ExPose: Reinforcing Video Generation Models for Extreme Pose Estimation

Supplementary Material

1. Overview

This supplementary material delves into the proposed method further, offering additional results and discussions that were not fully covered in the main paper:

- Implementation Details (Sec. 2).
- Additional Quantitative Results (Sec. 3).
- Limitations and Failure Cases (Sec. 4).
- Video Generation Results (Sec. 5).
- Pseudo Video Datasets for SFT (Sec. 6).

2. Implementation Details

2.1. Training and Evaluation Datasets

DL3DV We adopt the DL3DV-GS-960P [4] dataset as our primary training and evaluation source. Following the official train/test split, we construct 3,860 pseudo-video samples from the training set by forming triplets of views $(I_{\text{ref}}, I_{\text{aux}}, I_{\text{target}})$. For evaluation, we sample 300 $(I_{\text{ref}}, I_{\text{target}})$ pairs from the test split.

Cambridge Landmarks For outdoor scenes with large viewpoint variation, we use the Cambridge Landmarks dataset provided by the Extreme Rotations in the Wild benchmark [2]. From this dataset, we select 300 $(I_{\text{ref}}, I_{\text{target}})$ pairs. Because ground-truth translation values are not available, we report only rotation-related metrics.

NAVI We use the NAVI v1.5 dataset [7] and extract 300 $(I_{\text{ref}}, I_{\text{target}})$ evaluation pairs for object-centric evaluation.

ScanNet For indoor evaluation, we use the ScanNet-v2 official test split, following the subset configuration released by MonST3R [14]. From this split, we construct 300 $(I_{\text{ref}}, I_{\text{target}})$ evaluation pairs.

2.2. Training Protocol

Our overall training pipeline is organized into two stages: we initially establish a reliable supervised prior through SFT, and subsequently enhance the model via GRPO-based reinforcement learning.

Stage 1: Supervised Finetuning (SFT) We first train the model for 5,000 steps using the SFT objective alone. This stage optimizes the reconstruction loss $\mathcal{L}_{\text{SFT}}$ based on pseudo-video triplets.

Stage 2: GRPO with SFT We then continue training for 20,000 steps using GRPO while jointly applying the SFT loss. At each GRPO step, we generate a trajectory and compute the reward according to

$$\begin{aligned} r_{\text{total}} &= r_{\text{pose}} + r_{\text{pic}} \qquad \text{(two-frame condition)}, \\ r_{\text{total}} &= r_{\text{div}} + r_{\text{pic}} \qquad \text{(one-frame condition)}. \end{aligned} \tag{1}$$

where r_{two} conditions on both $(I_{\text{ref}}, I_{\text{target}})$ while r_{one} uses only I_{ref}. To encourage diversity during GRPO, at every training step we sample the reward mode stochastically: with probability 0.7 we apply the two-condition reward r_{two}, and with probability 0.3 we apply the one-condition reward r_{one}. This mixture stabilizes training while improving generalization across different supervision strengths.

2.3. Training Hyperparameters

We train the model with a batch size of 8 and generate 4 candidate videos per prompt, using a learning rate of 2×10^{-4} and trajectories of 73 frames. LoRA fine-tuning is applied with rank 128 and scaling factor $\alpha = 128$. During GRPO training, we control the exploration strength using

$$\sigma_t = a\sqrt{\frac{t}{1-t}}, \tag{2}$$

where $a = 0.5$. We adopt rectified flow sampling with 10 timesteps. The loss hyperparameters are set to $\lambda_{\text{rot}} = 0.25$, $\beta = 1$, $\lambda_{\text{KL}} = 0.01$, $\lambda_{\text{pic}} = 10$, $\lambda_{\text{div}} = 1$, and $\lambda_{\text{SFT}} = 0.2$. All experiments are conducted using two NVIDIA A100 GPUs with 80 GB memory each. For all experiments, we use a test prompt: *"A 3d consistent one-take video with a smooth and continuous camera trajectory, featuring a uniform viewpoint transition at a constant speed. No cuts, no jitter, fully stable motion."*

2.4. Dual frame selection for the test set

To construct a challenging evaluation set, we generate dual-frame $(I^{\text{ref}}, I^{\text{target}})$ pairs directly from the pose metadata provided in each dataset. For every scene, we first gather all frames with valid camera poses and compute the relative viewing-angle difference for every candidate pair. We then retain only pairs that exhibit clearly distinct viewpoints, while enforcing a minimum temporal spacing to avoid near-duplicate frames. This results in test pairs that capture meaningful geometric changes rather than minor or consecutive frame variations.

For datasets that include metric depth, we further suppress pairs that share substantial overlap. Each depth map is back-projected into world coordinates, and pairs whose

spatial support overlaps the least are preferred. The final test set thus consists of reference–target pairs with both wide viewpoint differences and minimal 3D overlap, providing a consistent and rigorous benchmark for evaluating two-frame conditioning and pose estimation performance.

2.5. Triplet frame selection for training set

To construct the SFT triplets $(I^{\text{ref}}, I^{\text{aux}}, I^{\text{target}})$, we first compute geometric overlap between COLMAP cameras using the sparse point cloud. For two cameras i and j, let V_i and V_j denote the sets of 3D points that project inside their image bounds with positive depth. We define the pairwise overlap using the Jaccard index

$$o_{ij} = \frac{|V_i \cap V_j|}{|V_i \cup V_j|}, \tag{3}$$

which measures the ratio of jointly visible points relative to the union of visible points.

Reference–target selection Among all camera pairs, we identify an *extreme* pair $(c_{\text{ref}}, c_{\text{target}})$ by selecting those that occupy the lowest-overlap region of the distribution $\{o_{ij}\}$, and then choosing the pair with the smallest o_{ij} while also exhibiting a large viewpoint change. These two cameras define I^{ref} and I^{target}.

Auxiliary view selection We then choose an auxiliary view c_{aux} by evaluating all remaining cameras and selecting one that (1) lies spatially between c_{ref} and c_{target}, (2) yields a smooth directional transition, and (3) has moderate and balanced overlaps with both endpoints:

$$\begin{aligned} 0 < o_{\text{ref,aux}},\, o_{\text{aux,target}} < 0.7, \\ |o_{\text{ref,aux}} - o_{\text{aux,target}}| < 0.3. \end{aligned} \tag{4}$$

This produces a geometrically natural triplet suitable for supervision.

Finally, $(I^{\text{ref}}, I^{\text{aux}}, I^{\text{target}})$ are placed at frames $\{0, 40, 72\}$ to form the pseudo video GT used for SFT.

2.6. Evaluation for pose estimation

For the evaluation of VGGT [10] and MapAnything [8], we use their official implementation.

2.7. Evaluation Setting

Our ref-target selection follows a two-step optimization to ensure extreme difficulty. First, we identify a candidate set of 5,000 pairs based on viewing angle and spatial distance. Second, we perform exhaustive overlap filtering using the Jaccard index from 3D point cloud visibility, prioritizing pairs with a target overlap of 0.0. This rigorous process is significantly more aggressive than random sampling or simple thresholding, highlighting our focus on extreme scenarios where conventional optimization methods fail completely.

Table 1. Comparison on the DL3DV dataset using VGGT.

Method	MRE ↓	MTE ↓	5°		15°		30°		
			R_{acc} ↑	T_{acc} ↑	R_{acc} ↑	T_{acc} ↑	R_{acc} ↑	T_{acc} ↑	AUC ↑
CogVideoX	43.78	25.52	48.33	31.67	61.67	48.33	66.67	63.33	42.22
Wan 2.1	43.00	27.64	51.67	31.67	65.00	46.67	66.67	61.67	42.44
ViewCrafter	39.00	25.06	45.00	33.33	63.33	46.67	68.33	63.33	41.67
Gen3R	<u>37.54</u>	<u>24.33</u>	48.33	30.00	<u>66.67</u>	<u>55.00</u>	<u>71.67</u>	61.67	44.22
InterPose (w/ Aether)	45.55	25.45	<u>52.33</u>	<u>35.00</u>	65.33	50.67	68.33	<u>67.00</u>	<u>45.42</u>
ExPose (Ours)	**33.78**	**20.50**	**60.67**	**42.67**	**73.67**	**59.67**	**75.67**	**74.00**	**53.64**

2.8. Training Protocol and Hyperparameters

The final optimization objective uses a total reward calculated as a weighted sum:

$$R_{total} = r_{pose} + w_{PIC} r_{PIC} + w_{div} r_{div} \tag{5}$$

This combined reward signal is then used within the reinforcement learning framework to guide the generator toward geometric consistency and trajectory diversity. To mitigate reward hacking, we exclude samples below a predefined reward threshold and apply the Pose Interpolation Constraint to ensure the continuity of the camera trajectory.

Training utilized two NVIDIA A100 GPUs, taking 24 hours for SFT and 48 hours for RL, resulting in a total of 72 hours. During the inference and generation process, ExPose generates 72 frames per sequence and uniformly samples 7 frames for the pose estimator. The inference latency is approximately 30 seconds per image pair.

3. Additional Quantitative Results

We provide further comparisons with recent generative baselines and pose-aware methods. As shown in Table 1, ExPose consistently outperforms recent video generation models such as Wan 2.1 [9] and CogVideoX [12] from the VideoX-Fun repository [1] across all pose estimation metrics.

Comparison with Pose-aware Methods. Our framework demonstrates superior robustness compared to pose-aware methods like ViewCrafter [13] and Gen3R [6]. ViewCrafter relies on a serial framework that is highly dependent on initial pose estimates. Under extreme-view trajectories, these initial estimates are often inaccurate, leading to geometric collapse in the synthesized results. While Gen3R applies reconstruction losses for geometric alignment, such standard geometric constraints are often insufficient and suffer from instability under extreme trajectories with large spatial gaps.

In contrast, ExPose explicitly recognizes the causal relationship between camera pose and video generation. By prioritizing structural alignment through RL, the model inherently recovers the geometric cues required for accurate pose estimation. This enables ExPose to maintain global consistency and structural fidelity even in extreme scenarios.

4. Limitations and Failure Cases

While semantic consistency is generally maintained by SFT and KL regularization, which constrain the generator to realistic video priors, our method has limitations. Typical failure cases include dynamic objects that violate the static scene assumptions, and extreme temporal shifts that create appearance ambiguities.

5. Video Generation Results

We present additional qualitative results that illustrate how ExPose improves the geometric consistency of generated videos across all four evaluation datasets. Shown in Fig. 1–8 are examples that consist of reference and target frames, and the corresponding sets of generated frames, comparing ExPose with recent video generation baselines (LTX-Video [5], Aether [15], DynamiCrafter [11], and InterPose [3]). Figures 1 and 2 shows results in ScanNet dataset, Fig. 3 and 4 shows results in NAVI dataset, Fig. 5 and 6 shows results in DL3DV dataset, and Fig. 7 and 8 shows results in Cambridge Landmarks dataset.

Across indoor, outdoor, and object-centric scenarios, ExPose produces video trajectories that more faithfully follow the camera motion representing a valid underlying geometry. Competing models often exhibit drift, abrupt viewpoint changes, or scene hallucination, especially under extreme-baseline settings. In contrast, ExPose maintains smooth transitions, stable scene geometry, and sharper alignment with the target frame. These improvements are particularly notable in scenes with strong parallax or large rotations.

6. Pseudo Video Datasets for SFT

We provide additional qualitative examples of the pseudo videos used for supervised finetuning (SFT). Each pseudo video is assembled from a reference frame, an auxiliary frame, and a target frame sampled from the training set. These triplets serve as short, 3-view trajectories that guide the model to learn smooth intermediate motion and consistent viewpoint transitions.

We visualize pseudo videos generated with and without the auxiliary frame. Without an auxiliary frame, as shown in Fig. 9, the interpolation between reference and target frames often lacks geometric stability and may introduce drift or distortions. In contrast, including an auxiliary frame as shown in Fig.10 produces significantly smoother trajectories and more stable scene structure, demonstrating the benefit of triplet-based supervision for establishing a reliable prior during SFT.

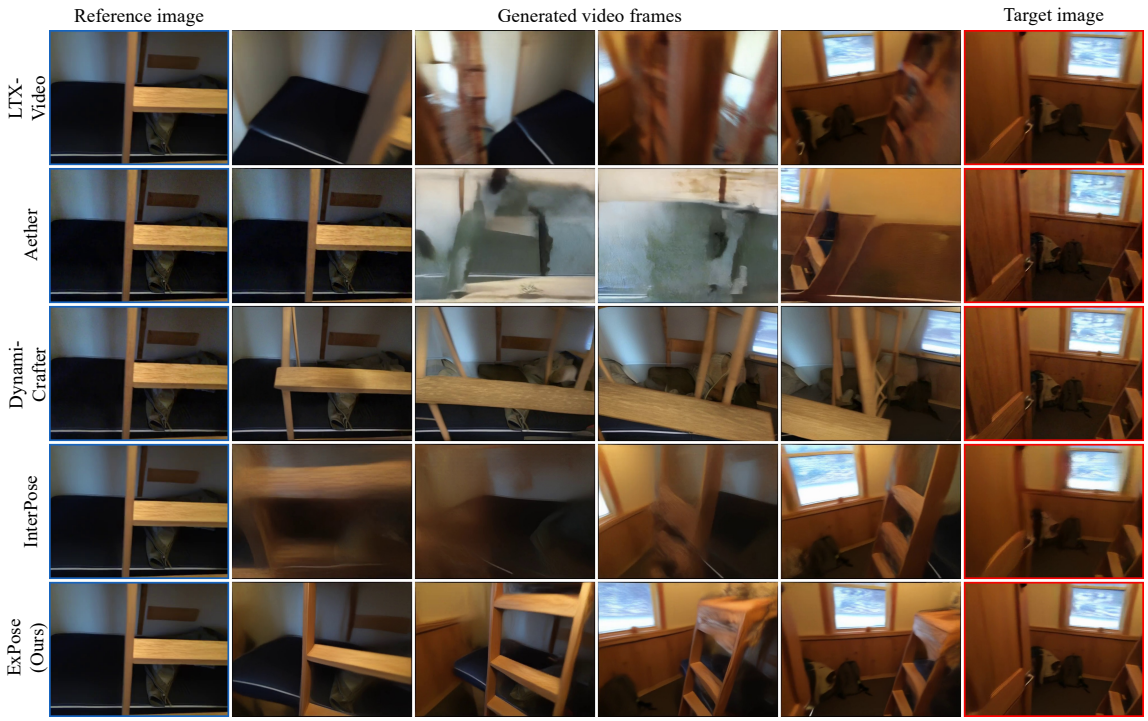

Figure 1. **Generated video frames on the ScanNet dataset (1).** Given a reference and a target frame, ExPose produces smoother camera trajectories and more stable indoor geometry compared to LTX-Video, Aether, DynamiCrafter, and InterPose. Each row shows the reference image, intermediate generated frames, and the target image.

Figure 2. **Generated video frames on the ScanNet dataset (2).** ExPose better preserves room layout and object structure, while baseline methods often exhibit drift, warping, or inconsistent geometry along the trajectory.

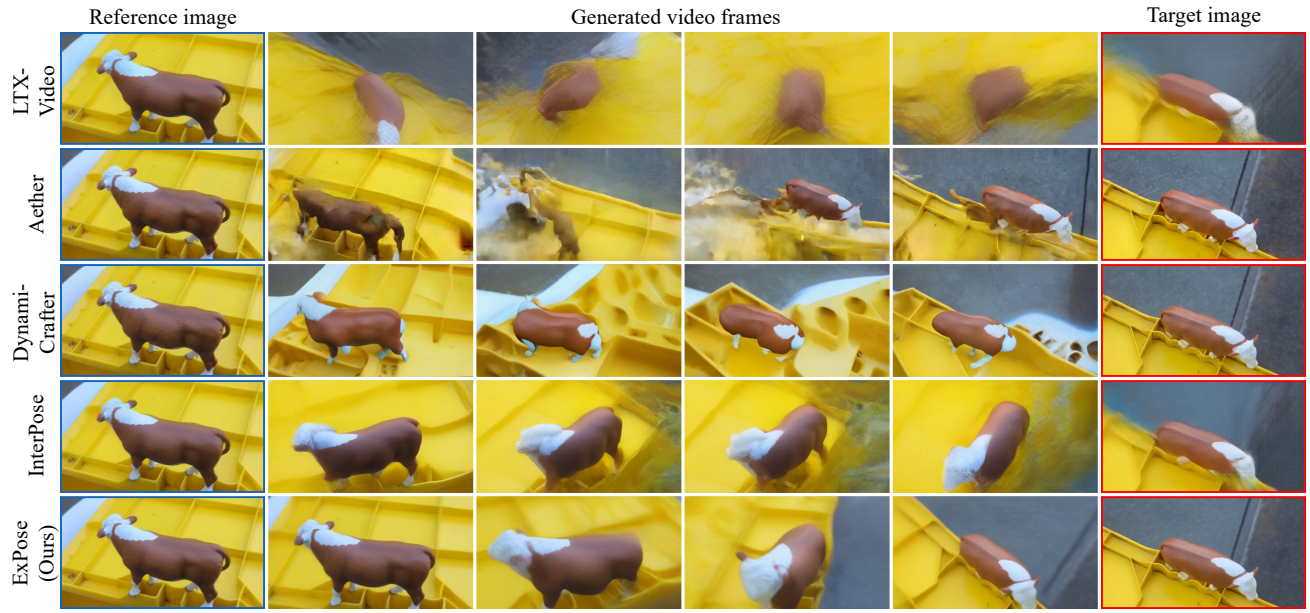

Figure 3. **Generated video frames on the NAVI dataset (1).** For object-centric rotations, ExPose yields 3D-consistent camera motion and preserves object identity across views, whereas baseline models tend to deform object shape or lose fine details under extreme-baseline configurations.

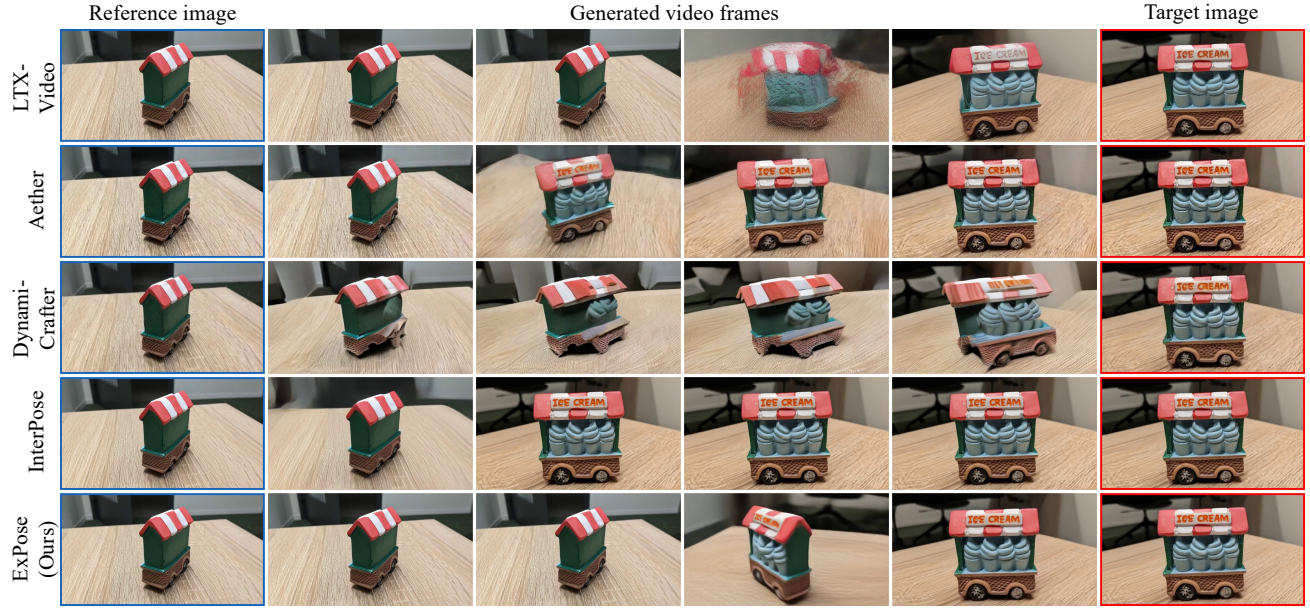

Figure 4. **Generated video frames on the NAVI dataset (2).** Even under large rotations, ExPose maintains coherent object geometry and background structure, while competing methods frequently introduce flickering, shape distortions, or inconsistent viewpoints.

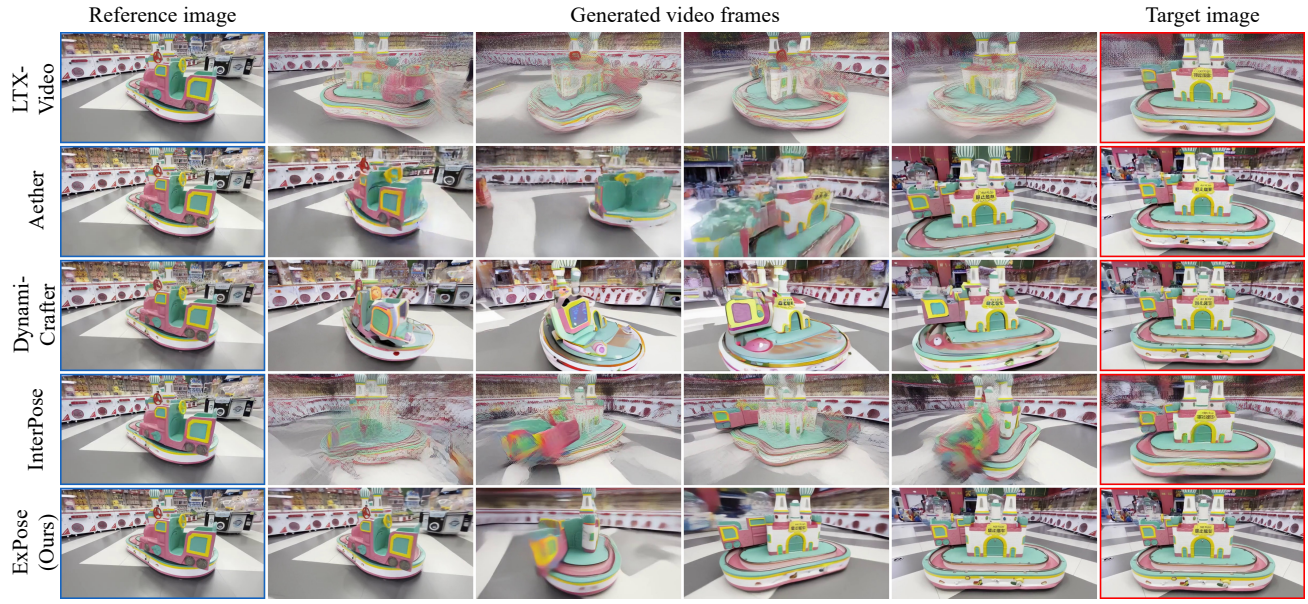

Figure 5. **Generated video frames on the DL3DV dataset (1).** In large-scale 3D scenes, ExPose produces trajectories that more faithfully follow a valid underlying geometry, keeping global structure intact and alleviating the severe distortions observed in other video generation baselines.

Figure 6. **Generated video frames on the DL3DV dataset (2).** ExPose better preserves geometries of both near and far distances over long trajectories, while baseline models often struggle to maintain consistent depth relationships and coherent scene layout across frames.

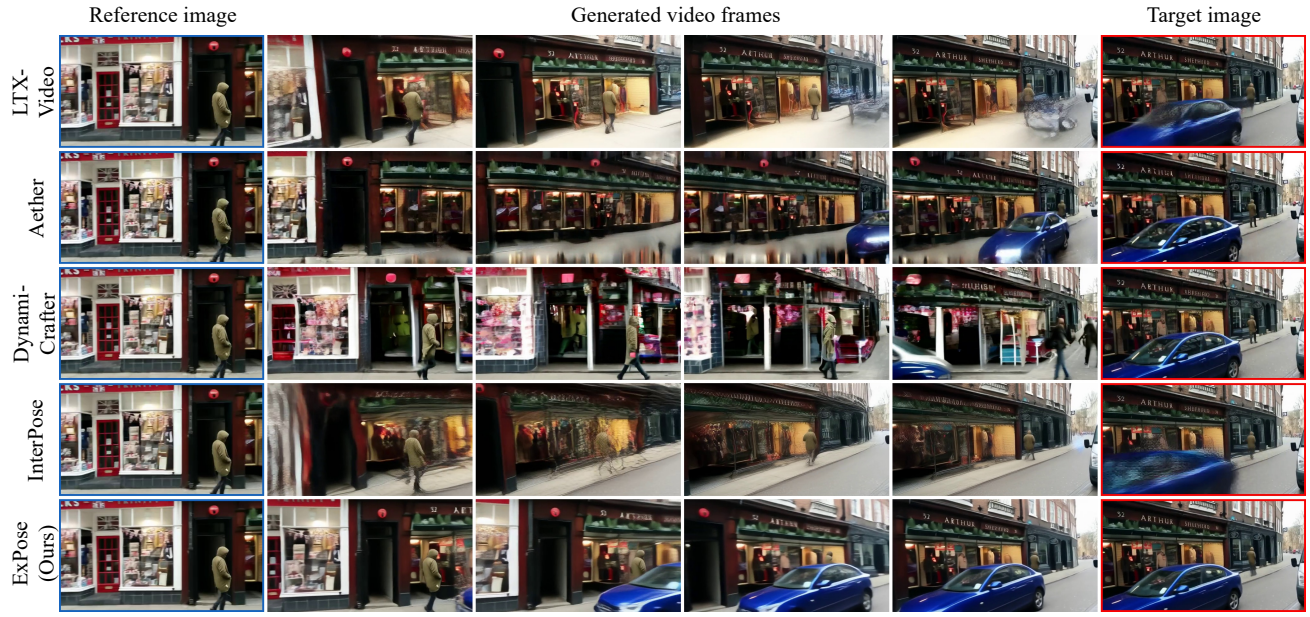

Figure 7. **Generated video frames on the Cambridge Landmarks dataset (1).** For outdoor landmark scenes, ExPose produces smooth camera orbits and maintains the global structure of buildings, whereas alternative methods frequently exhibit jitter, viewpoint jumps, or hallucinated facades under extreme-baseline conditions.

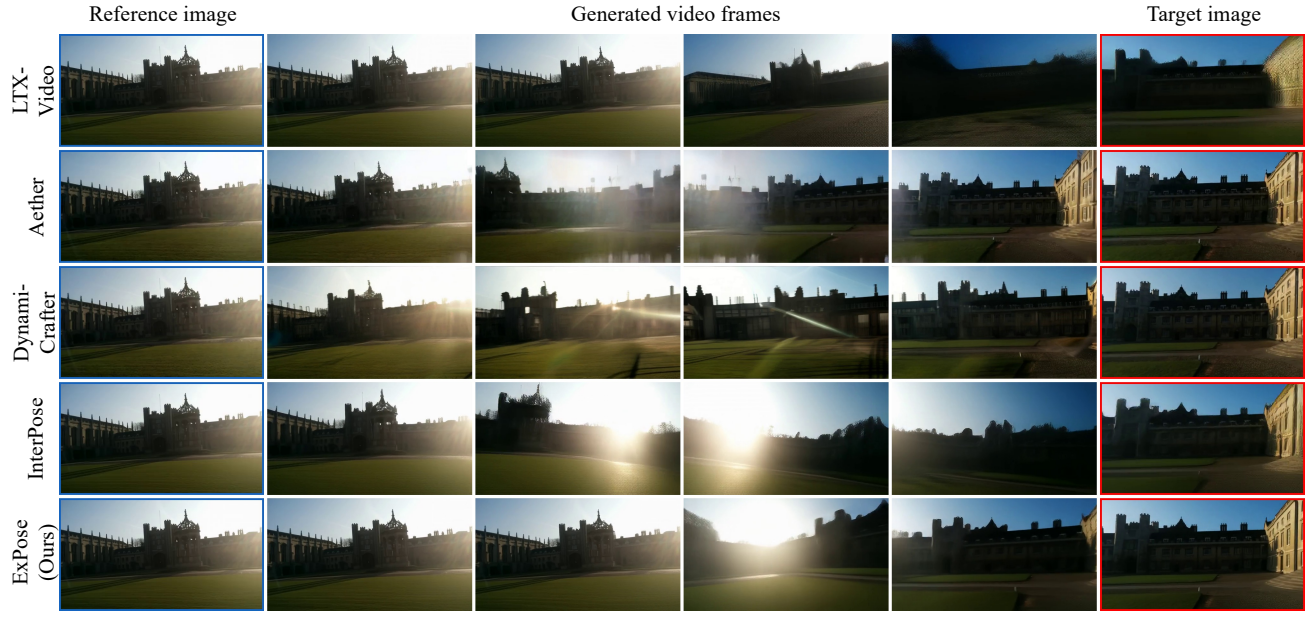

Figure 8. **Generated video frames on the Cambridge Landmarks dataset (2).** ExPose more reliably aligns the final generated frame with the target viewpoint, yielding sharper and more geometrically consistent landmarks compared to LTX-Video, Aether, DynamiCrafter, and InterPose.

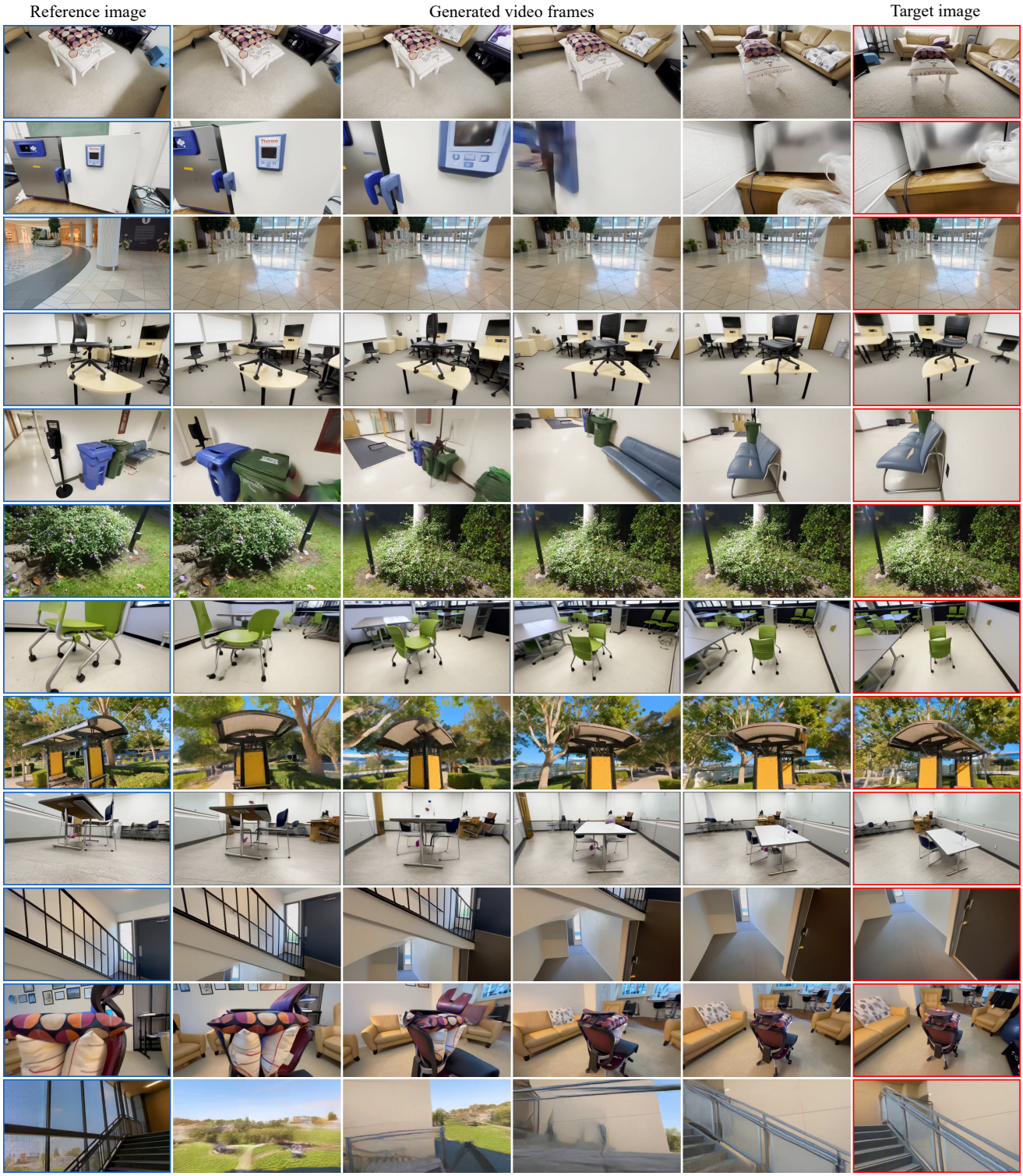

Figure 9. **Generated pseudo video without an auxiliary frame.** When prompted only with a pair of reference and target frame having extreme viewpoint changes, the interpolated video trajectory suffers from drift in geometry, unstable camera motion, and distortions of objects.

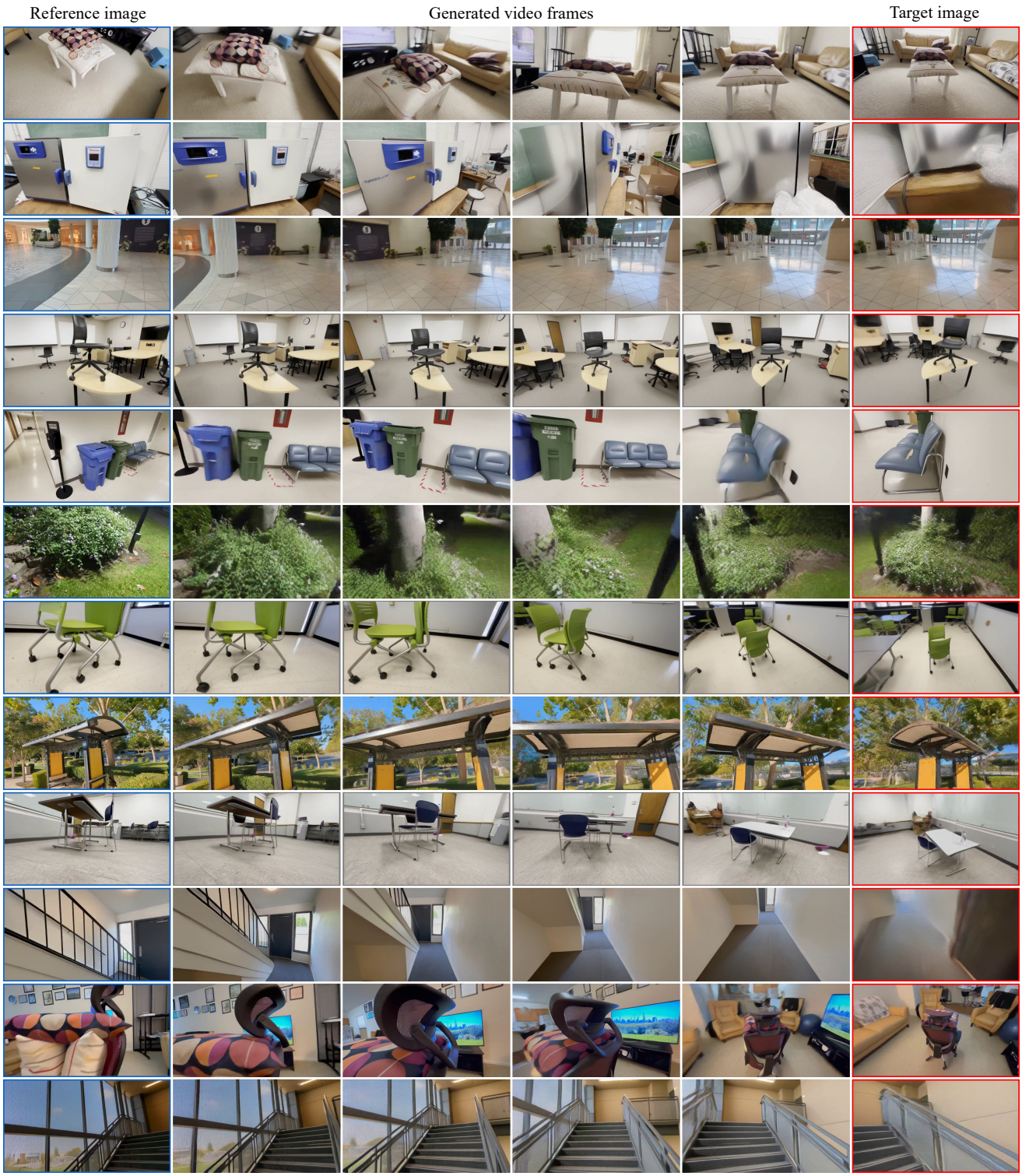

Figure 10. **Generated pseudo video with an auxiliary frame.** Adding an auxiliary view to the set of image prompts yields a much smoother and more geometrically stable pseudo trajectory, providing strong supervision target for SFT, which then serves to better shape the model's prior for interpolating intermediate camera poses when given only a pair of images.

References

[1] https://github.com/aigc-apps/VideoX-Fun. 2

[2] Hana Bezalel, Dotan Ankri, Ruojin Cai, and Hadar Averbach-Elor. Extreme rotation estimation in the wild. In *CVPR*, pages 1061–1070, 2025. 1

[3] Ruojin Cai, Jason Y Zhang, Philipp Henzler, Zhengqi Li, Noah Snavely, and Ricardo Martin-Brualla. Can generative video models help pose estimation? In *CVPR*, pages 16764–16773, 2025. 3

[4] Yihang Chen, Qianyi Wu, Mengyao Li, Weiyao Lin, Mehrtash Harandi, and Jianfei Cai. Fast feedforward 3d gaussian splatting compression. In *ICLR*, 2025. 1

[5] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 3

[6] Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction. *arXiv preprint arXiv:2601.04090*, 2026. 2

[7] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. NAVI: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *NeurIPS*, 2023. 1

[8] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 2

[9] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2

[10] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 2

[11] 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 *ECCV*, pages 399–417. Springer, 2024. 3

[12] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

[13] 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

[14] 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. 1

[15] Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *ICCV*, pages 8535–8546, 2025. 3