Reconstruct, Inpaint, Test-Time Finetune: Dynamic Novel-view Synthesis from Monocular Videos

Kaihua Chen*, Tarasha Khurana*, Deva Ramanan
The Robotics Institute, Carnegie Mellon University

Scene-level dynamic novel view synthesis

Scene-level dynamic novel view synthesis

Given a source view of a scene, generate its novel view from a target camera

Source views

Target view

Prior work fails at extreme novel view synthesis

Shape-of-Motion [1]

MoSca [2]

Optimization-based approaches cannot model dynamics well even for near novel views

[1] Wang, Q., Ye, V., Gao, H., Austin, J., Li, Z., & Kanazawa, A. (2025). Shape of motion: 4d reconstruction from a single video. ICCV.
[2] Lei, J., Weng, Y., Harley, A. W., Guibas, L., & Daniilidis, K. (2025). Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 6165-6177).
[3] Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., ... & Vondrick, C. (2024, September). Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision* (pp. 313-331).

Prior work fails at extreme novel view synthesis

Shape-of-Motion [1]

MoSca [2]

Optimization-based approaches cannot model dynamics well even for near novel views

Input view

Groundtruth novel-view

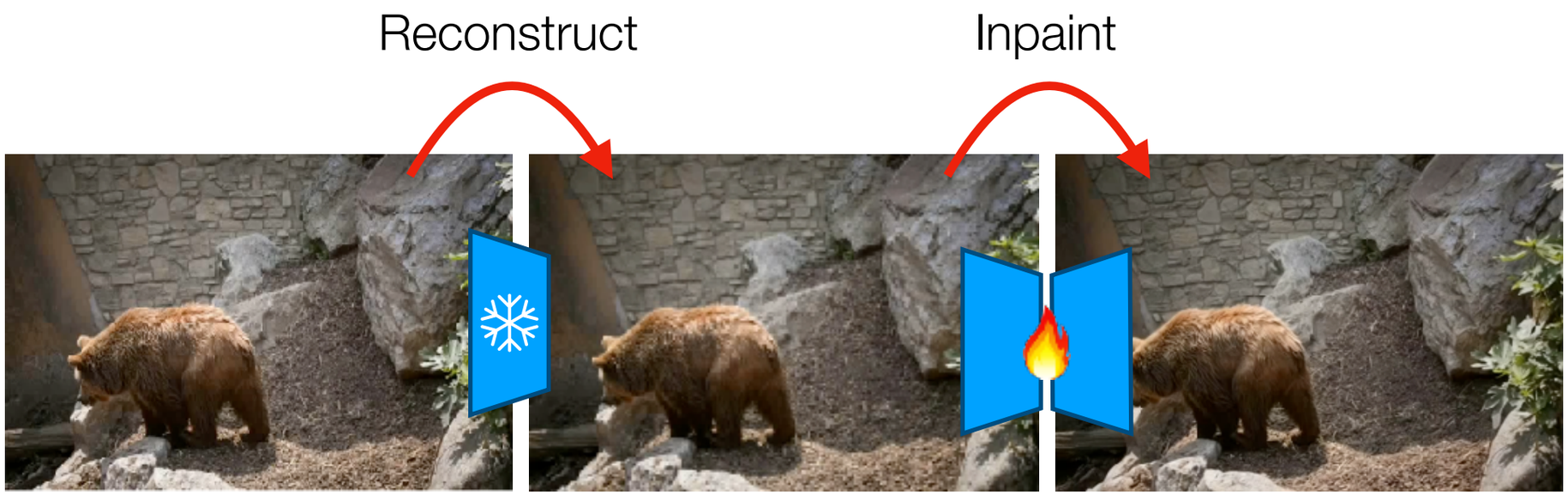
GCD [3] novel-view

Feed-forward methods are not 3D consistent because of implicit conditionings

[1] Wang, Q., Ye, V., Gao, H., Austin, J., Li, Z., & Kanazawa, A. (2025). Shape of motion: 4d reconstruction from a single video. ICCV.
[2] Lei, J., Weng, Y., Harley, A. W., Guibas, L., & Daniilidis, K. (2025). Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 6165-6177).
[3] Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., ... & Vondrick, C. (2024, September). Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision* (pp. 313-331).

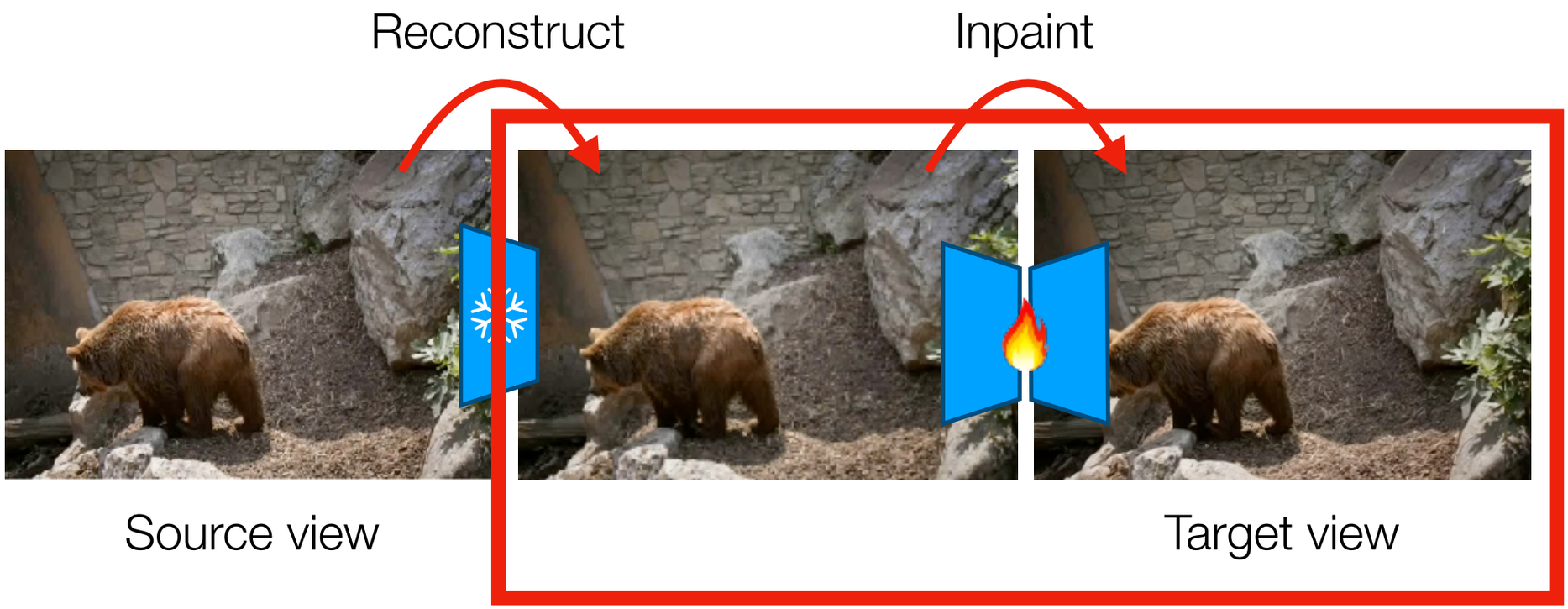
Novel view synthesis = Reconstruct + Inpaint

State-of-the-art reconstruction does much of the heavy-lifting …



Novel view synthesis = Reconstruct + Inpaint

State-of-the-art reconstruction does much of the heavy-lifting …

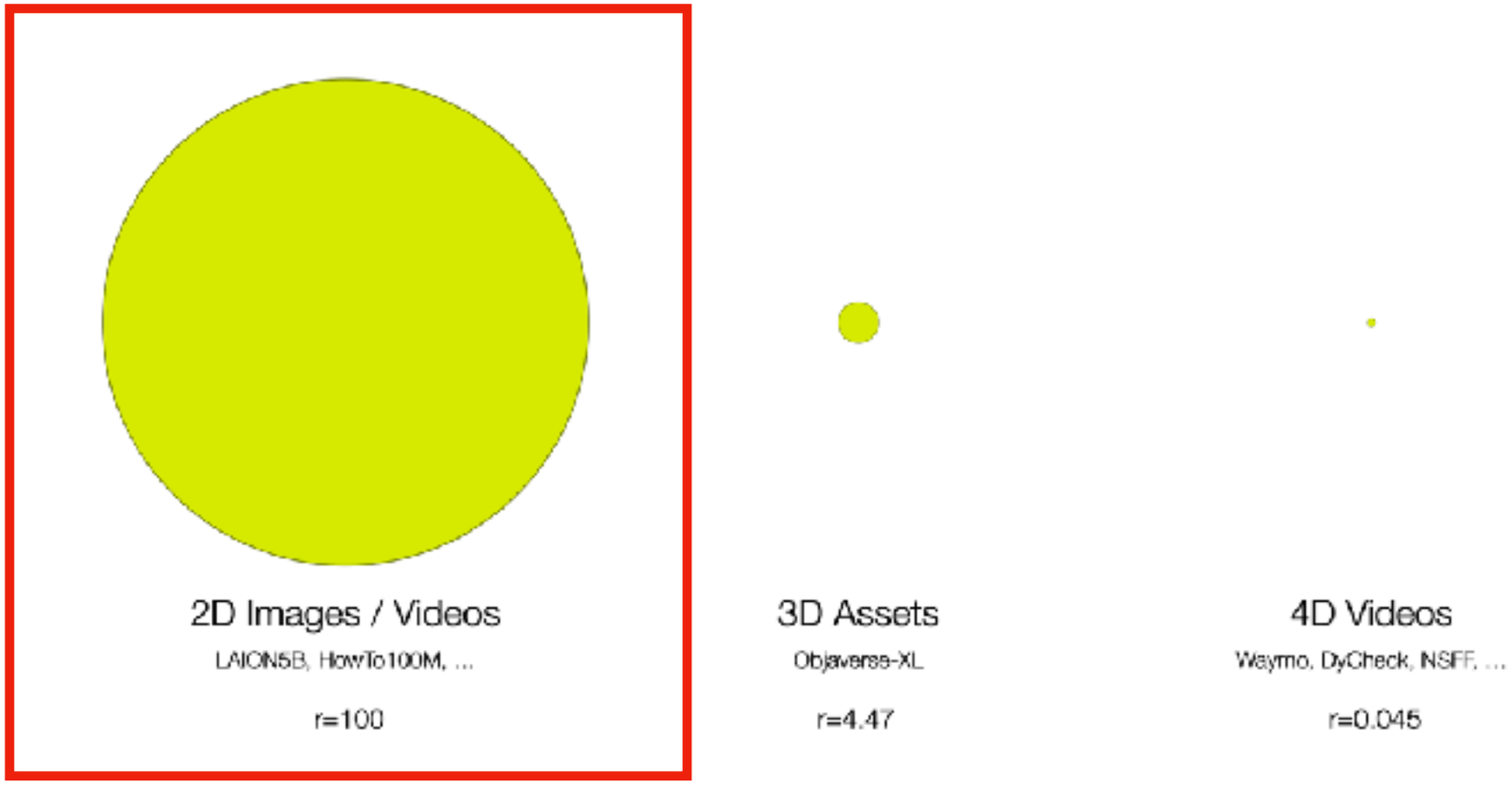


Building an inpainting engine for novel-view synthesis

How does one get the training data?

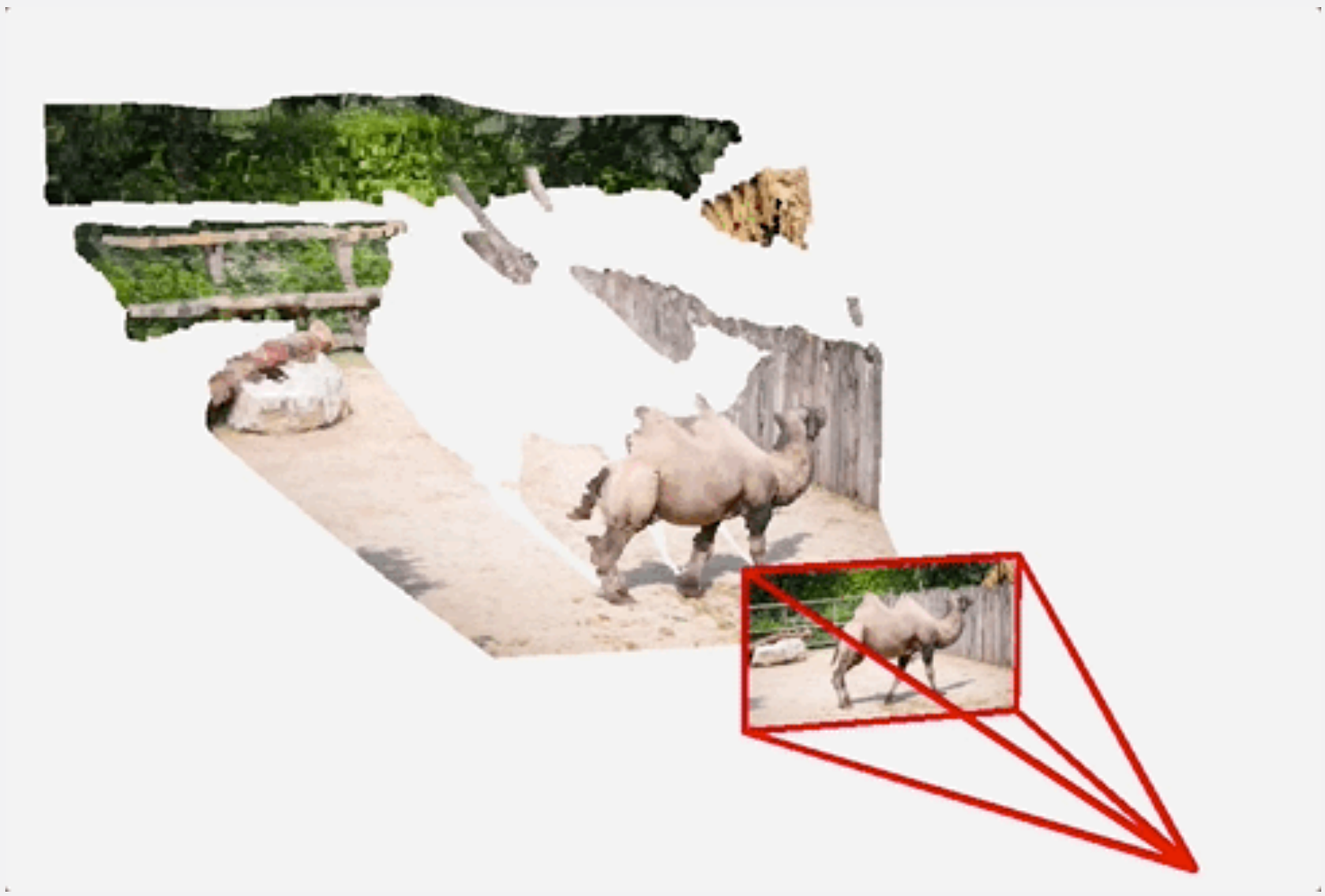
Easy fix: Multi-view datasets with groundtruth novel views

But …



Building an inpainting engine for novel-view synthesis

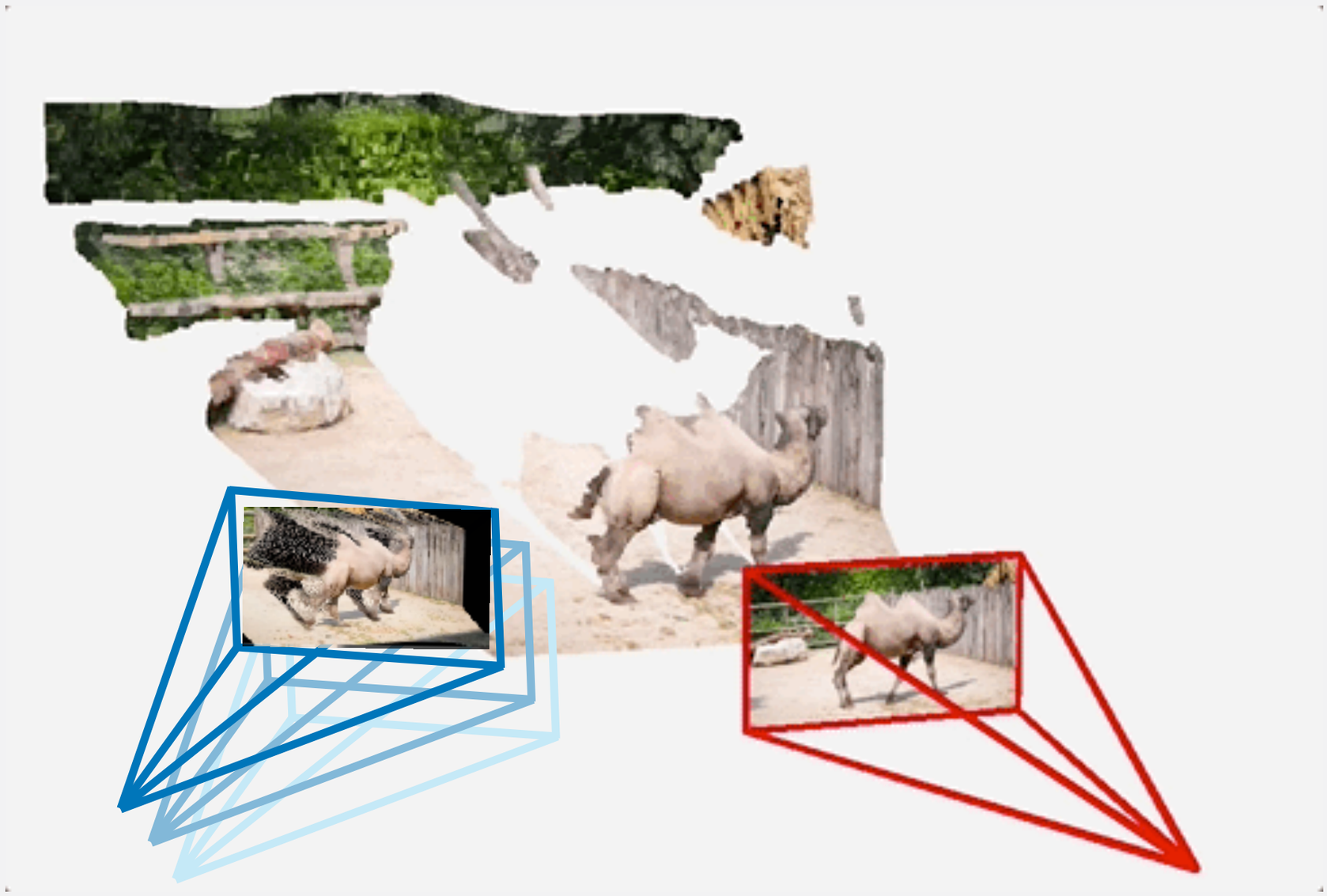
How does one get the training data?



Start with reconstruction with a dynamic SLAM framework like MegaSAM

Building an inpainting engine for novel-view synthesis

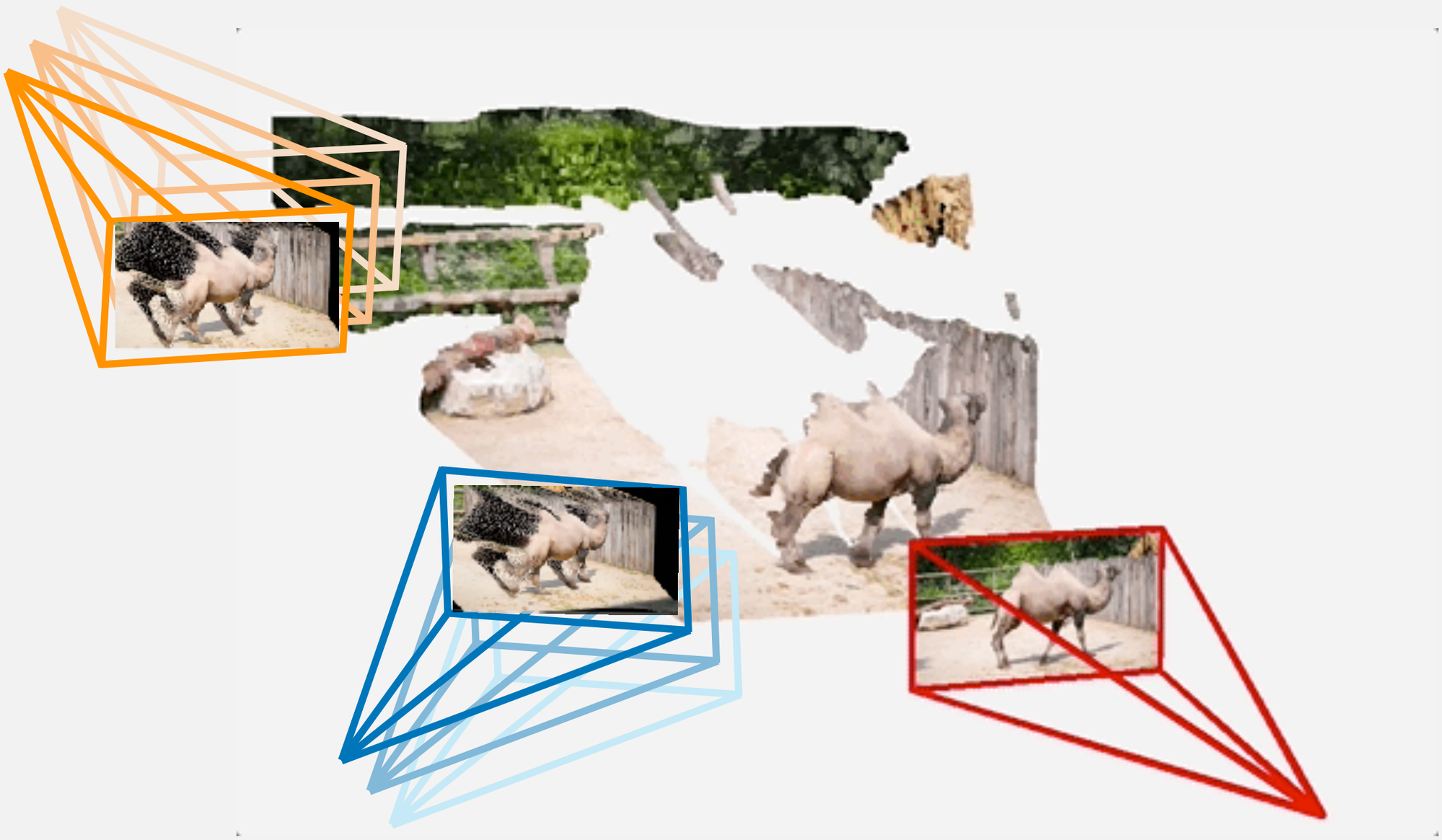
How does one get the training data?



Once we have the reconstruction, we can always render new views

Building an inpainting engine for novel-view synthesis

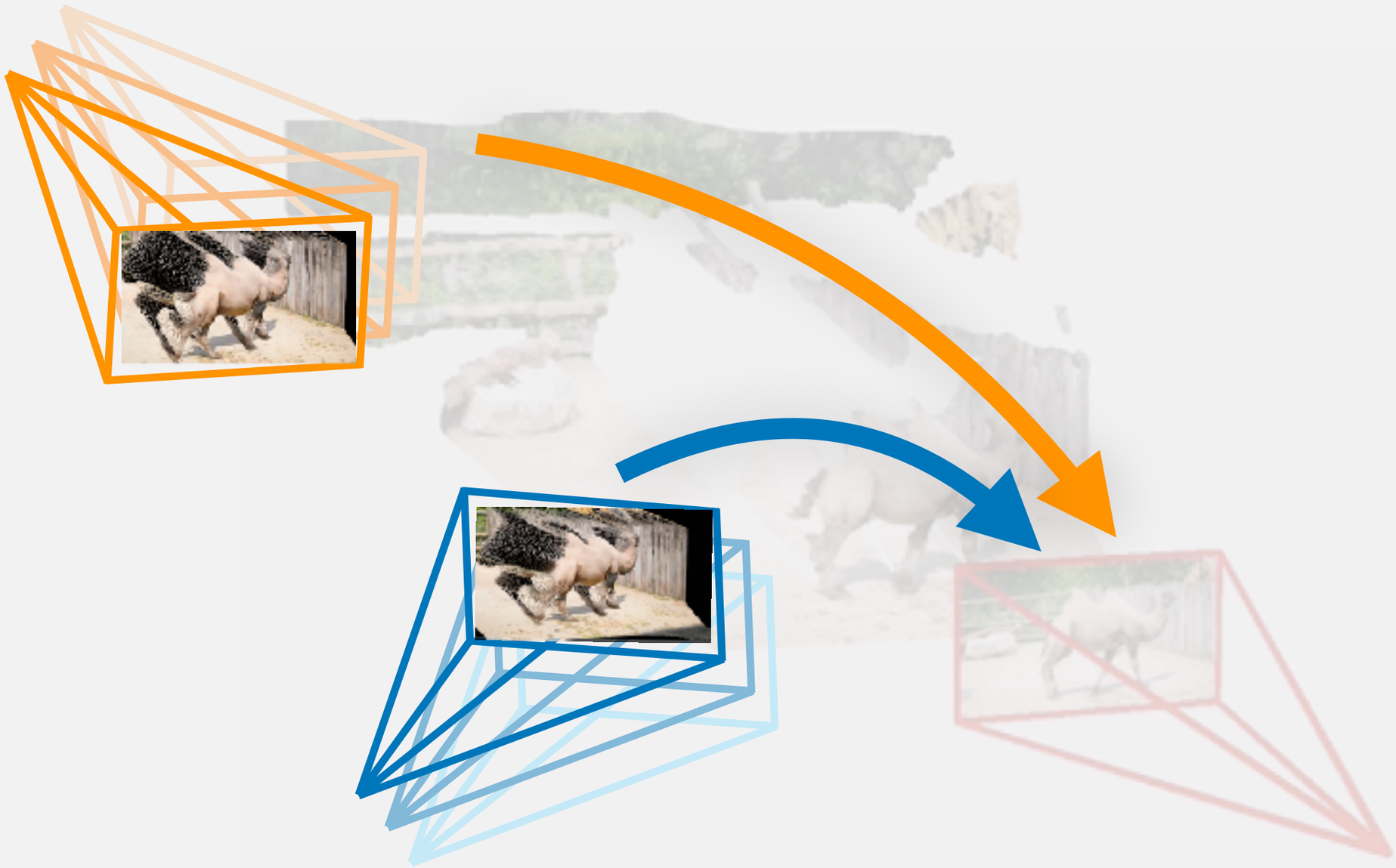
How does one get the training data?



Once we have the reconstruction, we can always render new views

Building an inpainting engine for novel-view synthesis

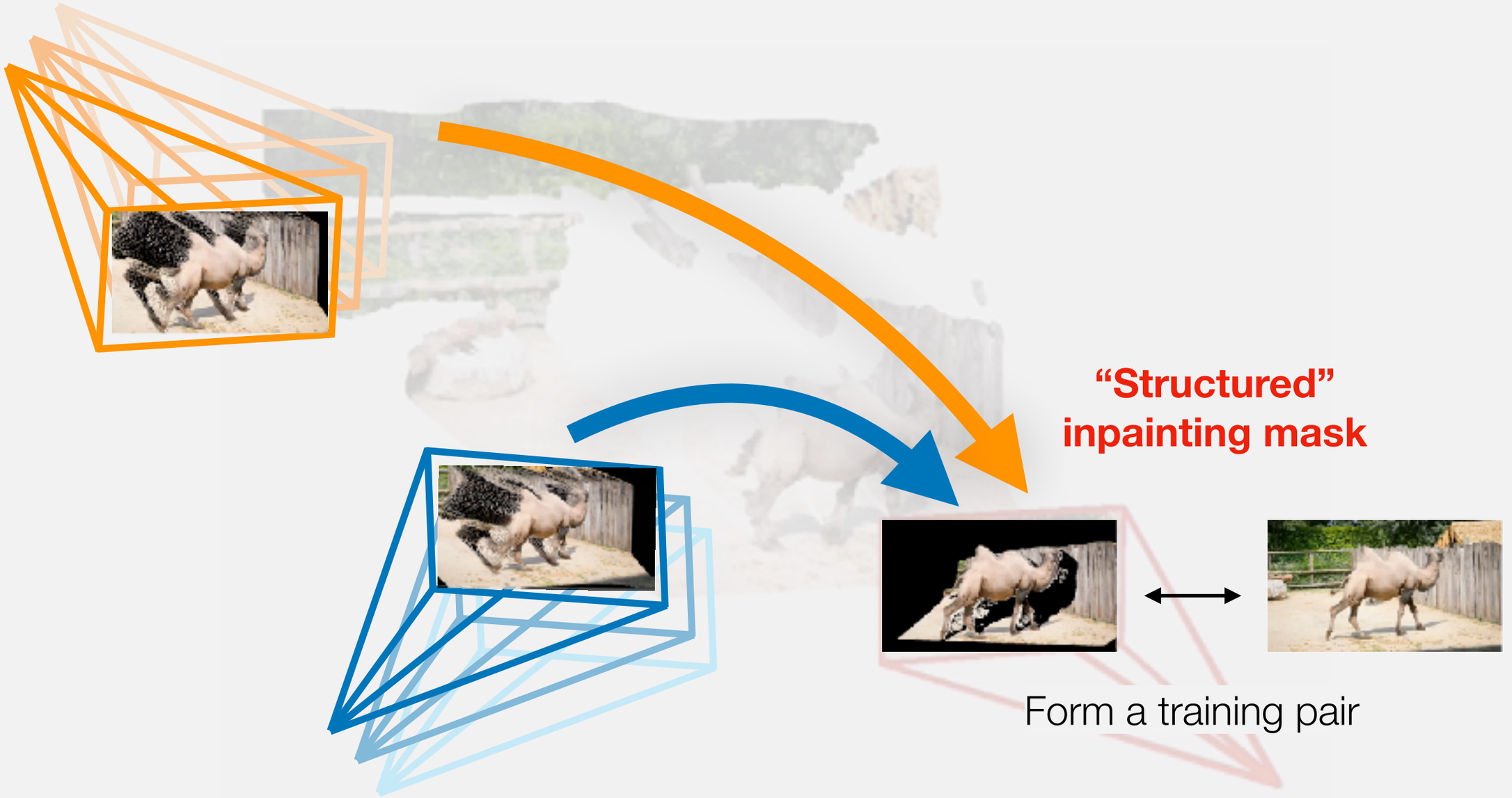
How does one get the training data?



Projecting these visible pixels back into source view gives us a set of covisible pixels

Building an inpainting engine for novel-view synthesis

How does one get the training data?

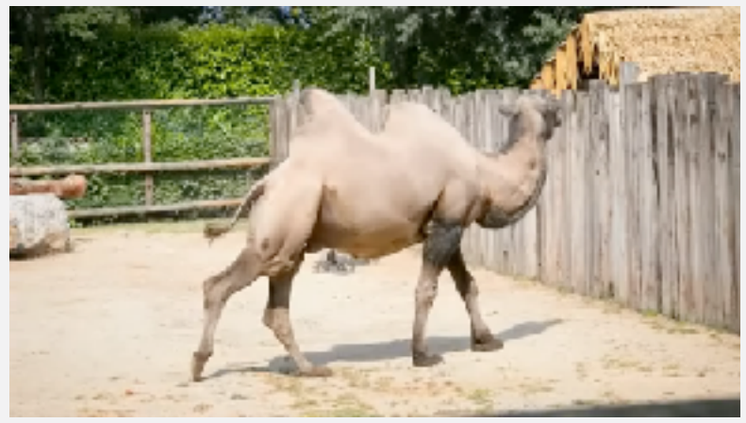


Projecting these visible pixels back into source view gives us a set of covisible pixels

Train purely on 2D videos via self-supervision

Can now rely on the large corpus of 2D videos for training “inpainter”

Training input



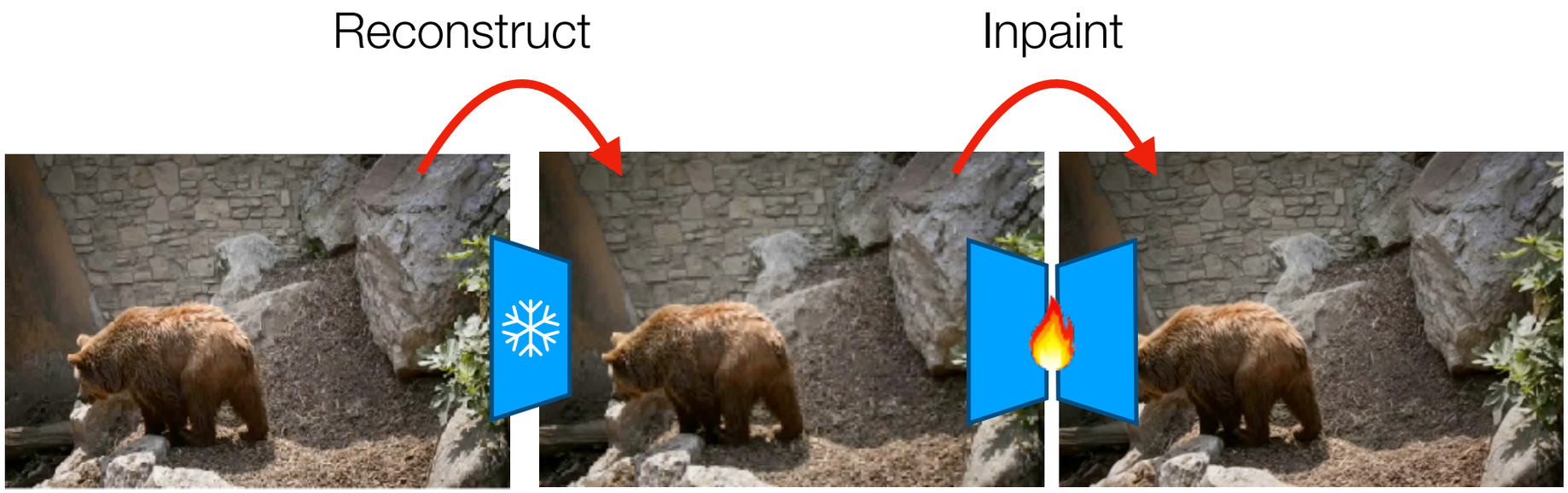
Target output

That’s not all! This self-supervision unlocks test-time adaptation

In fact, test-time fine-tuning is the most crucial component of our pipeline

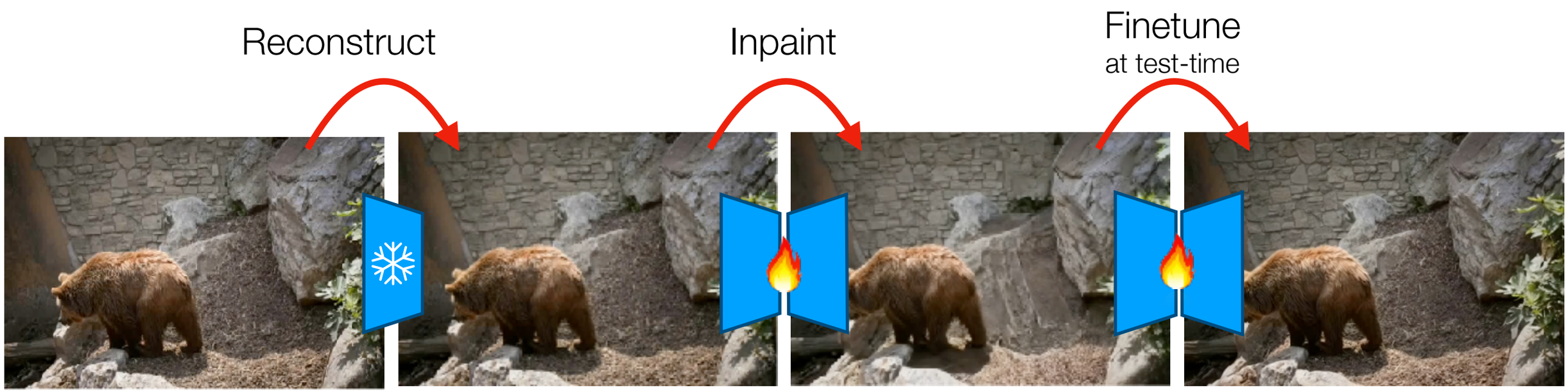
That’s not all! This self-supervision unlocks test-time adaptation

In fact, test-time fine-tuning is the most crucial component of our pipeline



That’s not all! This self-supervision unlocks test-time adaptation

In fact, test-time fine-tuning is the most crucial component of our pipeline



Novel-view synthesis = Reconstruct + Inpaint + Test-time finetune

Input view | Rendered novel view | Inpainted novel view

Synthetic video from SORA

Real-world dynamic scene

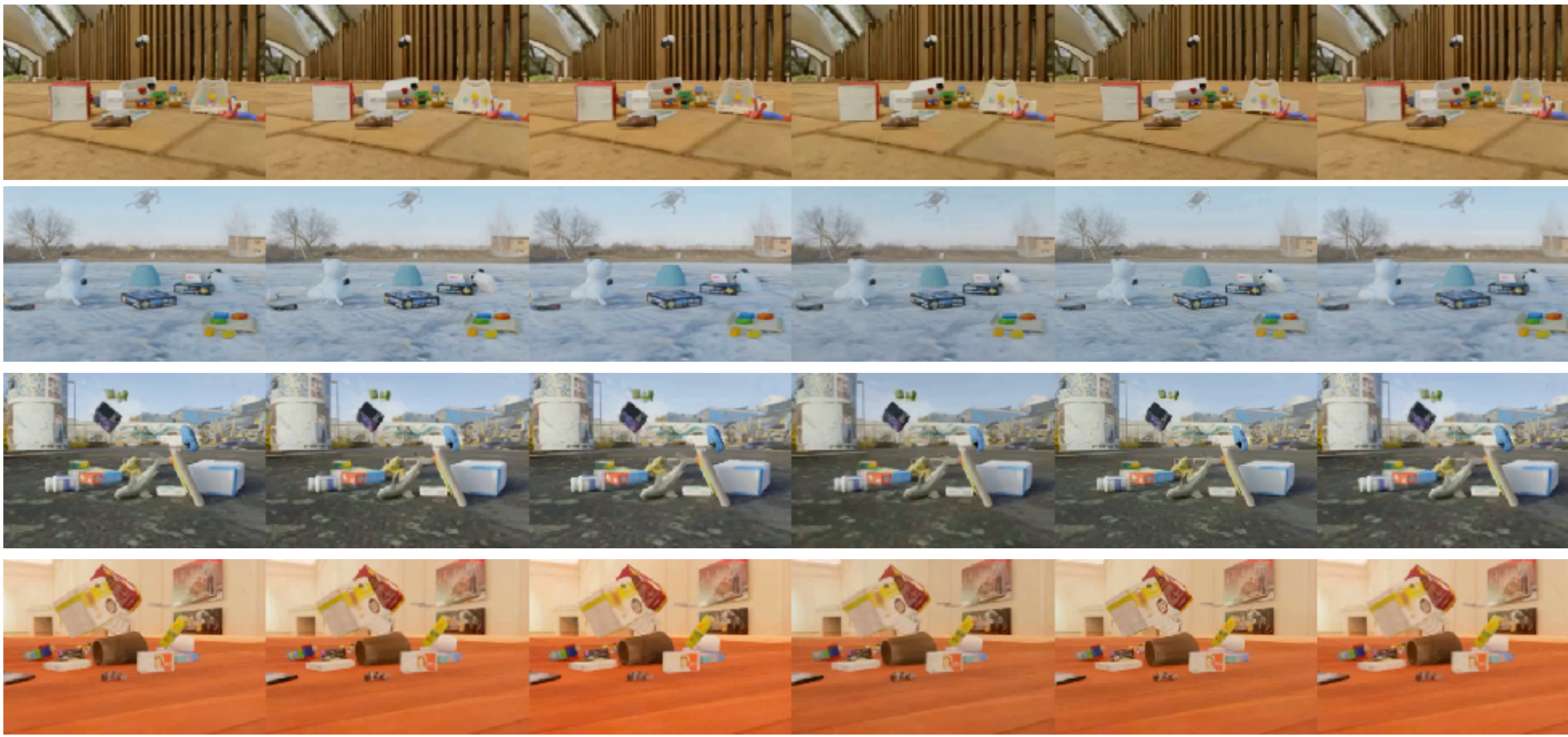
Real-world static scene

Zero-shot evaluation on ParallelDomain-4D

Given trajectory is an egocentric view, novel trajectory is camera panning to a birds'-eye-view

Input	GT point cloud	GCD	TrajCrafter	CogNVS	GT
					
PSNR	18.79	21.77	21.46	**24.34**	

Zero-shot evaluation on Kubric-4D



PSNR	15.12	18.59	20.93	**22.63**	

Zero-shot evaluation on DyCheck

FID	239.57	164.29	148.18	-	140.35	94.48	**92.83**

Thank you!