SimVS: Simulating World Inconsistencies for Robust View Synthesis

Supplementary Material

A. Evaluation Details

We use the pretrained CAT3D [4] model provided by the authors for the lighting benchmarks along with the default implementation of ZipNeRF with GLO [1]. For inflating the consistent views, we use the camera poses of the monocular video corresponding to an input condition, although interpolating the input camera poses would also suffice. For the dynamics benchmark, we use a CAT3D model finetuned to condition on 5 images and predict 3. The lower variance of the conditioning provides a slight benefit as seen in Tab. 1. For inflating the consistent views, we sample at the camera poses of the monocular video trajectory. It should be noted that interpolating the input cameras is a sufficient alternative.

For the comparison to WildGaussians, we use the public implementation in the NerfBaselines [5] package. Since the method is based on 3DGS, we rectify all images as input to the method, which is the default of the package. We also modify the testing function to use the appearance embedding of the target state.

Ablation	PSNR↑	SSIM↑	LPIPS↓
w/o 5 Cond. CAT3D	16.57	0.453	0.414
Our Complete Model	**16.60**	**0.462**	**0.409**

Table 1. Performance comparison of ablation conditions.

B. Comparison to Shape of Motion [7]

We include an additional baseline comparison to Shape of Motion [7], the current state-of-the-art method for 4D reconstruction. We consider this type of method to be slightly orthogonal to our approach; incorporating priors such as static masks and monocular depth may improve our results further.

As in our experiments in the main paper, we provide this baseline with a set of unordered sparse images from the DyCheck [3] dataset. We compare only on the scenes that Shape of Motion benchmarked, and therefore exclude Space-Out and Wheel.

We use the refined poses and aligned depth from the original paper and train the model to render the standard 360x480 images, center-cropped to a square aspect ratio as in the comparisons included in the main paper. As specified in their GitHub repository, we computed the video masks with Track Anything [8], which shows some robustness to the sparse inputs. However, TAPIR [2] seemed to struggle to compute reasonable tracks given sparse inputs. We show qualitative results in Fig. 1 and quantitative results in Tab. 2. Due to the inability of Shape of Motion to predict scene content outside of the frustums of the input images, we show results with covisibility masks as well. As evidenced by the metrics and qualitative results, Shape of Motion struggles to recover a cohesive representation under the sparse and unordered input setting of this paper.

	Condition	PSNR↑	SSIM↑	LPIPS↓
raw	Ours	**16.46**	**0.425**	**0.484**
	Shape of Motion [7]	14.10	0.396	0.485
mask	Ours	**17.03**	**0.557**	0.410
	Shape of Motion [7]	15.58	0.536	**0.391**

Table 2. Performance comparison of methods with and without covisibility masks from [3].

Figure 1. Qualitative comparison to Shape of Motion [7] on sparse input views from the DyCheck dataset.

C. Additional visualizations

We show the ability of our model to effectively and flexibly incorporate more information in Fig. 2, reducing the uncertainty in its prediction with larger context. We also show samples from the lighting dataset in Fig. 4. Due to privacy concerns, we do not show samples from the dynamics dataset, which consists of humans.

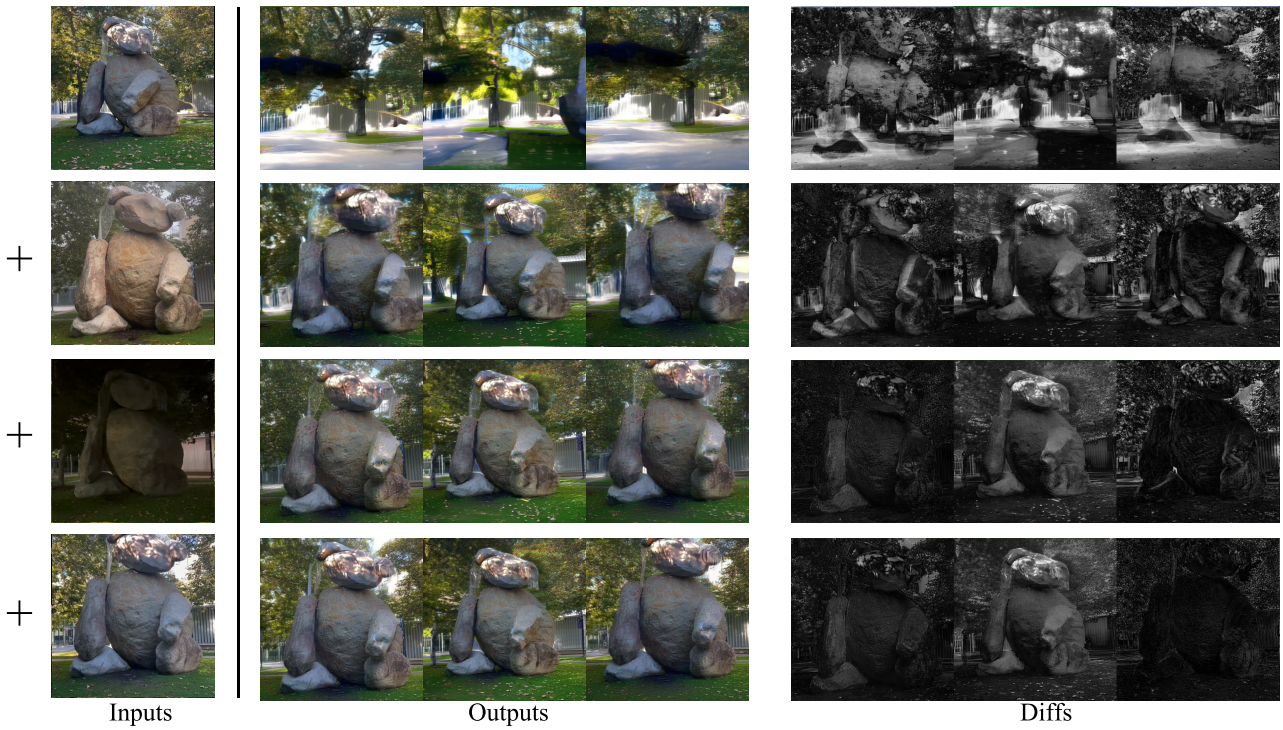

Figure 2. Our model incorporating more context given an increasing number of images. Given the (additional) inputs on the left, our model reduces uncertainty in its predictions and predicts more well-aligned images to three extra input images as seen in the difference map between additional inputs and the outputs.

D. Training Details

We finetune the pretrained CAT3D [4] model with 0 initialization for the input conditioning convolution layer to accept the inconsistent latents $\tilde{z}$. We train with a batch size of 64 (sets of multiview images) per gradient step. For the lighting model, we finetune for 36k iterations, and for dynamics, 48k iterations.

We train all ablations for the same amount of time as the corresponding model for the respective data types, except for the dynamics augmentation model which quickly overfits to copying; therefore, we train it for only 12k iterations, as this is where the loss on the held-out OOD data is minimized.

E. Video Model Prompt Details

In this section, we specify the details of the prompting for the video model including the meta-prompt, example prompts, and list of prompts for lighting.

E.1. Lighting prompts

For lighting, we sample the prompts from the following set:

1. `"a bright light casts shadows"`
2. `"the light slowly dims from bright to dark"`
3. `"an object flies around the room, casting hard shadows"`
4. `"a transition from a bright day to a dark night"`
5. `"the shadows and lights move"`
6. `"a strobe light flashes"`

E.2. Dynamics prompts

For dynamics, we sample about 10k total prompts using the meta-prompt given in Fig. 3. We include 20 examples below:

1. `"They walk quickly along the path, the child struggling to keep up while carrying the bottle."`
2. `"The boys playfully pose for a photo."`
3. `"The mechanics are actively repairing the car, with tools moving and parts being replaced."`
4. `"The girls are collaboratively typing on the laptop."`
5. `"The chef moves through the train serving food to passengers."`
6. `"Children run through the play tunnel and climb onto the boat."`
7. `"The children run around the line, crossing it repeatedly during the game."`
8. `"The girl walks past a classroom art`

display."
9. "Two people actively select books and papers from the table."
10. "The puppeteer manipulates the puppets, making them move and interact."
11. "The woman excitedly raises and lowers her arms."
12. "The woman gestures emphatically as the man adjusts a component on the truck door."
13. "The two assistants helped Santa adjust his position in the chair."
14. "The children reach for items on the table, some stand up and move to a different seat."
15. "The man gestures emphatically while speaking on the phone."
16. "The majorette tosses and catches the baton."
17. "The woman raises and lowers her mug as she drinks."
18. "The child reached for a cleaning supply."
19. "The woman dramatically throws her arms out in a wide arc."
20. "Someone rolled up the red fabric and placed it against the shelf."

F. Details of Lumiere sampling

For sampling from the Lumiere model, we utilize a random-frame variant where the input frame can be anywhere in the video (not just the first frame). This variant is trained by sampling a random frame for each training video and concatenating the input to every frame along the channel dimension, identically as the Lumiere inpainting model.

We use the following camera-based negative prompt to induce the desired characteristics in the output video and alleviate Lumiere's tendency to output still videos:

c_{negative} = "frozen, photograph, fixed lighting, moving camera, zoom in, zoom out, bird view, panning view, 360-degree shot, orbit shot, arch shot"

We use 250 DDPM sampling steps for the image- and text-conditioned Lumiere base model at a resolution of 128x128. We then upsample that video conditioned only the original prompt to a size of 1024x1024 with 250 sampling steps and resize to the desired size of 512x512. We set the guidance weight to 6 for both processes.

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields. *ICCV*, 2023. 1

[2] 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. 1

[3] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular Dynamic View Synthesis: A Reality Check. *NeurIPS*, 2022. 1

[4] 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

[5] Jonas Kulhanek and Torsten Sattler. Nerfbaselines: Consistent and reproducible evaluation of novel view synthesis methods. *arXiv*, 2024. 1

[6] 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. 4

[7] 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. 1

[8] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023. 1

meta_prompt = """I need you to generate prompts for a video model to create scene motion from a static camera perspective, using the above frames which occur at the end of the video.

Task:

- Describe Each Image: For each of the six images, provide a simple, concise description focusing on salient humans, their poses, and the overall 3D scene. Mention any key articulations or positions of objects or people. Descriptions should be exactly one sentence long.

- Describe Corresponding Motion: Imagine each image is a frame of a video shot from a stationary camera. What is that video about? For each image, provide a one-sentence description of significant motion that may have happened in that video.

Additional Requirements:

- The scene motion should be visually perceptible and significant.
- Avoid introducing new objects or content not present in the image.
- The motion description should not imply stillness or minimal movement (e.g., avoid words like "sitting").
- Do not specify rotation.

Example:

- Use the following format for each image and its corresponding motion. Make sure to provide exactly six pairs of descriptions:

  - Image 1: In a spacious studio, two young people dance in the foreground while others lie scattered on the carpeted floor.
  - Motion 1: The two children dance.

  - Image 2: In a modern hotel lobby, the woman holds a pillow mid-swing while another person lounges on a red chair.
  - Motion 2: The woman swings the pillow.
  (Continue this pattern through Image 6 and Motion 6.)

Guidelines:

- Provide descriptions for all six images.
- Do not mention camera movement or imply camera angles.
- Do not introduce new elements or actions not inferred from the scene.
- Avoid words that minimize the motion like "slowly" or "gently"
- Be specific and concise. Do not use similes or metaphors.
- Do not use slashes in your captions.
- Make sure that the motion can be seen WITHOUT moving the camera as the viewpoint is constant.
"""

Figure 3. The meta-prompt used to generate dynamics captions on the Mannequin Challenge dataset [6].

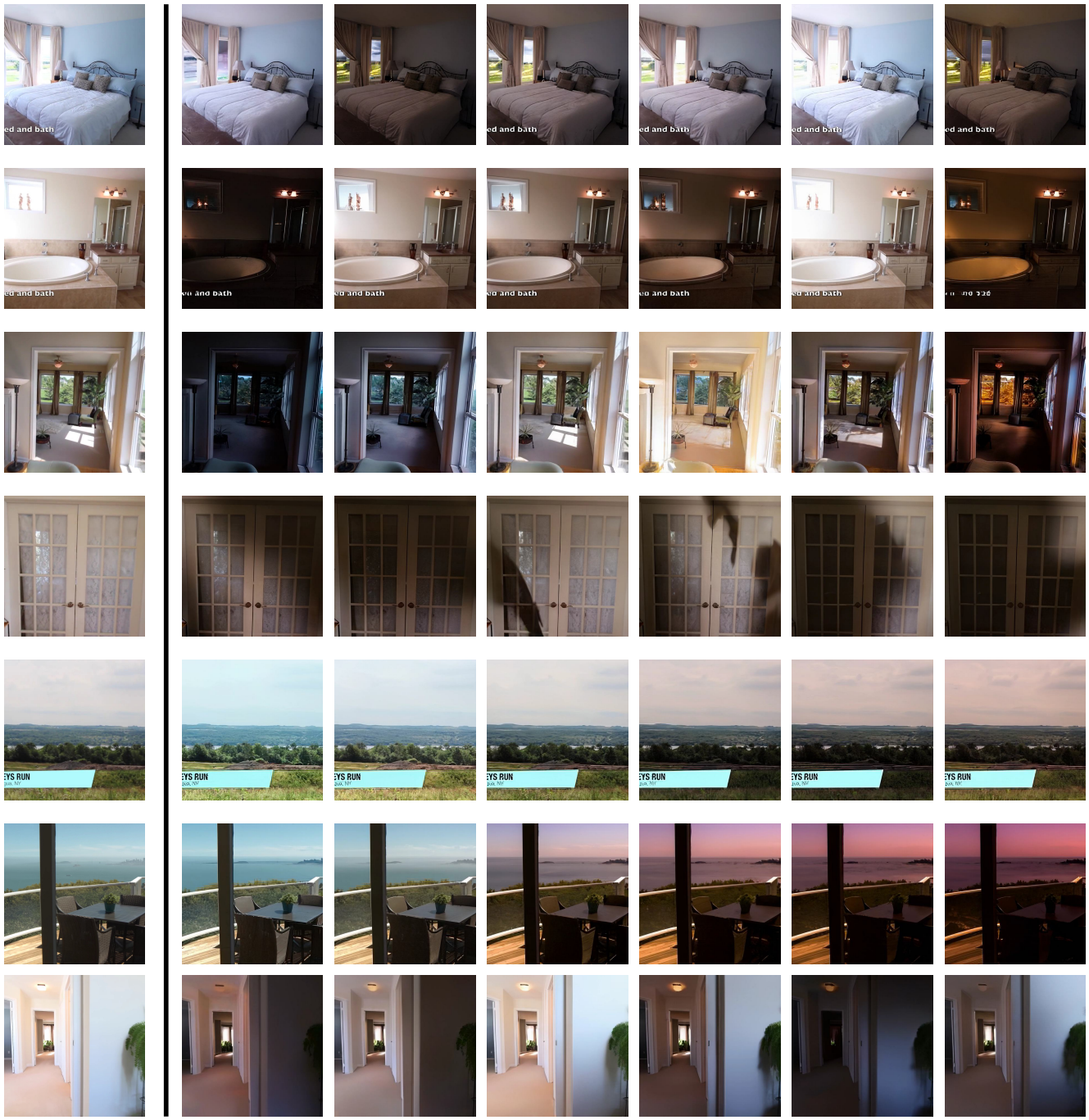

Figure 4. We show example samples from the lighting data we sampled.