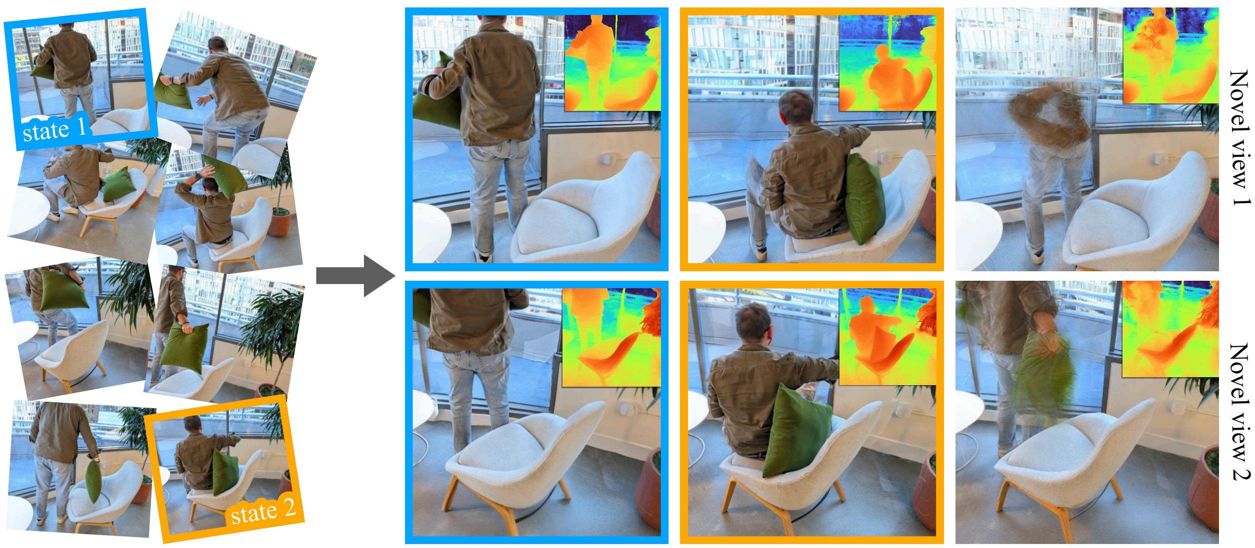
SimVS: Simulating World Inconsistencies for Robust View Synthesis

Alex Trevithick[1,2], Roni Paiss[2], Philipp Henzler[3], Dor Verbin[2], Rundi Wu[2,4], Hadi Alzayer[2,5], Ruiqi Gao[2], Ben Poole[2], Jonathan T. Barron[2], Aleksander Holynski[2], Ravi Ramamoorthi[1], Pratul P. Srinivasan[2]

[1]UC San Diego [2]Google DeepMind [3]Google Research [4]Columbia [5]Univ. Maryland



(a) 8 casually-captured posed images with scene motion (b) Our model frozen to state 1 (c) Our model frozen to state 2 (d) CAT3D (all images)

Figure 1. We show results of our model applied to a casual in-the-wild capture. (a) Given 8 unordered images of a scene with significant motion and desired states marked in blue and orange, our model generates a 3D representation for each desired state shown in corresponding colors in (b) and (c). The CAT3D baseline [15] in (d) cannot disentangle the different states, resulting in catastrophic failure.

Abstract

Novel-view synthesis techniques achieve impressive results for static scenes but struggle when faced with the inconsistencies inherent to casual capture settings: varying illumination, scene motion, and other unintended effects that are difficult to model explicitly. We present an approach for leveraging generative video models to simulate the inconsistencies in the world that can occur during capture. We use this process, along with existing multi-view datasets, to create synthetic data for training a multi-view harmonization network that is able to reconcile inconsistent observations into a consistent 3D scene. We demonstrate that our world-simulation strategy significantly outperforms traditional augmentation methods in handling real-world scene variations, thereby enabling highly accurate static 3D reconstructions in the presence of a variety of challenging inconsistencies. Project page: `https://alextrevithick.com/simvs`

1. Introduction

View synthesis, the task of creating images from unobserved camera viewpoints given a set of posed images, has seen remarkable progress in recent years. Current algorithms are able to render detailed photorealistic novel views of complicated 3D scenes. However, these techniques tend to assume that the provided input images are *consistent* — that the geometry and illumination of the scene is static during capture. Typical captures of real-world scenes seldom obey this constraint; people and objects may move and deform, and lights may move or change brightness.

Moreover, casual captures outside of tightly-controlled settings tend not only to be inconsistent but also *sparse*, containing only a small number of observed views. Methods for sparse view synthesis are usually trained on synthetic or captured multiview datasets that are consistent by design, and therefore fail to generalize to the inconsistencies seen in real-world casual captures (see Fig. 6 as an example).

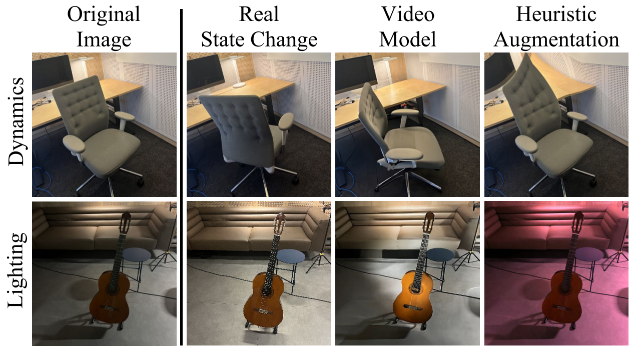


Figure 2. A comparison of real world state changes, those simulated through a video model, and heuristic augmentations (random sparse flow fields for dynamics and random color tints for lighting).

We address the problem of robust view synthesis from sparse captures in a new way by leveraging the ability of video diffusion models to simulate plausible world inconsistencies that could arise during capture. There has been considerable speculation about the usefulness of large video models as simulators or "world models" [29, 45] and in this work we demonstrate a new use case for their simulation capabilities.

Our approach augments existing multiview datasets with inconsistencies simulated by a video diffusion model and trains a multiview harmonization model to sample sets of consistent views of a scene conditioned on sparse inconsistent captures. We can then use existing 3D reconstruction and view synthesis techniques to synthesize novel viewpoints from these consistent images.

In summary, our key technical contributions are:

- A generative data augmentation strategy that leverages video diffusion models to sample world inconsistencies (*e.g*. scene motion and lighting changes) that could arise during capture (Section 3)
- A multiview harmonization model, trained on this generated data, that converts inconsistent sparse input images into a set of consistent images (Section 4)

We demonstrate that our *generative augmentation* strategy outperforms other alternatives such as using heuristic data augmentation or synthetic rendered data, and that novel views rendered from our harmonization model are superior to those from existing approaches for sparse and robust view synthesis. We encourage readers to view the video results on the project page.

2. Related Work

We address the task of view synthesis from sparse *and* inconsistent images of a scene. While existing techniques address view synthesis from densely-sampled inconsistent inputs or sparse consistent inputs, to our knowledge no existing method is capable of synthesizing novel views of full scenes from images that are both sparse and inconsistent.

2.1. Robust view synthesis

Prior methods for robust view synthesis typically require dense captures with hundreds of images. and focus on explicitly modeling a specific source of inconsistency (either motion or lighting) as part of reconstruction.

Scene dynamics In the case of scene dynamics, some works handle only transient occluders in dense, multiview captures [39–41]. Most existing methods start with a dense video and attempt to recover motion flows or trajectories to explain the observed motion. Early approaches based on Neural Radiance Fields [34] optimized time-varying flow fields to explain observed motion as deformations of an underlying consistent scene representation [12, 24, 35, 36, 38, 53]. Later NeRF-based methods improved quality further through prior integration [25, 30]. The most recent state-of-the-art methods have adopted 3D Gaussian representations and optimized explicit motion trajectories for this particle-based scene representation [22, 50, 56], often leveraging strong priors from pretrained monocular depth [13, 20, 64], optical flow [52], or tracking models [10, 19, 62]. These temporal priors have been crucial for rendering high-quality novel views in the dense capture setting, but tend to break down when applied to sparse or unordered captures.

Lighting inconsistencies Existing structure-from-motion pipelines display remarkable robustness to lighting variation [42–44], enabling 3D reconstruction from large-scale in-the-wild images [1]. To model inconsistencies due to changing scene lighting, 3D reconstruction and view synthesis techniques use per-image "appearance embeddings" that allow for the appearance of scene content to vary across observations [8, 21, 32, 33, 58, 63, 67]. This strategy can successfully model lighting inconsistencies given dense captures with smoothly-varying appearance changes, but is unable to reconcile large changes in appearance in sparsely-sampled captures. Some methods [37, 69] assume consistent captures and generate inconsistencies with generative models for relighting. Others [17, 66] can perform object-level relighting from one image, but cannot incorporate information from a sparse set of views.

2.2. Sparse view synthesis

In novel view synthesis settings with only a few captured views, most methods rely on strong priors learned from large multiview datasets. Some methods train feedforward models to directly predict 3D representations that can be used for view synthesis [7, 16, 18, 54, 65, 71]. Others rely on pretrained monocular depth, multiview stereo, or inpainting networks and rely on test-time optimization to fit a scene [11, 47, 48, 55, 59]. A recent class of methods has achieved high visual quality by directly generating images from novel viewpoints using diffusion models conditioned on observed image(s) and target camera poses [15, 27, 28, 31, 46, 61]. In particular, the multiview

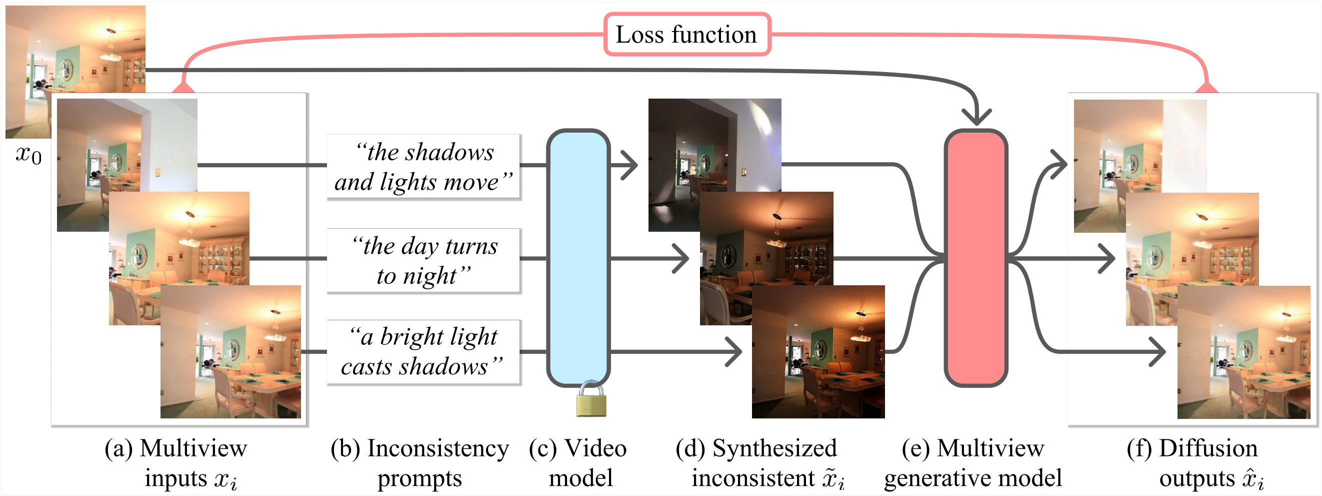


Figure 3. Our method's overall pipeline. (a) Given a dataset of multiview images x_i, we simulate inconsistencies by (b) prompting a (c) video model and then (d) selecting inconsistent frames $\tilde{x}_i$. We feed these images along with a held-out reference image x_0 under the original condition to a (e) multiview generative model to predict (f) a set of corresponding consistent outputs $\hat{x}_i$. This output is supervised by the original multiview images x_i.

diffusion model CAT3D [15] has emerged as the state-of-the-art in view synthesis from sparse image inputs. However, as they are trained only on consistent multiview image sets, CAT3D and other sparse view synthesis techniques are not robust to the inconsistencies observed in real-world casual captures. Concurrent work CAT4D [60] extends CAT3D with a temporal axis. Among other training data, it leverages the generative augmentation strategy proposed in this paper.

3. Simulating World Inconsistencies with Video Models

Training a robust view synthesis model is challenging due to the lack of paired training data of inconsistent captures and target consistent images. Most existing multiview datasets only contain captures of *consistent* scenes, so simply scaling such data is not sufficient for robust view synthesis. Gathering images from multiple viewpoints, each under multiple scene deformations or lighting settings, would be extremely onerous. Heuristic data augmentation strategies such as random transformations, tints, and sparse flow fields cannot adequately capture the diversity and 3D nature of scene motions and lighting changes, as displayed in Fig. 2. Conversely, synthetic datasets like Objaverse [9] only contain simple object-level motion and fail to enable generalization to real-world scenes.

The key idea in our work is to leverage generative video models to create a robust view synthesis dataset from existing consistent multiview image datasets. For each 3D scene, we desire a dataset that contains (1) a set of consistent multiview images x_i, (2) inconsistent images $\tilde{x}_i$ where the scene has undergone some transformation such as a deformation or lighting change, and (3) camera poses π_i for each image.

3.1. Video model augmentation

We propose to generate a realistic and diverse dataset of inconsistent conditioning images by simulating dynamic motion and lighting inconsistencies with pretrained image-to-video generative models. Starting with a multiview capture (taken from existing large-scale multiview image datasets), we first generate, for each view, a video from a static camera with simulated scene changes (motion or lighting). By sampling frames from these videos, we can obtain inconsistent observations for each captured viewpoint. Other image editing approaches such as InstructPix2Pix [6] could potentially be used to perform this inconsistency transformation, but these methods often fail to produce substantial variation in the layout of the image which are needed to simulate dynamic inconsistency.

To generate videos with simulated scene changes, we use an image- and text-conditioned video diffusion model that samples from $p(V|I, c)$, where V is a video, I is a conditioning frame that V should include, and c is a text caption. By setting I to an image from a multiview capture x_i and choosing c, we may simulate inconsistencies on top of the image. Note that the video must not contain camera motion in order to preserve the accuracy of existing camera parameters.

We simulate the two most prominent inconsistencies: dynamic motion and lighting changes. For dynamics, we use the Mannequin Challenge dataset [23]. This dataset is a natural choice as it includes static multiview captures of scenes with content that would typically be dynamic in casual captures. For our lighting-robust model, we simulate lighting changes on the RealEstate10k [70] dataset, which contains scenes in diverse indoor and outdoor illumination conditions.

We generate the captions c with a multimodal large lan-

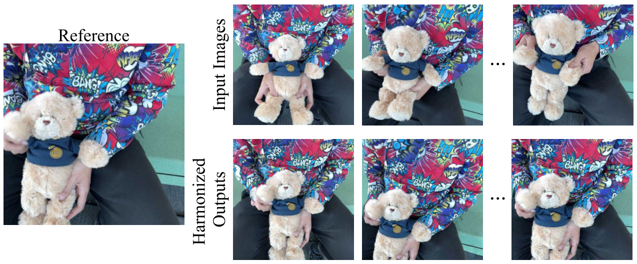


Figure 4. Samples from our multiview diffusion harmonization model, visualized for scene dynamics. Given the reference image and inconsistent input image, our model directly generates multiview images consistent with the state of the reference.

guage model, Gemini [51]. For each clip in the dataset, we randomly choose a representative frame x_i. We prompt Gemini with this frame and a meta-prompt m, designed to elicit simple but specific prompts, e.g., "the woman swings the pillow" or "the two children dance." We also ensure the generated prompts are sufficiently specific and concise through m (see the supplement for the meta-prompt in its entirety). We generate the complete inconsistency prompt as:

$$c = \text{“static shot, ”} \oplus \text{Gemini}(x_i, m)\,, \tag{1}$$

where $\oplus$ denotes string concatenation. Since the inconsistencies observed in casual captures of dynamic content are highly correlated across views, we use the same inconsistency prompt for all frames in the corresponding clip.

We find that incorporating a negative prompt [2] c_{negative} is extremely important for generating the desired inconsistencies without changing the camera viewpoint. We include phrases such as "panning view" and "orbit shot" in c_{negative}.

Given a multiview image x_i and generated inconsistency prompt c, we sample a video:

$$V = \text{VideoModel}(x_i, c, c_{\text{negative}})\,. \tag{2}$$

We assume all frames in the sampled video are inconsistent with respect to the original image x_i. Therefore, we get a set of T inconsistent frames corresponding to x_i for each sample from the video model, where T is the number of frames in the output video. At training time, we randomly sample one of the T frames as inconsistent conditioning for a given x_i. In our experiments, we generate 640 total "inconsistent" frames per multiview capture, giving us a total of about 6 million frames for the dynamics dataset and about 12 million frames for the lighting dataset. Figs. 2 and 3 visualize example frames from our synthesized videos.

3.2. Video model details

For all experiments in this paper, we use Lumiere [3], a pixel-space video diffusion model which operates in two-stages for high-resolution generation. We find that the Lumiere model struggles to generate lighting changes when given non-generic prompts, so for our lighting-robust model, we sample c uniformly from a set of predetermined lighting prompts found to generate large lighting variations instead of using a large language model. Please refer to Fig. 3 and the supplement for example prompts.

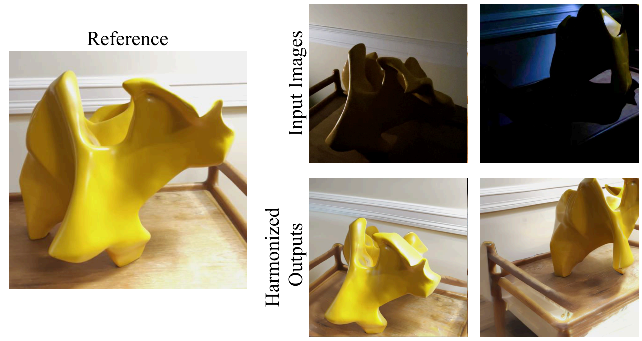


Figure 5. Samples from our multiview diffusion harmonization model, visualized for lighting. Given the reference image and inconsistent input image, our model directly generates multiview images consistent with the state of the reference.

4. Harmonization through multiview diffusion

We use our multiview simulated world inconsistencies dataset $(x, \tilde{x}, \pi)$ to learn a generative model that can map from sparse inconsistent captures to a consistent set of images, as displayed in Figs. 4 and 5. We call this model a "harmonization" model as it brings the inconsistent input images into harmony.

4.1. Architecture

We build our harmonization model on top of CAT3D [15], a latent multiview diffusion model that directly predicts target images conditioned on posed input images and target camera poses. To incorporate inconsistent observed images as conditioning, we simply concatenate latents of the inconsistent images $\tilde{z}_i = \mathcal{E}(\tilde{x}_i)$, encoded by the VAE encoder $\mathcal{E}$, to the target raymaps and noisy latents. Additionally, we concatenate a binary image mask (either all ones or all zeros) to each input to denote the reference image, i.e., the "desired state" with which all other outputs should be consistent.

4.2. Training

Our goal is to learn a generative model that produces consistent output image sets with N images, given a reference image latent z_0 signifying the desired scene state and $n \leq N$ observed inconsistent image latents $\tilde{z}_i$:

$$p(z_{1:N} \mid z_0, \tilde{z}_{1:n}, \pi_{0:N})\,. \tag{3}$$

Given a reference conditioning latent $\mathcal{E}(x_0) = z_0$ and up to 7 inconsistent posed latents $\tilde{z}_i$, our model predicts latents $z_{1:7}$ corresponding to the ground-truth consistent image latents. We finetune our model parameters from CAT3D [15], with additional parameters in the first layer to account for the

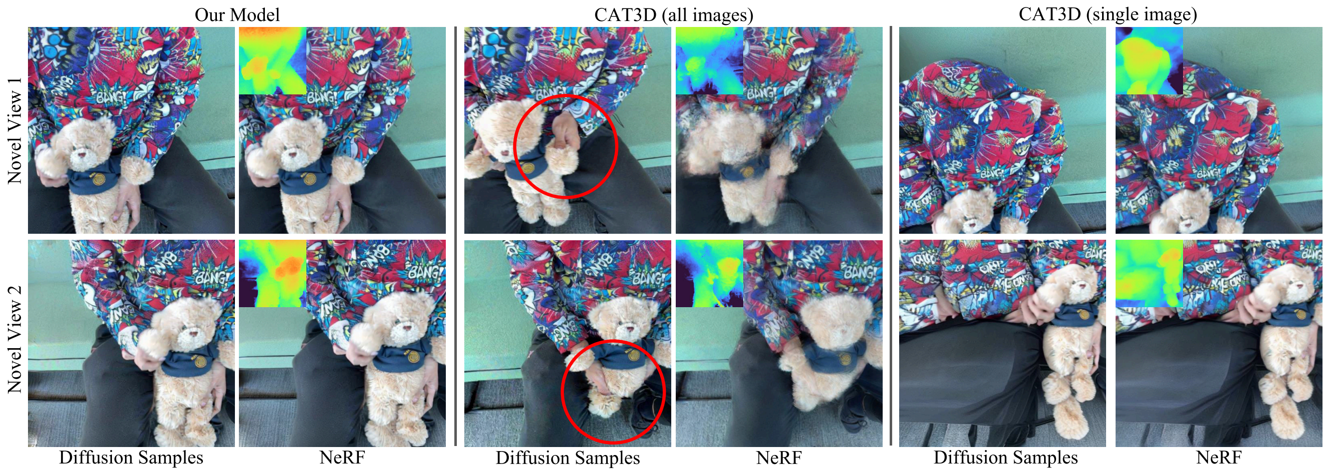


Figure 6. Given the reference and inputs in Fig. 4, we show the outputs of our model versus the CAT3D baselines. We display both the diffusion samples and learned NeRF representations with the depth maps inset. Note the 3D consistency of our samples in comparison to the changing articulation of the scene displayed by CAT3D taking all images as input. The single-image conditional CAT3D has no notion of scene scale, cannot use multiview cues to reason about the static parts, and must hallucinate all scene content outside of the input image's frustum.

additional conditioning channels. We train the model with the same diffusion loss and weighting as [15]:

$$\mathbb{E}_{t,\epsilon,z_{0:7},\tilde{z}_{1:7}}\left[w(t)\|f(\alpha_t z_{1:7}+\sigma_t\epsilon;z_0,\tilde{z}_{1:7})-z_{1:7}\|^2\right], \quad (4)$$

where f is our multiview diffusion "harmonization" model. Given noisy versions of the target consistent latents $z_{1:7}$, the target conditioning latent z_0, and the inconsistent inputs $\tilde{z}$, we aim to produce a denoised estimate output by f that is as close as possible to the consistent latents $z_{1:7}$. We additionally uniformly drop out the number of conditioning frames $\tilde{z}$ to allow the model to handle between 1 and 8 input images at test time.

4.3. 3D reconstruction

Having trained the harmonization model, we can sample consistent latents $\hat{z}_{1:7}$ and decode them into images $\hat{x}_{1:7}$ with the VAE decoder (visualized in Figs. 4, 5, and 6). We then have a total of 8 consistent images: the initial observed target x_0 and model outputs $\hat{x}_{1:7}$. While 3D reconstruction from such a small image collection is infeasible, we can use multiview diffusion models trained on consistent images such as CAT3D [15] to "densify" the sparse consistent capture into a dense consistent capture with enough views to train a NeRF. Instead of directly sampling the original 3-image conditional CAT3D model, we finetune it to condition on 5-frames, finding that the additional context improves the original 3-image conditional model in our setting.

5. Experiments

We evaluate our method for the two most common sources of inconsistency during casual multiview capture: scene dynamics and lighting changes. For all generative methods, we sample the models exactly once.

Method	PSNR↑	SSIM↑	LPIPS↓
CAT3D (single image)	14.61	0.382	0.473
CAT3D (all images)	15.59	0.448	0.462
Our Model	**16.73**	**0.463**	**0.413**

Table 1. View synthesis results on the DyCheck dataset [14] comparing our model to CAT3D taking one or all conditioning images. Our model outperforms CAT3D by all metrics.

5.1. Scene Dynamics

Dataset For scene dynamics, we evaluate our method on DyCheck [14], a dataset of 7 multiview videos where the assumed input is a monocular video with significant scene and camera motion. In this setting, we select 7 sparse frames uniformly in time as a consistent conditioning set, and uniformly select 4 target time images (top left of Fig. 4) per scene for which to compute metrics. Note that prior works which handle scene dynamics assume an ordered dense capture [22, 30, 56].

Baselines Considering this task as view synthesis from 8 inputs, we compare our performance to the state-of-the-art method for sparse view synthesis, CAT3D [15]. We evaluate variants of CAT3D which take all of the images as input CAT3D (all images), and only one of the images as input CAT3D (single image). For CAT3D (all images), we find that a finetuned model which conditions on 5 images and predicts 3 instead of conditioning on 3 and predicting 5 works slightly better for this setting. When sampling target views, we always include the reference image in the conditioning

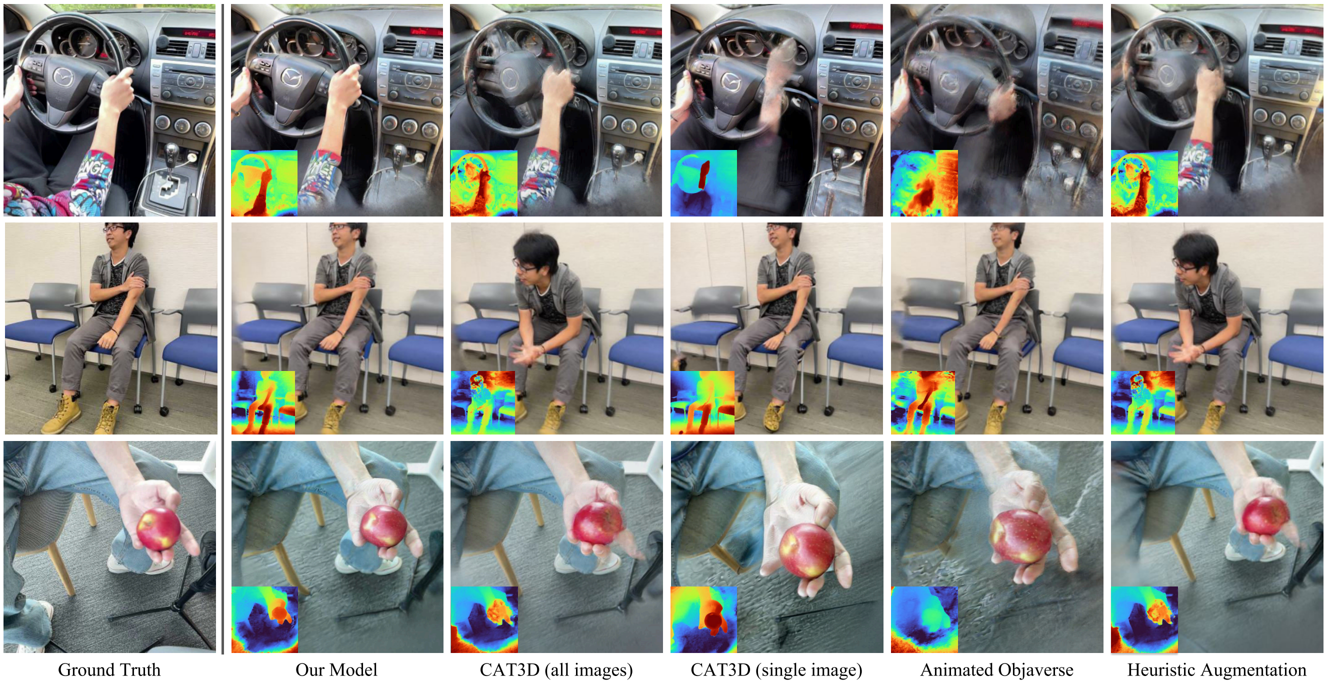


Figure 7. Qualitative results for the DyCheck [14] dataset for our model, two CAT3D baselines, and two of our ablations. The depth maps are inset on the bottom left. Images are cropped for visualization. Compared to CAT3D (all images), our method generates coherent 3D scenes despite the scene motion, while leveraging the information from multiple input views unlike CAT3D (single image). In comparison to the ablations, the quality of our approach is superior.

set, along with the 4 closest of the 7 views to the current target camera set. CAT3D (single image) simply receives only the ground truth reference image.

Due to noisy camera poses and the underdetermined nature of our task, we recompute the poses per method using COLMAP on the samples and train a Zip-NeRF [4] to evaluate novel view synthesis quality. In 4 of the 28 timesteps, COLMAP was unable to register the test images for at least one of the baselines; we discard those scenes from the calculation. Note that COLMAP never fails to register the test images for our method's results. See the supplement for a comparison to Shape of Motion [56].

Results The quantitative results shown in Tab. 1 demonstrate that we significantly outperform CAT3D across all metrics. Qualitatively, we can see in Figs. 1, 6 and 7 that CAT3D simply cannot handle inconsistencies. Their diffusion samples display high variance, typically changing state based on proximity to the input views. Training a 3D representation such as NeRF from these samples leads to undesirable averaging over all of the states and significant blur in inconsistent regions.

5.2. Lighting Changes

Dataset For lighting inconsistencies, we are not aware of an existing dataset of posed images that contains multiple illumination conditions and multiple "ground truth" images under a consistent lighting. Note that the widely-used Phototourism dataset [49] contains only one image under each lighting. Therefore, we collect a new dataset of 5 real-world scenes each captured under 3 separate lighting conditions. To construct this dataset, we take 3 monocular videos of a scene in 3 different lighting conditions, using approximately the same camera trajectory for each.

Method	PSNR↑	SSIM↑	LPIPS↓
WildGaussians	14.80*	0.494*	0.463*
CAT3D (single image)	15.06	0.526	0.552
CAT3D (all images)	18.26	0.625	0.419
Our Model	**20.98**	**0.707**	**0.357**

Table 2. View synthesis results on our lighting dataset. We compare to WildGaussians [21] and CAT3D [15] taking all images as input, along with CAT3D taking one image as input. Our model outperforms all baselines. * indicates comparison on rectified images.

Using 3 frames (one from each inconsistent video) as input, we evaluate renderings of the held-out images from one of the lighting conditions. For each collected scene, we select a target illumination condition and evaluate the method's ability to do novel-view synthesis for that illumination (from the three images). See Fig. 8 for example inputs and targets. We use Hierarchical Localization [42] with SuperGlue [43] feature matching to jointly pose all images.

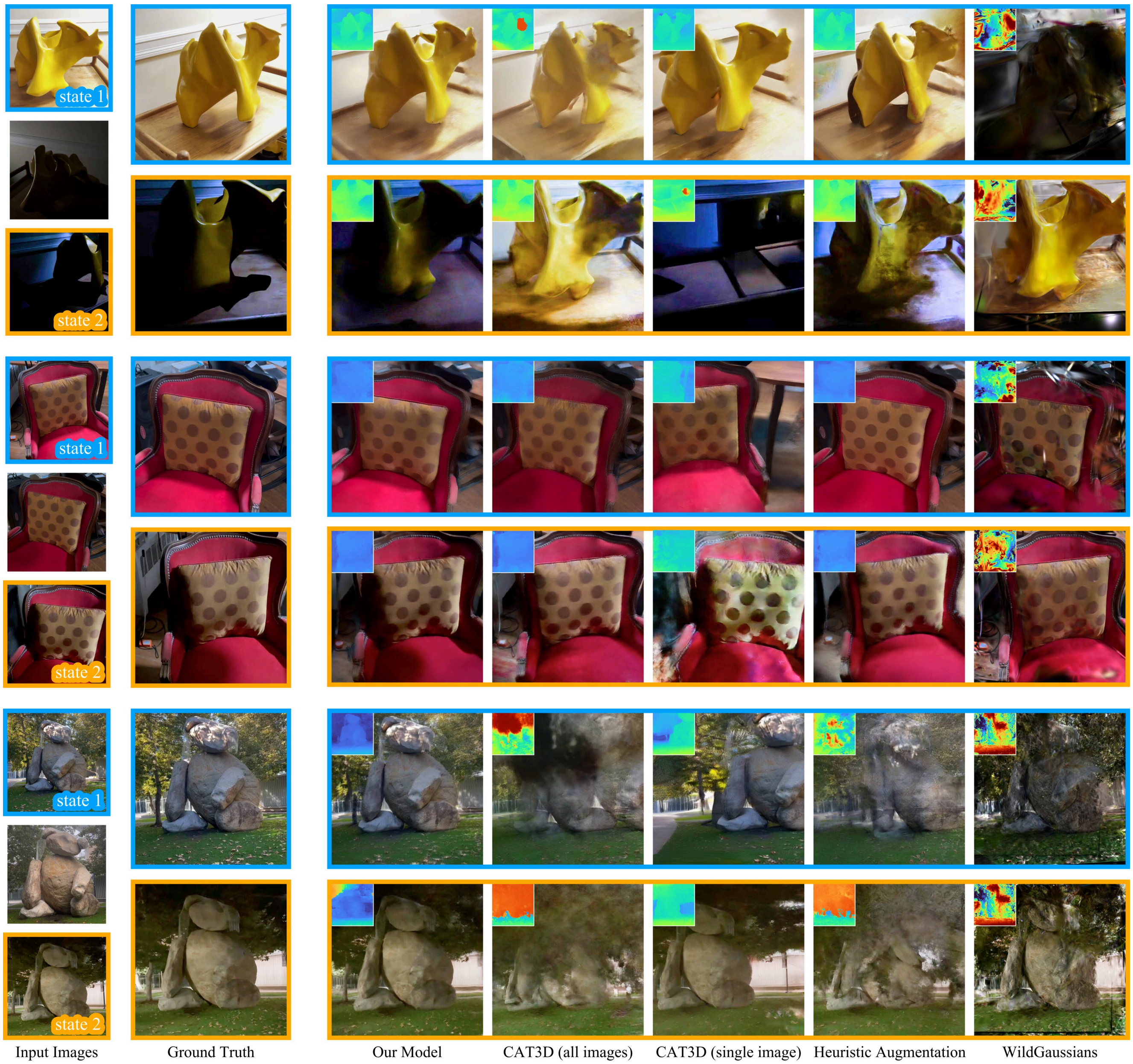


Figure 8. Qualitative results for 3 scenes from our captured lighting dataset. For each scene, we display the renders from the learned NeRFs given the 3 input images on the left. We show two states for each scene, with renders outlined in blue corresponding to the upper input image, and renders outlined in orange corresponding to the bottom input image. Although all methods are rendered with the appearance embedding of the corresponding states on the left, baselines struggle to generate plausible novel views, and often generate completely degenerate geometry. The bottom rows are brightened for visualization.

Baselines Methods which handle lighting changes typically assume a large number of captured images [32] and rely on latent embeddings [5] to parameterize per-image variations in appearance. In this line, we compare to WildGaussians [21]. For other baselines, we first generate a large number of novel views using CAT3D conditioned on the 3 inconsistent images, and then train a Zip-NeRF with latent appearance embeddings. At test-time, for all methods, we use the embedding of the reference image with the target illumination. We again evaluate against CAT3D (all images) and CAT3D (single image). Since there are only 3 input images, CAT3D (all images) is the original CAT3D model conditioned on the three input images.

Results The quantitative results in Tab. 2 show that our method significantly outperforms WildGaussians and

CAT3D in all metrics. Qualitatively, Fig. 8 displays our method's superior visual results. In some scenes, such as the stone bear shown in the bottom two rows, baselines fail to reconcile inconsistent input images into any coherent 3D scene. In other cases, the baselines reconstruct inaccurate "cloudy" scene geometry attempting to explain away changes in lighting. In contrast, our method reconciles highly disparate and sparse observations into a consistent 3D scene, allowing the generation of a high-fidelity NeRF with coherent geometry as demonstrated in the inset depth maps.

5.3. Ablations

In this section, we ablate key design decisions. Specifically, we demonstrate the importance of using our simulated inconsistency data by evaluating against heuristic augmentations and a synthetic data alternative. For dynamics, we compare to a warping-based heuristic augmentation where we apply sparse flow fields; an example can be seen in Fig. 2. The resultant model simply copies all "real-looking" pixels, indicating that such warping does not bridge the domain gap.

We also compare to generating a synthetic training dataset by animating 40k+ Objaverse assets [9, 26] with associated motions. Due to the small motion magnitude and the domain gap from object-level renderings to real scene-level data, the method significantly underperforms. The quantitative ablation results can be seen in the top of Tab. 3, where our method outperforms all ablated methodologies. For dynamics, a qualitative comparison is provided on the right of Fig. 7.

For lighting, we compare against heuristic augmentation whereby the input images are tinted inconsistently as seen in Fig. 2, and the targets images are tinted consistently. This method slightly outperforms the vanilla CAT3D as it requires the model to get the mean color correct; however, it cannot resolve lighting phenomena like shadows, nor localized changes in lighting. Results can be seen quantitatively in Tab. 3 and qualitatively on the right of Fig. 8.

	Ablation	PSNR↑	SSIM↑	LPIPS↓
Dynamic	Heuristic Augmentation	15.52	0.448	0.466
	Animated Objaverse	14.92	0.380	0.524
	Our Complete Model	**16.60**	**0.462**	**0.409**
Light	Heuristic Augmentation	18.96	0.645	0.406
	Our Complete Model	**20.98**	**0.707**	**0.357**

Table 3. Ablations. Heuristic augmentation and synthetic datasets lead to significantly worse performance for robust view synthesis. For both inconsistencies in dynamics and lighting, our complete model vastly outperforms the baselines due to the underlying video model's ability to simulate physics.

6. Discussion

We have proposed SimVS, a method for high-quality 3D generation from casual captures even in the presence of severe illumination changes and significant scene motion. We believe this represents a step forward in simplifying the capture and creation of 3D scenes.

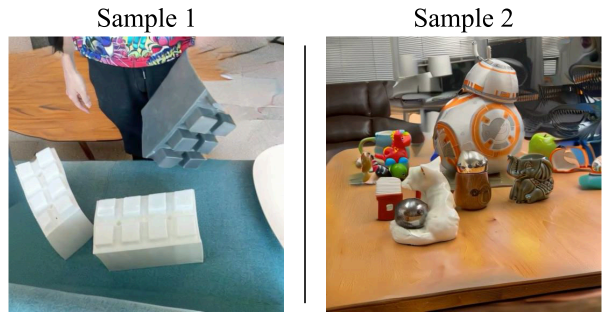


Figure 9. Two representative samples showing limitations of our method. On the left, in the presence of the large dynamic motions of the block, our model may synthesize degenerate scene content. On the right, for areas observed only once such as the background on the top right, our model sometimes outputs poor quality hallucinations.

Limitations Our method requires accurate camera poses, which can be difficult to compute for sparse captures using traditional techniques such as COLMAP. However, recent methods such as DUST3R [57] and the dynamics-robust follow-up MonST3R [68] have shown tremendous promise for camera pose estimation. Also, when there is significant inconsistency or little overlap between views, our method can struggle to reconcile the observations as seen in the samples in Fig. 9.

Conclusion Our work demonstrates the power of using video models to generate data for challenging tasks where collection is expensive and challenging. We believe the approach proposed here will scale well with the ever-improving quality of video models. Moreover, our method is not specific to a particular architecture or task: our method may be applied to make DUSt3R [57]-style models more robust to inconsistency and our harmonization network could be implemented with a camera-controlled video model to directly synthesize multiview-consistent videos in one sampling pass.

Acknowledgements

We would like to thank Paul-Edouard Sarlin, Jiamu Sun, Songyou Peng, Richard Tucker, Linyi Jin, Rick Szeliski and Stan Szymanowicz for insightful conversations and help. We also extend our gratitude to Shlomi Fruchter, Kevin Murphy, Mohammad Babaeizadeh, Han Zhang and Amir Hertz for training the base text-to-image latent diffusion model. This work was supported in part by an NSF Fellowship, ONR grant N00014-23-1-2526, gifts from Google, Adobe, Qualcomm and Rembrand, the Ronald L. Graham Chair, and the UC San Diego Center for Visual Computing.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10): 105–112, 2011. 2

[2] Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? *arXiv preprint arXiv:2406.02965*, 2024. 4

[3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A Space-Time Diffusion Model for Video Generation. *arXiv:2401.12945*, 2024. 4

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields. *ICCV*, 2023. 6

[5] Piotr Bojanowski, Armand Joulin, David Lopez-Pas, and Arthur Szlam. Optimizing the Latent Space of Generative Networks. *ICML*, 2018. 7

[6] 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*, pages 18392–18402, 2023. 3

[7] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14124–14133, 2021. 2

[8] Hiba Dahmani, Moussab Bennehar, Nathan Piasco, Luis Roldao, and Dzmitry Tsishkou. Swag: Splatting in the wild images with appearance-conditioned gaussians. In *European Conference on Computer Vision*, pages 325–340. Springer, 2025. 2

[9] 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. *CVPR*, 2022. 3, 8

[10] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10061–10072, 2023. 2

[11] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2, 2024. 2

[12] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[13] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2025. 2

[14] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular Dynamic View Synthesis: A Reality Check. *NeurIPS*, 2022. 5, 6

[15] 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. *NeurIPS*, 2024. 1, 2, 3, 4, 5, 6

[16] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[17] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[18] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022. 2

[19] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 2

[20] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 2

[21] Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. WildGaussians: 3D Gaussian Splatting in the Wild. *NeurIPS*, 2024. 2, 6, 7

[22] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. MoSca: Dynamic Gaussian Fusion from Casual Videos via 4D Motion Scaffolds. *arXiv:2405.17421*, 2024. 2, 5

[23] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the Depths of Moving People by Watching Frozen People. *CVPR*, 2019. 3

[24] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 2

[25] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4284, 2023. 2

[26] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024. 8

[27] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single

image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[28] 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. 2

[29] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 2

[30] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust Dynamic Radiance Fields. *CVPR*, 2023. 2, 5

[31] 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. *CVPR*, 2024. 2

[32] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. *CVPR*, 2021. 2, 7

[33] Moustafa Meshry, Dan B Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6878–6887, 2019. 2

[34] 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. *ECCV*, 2020. 2

[35] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable Neural Radiance Fields. *ICCV*, 2021. 2

[36] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 2

[37] Yohan Poirier-Ginter, Alban Gauthier, Julien Phillip, J-F Lalonde, and George Drettakis. A diffusion approach to radiance field relighting using multi-illumination synthesis. In *Computer Graphics Forum*, page e15147. Wiley Online Library, 2024. 2

[38] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[39] Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. Nerf on-the-go: Exploiting uncertainty for distractor-free nerfs in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8931–8940, 2024. 2

[40] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20626–20636, 2023.

[41] Sara Sabour, Lily Goli, George Kopanas, Mark Matthews, Dmitry Lagun, Leonidas Guibas, Alec Jacobson, David J Fleet, and Andrea Tagliasacchi. Spotlesssplats: Ignoring distractors in 3d gaussian splatting. *arXiv preprint arXiv:2406.20055*, 2024. 2

[42] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 2, 6

[43] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 6

[44] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[45] Abhishek Sharma, Adams Yu, Ali Razavi, Andeep Toor, Andrew Pierson, Ankush Gupta, Austin Waters, Aäron van den Oord, Daniel Tanis, Dumitru Erhan, Eric Lau, Eleni Shaw, Gabe Barth-Maron, Greg Shaw, Han Zhang, Henna Nandwani, Hernan Moraldo, Hyunjik Kim, Irina Blok, Jakob Bauer, Jeff Donahue, Junyoung Chung, Kory Mathewson, Kurtis David, Lasse Espeholt, Marc van Zee, Matt McGill, Medhini Narasimhan, Miaosen Wang, Mikołaj Bińkowski, Mohammad Babaeizadeh, Mohammad Taghi Saffar, Nando de Freitas, Nick Pezzotti, Pieter-Jan Kindermans, Poorva Rane, Rachel Hornung, Robert Riachi, Ruben Villegas, Rui Qian, Sander Dieleman, Serena Zhang, Serkan Cabi, Shixin Luo, Shlomi Fruchter, Signe Nørly, Srivatsan Srinivasan, Tobias Pfaff, Tom Hume, Vikas Verma, Weizhe Hua, William Zhu, Xinchen Yan, Xinyu Wang, Yelin Kim, Yuqing Du, and Yutian Chen. Veo. 2024. 2

[46] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view Diffusion for 3D Generation. *arXiv:2308.16512*, 2023. 2

[47] Meng-Li Shih, Wei-Chiu Ma, Lorenzo Boyice, Aleksander Holynski, Forrester Cole, Brian Curless, and Janne Kontkanen. Extranerf: Visibility-aware view extrapolation of neural radiance fields with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20385–20395, 2024. 2

[48] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*, 2024. 2

[49] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3D. *SIGGRAPH*, 2006. 6

[50] Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. *arXiv preprint arXiv:2406.18717*, 2024. 2

[51] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023. 4

[52] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[53] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 2

[54] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 2

[55] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4190–4200, 2023. 2

[56] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of Motion: 4D Reconstruction from a Single Video. *arXiv:2407.13764*, 2024. 2, 5, 6

[57] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. *CVPR*, 2024. 8

[58] Yuze Wang, Junyi Wang, and Yue Qi. We-gs: An in-the-wild efficient 3d gaussian representation for unconstrained photo collections. *arXiv preprint arXiv:2406.02407*, 2024. 2

[59] Ethan Weber, Aleksander Holynski, Varun Jampani, Saurabh Saxena, Noah Snavely, Abhishek Kar, and Angjoo Kanazawa. Nerfiller: Completing scenes via generative 3d inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20731–20741, 2024. 2

[60] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613*, 2024. 3

[61] 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. 2

[62] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024. 2

[63] Jiacong Xu, Yiqun Mei, and Vishal M Patel. Wild-gs: Real-time novel view synthesis from unconstrained photo collections. *arXiv preprint arXiv:2406.10373*, 2024. 2

[64] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 2

[65] 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. 2

[66] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2

[67] Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. *arXiv preprint arXiv:2403.15704*, 2024. 2

[68] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 8

[69] Xiaoming Zhao, Pratul Srinivasan, Dor Verbin, Keunhong Park, Ricardo Martin Brualla, and Philipp Henzler. Illuminerf: 3d relighting without inverse rendering. *Advances in Neural Information Processing Systems*, 37:42593–42617, 2024. 2

[70] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo Magnification: Learning View Synthesis using Multiplane Images. *SIGGRAPH*, 2018. 3

[71] Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. *arXiv preprint arXiv:2410.12781*, 2024. 2