NVComposer: Boosting Generative Novel View Synthesis with Multiple Sparse and Unposed Images

Supplementary Material

A. Implementation Details

Training-time Data Augmentation. To improve robustness against low-quality or blurry images, we incorporate random degradation into the training process of NVComposer. For each input sample, there is a 10% chance of applying Gaussian blur with a kernel size of 7, and a separate 10% chance of performing bilinear downsampling by a factor of 2, followed by upsampling by the same factor. This augmentation strategy helps NVComposer produce clear and sharp novel views, even when the input images are of low quality.

Camera Trajectory Normalization. In the main paper, we describe a method for normalizing camera trajectories by designating the first view in both the target and condition segments as the anchor view. This anchor view is assigned an identity transformation for its camera pose, represented by the camera-to-world matrix. This normalization process is applied to pose annotations across multi-view sequences. For camera translations, which can differ across datasets and lack a consistent physical reference, we implement a scaling normalization technique. This approach addresses variations in translation scales between different scenes. Specifically, the translation vector for each view, τ_t, is normalized as:

$$\tau_t := \tau_t / \big[max(\|\tau_1\|_2, ..., \|\tau_T\|_2) + 1\big], \tag{4}$$

where τ_t is the translation vector for the t-th view. The added constant 1 ensures that the relative scale differences across samples are preserved while confining the translation scale of all samples to the interval $[0, 1)$. This approach empirically improves the model's ability to adapt to varying translation scales across datasets.

Dataset Details. We train and evaluate our model using a mixed dataset comprising RealEstate10K [53], DL3DV [21], CO3D [27], and Objaverse [5]. The number of samples used from each sub-dataset is provided in Tab. A1. For Objaverse, we manually curate a subset of 90,000 high-quality objects for training. We disable the geometry-aware feature alignment for samples from Objaverse to avoid interference, as the DUSt3R shows instability in rendered images with pure-white backgrounds. From DL3DV, we utilize samples from the 1K to 7K subsets (a total of 7,000 samples). During training, two frame sequences are randomly selected from each sample (representing a scene) as target and condition views, following the sampling rules outlined in Sec. 4.1 of the main paper.

Sub-Dataset	Number of Samples
RealEstate10K [53]	62,992
DL3DV [21]	7,000
CO3D [27]	34,474
Objaverse [5]	90,000
Total	194,466

Table A1. Number of training samples used from each sub-dataset in our experiments, along with the overall total.

Weight Initialization. The weights of the dual-stream diffusion model are initialized as follows:

- For inherited structures, such as the original residual blocks and spatial-temporal attention layers, we initialize weights using those from the pretrained video diffusion model [44].
- For newly introduced layers, including input/output convolution layers, additional spatial-temporal attention layers, and the pose decoding head, weights are initialized using a bounded uniform distribution [11]. Additionally, the final sub-layer of each new component is initialized to zero. This strategy ensures compatibility with the original model's inference behavior while preventing instability during initialization, even with structural modifications.

Classifier-Free Guidance Settings. During training, we randomly mask the image CLIP features, target poses, and condition images with a probability of 10%. We observe that a proper classifier-free guidance scale enhances textures in generated views. At inference time, we empirically set the guidance scale to 3.0, balancing diversity and fidelity in the generated results.

B. Additional Visualization Results

Feature Correspondence Visualization. To enhance our understanding of the dual-stream diffusion model's effectiveness in NVComposer for managing multiple unposed condition images, we visualize the spatial-temporal feature correspondences within the model. Specifically, we focus on features extracted from the sixth layer during the decoding stage of the U-Net. Utilizing the method proposed in

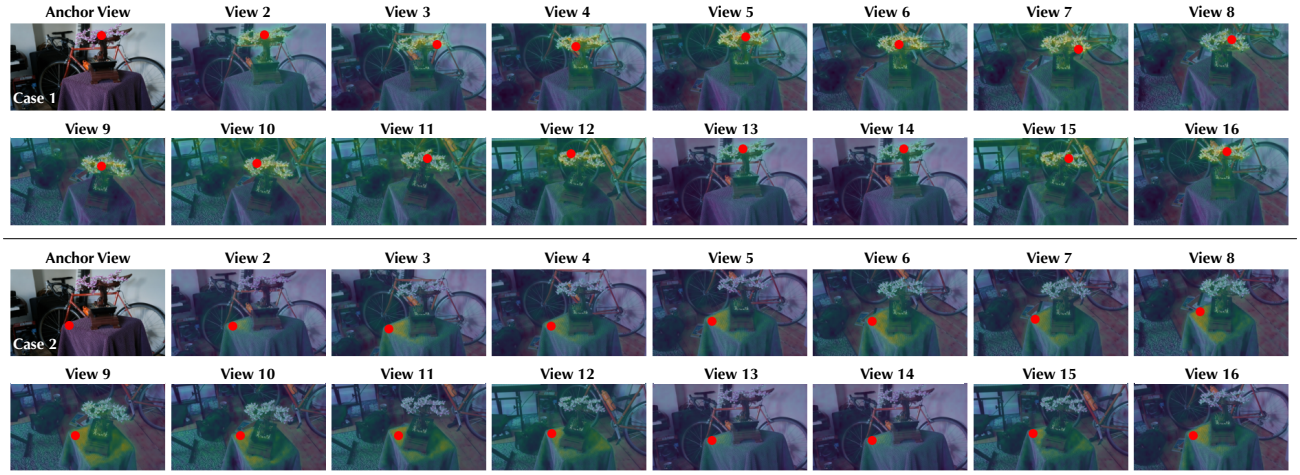

Figure B1. The cross-view feature correspondence using features extracted from the sixth layer of the U-Net decoder in the dual-stream diffusion model. The similarity map is computed between the red point in the anchor view and all spatial locations in the other views, visualized as heatmaps. The points with the highest correspondence are also marked with red dots. Please zoom in for a clearer view of the details.

DIFT [33], we infer the diffusion model at timestep 10, using it as a feature extractor. We calculate the cosine similarity between a feature at a red-marked location in the anchor view and all spatial locations in other views. The resulting heatmaps are visualized, with the highest similarity points marked by red dots.

As illustrated in Fig. B1, our model exhibits strong cross-view feature correspondence. In Case 1, where we analyze the feature on the flower in the anchor view, corresponding features in other views show a high response near the flower. In Case 2, the anchor feature is taken from the right corner of a table, and the corresponding table region in all other views displays a strong response. These visualizations underscore the model's internal understanding of spatial relationships, which is crucial for generative NVS tasks.

Additional Novel View Synthesis Results. To further support the results presented in Tab. 1 of the main paper, we provide additional visualizations in Fig. B2. For clearer comparison, we include zoomed-in patches of the generated novel views alongside the corresponding ground truth references.

As illustrated in Fig. B2, our method consistently produces results that are visually closer to the real references. In contrast, DUSt3R struggles with filling missing regions, as it merely stitches the input condition images without generating plausible completions. ViewCrafter [49], which relies on DUSt3R's pre-reconstruction, inherits artifacts and fails to achieve seamless outputs. MotionCtrl [40] and CameraCtrl [10], designed for controllable video generation, exhibit poor camera control accuracy, often rendering views from incorrect camera positions. Please refer to the supplementary video for more results.

3D Reconstruction from Generated Views. The ability of NVComposer to generate novel views from unposed conditional images makes it an ideal candidate for 3D generation applications. To illustrate this capability, we employ the NeRF version of InstantMesh [47], a multi-view large reconstruction model, to create NeRF representations based on the multi-view images and poses generated by NVComposer, without utilizing its generation pipeline. The target poses used for reconstruction are identical to those required for generating views with NVComposer. For comparison, we also generate 3D objects using multi-view outputs from SV3D [36].

Figure B3 showcases the reconstruction results of our method alongside those from SV3D [36]. The results demonstrate that NVComposer produces 3D-consistent novel views that enable high-quality 3D reconstructions. Furthermore, our model effectively leverages information from randomly unposed views, producing outputs that more closely resemble the real reference. Additional results are available in the supplementary video.

C. Discussion

Scalability in Number of Views. Our model is trained with a random selection of 1 to 4 views as conditions. Adjusting the training settings allows for training with more condition views. Additionally, extending the temporal length of the model also enables support for more target and condition views.

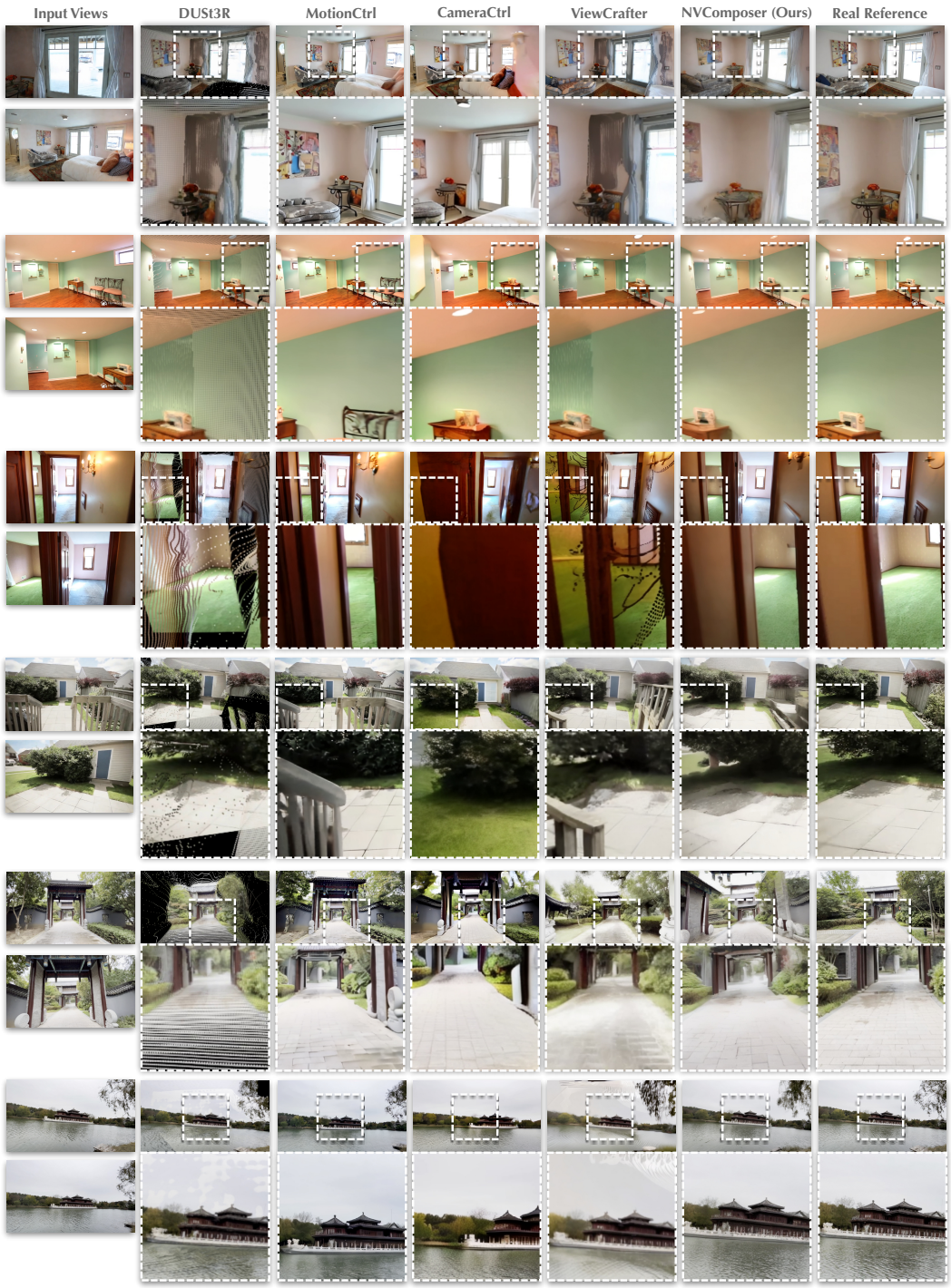

Figure B2. More visualization of generative NVS results on RealEstate10K [53] and DL3DV [21] test set. For a better view, we zoom in on patches for each method (in white square boxes). Dynamic comparison can be found in the supplementary video.

Limitation. As our model leverages the video priors, failure cases may occur when target camera poses are too discrete. In addition, as a diffusion-based model, NVComposer generates content in an iterative denoising process. This iterative nature results in extended generation times, often spanning several minutes. Such delays can limit its applicability in real-time contexts like interactive applications and live simulations, where rapid response is essential.

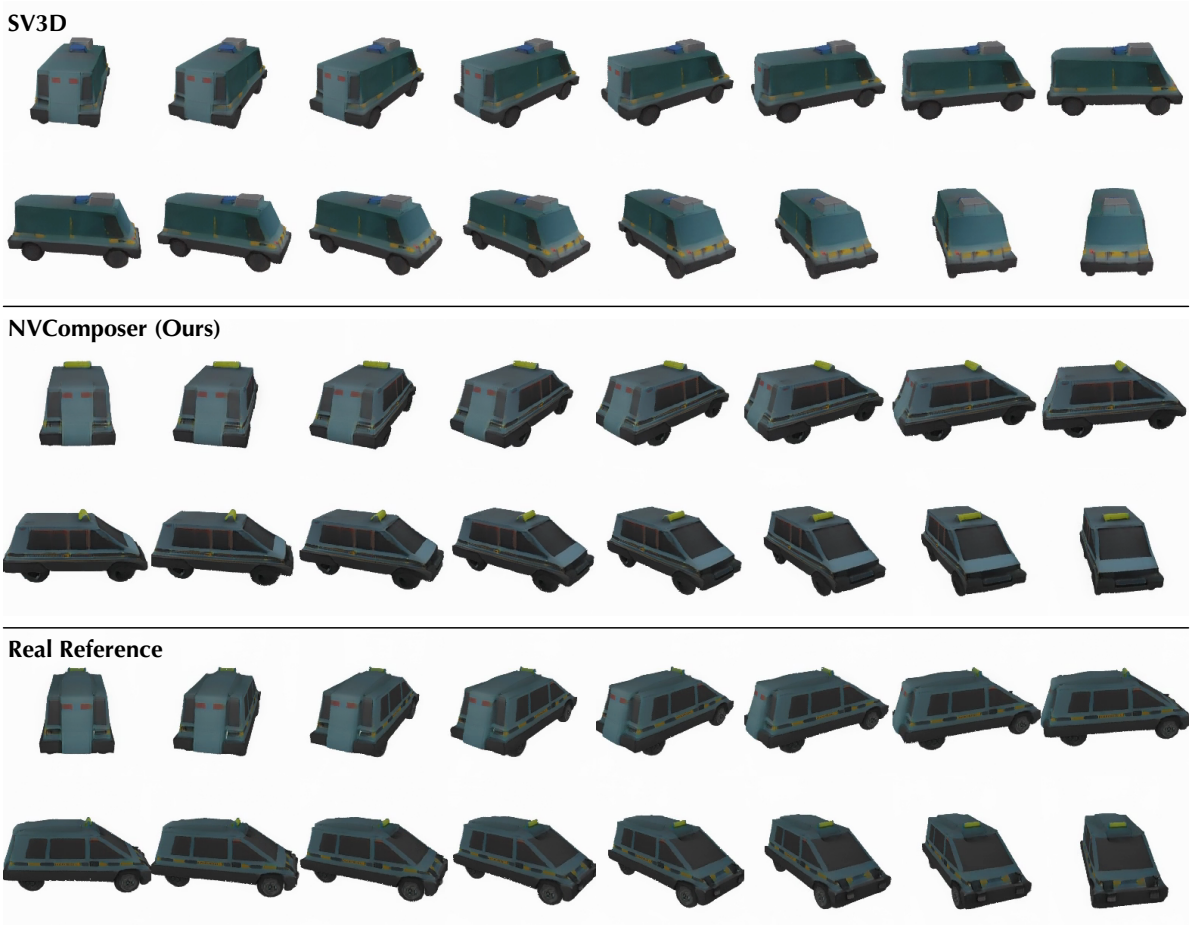

Figure B3. Results of 3D reconstruction on generated views from SV3D [36] and NVComposer. We utilize the large reconstruction model from InstantMesh [47] (without additional generation) to reconstruct NeRF representations from the multi-view outputs of SV3D and NVComposer. For animated results, please refer to the supplementary video.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 3

[2] 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. *arXiv preprint arXiv:2311.15127*, 2023. 3

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[4] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023. 2

[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. 2, 5, 7, 8, 1

[6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 5

[7] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 3

[8] 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. 1, 3, 4

[9] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[10] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 3, 5, 6, 7, 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 1

[12] 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*, 30, 2017. 6

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

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3

[15] 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. *arXiv preprint arXiv:2106.09685*, 2021. 3

[16] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10026–10038, 2024. 2

[17] 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. 1

[18] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *Advances in Neural Information Processing Systems*, 37:16240–16271, 2024. 3

[19] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 1

[20] 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*, pages 19730–19742. PMLR, 2023. 4

[21] 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. 2, 5, 6, 7, 1, 3

[22] 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. 1, 2, 3, 4

[23] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 1, 2, 3

[24] 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*, pages 9970–9980, 2024. 1, 2, 3

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[26] 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*, pages 8748–8763. PMLR, 2021. 4, 5

[27] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 2, 5, 1

[28] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14104–14113, 2021. 1, 2

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[30] 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

[31] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34: 19313–19325, 2021. 3

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[33] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 2

[34] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 2

[35] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[36] 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. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 2, 3, 6, 7, 4

[37] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2021. 2

[38] 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. 1, 2, 3, 4, 5, 6, 8

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[40] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3, 5, 6, 7, 2

[41] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7467–7477, 2020. 2

[42] Chin-Hsuan Wu, Yen-Chun Chen, Bolivar Solarte, Lu Yuan, and Min Sun. ifusion: Inverting diffusion for pose-free reconstruction from sparse views. *arXiv preprint arXiv:2312.17250*, 2023. 3

[43] 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. 1

[44] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 3, 4, 1

[45] 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. In *European Conference on Computer Vision*, pages 143–163. Springer, 2025. 1, 2

[46] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 3

[47] 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, 4

[48] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 1

[49] 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. 1, 2, 3, 4, 5, 6

[50] David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E. Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. *arXiv preprint arXiv:2411.05003*, 2024. 3

[51] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024. 8

[52] 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. 5

[53] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph*, 37, 2018. 2, 5, 6, 7, 8, 1, 3