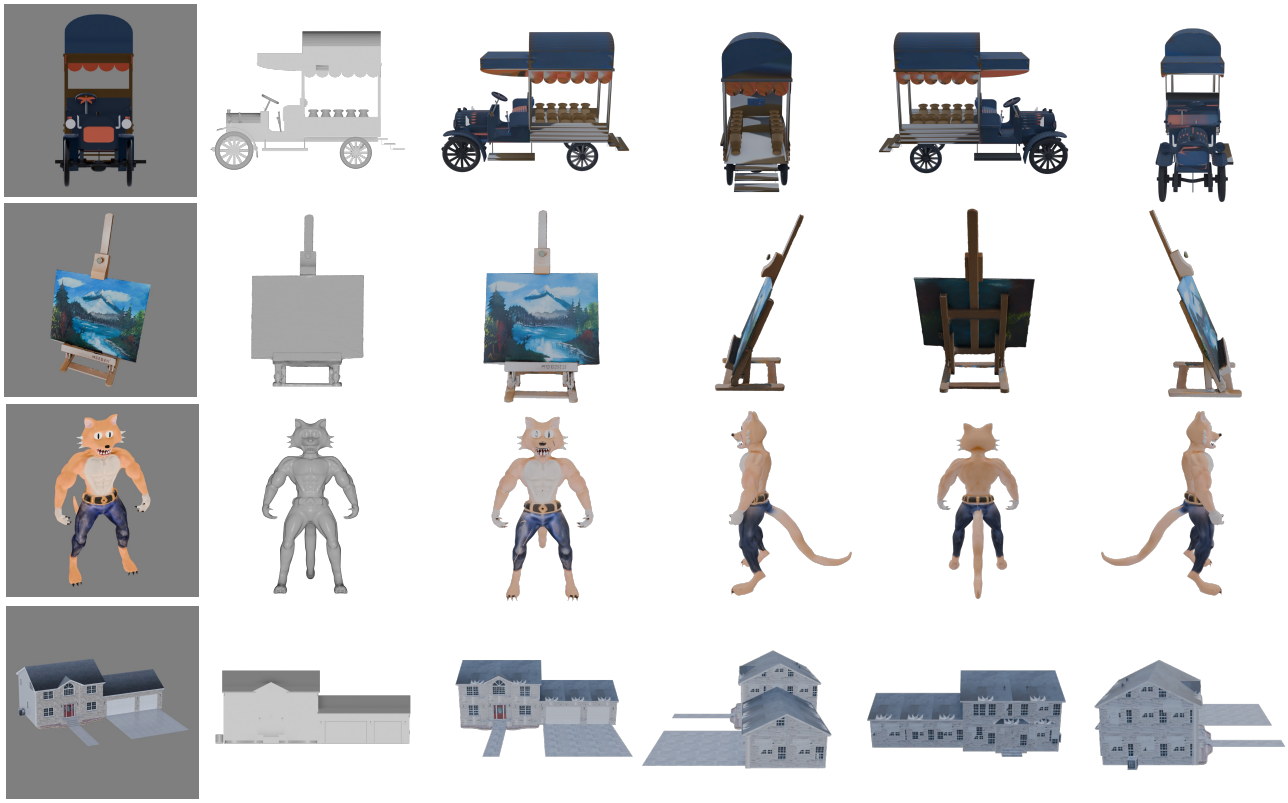

Figure 14. Additional results on geometry-guided image-to-texture generation.

[31] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 8

[32] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 7