Supplementary of "MVGenMaster: Scaling Multi-View Generation from Any Image via 3D Priors Enhanced Diffusion Model"

Chenjie Cao[1,2,3], Chaohui Yu[2,3], Shang Liu[2,3], Fan Wang[2], Xiangyang Xue[1], Yanwei Fu[1]
[1]Fudan University, [2]DAMO Academy, Alibaba Group, [3]Hupan Lab
{caochenjie.ccj,huakun.ych,liushang.ls,fan.w}@alibaba-inc.com, {xyxue,yanweifu}@fudan.edu.cn

A. Dataset Details

We discuss more details about the MvD-1M used in this work and the way to get metric depth of them. The visualization of some examples from MvD-1M is shown in Figure 1.

Co3Dv2 [16]: Co3Dv2 is a widely used object-centric dataset with diverse classes, backgrounds, and image resolutions. We filtered the official Co3Dv2 dataset, retaining the scenes whose short sides are larger than 256. Furthermore, we ensure that all categories in Co3Dv2 are sampled in balance during training. Since Co3Dv2 contained sparse depth maps from SfM, which were used to align metric depth.

MVImgNet [24]: MVImgNet contains high-quality object-centric images, while the viewpoint changes are not as large as Co3Dv2. Similarly, we maintained category balance for each training epoch in MVImgNet. MVImgNet has sparse Colmap points which could be utilized to align metric depth.

DL3DV [10]: DL3DV is a large-scale, scene-level dataset with a variety of scenarios, and it serves as the primary source for our model's training. Note that only some DL3DV scenes contain sparse Colmap points, others do not. For scenes lacking SfM results, we employed the MVS [3] to achieve metric depths, applying a confidence filter of > 0.5.

GL3D [17]: GL3D primarily contains aerial images, alongside a limited number of scene-level multi-view images. Given that the authors did not provide detailed camera poses for the undistorted images, we utilized Colmap to estimate the camera poses and extract sparse SfM points.

Scannet++ [22]: Scannet++ comprises indoor scenes. To ensure the image quality, we use images captured by DSLRs rather than RGBD from cell phones. We first undistorted the DSLR images, and then applied the officially provided sparse Colmap points to align the metric depth.

3D-Front [5]: 3D-Front includes numerous indoor scenes rendered using Blender. For this dataset, we directly utilized the rendered ground truth depth for geometric warping.

RealEstate10K [26]: RealEstate10K is a widely used scene-level dataset. We combined the training and test sets to maximize its utility, retaining 50 scenes for validation, as detailed in Table 1 of the main paper. Because RealEstate10k only contains camera poses, we employed MVS [3] to get metric depths.

ACID [11]: ACID is built with aerial images captured in natural landscapes. We filtered out samples with minimal motion and used MVS [3] to obtain the metric depths.

Objaverse [4]: All images from Objaverse are rendered in 512×512. To improve the diversity, we randomly altered the background colors for different objects in this dataset. Importantly, we did not apply 3D priors in this relatively simpler dataset to increase the challenge.

Megascenes [18]: Megascenes contains diverse places collected from the wiki, captured at various times, using a range of devices. We retained scenes with a valid view count of at least 8 and aligned the metric depth using sparse SfM points.

Aerial: The Aerial dataset was collected from Google Earth [1], capturing images below 500 meters above the ground. We retain the images with top-3 heights to ensure superior visual quality. To derive metric depths, we trained Instant-NGP [14] for each scene and rendered the metric depth through the reconstructed mesh.

Streetview: The Streetview dataset was also collected from Google Earth [1], which comprises both indoor and outdoor scenes from New York. All Streetview scenes are panoramic images. We utilized Equirectangular projection to derive standard images and camera poses with random intrinsic and extrinsic matrices to enhance diversity. Notably, as depth data is essentially meaningless for geometric warping for panoramic images, we opted not to employ geometric warping for this dataset, including only the camera poses.

B. More Implementation Details

B.1. Canonical Coordinate Map (CCM)

We provide more details about CCM [9] mentioned in the main paper. To incorporate precise positional information to MVGenMaster, we begin by unprojecting 2D coordinates x_r from reference views into the 3D world space (canonical space) as:

$$C_r^{world} \simeq P_r^{c\rightarrow w} \hat{D}_r(x_r) K_r^{-1} x_r, \tag{1}$$

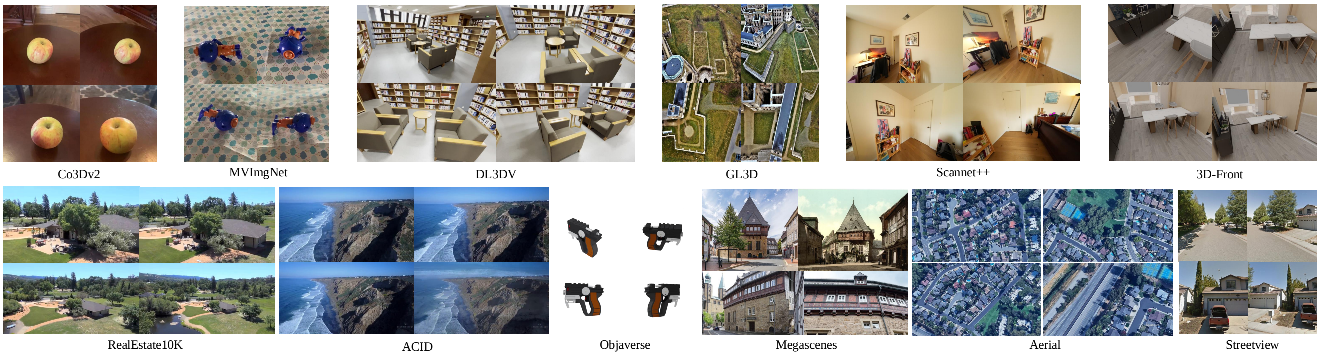


Figure 1. **Illustration of MvD-1M used in MVGenMaster.**

where C_r^{world} denotes the 3D coordinates in the world space, *i.e.*, CCM; $P_r^{c\rightarrow w}$ is the transformation matrix converting coordinates from camera to world space; $\hat{D}_r(x_r)$ and K_r^{-1} indicate metric depth and camera intrinsic's inversion of reference views, respectively. Subsequently, C_r^{world} is further warped into target views, *i.e.*, $C_{r\rightarrow t}^{warp}$ following Eq.(3) in the main paper. Note that we further normalize $C_{r\rightarrow t}^{warp}$ into 0 to 1 before send it to the Fourier embedding.

B.2. Multi-Scale Training

All resolutions used in our multi-scale training are listed in Table 1. And we ensure that at least one batch samples are allocated for each resolution group per-epoch.

Resolution (height × width)		
320×768	384×448	576×448
320×704	448×576	576×384
320×640	448×512	576×320
320×576	512×512	640×384
320×512	448×384	640×320
384×640	512×448	704×320
384×576	512×384	768×320
384×512	512×320	

Table 1. **Resolutions used in the multi-scale training.**

B.3. Normalization of Camera Poses

The Plücker ray [21] enjoys dense and good camera presentations for NVS, which can be denoted as $(\mathbf{o}\times\mathbf{d},\mathbf{d})$, where $\mathbf{o},\mathbf{d}$ indicate the origin and direction of pixel-aligned rays respectively. We clarify that the cross-product result $\mathbf{o}\times\mathbf{d}$ is not scaling invariable due to different scales of $\mathbf{o}$. Since the SfM results of different datasets contain different camera scales, we need to normalize them before the training and inference to avoid unseen camera scales during the inference. Specifically, we analyzed all camera positions within each scene and saved the longest distance between the two farthest cameras. If this distance exceeded 5, we normalized the global scene to ensure the longest side was 5, while the metric depth should also be re-scaled accordingly. Such normalization allows the model to effectively adapt small camera motions instead of straightforwardly normalizing all cameras into the same scale.

B.4. Settings of 3DGS Reconstruction

We follow the vanilla 3DGS implementation of [8] with L1, SSIM, and LPIPS [25] losses as many other works [12, 13, 23]. The point cloud from Dust3R [19] is leveraged to initialize the 3DGS training. The dynamic LPIPS weighting strategy [6, 12] is also incorporated according to the distance to the nearest reference view, and the LPIPS weight changes from 0 (reference view) to 0.25 (the farthest view). The training is accomplished within 2000 steps in our work.

B.5. Efficiency and Memory Costs

Our model enjoys a similar efficiency as CAT3D [6], while the cost of 3D priors incorporation is negligible. Inference efficiency and GPU memory costs of the 50-step DDIM scheduler with different target view numbers are listed in Table 2. The experimental device is one 80GB A800 NVIDIA GPU; the image resolution is 384×576. Note that our model can synthesize all views at once instead of cumbersome iterative generations.

Views	3D priors	Time	Memory
8	×	7.0085s	6420M
8	✓	7.1768s	6440M
25	✓	21.272s	14048M
50	✓	53.055s	27762M
100	✓	132.67s	50542M
158	✓	262.86s	80738M

Table 2. **Efficiency and memory costs of our model.** The experimental device is one 80GB NVIDIA A800 GPU, while the image resolution is 384x576. The setting of 8-view without 3D priors can be regarded as the baseline CAT3D [6].

C. Supplementary Experiment Results

We show the qualitative comparison of 3DGS reconstruction in Figure 2. More NVS results are listed in Figure 5, Figure 6, and Figure 8. Additionally, more reference views further strengthen the performance as shown in Table 4 and Figure 9.

		3-view		6-view		9-view	
		PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
Real10k	Recon[20]	25.84	0.144	29.99	0.103	31.82	0.092
	CAT3D[6]	**26.78**	0.132	**31.07**	0.092	**32.20**	0.082
	Ours	25.15	**0.091**	27.28	**0.067**	27.54	**0.062**
CO3D	Recon[20]	19.59	0.398	21.84	0.342	22.95	0.318
	CAT3D[6]	**20.57**	0.351	**22.79**	0.292	**23.58**	0.273
	Ours	19.99	**0.278**	21.96	**0.214**	22.77	**0.195**
Mip360	Recon[20]	15.50	0.585	16.93	0.544	18.19	0.511
	CAT3D[6]	**16.62**	0.515	**17.72**	0.482	**18.67**	0.460
	Ours	16.29	**0.489**	17.51	**0.405**	18.21	**0.354**

Table 3. **Reconstruction results of the official Reconfusion benchmark.** Results of reconfusion [20] and CAT3D [6] are from their papers. Our results are under once NVS inference and 3DGS.

Reconstruction Benchmark. To ensure the effectiveness of MVGenMaster, we compare our method with the official data splits from the benchmark of Reconfusion [20] as in Table 3. Results of Reconfusion [20] and CAT3D [6] are from their papers. Our method enjoys a much simpler pipeline to produce all views with one forward process without iterative generation (CAT3D) and distillation (Reconfusion), only taking 3 minutes for each scene with one A800 (NVS+3DGS). Besides, Reconfusion and CAT3D use ZipNeRF [2], which costs much more training times for each instance. Moreover, we should clarify that the results of MipNeRF-360 are slightly different from Table.3 of the main paper. Because the data splits of Reconfusion are different from ours. They use a heuristic strategy to encourage more reasonable camera placing for the object centric reconstruction [20]. Even under such a challenging setting, our method achieves better human perception (lower LPIPS) and comparable PSNR.

Datasets Reference View	PSNR↑	SSIM↑	LPIPS↓
CO3D+MVImgNet			
Ours (1-view)	18.619	0.573	0.316
Ours (3-view)	21.466	0.653	0.220
Ours (5-view)	22.594	0.682	0.188
Ours (7-view)	**23.299**	**0.697**	**0.169**
DL3DV+Real10k			
Ours (1-view)	15.729	0.468	0.376
Ours (3-view)	18.296	0.552	0.266
Ours (5-view)	19.366	0.585	0.224
Ours (7-view)	**20.039**	**0.604**	**0.204**
Zero-shot Datasets			
Ours (1-view)	12.879	0.394	0.466
Ours (3-view)	15.533	0.491	0.319
Ours (5-view)	16.826	0.531	0.263
Ours (7-view)	**17.559**	**0.550**	**0.233**

Table 4. **Results of MVGenMaster with different reference views.** The total view number $N + M = 25$.

Extending to More Target Views. To explore the robustness of the proposed key-rescaling, we conduct an exploratory ablation study in Table 5. In this study, we increase the total number of views to 158, which represents the memory limit of an 80GB GPU. Key-rescaling with $\gamma = 1.2$ is still effective and robust in persisting the reference guidance for such an extremely long sequence. We should clarify that this memory limit is merely a hardware constraint of our method, which could potentially be overcome with efficient engineering optimizations. We regard these engineering optimizations as interesting future work to further extend the target view numbers of MVGenMaster. Furthermore, our method's capability to generate 100 views is sufficient to address most NVS scenarios, as verified in our main paper.

views at once	key-rescaling	PSNR↑	SSIM↑	LPIPS↓
8/158	–	15.193	0.378	0.362
28/158	–	14.913	0.386	0.361
158/158	1.15	15.439	**0.410**	0.358
158/158	1.2	15.565	0.409	**0.355**
158/158	1.25	**15.597**	0.403	0.356
158/158	1.3	15.538	0.397	0.359

Table 5. **Results of extremely long sequence generation of MVGenMaster.** The total view number $N + M = 158$ reaches the upper bound of an 80GB GPU.

N/M	γ	PSNR↑	LPIPS↓	N/M	γ	PSNR↑	LPIPS↓
1/97	1.0	12.53	0.541	2/97	1.0	13.69	0.468
	1.1	**12.56**	0.528		1.1	14.31	0.432
	1.2	12.52	**0.523**		1.2	**14.46**	**0.426**
	1.3	12.50	0.526		1.3	14.35	0.438
3/50	1.0	14.82	0.386	3/97	1.0	14.75	0.402
	1.1	15.47	0.359		**A1**	15.59	0.360
	1.2	**15.55**	**0.358**		1.2	**15.64**	**0.358**
	1.3	15.28	0.371		**A2**	15.48	0.364

Table 6. **Detailed Ablation of Key-rescaling.** Following the adaptive γ settings based on attention-entropy [7], **A1** is set based on the length of $N + M$, while **A2** further considers the various sequential lengths across different layers.

Detailed Ablation of Key-rescaling. We clarified that $\gamma = 1.2$ defined in the main paper works robustly enough for our method across various N, M (up to upper bound $N + M$=158). We further provide detailed ablation studies and discussions about key-rescaling in Table 6. i) $N = 1$ suffers from ambiguous NVS by generating vast number of views (usually set $M \leq 25$ for $N = 1$). As exploratory results, key-rescaling's effect of $N = 1$ is a little weaker than $N > 1$, but it still achieves the best quality under $\gamma = 1.2$ (best LPIPS). ii) Longer sequence enjoys larger γ, while shorter one requires smaller γ. Since $N + M = 100$ addresses most scenarios, $\gamma = 1.2$ is robust enough for our method. iii) We show adaptive γ settings based on attention-entropy [7] (**A1**, **A2**). They fail to get superior results compared to constant γ. Further exploring the key-rescaling theory and finding the optimal γ is interesting future work. Our work provides this insight to the community and verifies the proposed key-rescaling is generalized enough to address our problems based on Occam's Razor.

Results Compared on Megascenes. We further compare

Methods	PSNR↑	SSIM↑	LPIPS↓
Warp+Pose [18] (1-view)	9.808	0.201	0.597
MVGenMaster (1-view)	11.188	0.299	0.575
MVGenMaster (3-view)	**12.374**	**0.347**	**0.484**

Table 7. **Results on the challenging Megascenes dataset.** Our method is compared with the warp+pose based model in [18].

the NVS results on the challenging Megascenes dataset [18] in Table 7 and Figure 4. Our method outperforms the Warp+Pose baseline in [18] with 1-view reference. Moreover, our 3-view-based results achieve better qualities. Although both methods of MVGenMaster and [18] used 3D priors from depth estimation, MVGenMaster could handle high-resolution, multi-view consistent, and impressive generations with varying reference and target views. In contrast, the baseline in [18] only performs with restricted low-resolution (256x256) and 1-view conditions.

Robustness of 3D Priors. Benefiting from the 3D prior dropout and suboptimal monocular depth alignment, MVGenMaster enjoys good robustness to 3D priors as illustrated in Figure 7. In this case, the depth alignment is failed, leading to poor metric depth maps. However, our method still achieves proper results with camera pose conditions.

Limitations. In Figure 3, we show some ambiguous artifacts in the background generated by our model. Such artifacts are mainly caused by the unclear and ambiguous background regions of the reference view. Improving the model backbone with superior capacity would refine this issue as mentioned in our main paper.

D. Discussion and Future Work

Although MVGenMaster is a powerful NVS model, an interesting future work involves modifying the base model design, *e.g.*, replacing the backbone with Diffusion Transformer (DiT) [15] for superior capacity. Besides, unifying both rendering and multi-view generation is a promising way to improve consistency and stability of NVS.

E. Broader Impacts

This paper delves into the realm of image-based multi-view generation. Because of the powerful generative capacity, these models pose risks such as the potential for misinformation and the creation of fake images. We sincerely remind users to pay attention to generated content. Besides, it is crucial to prioritize privacy and consent, as generative models frequently rely on vast datasets that may include sensitive information. Users must remain vigilant about these considerations to uphold ethical standards in their applications. Note that our method only focuses on technical aspects. Both images and pre-trained models used in this paper will be open-released.

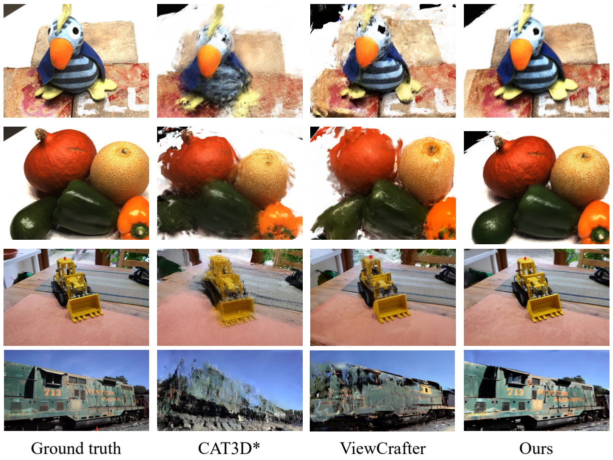


Figure 2. **Visualization of the novel views from 3DGS reconstruction on DTU, MipNeRF-360, and Tanks-and-Temples.** 3DGS results are trained with 21 frames, while the other 4 frames are validated. Given the first and last frames, other views are generated by related methods.

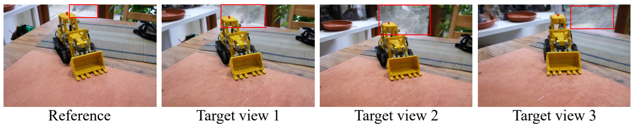


Figure 3. **Limitation.** The artifacts in the background caused by unclear and ambiguous background regions in the reference view.

References

[1] Google earth. https://www.google.com/earth/studio/. 1

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 3

[3] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer++: Revealing the devil in transformer's details for multi-view stereo. In *International Conference on Learning Representations*, 2024. 1

[4] 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. 1

[5] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, pages 1–25, 2021. 1

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

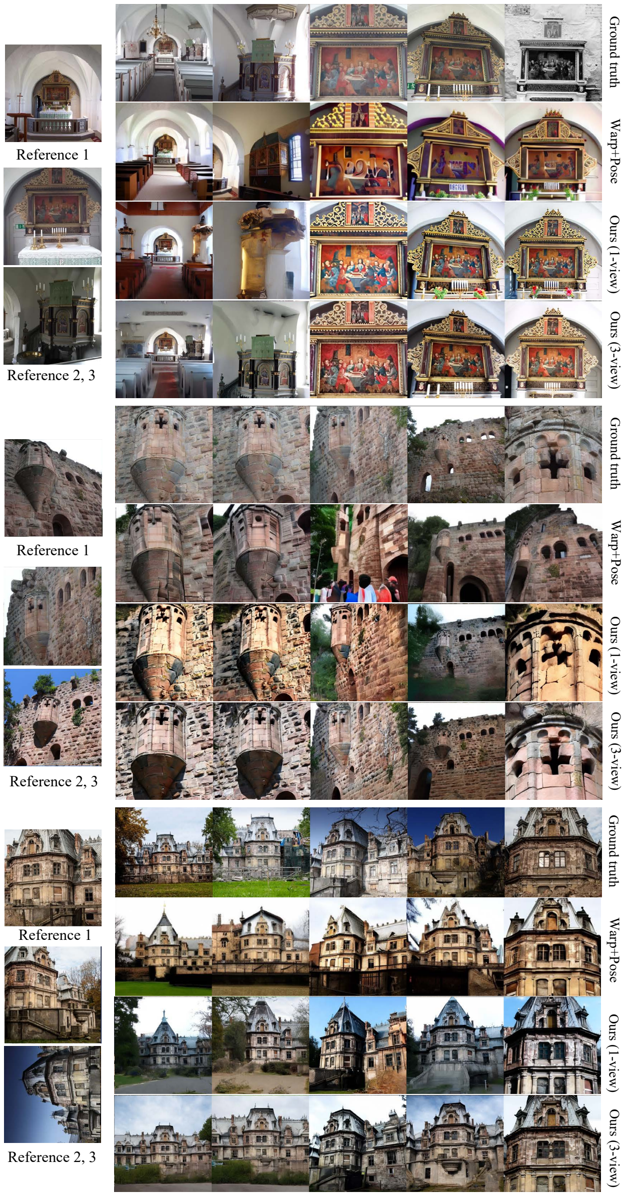


Figure 4. **Qualitative NVS results compared with the Warp+Pose baseline in [18] on Megascenes.**

Neural Information Processing Systems, 37:75468–75494, 2024. 2, 3

[7] Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems*, 36:70847–70860, 2023. 3

[8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[9] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. In *International Conference on Learning Representations*, 2024. 1

[10] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 1

[11] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021. 1

[12] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024. 2

[13] 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. 2

[14] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 1

[15] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 4

[16] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1

[17] Tianwei Shen, Zixin Luo, Lei Zhou, Runze Zhang, Siyu Zhu, Tian Fang, and Long Quan. Matchable image retrieval by learning from surface reconstruction. In *The Asian Conference on Computer Vision (ACCV*, 2018. 1

[18] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *European conference on computer vision*, 2024. 1, 4, 5

[19] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2

[20] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21551–21561, 2024. 3

[21] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. In *International Conference on Learning Representations*, 2024. 2

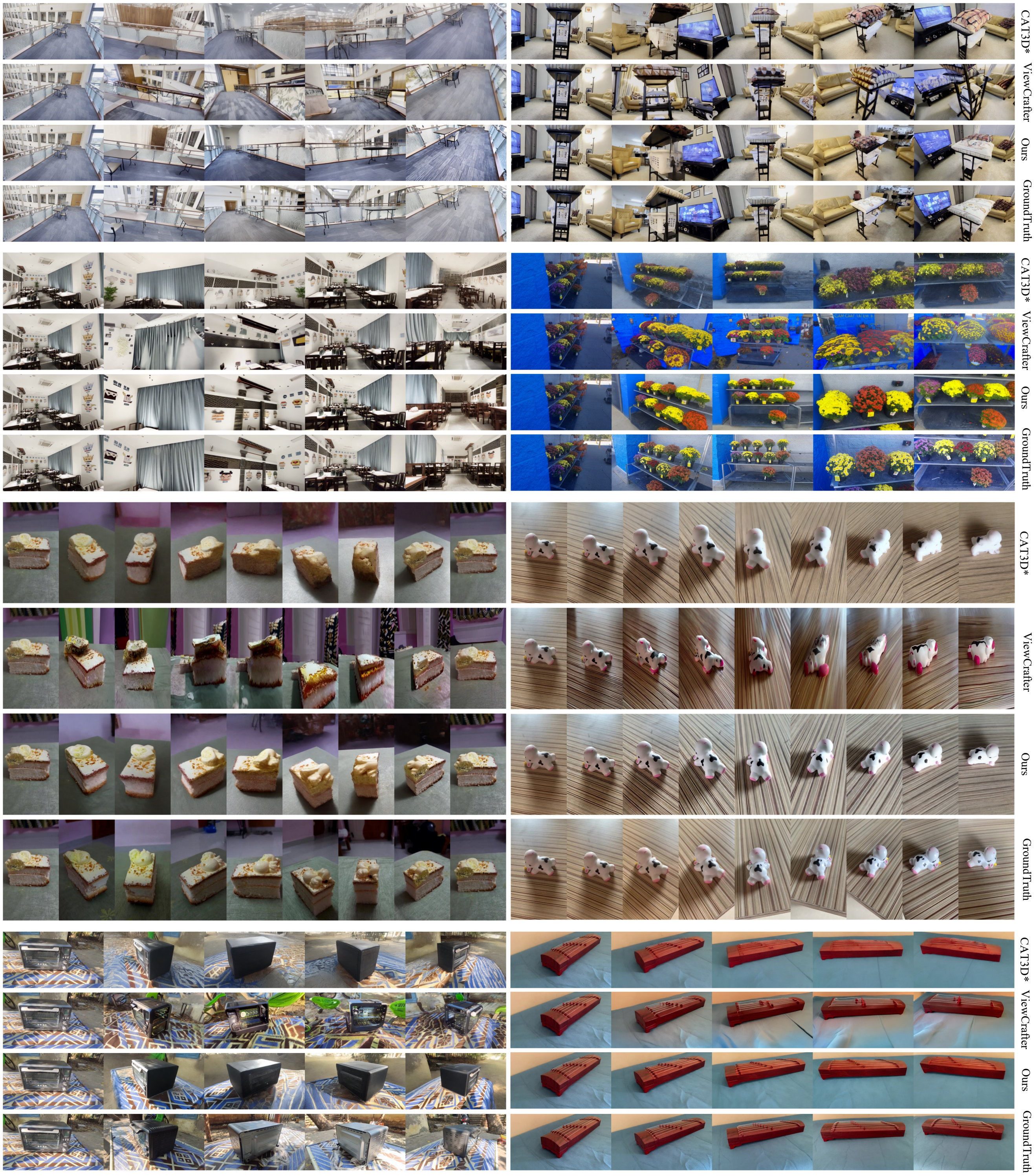
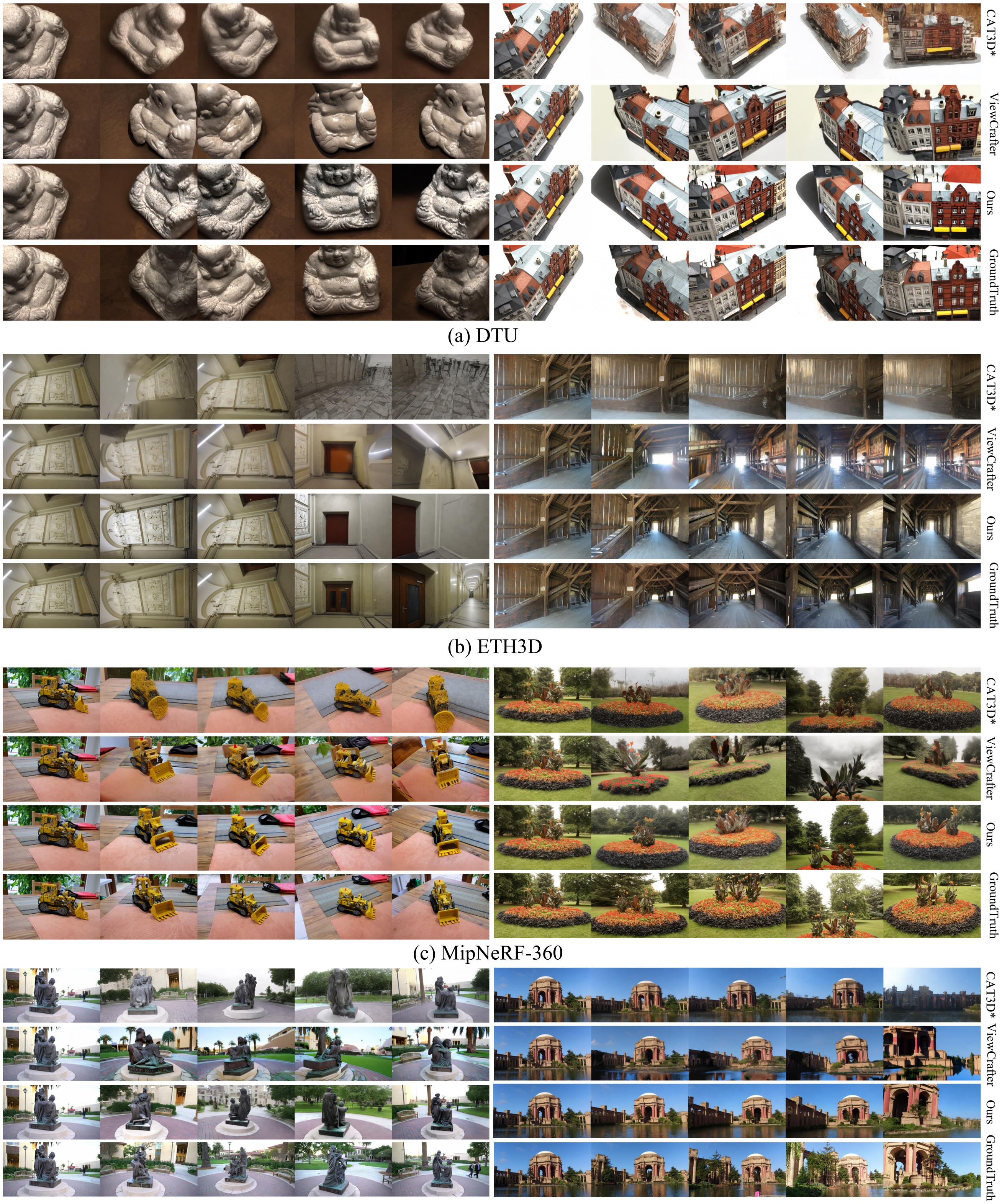


Figure 5. **Qualitative NVS results compared among CAT3D*, ViewCrafter, and our MVGenMaster from DL3DV, MVImgNet, and Co3Dv2.** The synthesis is based on ($N = 1$) reference view and ($M = 24$) target views. The leftmost column displays the reference view, while the remaining visualizations are uniformly sampled from the 24-frame generation due to page limitation.

[22] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor



(a) DTU

(b) ETH3D

(c) MipNeRF-360

(d) Tanks-and-Temples

Figure 6. **Qualitative NVS results compared among CAT3D*, ViewCrafter, and our MVGenMaster from zero-shot datasets.** The synthesis is based on ($N = 1$) reference view and ($M = 24$) target views. The leftmost column displays the reference view, while the remaining visualizations are uniformly sampled from the 24-frame generation due to page limitation.

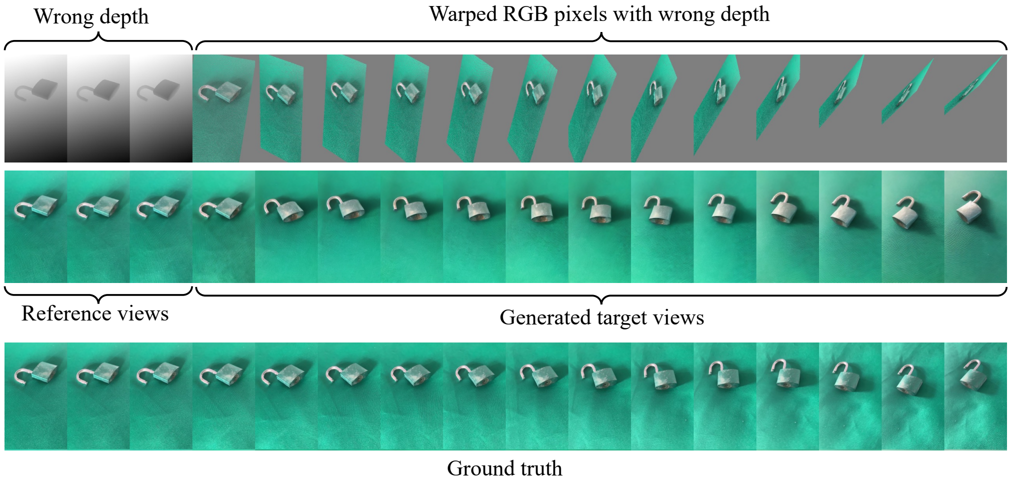


Figure 7. **Visualization of NVS results with inaccurate metric depth.** Our method still shows good robustness.

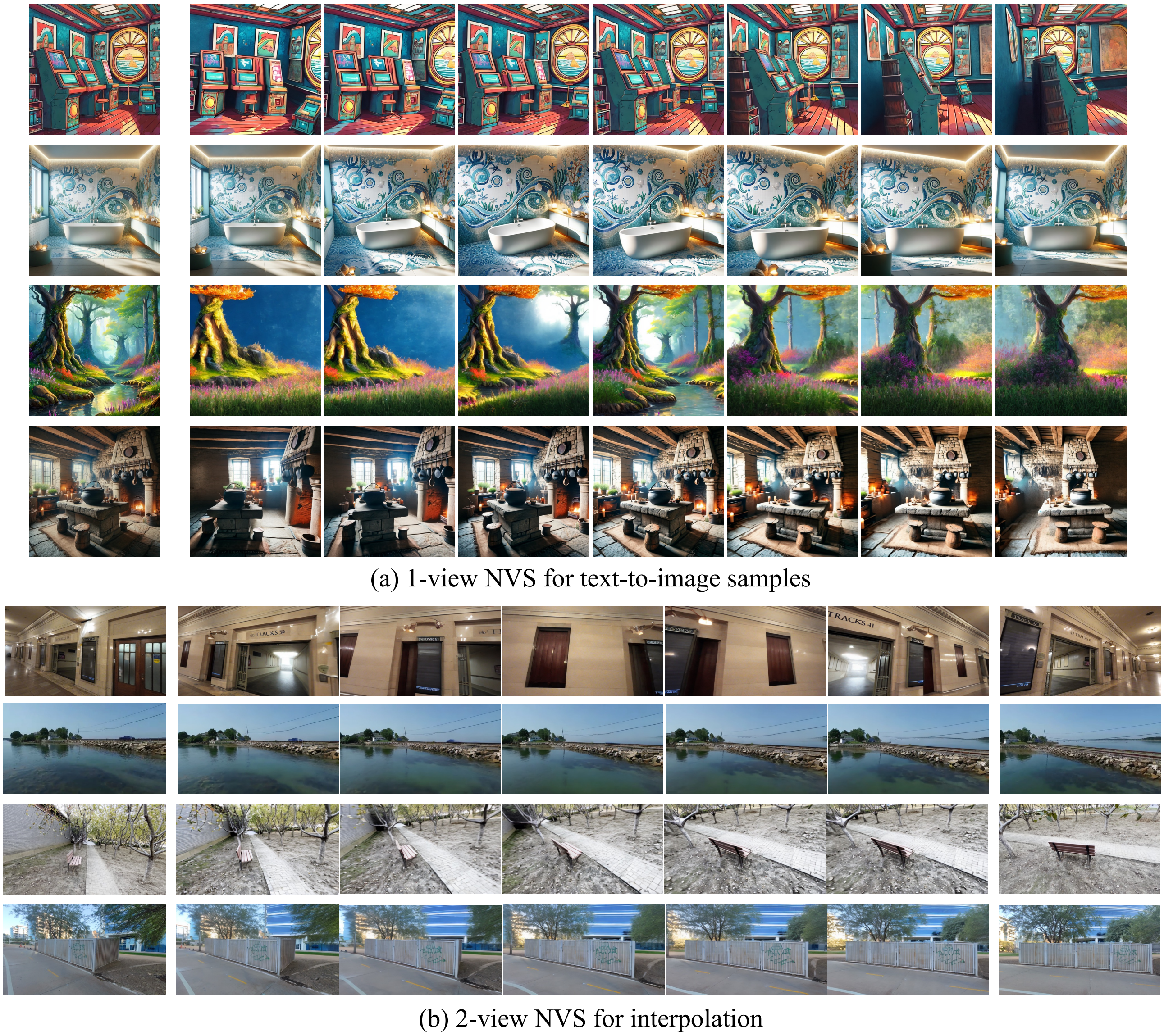

(a) 1-view NVS for text-to-image samples
(b) 2-view NVS for interpolation

Figure 8. **More results of MVGenMaster based on 1-view NVS and 2-view interpolation.**

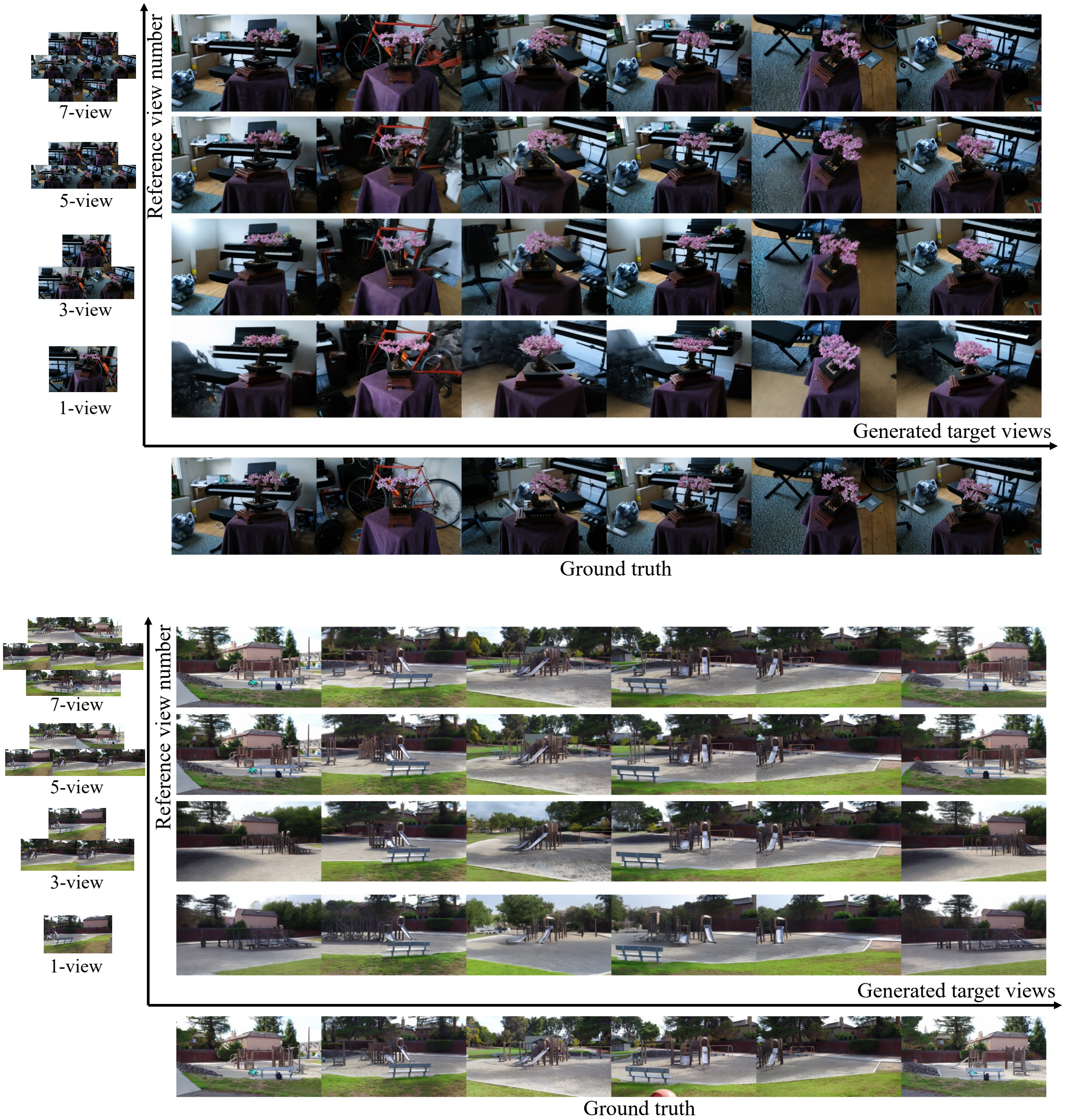


Figure 9. **NVS results with varying reference views.**

scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 1

[23] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 2

[24] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 1

[25] 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*, pages 586–595, 2018. 2

[26] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4), 2018. 1