ROVER: INTERACTION-PRESERVING VIDEO RESHOOTING

Anonymous authors
Paper under double-blind review

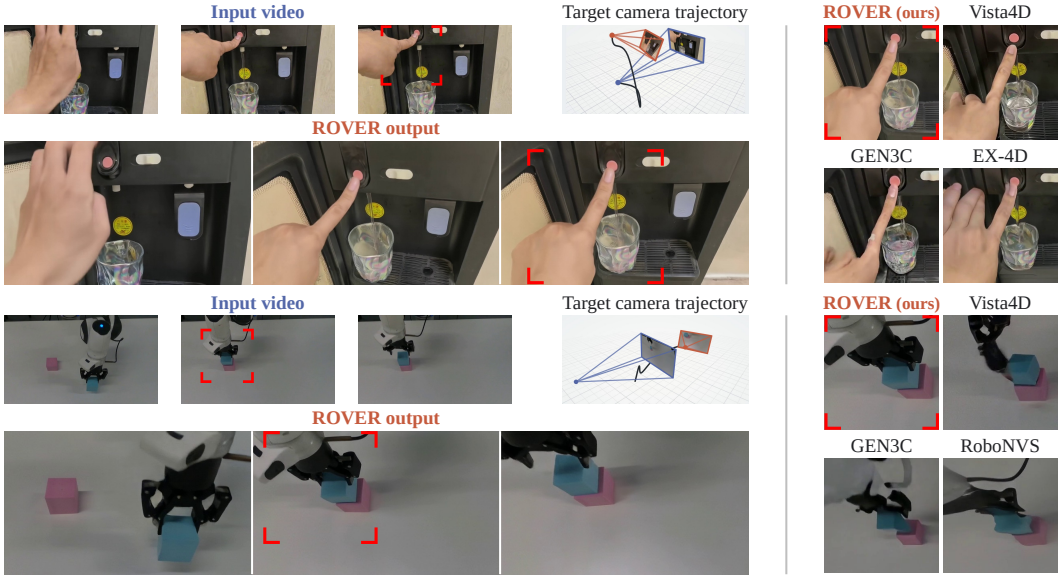

Figure 1: **Interaction-preserving video reshooting with ROVER.** Given an input video and a target camera trajectory, ROVER synthesizes novel views while preserving human–object and robot–object interactions. Blue and orange denote source and target cameras, respectively. Video results and comparisons are available on our **project page**.

ABSTRACT

Extending video reshooting from general scenes to object interactions requires preserving contact, motion, and object state across viewpoints. However, existing synchronized multi-view datasets offer limited coverage of interaction dynamics under diverse camera trajectories. We present **ROVER** for interaction-preserving video reshooting. Our approach distills interaction and dynamics priors from a pre-trained video generator through **OBJALIVE**, a collection of 13.5K filtered videos produced via image-to-video generation. The collection covers contact-rich manipulation, fluid motion, deformation, topology changes, articulation, and appearance changes. We combine it with synchronized multi-view and real single-view videos to fine-tune a video diffusion model conditioned on source videos and target-view warps. Blinded four-way human evaluation favors ROVER, which receives 71.6% and 63.6% of interaction-fidelity votes on held-out human–object interaction and robotic manipulation videos, respectively. These results indicate generalization to robotic manipulation without robot-specific fine-tuning. A controlled, step-matched ablation demonstrates the benefit of OBJALIVE for preserving contact and non-rigid dynamics.

1 INTRODUCTION

Generative video models open a path to embodied simulation, from creating 3D environments (Shen et al., 2026) to synthesizing observations for closed-loop driving (NVIDIA, 2026). Extending these capabilities toward manipulation requires consistent contact, deformation, and event timing across

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

viewpoints. We study *interaction-preserving video reshooting*: given a source video and a user-specified camera trajectory, synthesize new views that preserve the observed interaction (Figure 1).

Despite progress on general dynamic scenes (Bai et al., 2025a; Yu et al., 2025; Ren et al., 2025; Lin et al., 2026b), fine-grained interactions remain challenging. Hands, tools, and robot arms occlude objects and contact regions that become visible from new viewpoints. Plausible outputs can still misrepresent hand geometry, contact relationships, or object-state changes. In geometry-guided methods, inaccurate depth further distorts thin structures and contact regions, requiring the model to correct projections while synthesizing newly visible content (Figure 2).

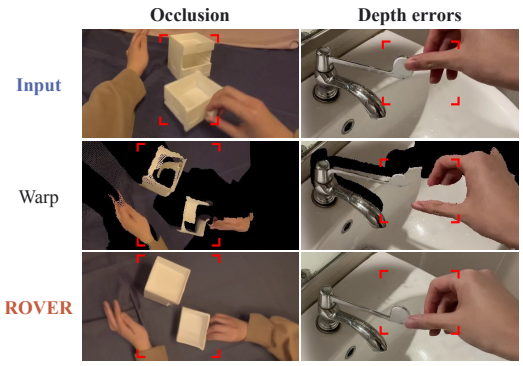

Figure 2: ROVER fills missing regions and restores finger–handle contact.

Synchronized multi-view videos, captured in real environments or rendered in simulation (Chao et al., 2021; Wang et al., 2025a; Bai et al., 2025a; Jiang et al., 2026b; Zhou et al., 2026), provide direct cross-view supervision. However, general dynamic-scene datasets offer varied camera motion but few contact-rich events, while specialized interaction datasets often use fixed cameras and cover limited objects and actions. Real single-view videos (Liu et al., 2022) add diversity, but covering specific object–event combinations still requires finding or recording suitable examples.

Our key finding is that *interaction and dynamics priors learned by a pretrained video generator can improve reshooting through generated training videos*. We distill these priors by training a reshooting model to reconstruct generated clips from incomplete geometric observations, using pseudo-view pairs constructed through double reprojection (Yu et al., 2025). The generator supplies plausible events, while the reshooting model learns to preserve them under viewpoint changes.

We instantiate this strategy with **OBJALIVE**, a collection of 13.5K filtered I2V-generated videos. Starting from object-centric images (Han et al., 2024; Reizenstein et al., 2021; Liu et al., 2025), we generate videos with event-specific prompts using a pretrained I2V model (Wan et al., 2025), filtering out implausible dynamics and visual artifacts. The collection covers contact-rich manipulation, articulation, non-rigid deformation, topology changes, fluid motion, and appearance changes associated with object-state transitions. These clips provide training examples without requiring synchronized multi-view capture. Camera trajectories can be sampled independently of the generated events. We combine these generated videos with synchronized multi-view and supplementary real single-view videos, retaining direct cross-view supervision while broadening interaction coverage.

Using this training mixture, we develop **ROVER** for interaction-preserving video reshooting by fine-tuning a pretrained video diffusion model (Wan et al., 2025). We retain the unwarped source as a separate condition to anchor appearance and interaction dynamics, while target-view warps provide spatially aligned but incomplete observations. Camera rays specify viewpoints independently of warp visibility, while timestamp-aligned coordinate maps associate warped observations with source pixels. We inject these encodings before query and key projections to guide attention between source and target tokens. We further extend reshooting to long videos through history conditioning and static background memory.

Blinded four-way human evaluation favors ROVER in interaction fidelity, visual quality, and camera control, with 71.6% and 63.6% of interaction-fidelity votes on held-out human–object and robotic manipulation videos, respectively (Table 1). These results indicate generalization to robotic manipulation without robot-specific fine-tuning. In a controlled, step-matched ablation, 69.0% of interaction-fidelity judgments favor the model trained with OBJALIVE. Qualitative comparisons show better preservation of contact and non-rigid dynamics (Figure 8 and Appendix D.1).

Our contributions are: (i) we distill a pretrained video generator's interaction and dynamics priors into reshooting through generated single-view videos; (ii) we construct OBJALIVE, 13.5K filtered I2V-generated videos that broaden interaction and dynamics coverage without additional synchronized

capture; and (iii) we develop ROVER for interaction-preserving reshooting, including long videos, and demonstrate its generalization to robotic interactions without robot-specific fine-tuning.

2 RELATED WORK

Camera-controlled video generation and reshooting. Camera-controlled video generation spans text-to-video (Wang et al., 2024; He et al., 2024; Bahmani et al., 2025b; Yang et al., 2024; Bahmani et al., 2025a; Kuang et al., 2024; Bai et al., 2025b) and image-to-video (Xu et al., 2024; Zheng et al., 2024; Li et al., 2025b; Feng et al., 2025; Watson et al., 2025; He et al., 2025). Video reshooting additionally conditions on a source sequence to preserve its appearance and dynamics under a new camera trajectory. Approaches without explicit scene reconstruction learn view transformations using camera embeddings (Van Hoorick et al., 2024; Bai et al., 2025a; Fu et al., 2026; Li et al., 2026) or camera motion from reference videos (Luo et al., 2025). Geometry-conditioned methods instead render reconstructed scene geometry into target view (Yu et al., 2025; Hu et al., 2025; Ren et al., 2025; Lin et al., 2026b; Chen et al., 2025a; Cao et al., 2026; Zhu et al., 2026). Other methods explore trajectory-based attention (Xiao et al., 2025), masked per-video fine-tuning (Jeong et al., 2025; Zhang et al., 2025b), or training-free sampling guided by warped observations (You et al., 2024). Our model combines target-view warps, Plücker camera embeddings (Sitzmann et al., 2021), and coordinate-based source–target correspondences. Building on these conditioning mechanisms, we focus on interaction-preserving supervision, complementing synchronized multi-view and real single-view videos with I2V-generated examples to broaden interaction and dynamics coverage.

Novel-view synthesis for human and robotic interactions. Novel-view synthesis of human–object interactions has been explored through neural reconstruction (Liu et al., 2023; Qu et al., 2023; Choi et al., 2024) and Gaussian representations (On et al., 2025; Zhang et al., 2025a). In robotics, view synthesis and controllable video generation support demonstration augmentation and manipulation policy learning (Tian et al., 2024; Cai et al., 2026; Gu et al., 2026; Chen et al., 2026), while multi-view video generation provides synchronized views of robotic manipulation (Fu et al., 2026). Related work targets human egocentric video generation from exocentric observations (Luo et al., 2024; Kang et al., 2026), or robot wrist-view generation from external views (Qian et al., 2025). We study object-centric camera control across human and robotic interactions, preserving action timing, contact, and object state. We complement synchronized captures such as DexYCB (Chao et al., 2021) and HO-Cap (Wang et al., 2025a) with generated examples to expand interaction coverage without requiring additional synchronized captures for every event.

3 METHOD

3.1 PROBLEM SETUP AND OVERVIEW

Given a source video $\mathbf{V}^{\text{src}} = \{V_t^{\text{src}}\}_{t=0}^{N-1}$ and a target camera trajectory $\mathbf{C}^{\text{tgt}} = \{C_t^{\text{tgt}}\}_{t=0}^{N-1}$, we synthesize a target-view video $\mathbf{V}^{\text{tgt}} = \{V_t^{\text{tgt}}\}_{t=0}^{N-1}$ while preserving the observed contact, motion, deformation, and object-state transitions at matched timestamps.

Our key idea is to distill interaction and dynamics priors from a pretrained video generator into a camera-controlled video reshooting model through reconstruction. OBJALIVE supplies generated single-view videos, and ROVER learns to reconstruct the original clips from incomplete geometric observations produced by virtual-view reprojection. We complement this reconstruction training with real single-view videos and direct cross-view supervision from synchronized multi-view recordings.

3.2 OBJALIVE: INTERACTIONS AND OBJECT DYNAMICS

To broaden the interactions and object dynamics available for training, we construct OBJALIVE, a collection of 13.5K filtered videos produced through image-to-video generation. Starting from object-centric images, we specify events involving contact, articulated motion, deformation, and changes in object state to generate examples for interaction-preserving reshooting.

Selection and generation. We begin with real images from object-centric multi-view collections, including MVImgNet2.0 (Yu et al., 2023; Han et al., 2024), CO3D (Reizenstein et al., 2021), and uCO3D (Liu et al., 2025). Gemini (Team et al., 2023) identifies object categories whose instances can exhibit meaningful dynamics, and we sample instances from each selected category. For each instance, we select the image that maximizes mask-based object visibility. We additionally generate

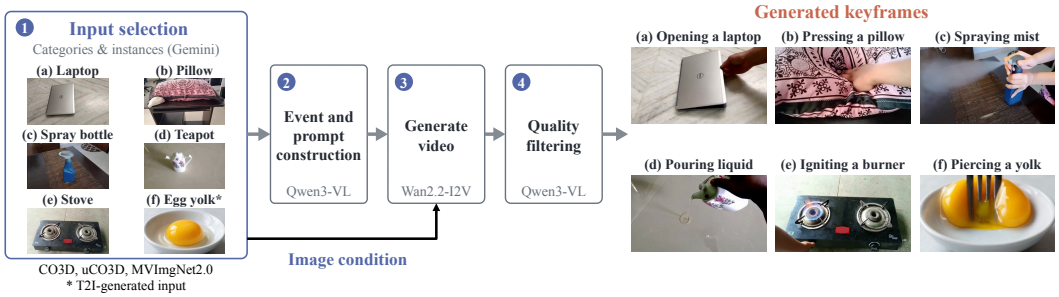

Figure 3: **Construction of OBJALIVE.** Selected object images and event prompts guide I2V generation, followed by quality filtering to obtain 13.5K videos. Representative keyframes illustrate the resulting object interactions and diverse dynamics.

object-centric images using Qwen-Image-2512 (Wu et al., 2025) to further diversify the inputs. For each input image, Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025c) proposes a plausible dynamic event and refines its description into a detailed I2V prompt. Wan2.2-I2V-A14B (Wan et al., 2025) then generates a video conditioned on both the image and prompt. Figure 3 summarizes the pipeline.

Filtering. We first remove clips with insufficient dynamics using motion-vector statistics and frame differences. We then use Qwen3-VL to discard clips with implausible dynamics or obvious visual artifacts, while preserving plausible changes in shape, appearance, and physical state. We additionally perform manual review to identify errors missed by the automated filtering. These steps help reduce generation errors in the training supervision.

Coverage. Figure 4 summarizes seven interaction and dynamics categories in OBJALIVE. Alongside hand–object interaction and rigid motion, the collection covers non-rigid deformation, articulated motion, fluid dynamics, appearance changes, and topology changes. These videos include both contact-rich human–object manipulation and object dynamics without human involvement. OBJALIVE contains no robotic manipulation videos. For training, we construct pseudo-view pairs under sampled interaction-centric camera trajectories, independently of the generated events, as described in Section 3.3.2.

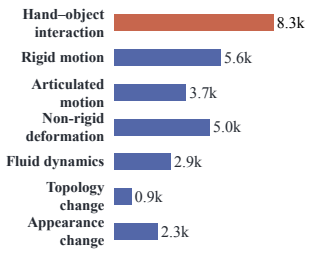

Figure 4: **OBJALIVE coverage.**

3.3 ROVER: INTERACTION-PRESERVING VIDEO RESHOOTING

ROVER uses target-view warps for spatially aligned guidance and retains the unwarped source as a separate condition to anchor appearance and interaction dynamics when reprojections are incomplete or distorted. Camera and coordinate encodings guide matching between source and target tokens (Figure 5). We first describe the conditioning inputs and model architecture, then present reconstruction-based training and the extension to long-video generation.

3.3.1 CONDITIONING AND ARCHITECTURE

Source and warp conditions. We denote source and target cameras by $C_t^a = (K_t^a, P_t^a)$, where $a \in \{\text{src}, \text{tgt}\}$, K_t^a contains the intrinsics, and P_t^a is the camera-to-world transformation. Source depths $\mathbf{D}^{\text{src}}$ and source camera parameters are obtained from dataset annotations when available, or estimated using DA3 (Lin et al., 2025). Depth-aware forward reprojection $\mathcal{R}$ (Ren et al., 2025; Yu et al., 2025) maps source observations to the target cameras at matched timestamps:

$$(\mathbf{W}, \mathbf{M}^w) = \mathcal{R}\big(\mathbf{V}^{\text{src}}, \mathbf{D}^{\text{src}}, \mathbf{C}^{\text{src}}, \mathbf{C}^{\text{tgt}}\big). \tag{1}$$

The warp $\mathbf{W}$ provides spatially aligned appearance, while $\mathbf{M}^w$ marks observed pixels with one and holes with zero. Since observed warp pixels can also contain depth-induced distortions, we retain the original source video as a separate condition providing appearance and interaction dynamics.

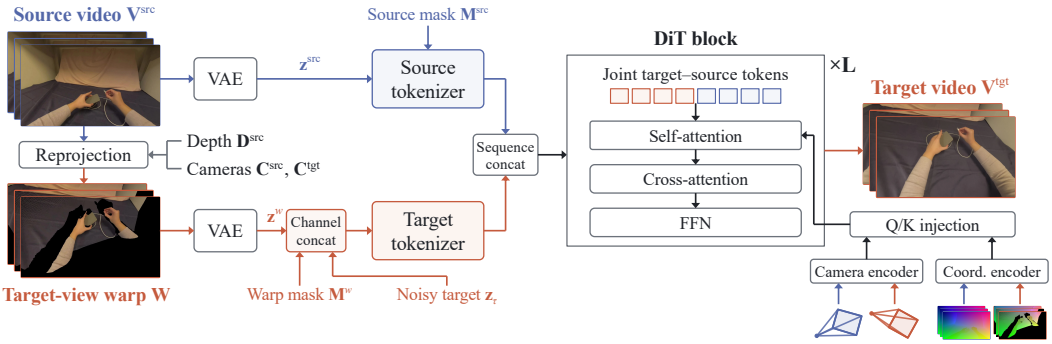

Figure 5: **Overview of ROVER.** We reproject the source video using depth and cameras to obtain a target-view warp and its validity mask. The noisy target latent, mask, and warp latent are concatenated along the channel dimension and tokenized, then processed jointly with source tokens. In each DiT block, camera and coordinate encoders embed source/target Plücker rays and source/warped coordinate maps into the query/key inputs of self-attention, guiding source–target matching.

Camera and coordinate encodings. We encode source and target cameras using per-pixel Plücker rays (He et al., 2024; Sitzmann et al., 2021), providing viewpoint information independently of warp visibility. All cameras are expressed relative to the first target camera, $\widetilde{P}_t^a = (P_0^{\mathrm{tgt}})^{-1} P_t^a$.

To complement these viewpoint cues with pixel correspondences, we adapt coordinate-based conditioning to dynamic video view synthesis. Rather than deriving coordinate conditions from a single image (Seo et al., 2024) or retrieved history frames (Shen et al., 2026), we construct source and warped coordinate maps at matched timestamps throughout the input video. Each source pixel receives the descriptor $\mathbf{c}_t^{\mathrm{src}}(u,v) = (\bar{u}, \bar{v}, \bar{D}_t^{\mathrm{src}}(u,v))$, where image coordinates are normalized to $[-1,1]$ and depth is min–max normalized over the source clip. Reprojecting these descriptors alongside RGB produces $\mathbf{c}^{\mathrm{tgt}}$, whose valid locations retain the corresponding source coordinates and depth. We pack both ray and coordinate descriptors onto the DiT token grid to guide attention.

Joint processing and attention. We initialize from Wan2.1-I2V-14B (Wan et al., 2025) and encode the source, warp, and target videos with its frozen VAE $\mathcal{E}$. Noise is added only to the target latent, while the source and warp latents remain clean conditioning inputs. The noisy target latent, warp latent, and warp-validity mask are concatenated along the channel dimension and projected into target tokens. Source latents are tokenized separately, with a learned mask embedding (Lin et al., 2026b) indicating missing or dropped source observations. We concatenate target and source tokens into a single sequence (Bai et al., 2025a) and process them jointly through the DiT (Peebles & Xie, 2023), predicting only the target segment.

In each self-attention block, we separately encode Plücker rays and coordinate maps and sum their embeddings to obtain $\mathbf{e}_i$ for token i. We add $\mathbf{e}_i$ to the attention-input feature $\mathbf{h}_i$ before the query and key projections, leaving the value projection unchanged (Shen et al., 2026):

$$\widehat{\mathbf{q}}_i = W_Q(\mathbf{h}_i + \mathbf{e}_i), \qquad \widehat{\mathbf{k}}_i = W_K(\mathbf{h}_i + \mathbf{e}_i), \qquad \mathbf{v}_i = W_V \mathbf{h}_i. \tag{2}$$

This allows camera geometry and pixel correspondences to guide attention between source and target tokens. The ray projections and final layers of the coordinate encoders are zero-initialized, so the added embeddings do not affect attention at initialization.

3.3.2 DISTILLING INTERACTION PRIORS VIA RECONSTRUCTION

Single-view reconstruction. For a clip $\mathbf{V}$ from OBJALIVE with cameras $\mathbf{C}$, we sample a virtual interaction-centric trajectory $\mathbf{C}^{\mathrm{virt}}$ and apply double reprojection (Yu et al., 2025):

$$\mathbf{V} \xrightarrow{\mathbf{C}^{\mathrm{virt}}} \widetilde{\mathbf{V}}^{\mathrm{src}} \xrightarrow{\mathbf{C}} \widetilde{\mathbf{W}}. \tag{3}$$

The model reconstructs $\mathbf{V}$ from the pseudo-source $\widetilde{\mathbf{V}}^{\mathrm{src}}$ and its target-view warp $\widetilde{\mathbf{W}}$, using validity masks from the two reprojections. Reconstructing generated clips from incomplete observations distills the generator's interaction and dynamics priors into the reshooting model. The same procedure applies to real single-view clips (Liu et al., 2022).

Complementary multi-view supervision. For synchronized multi-view videos (Bai et al., 2025a; Jiang et al., 2026b; Zhou et al., 2026; Chao et al., 2021; Wang et al., 2025a), we select source and target clips from different cameras over the same temporal window. The source clip provides the conditioning inputs, while the synchronized target clip supplies direct cross-view supervision. Combining these pairs with single-view reconstruction broadens the coverage of interactions and dynamics while retaining supervision from observed target views. Table 2 in Appendix B.1 summarizes the full training mixture.

Training objective. We optimize both types of training pairs using flow matching (Lipman et al., 2022). Given the target latent $\mathbf{z}_0 = \mathcal{E}(\mathbf{V}^{\text{tgt}})$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we construct the noisy latent $\mathbf{z}_\tau = (1-\tau)\mathbf{z}_0 + \tau\boldsymbol{\epsilon}$ at timestep $\tau \in [0, 1]$ and minimize

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{(\mathbf{z}_0, \mathcal{C}),\, \tau,\, \boldsymbol{\epsilon}} \left[\| v_\theta(\mathbf{z}_\tau, \tau; \mathcal{C}) - (\boldsymbol{\epsilon} - \mathbf{z}_0) \|_2^2 \right], \tag{4}$$

where v_θ denotes the DiT model and $\mathcal{C}$ collects the source video, warp, validity masks, and camera and coordinate conditions. The target video is the original clip for single-view reconstruction and the synchronized target view for multi-view training.

3.3.3 Long-Video Generation

We extend generation to long videos through overlapping windows and 3D background memory. We fine-tune a history-conditioned model that generates the first window without history and subsequent windows conditioned on overlapping frames from the preceding output. To reuse content beyond the overlap, we reconstruct static background from generated views and accumulate it in a shared 3D memory, excluding dynamic objects. We render this memory at subsequent target viewpoints to fill missing regions in the source-to-target warp while preserving valid source observations. Appendix A.4 details history-conditioned training, the window schedule, and memory construction.

4 Experiments

4.1 Experimental Setup

Implementation details. We fine-tune the self-attention layers and conditioning modules of Wan2.1-I2V-14B (Wan et al., 2025), while freezing the video VAE, cross-attention layers, and feed-forward layers. The final model generates 49-frame videos at a resolution of 832×480. During training, source and warp conditions are independently dropped with probability 0.1. At inference, we use 50 sampling steps and empty text prompts. Appendix A provides further details.

Baselines. For human–object interactions, we compare with EX-4D (Hu et al., 2025), GEN3C (Ren et al., 2025), and Vista4D (Lin et al., 2026b), which use depth-derived meshes, 3D caches, and 4D point clouds as geometric guidance, respectively. Robotic comparisons include GEN3C, Vista4D, and RoboNVS (Cai et al., 2026), which uses depth-based guidance. Additional comparisons, including TrajectoryCrafter (Yu et al., 2025), ReCamMaster (Bai et al., 2025a), and CamCloneMaster (Luo et al., 2025), are provided in the appendix.

Table 1: **Human evaluation of method comparisons and ablations.** We report preference percentages from 21 participants on 25 videos per domain and 10 per ablation. (a,b) Four-way preference shares. (c) Preference for training with ObjAlive. (d) No clear difference with versus without robotic fine-tuning. Ablation percentages include ties in the denominator.

(a) Human–object interactions				*(b) Robotic manipulation*			
Method	Interaction	Quality	Camera	Method	Interaction	Quality	Camera
EX-4D	6.1	5.3	8.2	RoboNVS	8.6	7.6	11.2
GEN3C	8.8	7.8	9.3	GEN3C	8.6	8.0	10.1
Vista4D	13.5	17.7	16.8	Vista4D	19.2	20.4	19.2
Rover (ours)	**71.6**	**69.1**	**65.7**	Rover (ours)	**63.6**	**64.0**	**59.4**
(c) ObjAlive *ablation*				*(d) Robotic fine-tuning ablation*			
Response	Interaction	Quality	Camera	Response	Interaction	Quality	Camera
Prefer w/	69.0	72.4	57.1	No clear diff.	74.3	71.9	81.0

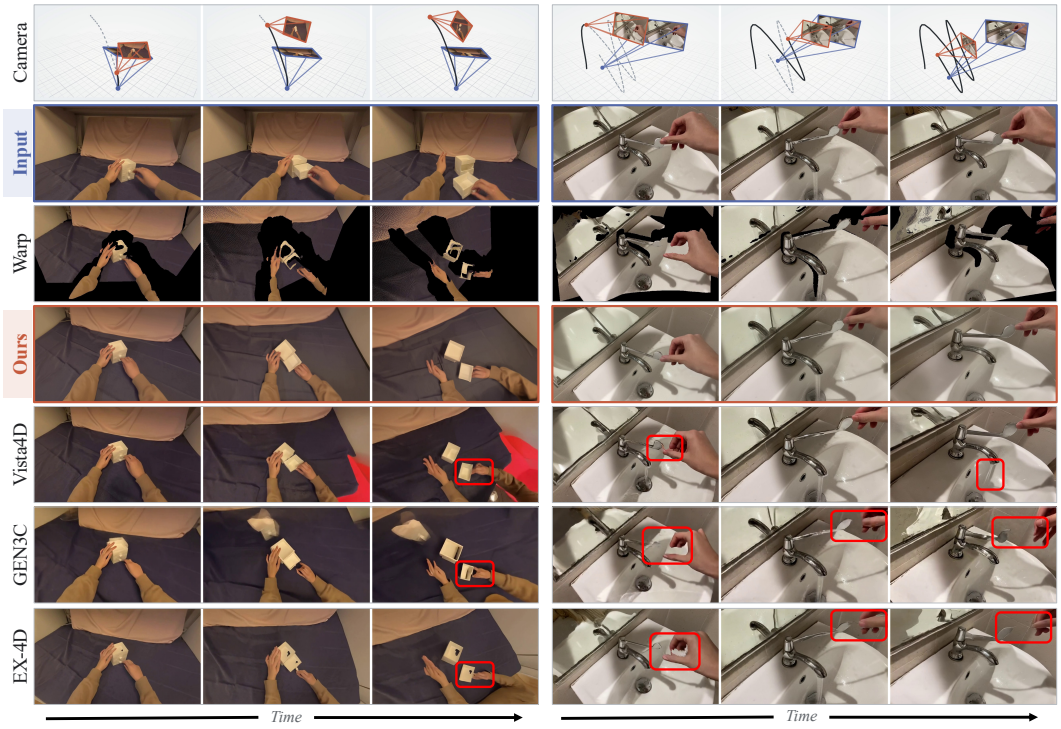

Figure 6: **Qualitative comparison on human–object interaction videos.** On *drawer opening* (Hoque et al., 2026) and *turning a faucet on and off* (Song et al., 2024), our method better preserves hand–object contact, object structure, and interaction state. Camera diagrams: blue for source, orange for target cameras. Additional results are in Appendix E.1, with video comparisons on the project page.

Evaluation datasets. We curate held-out human–object interaction clips from EgoDex (Hoque et al., 2026), M^3-VOS (Chen et al., 2025b), SWITCH (Cheng et al., 2025), WhereToChange (Mandikal et al., 2025), and real-world articulated-object videos (Song et al., 2024). These clips cover dexterous hand–object manipulation, non-rigid deformation, topology changes, articulated motion, fluid dynamics, and changes in appearance and object state. Small or thin objects pose additional challenges under large viewpoint changes.

Robotic evaluation uses held-out videos from DROID (Khazatsky et al., 2024), VITRA (Li et al., 2025a), ArmNetBench (Selvaraj et al., 2026), RoboMIND (Wu et al., 2024), and SynthVerse (Zhao et al., 2026). We use source videos from fixed external viewpoints. The evaluation covers different embodiments and both single- and dual-arm manipulation, including pick-and-place, insertion, and tool use. We include full-body robot manipulation videos from prior work. Their sources and qualitative results appear in Appendix E.2, with videos available on the project page.

For both domains, we use single-view inputs and object-centric target trajectories without synchronized target-view ground truth. We estimate source depth and cameras with DA3 (Lin et al., 2025), except for SynthVerse (rendered depth and provided cameras) and RoboMIND (sensor depth and reference intrinsics, with an identity source pose). Complementary quantitative results on DyCheck (Gao et al., 2022), which provides synchronized target-view ground truth, are reported in Appendix C.2.

4.2 Evaluation on Human and Robotic Interactions

Human–object interactions. We evaluate whether methods can synthesize newly visible content while preserving hand–object contact, object geometry, and interaction state under object-centric camera motion. Since synchronized target-view ground truth is unavailable, we conduct blinded four-way human evaluation on 25 held-out HOI videos. ROVER receives the highest preference shares for interaction fidelity (71.6%), visual quality (69.1%), and camera control (65.7%) (Table 1(a)).

Figure 6 highlights failures under disocclusion and inaccurate depth guidance. Competing methods distort the hands, drawer housing, or pulled-out drawer; ROVER better preserves their geometry and spatial relationships. In the faucet sequence, Vista4D generates flowing water after shutoff, while

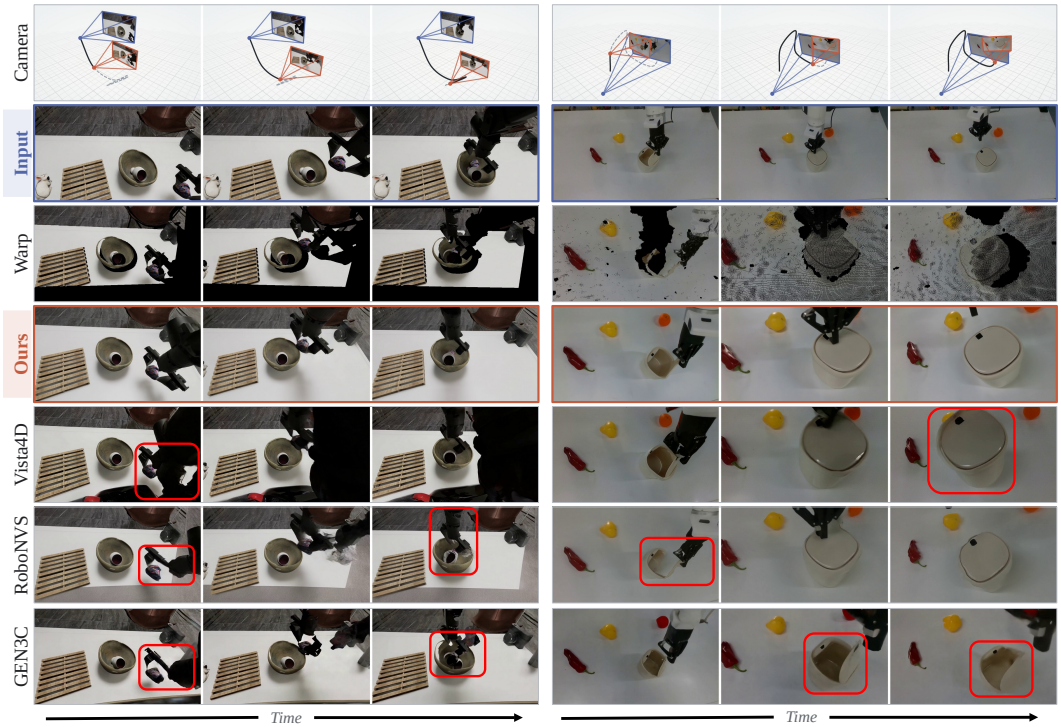

Figure 7: **Qualitative comparison on robotic manipulation videos.** The examples show *object placement* (Zhao et al., 2026) and *bin closing* (Wu et al., 2024) under interaction-centric camera trajectories. Our method better preserves robot and object geometry, while red boxes highlight artifacts in baselines. See Appendix E.2 for more comparisons and the project page for videos.

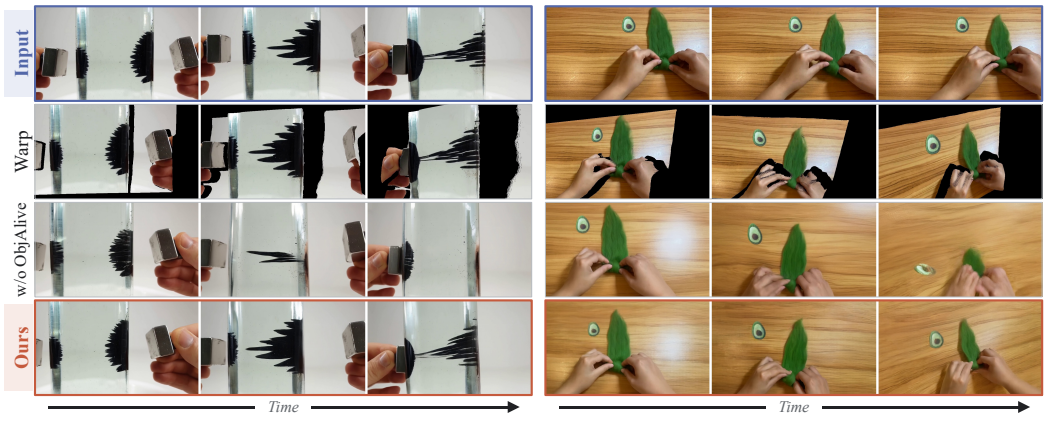

Figure 8: **Ablation of ObjAlive.** Training with ObjAlive improves preservation of evolving fluid structures in *manipulating magnetic fluid* (Chen et al., 2025b), and non-rigid deformation and hand–object contact in *rolling a fiber bundle* (Chen et al., 2025b).

GEN3C and EX-4D retain finger–handle gaps introduced by depth errors in the geometric warp. ROVER maintains finger–handle contact and shows the water stopping when the faucet is turned off.

Robotic manipulation. We further evaluate preservation of robot and object geometry and gripper–object contact under viewpoint changes. Without robot-specific fine-tuning, ROVER receives the highest preference shares in blinded four-way evaluation on 25 held-out robotic videos: 63.6% for interaction fidelity, 64.0% for visual quality, and 59.4% for camera control (Table 1(b)). Figure 7 compares ROVER with Vista4D, RoboNVS, and GEN3C. In *object placement*, competing methods distort the robot arm, gripper, or held object; in *bin closing*, they deform the lid or container. ROVER more consistently preserves their geometry and spatial relationships during manipulation.

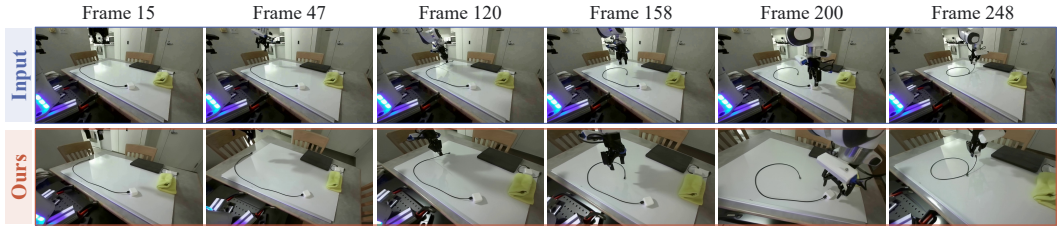

Figure 9: **Long-video camera control on *coiling a charging cable* (Khazatsky et al., 2024).** Source and generated frames are shown at six matched timestamps. Videos are on the project page.

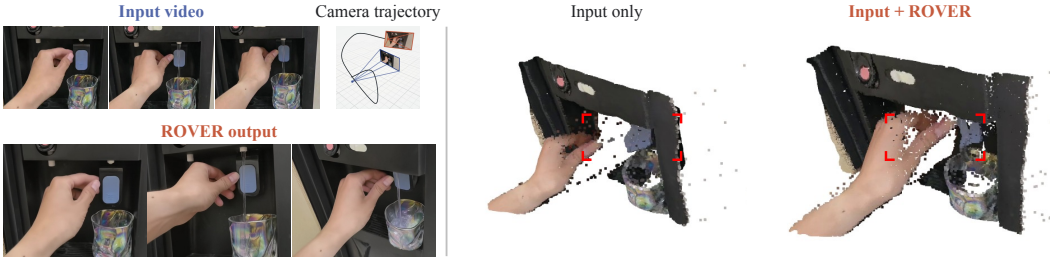

Figure 10: **Generated views for 4D reconstruction.** Joint reconstruction with a ROVER-generated video reduces the apparent hand–lever gap compared with source-only reconstruction. Both visualizations show only source-view points from the same inspection camera, without static merging.

To assess the benefit of robotic supervision, we compare ROVER with a variant further fine-tuned on a robotic training set. In blinded evaluation on 10 held-out robotic videos, over 70% of responses indicate no clear difference for each criterion (Table 1(d)). Larger or more diverse robotic datasets may still provide benefits. Appendix D.2 provides training details and qualitative comparisons.

4.3 Ablation: Contribution of ObjAlive

We compare the full model with a step-matched variant trained without ObjAlive, keeping the architecture, initialization, multi-view training data, and real single-view HOI4D videos fixed. In a blinded study on 10 held-out HOI videos with identical target trajectories for both variants, the full model is preferred in 69.0% of responses for interaction fidelity, 72.4% for visual quality, and 57.1% for camera control (Table 1(c)). Figure 8 illustrates these improvements. In *manipulating magnetic fluid*, the full model preserves extended fluid structures and fine detail that the variant without ObjAlive loses. In *rolling a fiber bundle*, it more faithfully captures the bundle's deformation while maintaining hand–object contact.

4.4 Applications

Long-video reshooting. Using the history-conditioned model (Section 3.3.3), we reshoot an 259-frame video across overlapping windows. In Figure 9, generated views preserve robot-arm geometry and cable structure while completing newly revealed background regions under changing viewpoints.

Qualitative 4D reconstruction. We explore generated views for 4D reconstruction by applying the same model (Jiang et al., 2026a) to the source video alone or jointly with a ROVER-generated view. Both visualizations render only source-view points under the same inspection camera, with static merging disabled. Joint reconstruction shows a smaller apparent hand–lever gap, more consistent with observed contact (Figure 10). Appendix E.4 provides implementation details and more examples.

5 Conclusion

We present ROVER for interaction-preserving video reshooting along specified camera trajectories. Through ObjAlive, we distill interaction and dynamics priors from a pretrained video generator using I2V-generated training videos. Qualitative comparisons and a controlled ablation demonstrate improved preservation of contact and non-rigid dynamics. Our model also generalizes to robotic manipulation without robot-specific fine-tuning. These findings support generated videos as a practical source of training data for reshooting without synchronized capture of each new event.

AI USE STATEMENT

We used AI tools to improve the language and readability of the manuscript. Generative AI was also used to construct OBJALIVE, as described in Section 3.2. The authors reviewed the AI-assisted text and take responsibility for the final manuscript, including its claims, results, and artifacts.

REFERENCES

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22875–22889. IEEE, 2025a.

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. In *International Conference on Learning Representations*, volume 2025, pp. 66712–66737, 2025b.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025a.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. In *International Conference on Learning Representations*, volume 2025, pp. 58038–58060, 2025b.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025c.

Boyang Cai, Qiwei Liang, Jiawei Li, Shihang Weng, Zhaoxin Zhang, Tao Lin, Xiangyu Chen, Wenjie Zhang, Jiaqi Mao, Weisheng Xu, et al. Beyond viewpoint generalization: What multi-view demonstrations offer and how to synthesize them for robot manipulation? *arXiv preprint arXiv:2603.26757*, 2026.

Wei Cao, Hao Zhang, Fengrui Tian, Yulun Wu, Yingying Li, Shenlong Wang, Ning Yu, and Yaoyao Liu. Freeorbit4d: Training-free arbitrary camera redirection for monocular videos via foreground-complete 4d reconstruction. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9044–9053, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chao_DexYCB_A_Benchmark_for_Capturing_Hand_Grasping_of_Objects_CVPR_2021_paper.html.

Cheng Chen, Jerry Bai, Jiacheng Wei, Boyu Chen, Xiaoji Zheng, Fan Wu, Minghao Yang, Tianrun Chen, Ruibo Li, Xiaoyu Yue, et al. Anyworld: Factorized egocentric world models for cross-embodiment generalization. *arXiv preprint arXiv:2608.29242*, 2026.

Kaihua Chen, Tarasha Khurana, and Deva Ramanan. Reconstruct, inpaint, test-time finetune: Dynamic novel-view synthesis from monocular videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Zixuan Chen, Jiaxin Li, Liming Tan, Yejie Guo, Junxuan Liang, Cewu Lu, and Yong-Lu Li. M^3-vos: Multi-phase, multi-transition, and multi-scenery video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025b.

Juntao Cheng, Wanyue Zhang, Zhiwei Yu, Shuo Ren, Zheqi He, Shaoxuan Xie, Guocai Yao, Jieru Lin, Börje F Karlsson, and Jiajun Zhang. Switch: Benchmarking modeling and handling of tangible interfaces in long-horizon embodied scenarios. *arXiv preprint arXiv:2511.17649*, 2025.

Hongsuk Choi, Nikhil Chavan-Dafle, Jiacheng Yuan, Volkan Isler, and Hyunsoo Park. Handnerf: Learning to reconstruct hand-object interaction scene from a single rgb image. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13940–13946. IEEE, 2024.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 653–660. IEEE, 2024.

Wanquan Feng, Jiawei Liu, Pengqi Tu, Tianhao Qi, Mingzhen Sun, Tianxiang Ma, Songtao Zhao, Siyu Zhou, and Qian He. I2vcontrol-camera: Precise video camera control with adjustable motion strength. In *International Conference on Learning Representations*, volume 2025, pp. 49560–49572, 2025.

Xiao Fu, Shitao Tang, Min Shi, Xian Liu, Jinwei Gu, Ming-Yu Liu, Dahua Lin, and Chen-Hsuan Lin. Plenoptic video generation. *arXiv preprint arXiv:2601.05239*, 2026.

Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *NeurIPS*, 2022.

Songen Gu, Yuhang Zheng, Weize Li, Yupeng Zheng, Yating Feng, Xiang Li, Yilun Chen, Pengfei Li, and Wenchao Ding. Vistabot: View-robust robot manipulation via spatiotemporal-aware view synthesis. *arXiv preprint arXiv:2604.21914*, 2026.

Xiaoguang Han, Yushuang Wu, Luyue Shi, Haolin Liu, Hongjie Liao, Lingteng Qiu, Weihao Yuan, Xiaodong Gu, Zilong Dong, and Shuguang Cui. Mvimgnet2. 0: A larger-scale dataset of multi-view images. *arXiv preprint arXiv:2412.01430*, 2024.

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.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13416–13426. IEEE, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ryan Hoque, Peide Huang, David Yoon, Jian Zhang, et al. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, volume 2026, pp. 4218–4237, 2026.

Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025.

Hyeonho Jeong, Suhyeon Lee, and Jong Chul Ye. Reangle-a-video: 4d video generation as video-to-video translation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11164–11175. IEEE, 2025.

Yanqin Jiang, Tengfei Wang, Zhengwei Wang, Chenjie Cao, Junta Wu, Wenhan Luo, Weiming Hu, Jin Gao, and Chunchao Guo. OmniX: Any-view and any-time 4D reconstruction via feed-forward trajectory fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026a. URL https://omnix4d.github.io/.

Zeren Jiang, Yushi Lan, Yihang Luo, Yufan Deng, Zihang Lai, Edgar Sucar, Christian Rupprecht, Iro Laina, Diane Larlus, Chuanxia Zheng, et al. Syn4d: A multiview synthetic 4d dataset. *arXiv preprint arXiv:2605.05207*, 2026b.

Taewoong Kang, Kinam Kim, Dohyeon Kim, Minho Park, Junha Hyung, and Jaegul Choo. Egox: Egocentric video generation from a single exocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11116–11126, 2026.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

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.

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, Yizhong Zhang, Xi Chen, Hao Chen, Lily Sun, Dong Chen, Jiaolong Yang, and Baining Guo. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025a.

Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28785–28796. IEEE, 2025b.

Yixuan Li, Yutang Lin, Jieming Cui, Tengyu Liu, Wei Liang, Yixin Zhu, and Siyuan Huang. Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. In *9th Annual Conference on Robot Learning*, 2025c.

Yixuan Li, Yanhong Zeng, Ka Leong Cheng, Jiayi Zhu, Hanlin Wang, Wen Wang, Yihao Meng, Hao Ouyang, Qiuyu Wang, Yue Yu, et al. Cameraanything: Refilming videos with arbitrary camera control. *arXiv preprint arXiv:2607.24591*, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Kevin Lin, Ajay Mandlekar, Caelan Reed Garrett, Nikita Chernyadev, Yu Fang, Runyu Ding, Yuqi Xie, Linxi Fan, and Yuke Zhu. Humanoidmimicgen: Data generation for loco-manipulation via whole-body planning and adaptation. In *ICRA 2026 Workshop on Synthetic Data for Robot Learning*, 2026a.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, et al. Vista4d: Video reshooting with 4d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32671–32682, 2026b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jia-Wei Liu, Yan-Pei Cao, Tianyuan Yang, Zhongcong Xu, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Hosnerf: Dynamic human-object-scene neural radiance fields from a single video. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18437–18448. IEEE, 2023.

Xingchen Liu, Piyush Tayal, Jianyuan Wang, Jesus Zarzar, Tom Monnier, Konstantinos Tertikas, Jiali Duan, Antoine Toisoul, Jason Y Zhang, Natalia Neverova, et al. Uncommon objects in 3d. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14102–14113. IEEE, 2025.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20981–20990. IEEE, 2022.

Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In *European Conference on Computer Vision*, pp. 407–425. Springer, 2024.

Yawen Luo, Xiaoyu Shi, Jianhong Bai, Menghan Xia, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Camclonemaster: Enabling reference-based camera control for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–10, 2025.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026.

Priyanka Mandikal, Tushar Nagarajan, Alex Stoken, Zihui Xue, and Kristen Grauman. Spoc: Spatially-progressing object state change segmentation in video. In *ArXiv*, 2025.

NVIDIA. NVIDIA OmniDreams: Real-time generative world model for closed-loop autonomous vehicle simulation. *arXiv preprint arXiv:2606.03159*, 2026. URL https://arxiv.org/abs/2606.03159.

Jeongwan On, Kyeonghwan Gwak, Gunyoung Kang, Junuk Cha, Soohyun Hwang, Hyein Hwang, and Seungryul Baek. Bigs: Bimanual category-agnostic interaction reconstruction from monocular videos via 3d gaussian splatting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17437–17447. IEEE, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Zezhong Qian, Xiaowei Chi, Yuming Li, Shizun Wang, Zhiyuan Qin, Xiaozhu Ju, Sirui Han, and Shanghang Zhang. Wristworld: Generating wrist-views via 4d world models for robotic manipulation. *arXiv preprint arXiv:2510.07313*, 2025.

Wentian Qu, Zhaopeng Cui, Yinda Zhang, Chenyu Meng, Cuixia Ma, Xiaoming Deng, and Hongan Wang. Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15054–15065. IEEE, 2023.

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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10881–10891. IEEE, 2021.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Praveen Selvaraj, Lorenzo Uttini, and Ville Kuosmanen. Armnetbench v0. 1: Parallel real-world evaluation of manipulation policies on a low-cost arm farm. *arXiv preprint arXiv:2607.24481*, 2026.

Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. *Advances in Neural Information Processing Systems*, 37:80220–80243, 2024.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026.

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.

Chaoyue Song, Jiacheng Wei, Chuan Sheng Foo, Guosheng Lin, and Fayao Liu. Reacto: Reconstructing articulated objects from a single video. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5384–5395. IEEE, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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 preprint arXiv:2312.11805*, 2023.

Stephen Tian, Blake Wulfe, Kyle Sargent, Katherine Liu, Sergey Zakharov, Vitor Guizilini, and Jiajun Wu. View-invariant policy learning via zero-shot novel view synthesis. *arXiv preprint arXiv:2409.03685*, 2024.

Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision*, pp. 313–331. Springer, 2024.

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.

Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. HO-Cap: A capture system and dataset for 3D reconstruction and pose tracking of hand-object interaction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025a. URL https://openreview.net/forum?id=hpu6r8oLw9.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. pi3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025b.

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*, pp. 1–11, 2024.

Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David Fleet. Controlling space and time with diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 95989–96012, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.

Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2z1HT5lw5M.

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.

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364*, 2024.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111. IEEE, 2025.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9150–9161. IEEE, 2023.

Daiwei Zhang, Gengyan Li, Jiajie Li, Mickaël Bressieux, Otmar Hilliges, Marc Pollefeys, Luc Van Gool, and Xi Wang. Egogaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting. In *2025 International Conference on 3D Vision (3DV)*, pp. 1091–1102. IEEE, 2025a.

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. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2050–2062. IEEE, 2025b.

Weiguang Zhao, Haoran Xu, Xingyu Miao, Qin Zhao, Rui Zhang, Kaizhu Huang, Ning Gao, Peizhou Cao, Mingze Sun, Mulin Yu, et al. Synthverse: A large-scale diverse synthetic dataset for point tracking. *arXiv preprint arXiv:2602.04441*, 2026.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

Rui Zhong, Yizhe Sun, Junjie Wen, Jinming Li, Chuang Cheng, Wei Dai, Zhiwen Zeng, Huimin Lu, Yichen Zhu, and Yi Xu. Humanoidexo: Scalable whole-body humanoid manipulation via wearable exoskeleton. *arXiv preprint arXiv:2510.03022*, 2025.

Siyuan Zhou, Hejun Wang, Hu Cheng, Jinxi Li, Dongsheng Wang, Junwei Jiang, Yixiao Jin, Jiayue Huang, Shiwei Mao, Shangjia Liu, et al. Physinone: Visual physics learning and reasoning in one suite. *arXiv preprint arXiv:2604.09415*, 2026.

Liyuan Zhu, Shengyu Huang, Amrita Mazumdar, Tianye Li, Zan Gojcic, Gordon Wetzstein, Iro Armeni, Shalini De Mello, and Alex Trevithick. World from motion: Generative dynamic gaussian reconstruction from monocular video. *arXiv preprint arXiv:2607.01202*, 2026.

APPENDIX CONTENTS

A Implementation Details **17**
A.1 Model Architecture . . . 17
A.2 Training Configuration . . . 17
A.3 Sampling and Guidance . . . 17
A.4 Long-Video Generation . . . 18

B Training Data and Preparation **18**
B.1 Training Data Mixture . . . 18
B.2 Depth and Camera Preparation . . . 19

C Evaluation Protocols and Quantitative Results **19**
C.1 Human Evaluation . . . 19
C.2 Novel-View Video Synthesis on DyCheck . . . 19
C.3 Inference Time . . . 19

D Additional Ablation Studies **21**
D.1 Contribution of OBJALIVE . . . 21
D.2 Robotic Fine-Tuning . . . 21
D.3 Source and Warp Guidance . . . 22
D.4 Static Background Memory . . . 22

E Additional Qualitative Results **23**
E.1 More Results on Human–Object Interactions . . . 23
E.2 More Results on Robotic Manipulation . . . 24
E.3 Long-Video Camera Control . . . 25
E.4 Additional 4D Reconstruction Examples . . . 25

F Limitations and Future Work **26**

A Implementation Details

A.1 Model Architecture

Source video and warp conditioning. The frozen video VAE produces 16-channel latents with a spatial downsampling factor of eight and a temporal length of $N_\ell = 1 + (N-1)/4$. A 49-frame clip therefore yields 13 latent frames. The target input concatenates 16 noisy-target channels, four warp-validity mask channels, and 16 warp channels, giving 36 channels in total. The validity mask marks observed warp pixels with one and missing regions with zero. It is resampled to the latent grid and replicated across four channels. The source input projection is initialized by copying the pretrained target-projection weights associated with the 16 noisy-target latent channels. It tokenizes the 16-channel source latent and remains frozen, while a learned mask embedding encodes source validity. The first source frame is encoded by a frozen CLIP image encoder, and its embeddings are supplied to the pretrained image cross-attention branch.

Camera and coordinate encoders. We express source and target camera-to-world poses relative to the first target camera: $\widetilde{P}_t^a = (P_0^{\mathrm{tgt}})^{-1} P_t^a$, where $a \in \{\mathrm{src}, \mathrm{tgt}\}$. For each pixel, we compute the Plücker ray (He et al., 2024; Sitzmann et al., 2021) $\mathbf{r} = (\mathbf{o} \times \mathbf{d}, \mathbf{d})$, where $\mathbf{o}$ is the camera center and $\mathbf{d}$ is the normalized viewing direction in this shared frame. We compute rays at full image resolution, concatenate them within each spatial patch, and temporally subsample them to align with the VAE latent frames.

Source coordinate maps combine UV coordinates normalized to $[-1, 1]$ with source depth min–max normalized over each clip. We reproject these maps alongside RGB to obtain warped coordinate maps, retaining the source coordinates and depth at valid locations. Both maps are packed directly onto the token grid without passing through the video VAE. Each DiT block has its own linear layer for rays and sinusoidal encoding followed by two linear layers for coordinates. Within each block, the encoders are shared between source and target tokens. The ray projections and the final coordinate-encoder layers are zero-initialized, so both branches initially output zero embeddings.

Token sequence and positional encoding. DiT patchification uses $1 \times 2 \times 2$ latent patches. We concatenate target tokens followed by source tokens, with no separate warp segment. Compared with Vista4D's three-segment design (Lin et al., 2026b), our two-segment layout reduces sequence length and self-attention computation at matched clip lengths and token resolutions. The query and key projections in Equation 2 are followed by the backbone's query/key normalization and RoPE (Su et al., 2024). Source and target tokens share spatial RoPE indices, while the source stream's temporal indices are offset relative to the target stream to distinguish the two segments.

A.2 Training Configuration

Condition dropout. We independently drop the source and warp conditions with probability 0.1, replacing each dropped condition with its learned null embedding. The associated coordinate embeddings are also disabled, while Plücker ray conditioning is retained.

Optimization. We use AdamW with a two-stage training schedule: 3,500 optimizer steps on 17-frame clips, followed by 1,500 steps on 49-frame clips. Both stages use an effective global batch size of 128 across 128 NVIDIA A100 GPUs. The learning rate is linearly warmed up over the first 5% of each stage and then held constant at 10^{-5} and 5×10^{-6} for the 17- and 49-frame stages, respectively. We optimize the self-attention layers, source-validity and null embeddings, camera and coordinate encoders, and the weights associated with the conditioning channels of the target input projection. The source input projection and the target projection's noisy-latent weights remain frozen. The 49-frame stage starts from the 17-frame checkpoint, retaining all learned conditioning parameters without reinitialization.

A.3 Sampling and Guidance

We generate 49-frame videos at a resolution of 832×480 using 50 sampling steps and an empty text prompt. Source-video and warp classifier-free guidance (CFG) (Ho & Salimans, 2022) are controlled

independently. A guidance scale of 1 retains the corresponding condition without additional guidance. Appendix D.3 examines the effects of varying these scales.

A.4 Long-Video Generation

History-conditioned fine-tuning. We fine-tune our base model to generate target videos conditioned on a history prefix. During training, the prefix is taken from the ground-truth target video and encoded with the frozen video VAE. We then apply noise augmentation to the history latents to expose the model to imperfect history conditions. This augmentation affects only the history condition; the clean target latents used to define the flow-matching objective remain unchanged. The loss is applied only to target latents outside the prefix. Source frames, camera conditions, and warps remain aligned with the current training window. The same fine-tuned checkpoint is used for all inference windows.

Window schedule and sampling. We use 49-frame windows with a 17-frame overlap, advancing by 32 frames per window. The first window is generated without a history prefix. For each subsequent window, overlapping frames are selected from the preceding output by their global frame indices and encoded into five history latent frames. We retain the earlier RGB outputs at overlapping timestamps and append only newly generated frames.

Static background memory. We estimate geometry from generated views using Pi3X (Wang et al., 2025b), with source depth and the supplied camera parameters as priors. After aligning depth scales using static source regions, we accumulate the reconstructed background in a shared coordinate system. SAM3 (Carion et al., 2025) masks exclude dynamic objects from this memory. We render the memory at subsequent target viewpoints and use it only to fill missing regions in the current source warp, preserving valid source observations. The warp-validity mask includes these filled regions. Figure 16 illustrates how this memory helps preserve background details across generation windows through a comparison with and without memory.

B Training Data and Preparation

B.1 Training Data Mixture

Our training mixture combines synchronized multi-view videos with generated and real single-view videos. MultiCamVideo (Bai et al., 2025a), Syn4D (Jiang et al., 2026b), and PhysInOne (Zhou et al., 2026) provide synthetic dynamic scenes with varied camera configurations, while DexYCB (Chao et al., 2021) and HO-Cap (Wang et al., 2025a) provide real hand–object interactions. The single-view data combine ObjAlive with 2.9K real interaction clips from HOI4D (Liu et al., 2022). Multi-view videos provide synchronized source–target pairs, while single-view videos provide reconstruction pairs through double reprojection. Table 2 summarizes the processed scenes or sequences used for training and their geometry sources.

Table 2: **Video data used for training.** Multi-view videos provide direct cross-view supervision; single-view videos provide double-reprojection reconstruction tasks. Counts report processed scenes or sequences used for training, rather than full dataset sizes.

Dataset	Domain	Count	Views	Camera setup	Depth and camera source
Multi-view videos					
MultiCamVideo	Synthetic general scenes	13.6K	8–9	Moving	Predicted depth and cameras
Syn4D	Synthetic dynamic scenes	565	8	Static and moving	Rendered depth, GT cameras
PhysInOne	Synthetic physical interactions	5.3K	13	12 static + 1 moving	Rendered depth, GT cameras
DexYCB	Real hand–object interactions	1K	8	Static	Sensor depth, calibrated cameras
HO-Cap	Real hand–object interactions	64	8	Static	Sensor depth, calibrated cameras
Single-view videos					
ObjAlive	Generated dynamic interactions	13.5K	1	Unconstrained	Predicted depth and cameras
HOI4D	Real hand–object interactions	2.9K	1	Egocentric	Sensor depth, provided SLAM cameras

B.2 Depth and Camera Preparation

Depth and camera sources. We use rendered or sensor depth with the corresponding ground-truth, calibrated, or SLAM-estimated camera parameters for the datasets using provided geometry in Table 2. For ObjAlive, we estimate depth and camera parameters from the generated RGB videos using Depth Anything 3 (DA3) (Lin et al., 2025) before constructing double-reprojection pairs.

MultiCamVideo preprocessing. We estimate depth and camera parameters for MultiCamVideo using DA3, with RGB videos as the only input. Synchronized views are grouped by timestamp and processed in overlapping multi-camera windows. We align the window-wise reconstructions using DA3-Streaming, updating camera poses and depth scales to obtain a shared scene-level coordinate system. Finally, we apply temporal smoothing to the per-camera estimates to reduce jitter.

C Evaluation Protocols and Quantitative Results

C.1 Human Evaluation

Study design. We conduct blinded human evaluation with 21 participants. Method comparisons use 25 held-out human–object interaction videos and 25 held-out robotic manipulation videos. The ObjAlive and robotic fine-tuning ablations each use 10 held-out videos from their respective domains. We design the target camera trajectories based on Vista4D's camera-control interface (Lin et al., 2026b). For each example, participants view the source video, a camera-motion reference, and candidate results generated under the same target camera trajectory. Videos play synchronously and loop, with pause and replay controls available. The camera-motion reference serves only as a viewpoint guide; participants are instructed to ignore its holes and visual artifacts. Figure 11 shows the interface for the four-way comparisons.

Evaluation criteria. Participants make a separate choice for each of three questions:

1. **Interaction fidelity.** Which video more faithfully preserves the interaction in the input, including object motion, contact relationships, and changes in object state?
2. **Visual quality.** Which video has better visual quality, with clearer details, fewer artifacts, and less flickering?
3. **Camera control.** Which video better follows the requested camera motion?

Voting and reporting. Method comparisons present four candidate results per example, and participants select their preferred result for each criterion. Ablations present two variants and additionally allow an "About the same" response, reported as "No clear difference" in our results. We compute preference shares separately for each criterion and study group using the collected responses. For ablations, the denominator includes "No clear difference" responses. Table 1(a,b) reports four-way preference shares. Panel (c) reports preference for the model trained with ObjAlive, while panel (d) reports the share of responses indicating no clear difference with and without robotic fine-tuning.

C.2 Novel-View Video Synthesis on DyCheck

We complement the interaction-centric evaluation with DyCheck (Gao et al., 2022), which provides synchronized dynamic captures and ground-truth target views for direct quantitative evaluation of novel-view video synthesis. For a fair comparison, all methods use the same source and target cameras, evaluation frames, native-resolution ground truth, and co-visible masks. Table 3 reports PSNR, SSIM, and LPIPS for both the full frame and the co-visible region. Our method achieves the best performance across all reported metrics. Figure 12 further presents qualitative comparisons on challenging cases involving disocclusions, thin structures, and fast motion.

C.3 Inference Time

Table 4 reports runtimes from completed evaluation runs, with one NVIDIA A100 GPU allocated per video. All configurations use 50 sampling steps except GEN3C, which uses two generation calls with

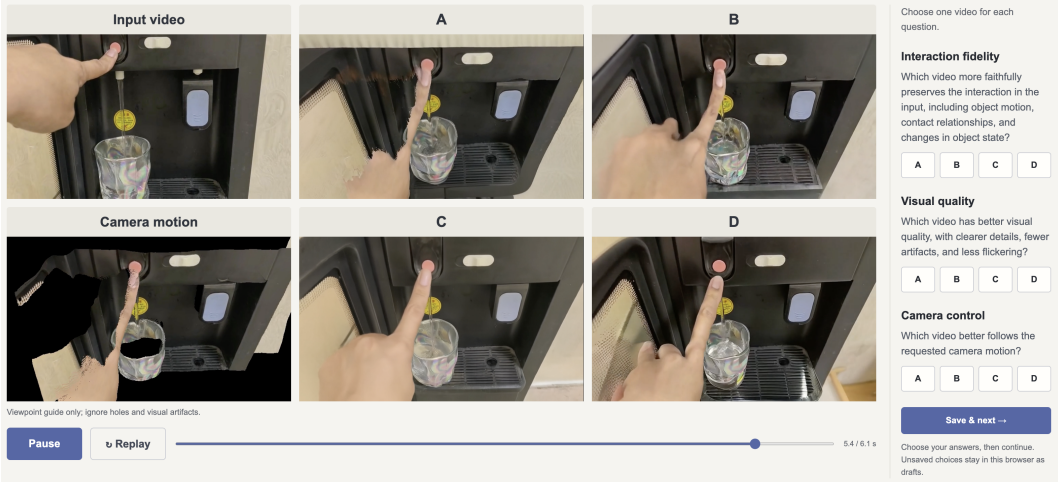

Figure 11: **Interface for blinded human evaluation.** Participants view the source video, a camera-motion reference, and four anonymized candidate videos (A–D), then select their preferred result separately for interaction fidelity, visual quality, and camera control.

Table 3: **Novel-view video synthesis on DyCheck (Gao et al., 2022).**

Method	mPSNR↑	mSSIM↑	mLPIPS↓	PSNR↑	SSIM↑	LPIPS↓
ReCamMaster (Bai et al., 2025a)	11.44	0.479	0.750	11.61	0.310	0.820
CamCloneMaster (Luo et al., 2025)	10.89	0.460	0.710	11.03	0.283	0.779
TrajectoryCrafter (Yu et al., 2025)	14.31	0.531	0.576	12.87	0.350	0.665
EX-4D (Hu et al., 2025)	14.50	0.508	0.580	13.82	0.338	0.649
GEN3C (Ren et al., 2025)	15.12	0.535	0.510	14.75	0.364	0.578
Vista4D (Lin et al., 2026b)	15.47	0.535	0.443	15.03	0.360	0.510
ROVER (ours)	**16.18**	**0.542**	**0.437**	**15.83**	**0.377**	**0.502**

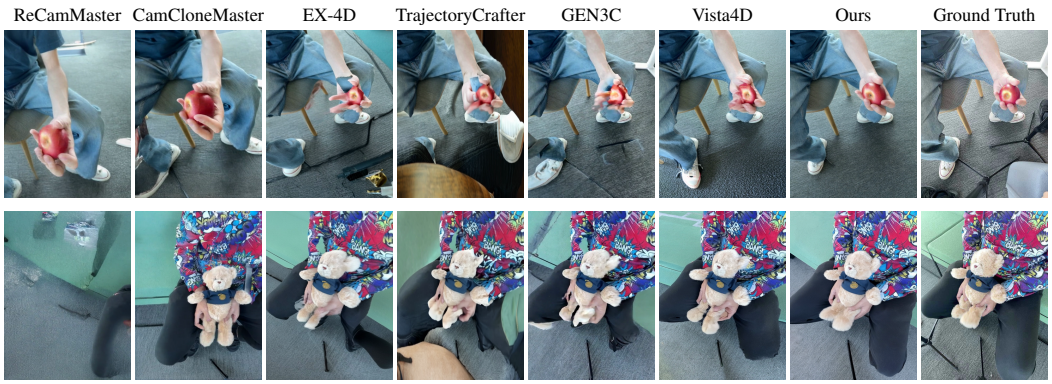

Figure 12: **Qualitative comparison on DyCheck.** Each row shows the same target view and evaluation frame for all methods. Our method produces better results.

35 steps each. For GEN3C, we report the combined time of these model calls. Other timings cover denoising only, excluding geometry preparation, model initialization, VAE encoding and decoding, and video export.

Table 4: **Recorded inference times.** Times are medians in minutes from completed evaluation runs.

Method	Base model	Resolution	Time (min)
ReCamMaster	Wan2.1-T2V-1.3B	832×480	8.57
CamCloneMaster	Wan2.1-T2V-1.3B	832×480	17.68
TrajectoryCrafter	CogVideoX-Fun-5B	672×384	2.43
EX-4D	Wan2.1-T2V-14B	672×384	7.68
GEN3C	Cosmos-1.0-Diffusion-7B-Video2World	1280×704	27.81
Vista4D	Wan2.1-T2V-14B	672×384	20.13
RoboNVS	Wan2.1-T2V-14B	832×480	6.85
ROVER (no CFG)	Wan2.1-I2V-14B	832×480	11.01
ROVER (source CFG 1.5)	Wan2.1-I2V-14B	832×480	22.00

Table 5: **Full human evaluation results for the ablations.** Each ablation uses 10 held-out videos. Values are percentages of responses preferring either variant or reporting no clear difference, evaluated separately for interaction fidelity, visual quality, and camera control.

Response	Interaction fidelity	Visual quality	Camera control
(a) OBJALIVE *ablation on HOI videos*			
w/ OBJALIVE	**69.0**	**72.4**	**57.1**
w/o OBJALIVE	15.2	11.4	17.1
No clear difference	15.8	16.2	25.8
(b) Robotic fine-tuning ablation on robotic videos			
w/ robotic fine-tuning	14.3	14.8	9.5
w/o robotic fine-tuning	11.4	13.3	9.5
No clear difference	**74.3**	**71.9**	**81.0**

D ADDITIONAL ABLATION STUDIES

D.1 CONTRIBUTION OF OBJALIVE

Figure 13 extends the comparison in Section 4.3 to four additional interactions. In *lighting a gas stove*, the model without OBJALIVE changes the stove's appearance and introduces a flame before it appears in the input, whereas the full model more faithfully preserves the ignition state. In *smashing fruit* and *knitting*, removing OBJALIVE produces blurred or distorted hands during manipulation. The full model retains clearer hand geometry and thin tool structures, including the knitting needles and yarn. In *pressing a nail clipper*, the full model more consistently preserves the clipper's structure and fingertip contact as the viewpoint changes. These results complement the non-rigid motion examples in the main paper with object-state changes and fine-grained hand–tool interactions.

On 10 held-out HOI videos, blinded human evaluation favors the model trained with OBJALIVE across all three criteria. Table 5 (a) reports the full response distribution, including preferences for the model without OBJALIVE and no-clear-difference responses.

D.2 ROBOTIC FINE-TUNING

We further fine-tune ROVER for 400 steps using approximately 9K multi-view sequences and 19K single-view videos from RH20T (Fang et al., 2024), together with 3K single-view videos from DROID (Khazatsky et al., 2024). The DROID fine-tuning split is disjoint from all evaluation episodes.

Table 5 (b) reports human preferences on 10 held-out robotic videos. Most responses indicate no clear difference across all three criteria, with relatively few responses favoring either variant. Figure 14 complements these results with qualitative comparisons under the same target camera trajectories.

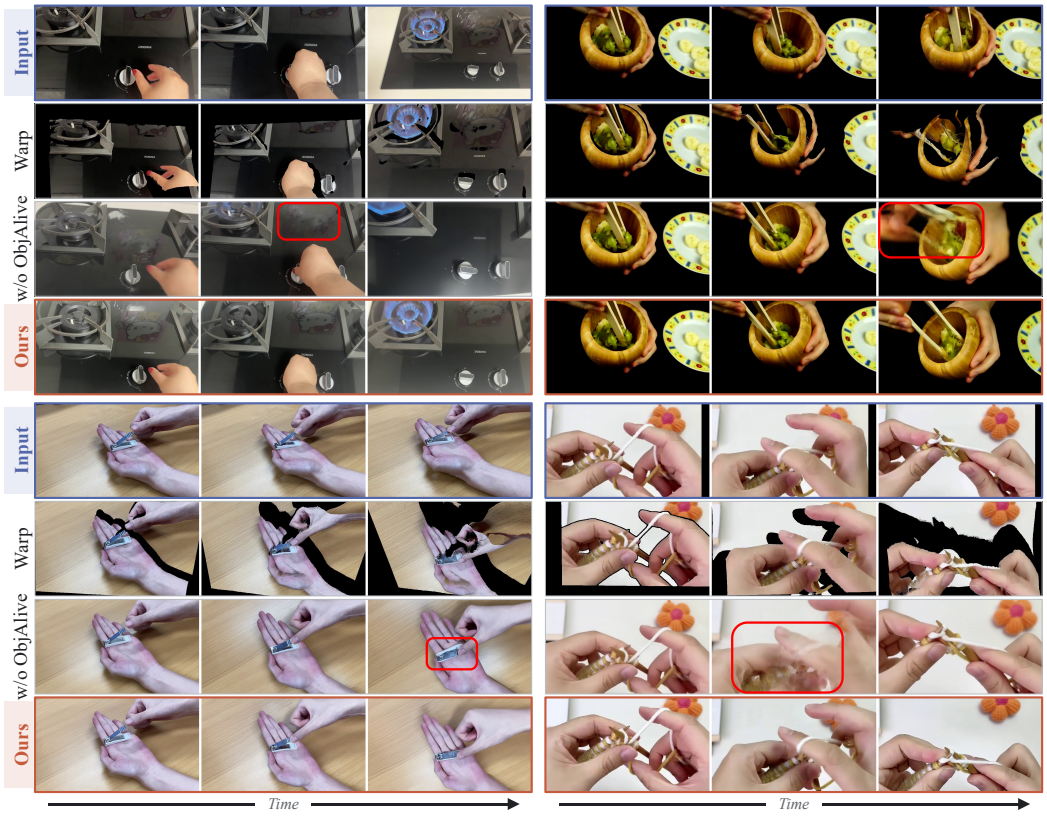

Figure 13: **Additional qualitative results for the OBJALIVE ablation.** From left to right, the top pair shows *lighting a gas stove* (Cheng et al., 2025) and *smashing fruit* (Chen et al., 2025b); the bottom pair shows *pressing a nail clipper* (Song et al., 2024) and *knitting* (Chen et al., 2025b). Including OBJALIVE reduces artifacts in hands and manipulated objects while better preserving interaction dynamics under viewpoint changes. Video comparisons are available on the project page.

D.3 SOURCE AND WARP GUIDANCE

We examine classifier-free guidance (CFG) (Ho & Salimans, 2022) applied separately to the source-video and warp conditions in Figure 15. Starting with both guidance scales set to 1, we increase one scale while keeping the other fixed. Without additional guidance, thin structures such as the knife and fork are poorly resolved, and the robot gripper appears disconnected from the manipulated object. Source CFG at 1.5 produces clearer thin structures and more plausible gripper–object contact. Increasing the source scale to 2.0, however, introduces exaggerated texture on the human hands. Warp CFG at either 1.5 or 2.0 leaves the thin-structure and contact artifacts unresolved. Overall, source CFG at 1.5 offers the most balanced trade-off among fine-detail preservation, interaction consistency, and natural appearance in these qualitative comparisons.

D.4 STATIC BACKGROUND MEMORY

Figure 16 compares long-video generation with and without 3D background memory. In the *pouring water* (Wu et al., 2024) sequence, both variants initially generate the drawer handle. Without memory, the handle disappears at frames 160 and 232 as the viewpoint changes. With memory, previously generated background observations supplement the warps, and the handle remains visible at both timestamps. This example illustrates how reusing background content helps preserve scene details across generation windows.

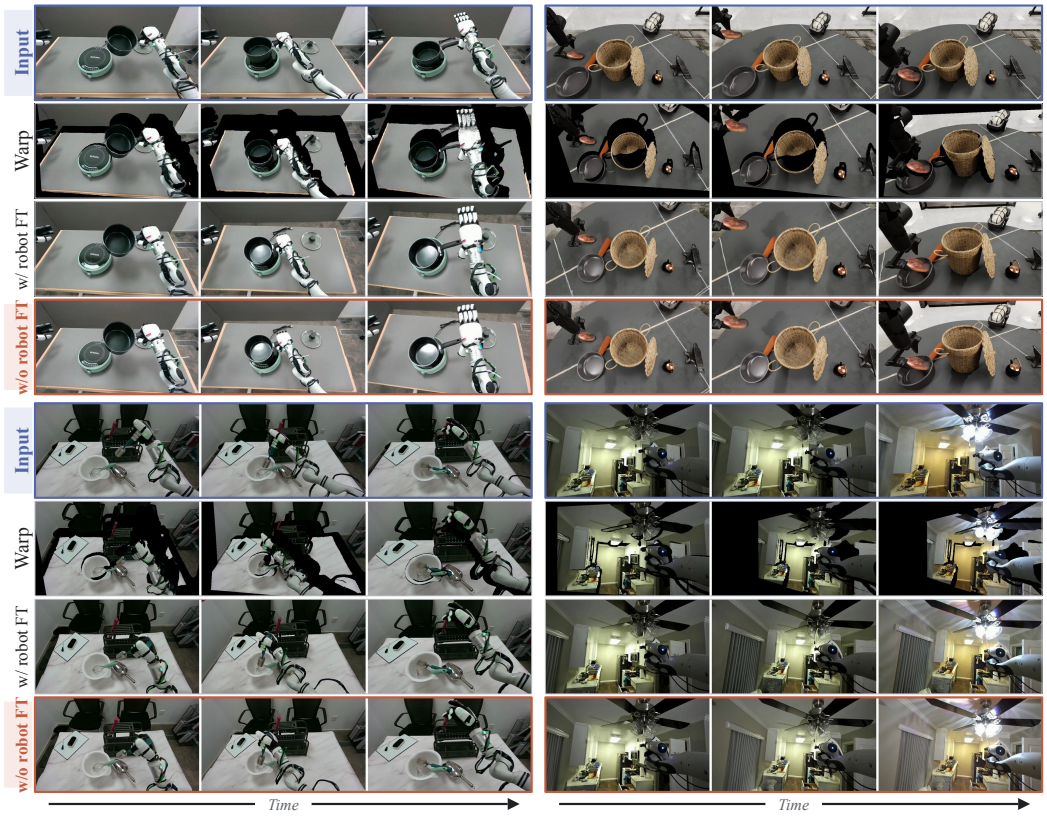

Figure 14: **Effect of additional robotic fine-tuning.** We compare ROVER, which uses no robot-specific fine-tuning, with a variant further fine-tuned on RH20T and DROID. Both variants use the same input videos and target camera trajectories, with frames shown at matched timestamps.

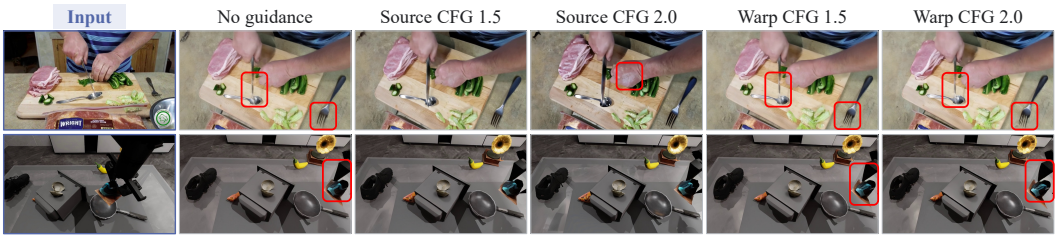

Figure 15: **Effect of source and warp classifier-free guidance.** Each row shows one example at a matched timestamp, with the target camera and random seed fixed across configurations. Red boxes highlight thin-structure artifacts, exaggerated hand texture, and disrupted gripper–object contact. Source CFG at 1.5 improves the highlighted details, while stronger source guidance introduces texture artifacts and warp guidance provides limited improvement.

E ADDITIONAL QUALITATIVE RESULTS

E.1 MORE RESULTS ON HUMAN–OBJECT INTERACTIONS

Qualitative comparisons. Figures 17 and 18 present additional comparisons on held-out human–object interaction videos. These sequences evaluate the model's ability to handle thin articulated structures, small contact regions, and visually subtle appearance changes associated with object-state transitions, such as the illuminated display on the heated towel rack. In *folding eyeglasses* and *opening a stapler*, competing methods often distort thin object parts or produce implausible articulations. In the *unplugging an AirPods case* sequence, competing methods introduce artifacts or implausible content in newly revealed regions. Our method completes these regions more coherently while

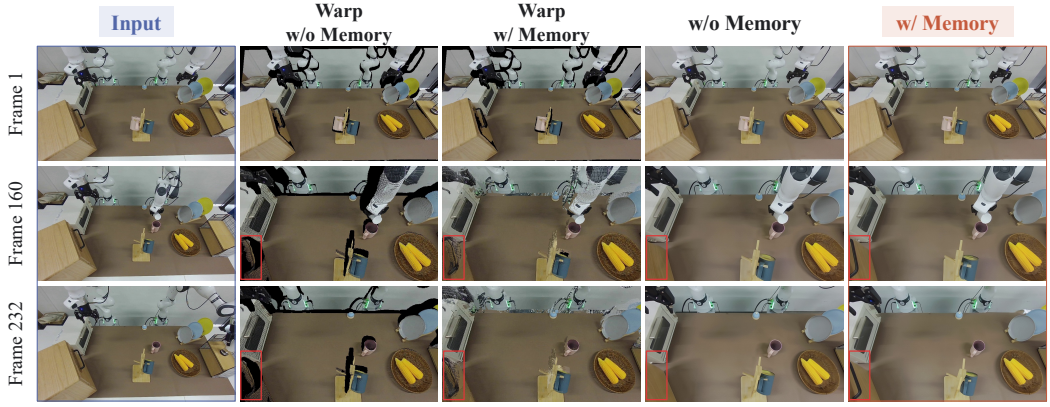

Figure 16: **Effect of 3D background memory.** Columns show the input, warps without and with memory, and the corresponding generated results. Rows show frames 1, 160, and 232 of the *pouring water* (Wu et al., 2024) sequence. Red boxes highlight the drawer-handle region. Without memory, the generated handle disappears at the two later timestamps; with memory, it is preserved.

preserving the unplugging action. The *heated towel rack* sequence further tests whether the control panel and illuminated display remain consistent as the viewpoint changes. Our method more reliably preserves the object structure and interaction state across these examples.

More results from our method. Figure 19 presents further results from our method across diverse human–object interactions. Peeling a kiwi and breaking a cookie involve non-rigid deformation and topology changes, while the laptop and scissors sequences require consistent articulated motion. The toaster sequence tests whether the model preserves a subtle state-dependent appearance change as the lever is lowered and the indicator light turns on. Together with frying an egg, these examples broaden the range of interactions covered by our evaluation.

E.2 More Results on Robotic Manipulation

Additional evaluation videos. We also evaluate on full-body robot manipulation videos from prior work (Li et al., 2025c; Lin et al., 2026a; Luo et al., 2026; Zhong et al., 2025). These videos complement the evaluation datasets described in the main paper by extending the evaluation to full-body robot embodiments.

Qualitative comparisons. Figures 20 and 21 present additional comparisons on held-out robotic manipulation videos. Alongside Vista4D (Lin et al., 2026b), RoboNVS (Cai et al., 2026), and GEN3C (Ren et al., 2025), we include EX-4D (Hu et al., 2025) and TrajectoryCrafter (Yu et al., 2025). These examples involve gripper–object contact, thin object parts, and substantial occlusions from robot arms. Competing methods can distort the manipulated objects, fragment gripper geometry, or alter the spatial relationship between the gripper and object. These errors are particularly visible around the gripper and blocks in the stacking sequence, the held brush, and the gripper operating the knob. Our method better preserves these local structures and contact relationships while completing newly visible regions.

More results from our method. Figure 22 presents further results across different robot embodiments and manipulation tasks. The pot, cup, eye-drop bottle, and twisting sequences involve maintaining grasp or contact as the camera moves around the interaction. Lifting a box extends the evaluation to a humanoid robot manipulating an object, while pulling a tissue introduces non-rigid deformation. In these examples, our method completes regions missing from the geometric warps while preserving robot and object structure, contact relationships, and interaction dynamics.

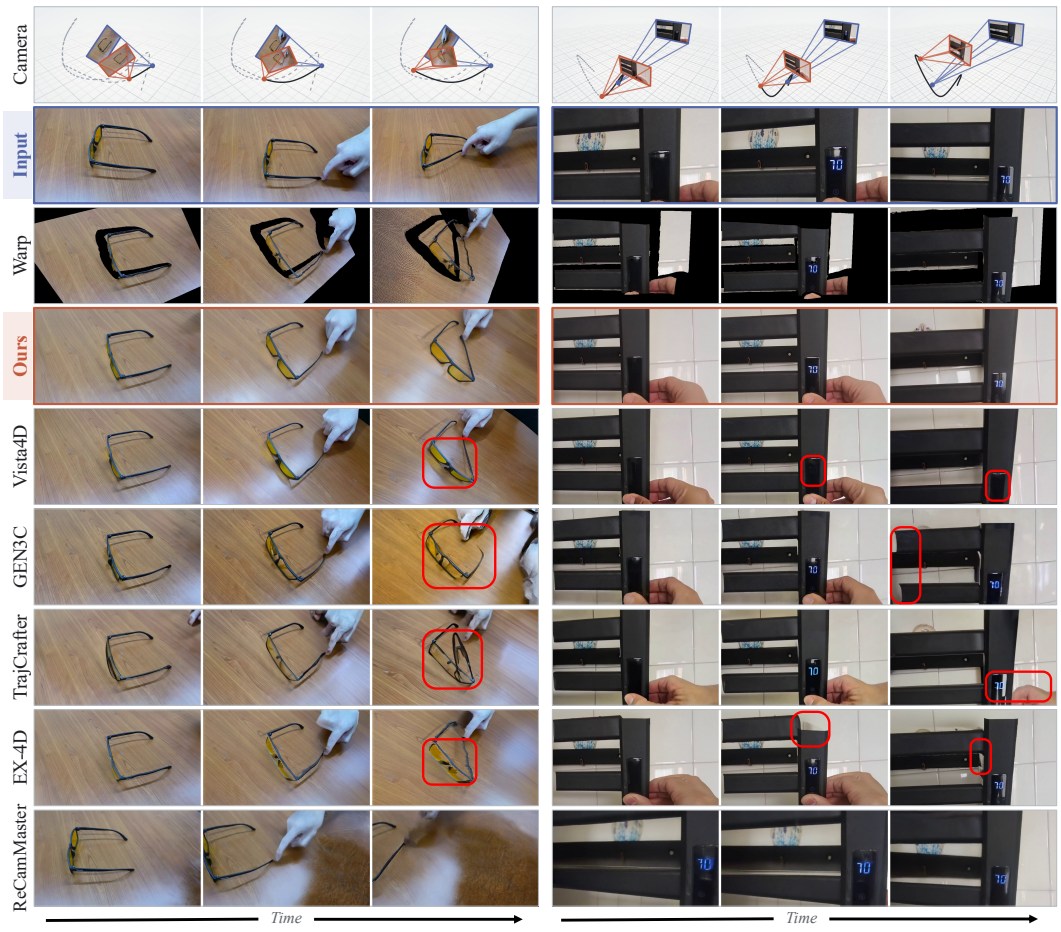

Figure 17: **Additional qualitative comparisons on human–object interaction videos.** In *folding eyeglasses* (Song et al., 2024), our method better preserves the thin frame, hinged temples, and fingertip contact under large viewpoint changes. In *activating a heated towel rack* (Cheng et al., 2025), it maintains the hand, control panel, and illuminated display more consistently than competing methods. Blue and orange denote the input and target cameras, while the black curve marks the traversed camera path.

E.3 Long-Video Camera Control

Figure 23 presents additional long-video results spanning multiple generation windows. In *wiping a table* (Khazatsky et al., 2024), ROVER maintains the robot structure and tabletop layout as the gripper moves across the surface. The *threading beads* (Hoque et al., 2026) sequence involves small objects and fine hand movements, while *wrapping food in foil* (Hoque et al., 2026) introduces deformation of thin, reflective material. In these examples, our method completes regions missing from the warps while retaining the interaction structure and surrounding scene under the changing viewpoint.

E.4 Additional 4D Reconstruction Examples

Reconstruction setup. We use frozen, released Pi3X (Wang et al., 2025b) and OmniX (Jiang et al., 2026a) checkpoints without fine-tuning. We uniformly sample 16 frames from each input video and select the same frame indices from its generated counterpart, assuming that the actions remain synchronized. Both streams undergo aspect-preserving resizing and center cropping to 504×280 pixels. For each model, we run inference twice: once on the 16 input frames, and once on those frames followed by the 16 generated frames. Each run processes all sampled frames jointly; OmniX additionally receives video and frame indices. No camera poses, intrinsics, or depths are supplied, including the camera trajectories used for generation.

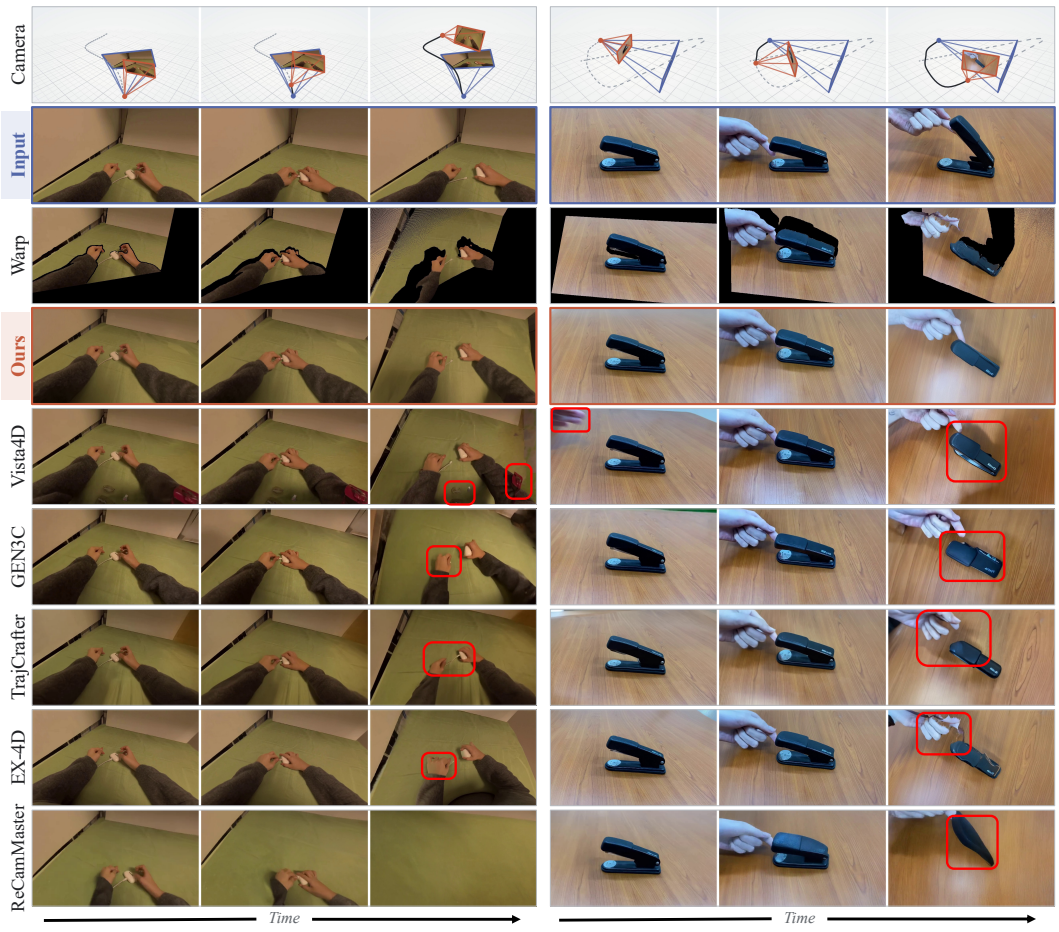

Figure 18: **Additional qualitative comparisons on human–object interaction videos.** In the *unplugging an AirPods case* (Hoque et al., 2026) sequence, competing methods introduce artifacts or implausible content in newly revealed regions. Our method completes these regions more coherently while preserving the unplugging action. In the *opening a stapler* (Song et al., 2024) sequence, it maintains the articulated structure and its interaction with the hand, while competing methods often deform or disconnect object parts.

Comparison visualization. We align the joint reconstruction to the input-only result with one robust, sequence-wide similarity transform fitted to confidence-screened corresponding source pixels across all 16 timesteps. The transform is fixed across time and uses no ground-truth calibration. We render predicted pointmaps using the same 50,000 source-pixel indices and colors in both conditions at every timestep, under a shared inspection camera and crop. Generated-view points are hidden, and displayed points undergo no confidence pruning, static merging, or temporal accumulation. This comparison isolates changes in estimated source-view geometry.

Additional examples. Figure 24 extends the qualitative study in Section 4.4 to four additional interactions. We use Pi3X for the first two rows and OmniX for the last two, keeping the reconstruction model fixed within each comparison. With the generated view, the hand and lighter are less stretched, the visible gripper–lid gap is smaller, and the robot arms appear less flattened toward the workspace. Surfaces remain incomplete.

F Limitations and Future Work

Limitations. ROVER improves interaction fidelity and visual quality under object-centric viewpoint changes. It also generalizes to robotic manipulation without robot-specific fine-tuning. However,

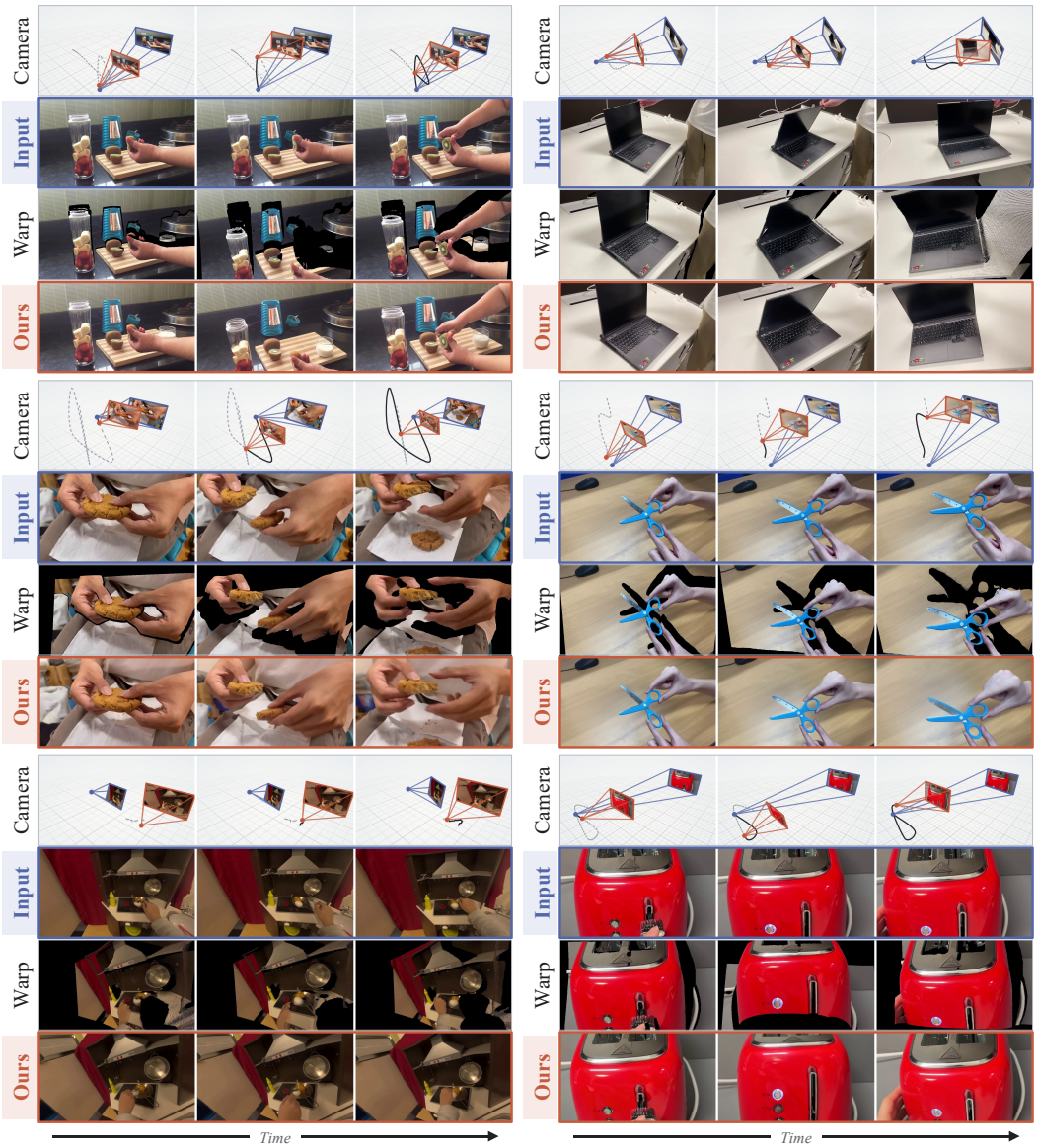

Figure 19: **Additional results from our method on diverse human–object interactions.** From top to bottom, the examples show *peeling a kiwi* (Mandikal et al., 2025) and *adjusting a laptop screen* (Song et al., 2024); *breaking a cookie* (Chen et al., 2025b) and *opening and closing scissors* (Song et al., 2024); and *frying an egg* (Hoque et al., 2026) and *lowering a toaster lever* (Cheng et al., 2025). These examples cover non-rigid deformation, topology changes, articulated motion, and changes in appearance and object state.

several limitations remain. The coverage and realism of OBJALIVE depend on the input images, event prompts, and underlying video generator; artifacts and implausible dynamics may remain despite filtering. For long videos, errors can accumulate through history conditioning, while inaccurate dynamic-object masks can introduce moving content into the static background memory. Inference remains time-consuming due to multi-step diffusion sampling and geometry processing, limiting interactive use. Finally, visual quality and human preference do not establish physical correctness or downstream robot-policy utility.

Failure cases. Figure 25 illustrates limitations in preserving fine visual details under viewpoint changes. In the laptop sequence, ROVER does not consistently preserve the keyboard structure and key markings. In the pencil-sharpener sequence, the printed label is distorted and its characters vary

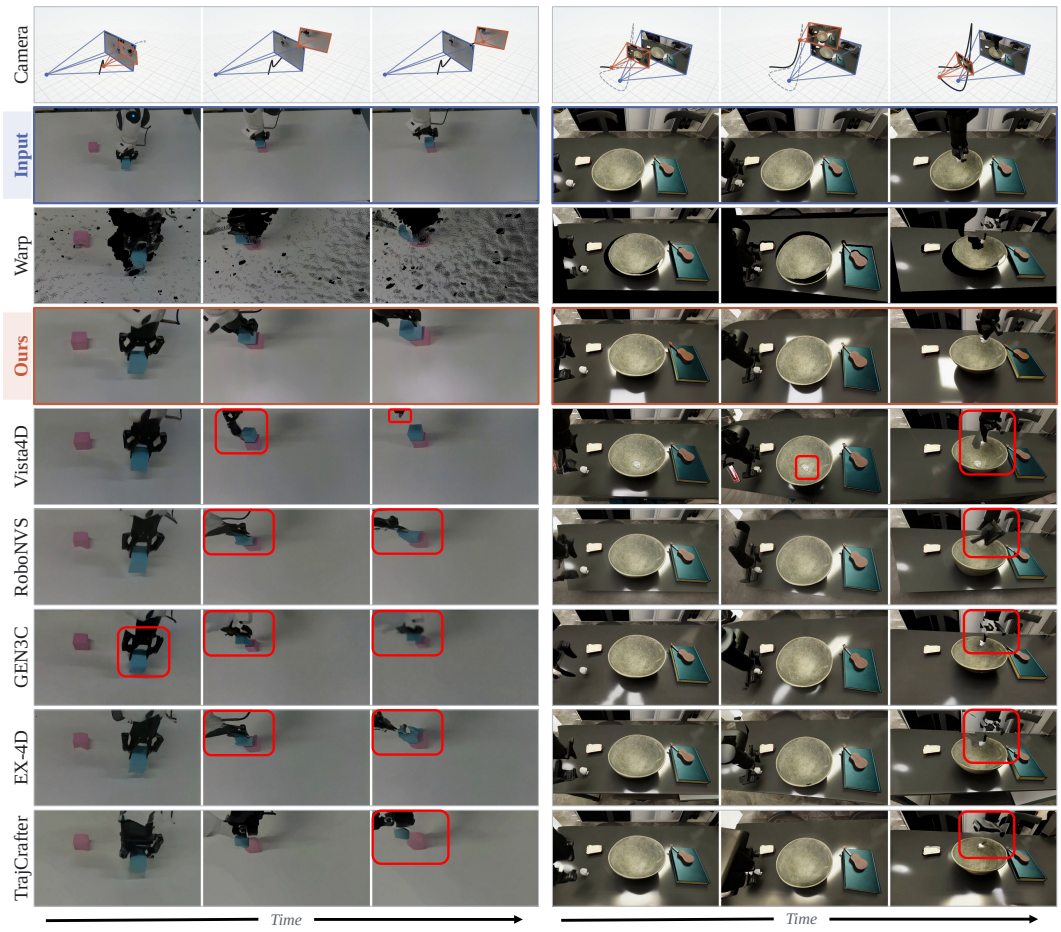

Figure 20: **Additional qualitative comparisons on robotic manipulation videos.** In *block stacking* (Wu et al., 2024), competing methods produce malformed grippers, blurred or deformed blocks, or incorrect spatial relationships between the blocks. Our method better preserves gripper geometry and the relative positions of the blocks during stacking. In *placing an object in a bowl* (Zhao et al., 2026), competing methods produce distorted or fragmented robot arms, while our method better preserves the arm structure throughout the manipulation.

across frames. These examples show that plausible overall appearance does not guarantee faithful reproduction of small text and repetitive structures.

Future work. Future directions include broadening object–event coverage and improving quality control for generated training videos. More robust history conditioning and more reliable separation of static and dynamic content could improve long-video consistency. Faster sampling and more efficient geometry processing could make reshooting more practical for interactive applications. Contact-aware evaluation and downstream robotic tasks would further assess physical consistency and task utility beyond visual quality and human preference.



Figure 21: **Additional qualitative comparisons on robotic manipulation videos.** In *placing a brush in a bowl* (Li et al., 2025a) and *turning a knob* (Wu et al., 2024), competing methods produce distorted grippers, fragmented object boundaries, or artifacts around the interaction region. Our method better preserves the robot and object structure under the requested camera motion.

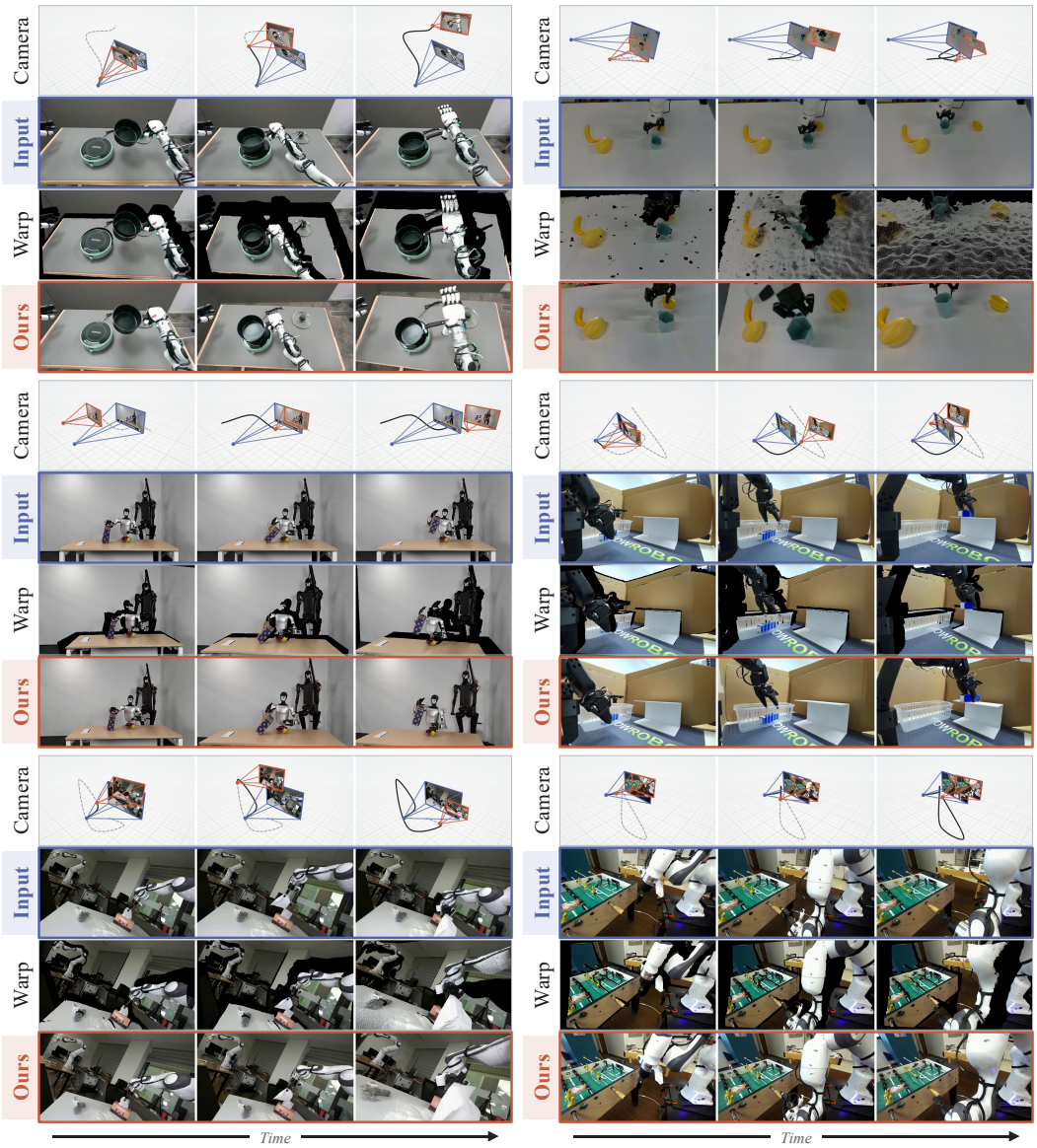

Figure 22: **Additional results from our method on diverse robotic manipulation videos.** From top to bottom, the examples show *placing a pot on a cooker* (Li et al., 2025a) and *moving a cup* (Wu et al., 2024); *lifting a box* (Li et al., 2025c) and *manipulating an eye-drop bottle* (Selvaraj et al., 2026); and *pulling a tissue* (Khazatsky et al., 2024) and *twisting a foosball handle* (Khazatsky et al., 2024). Our results preserve grasp relationships, robot and object structure, and tissue deformation while completing regions missing from the geometric warps.

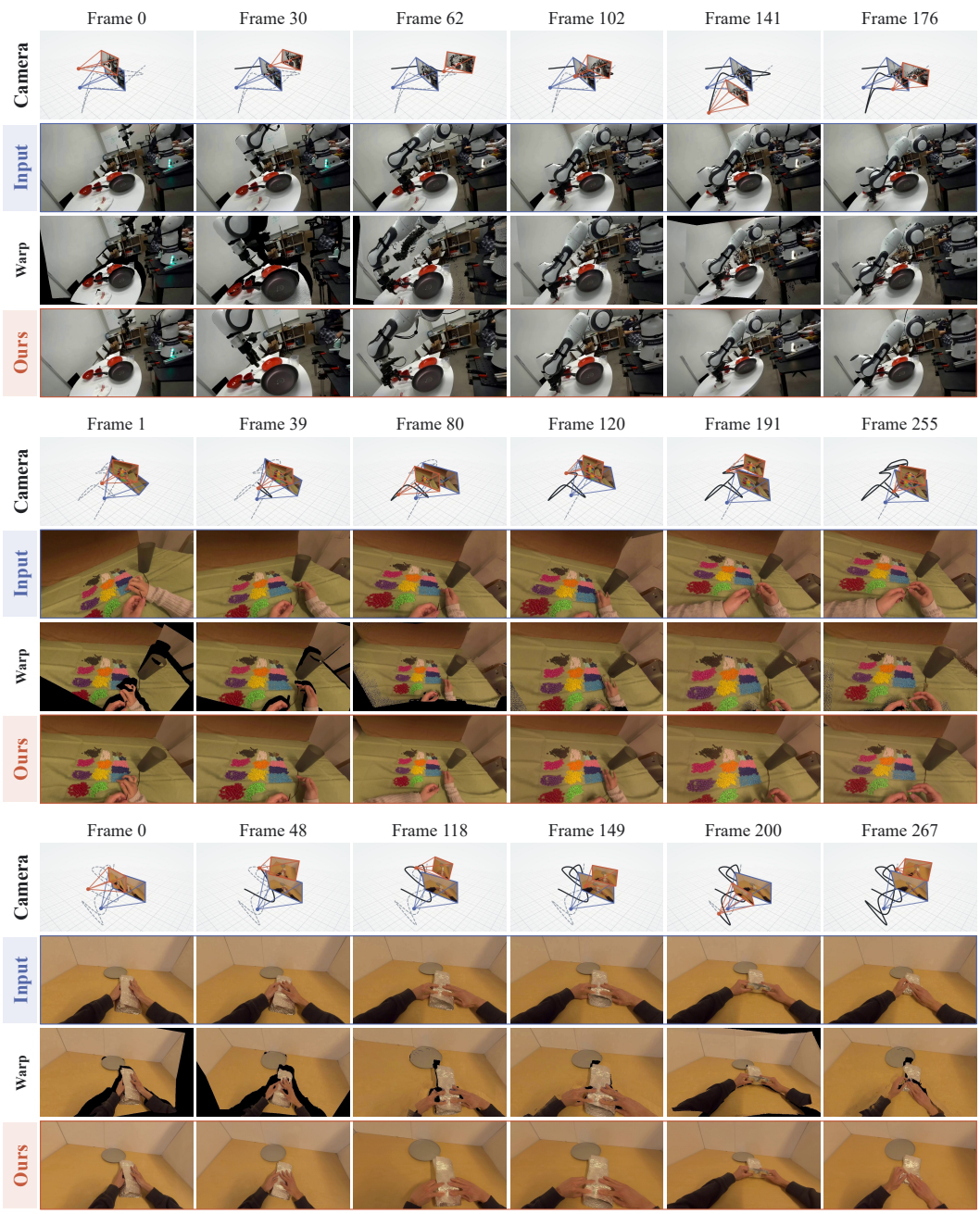

Figure 23: **Additional results on long-video camera control.** From top to bottom, the examples show *wiping a table* (Khazatsky et al., 2024), *threading beads* (Hoque et al., 2026), and *wrapping food in foil* (Hoque et al., 2026). Each group includes camera diagrams, input frames, warps, and our results at six labeled frame indices. Our results preserve interaction structure while completing missing content across multiple generation windows.

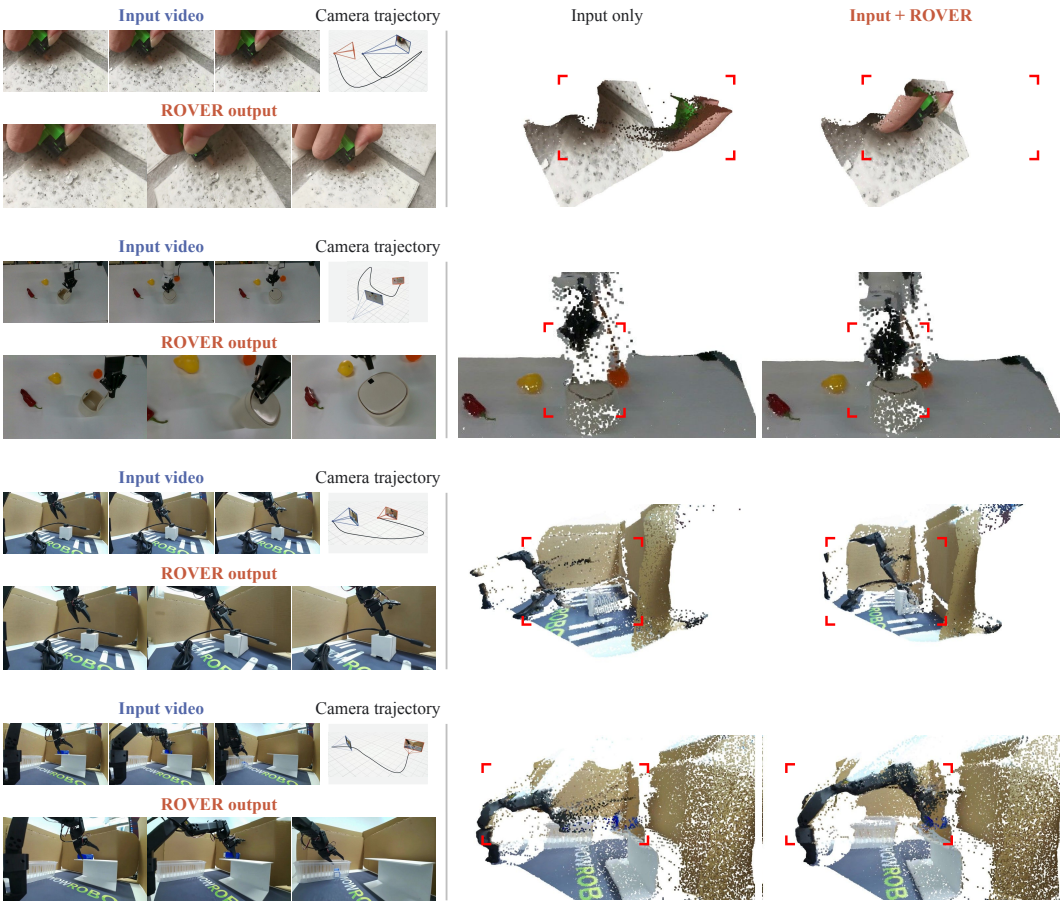

Figure 24: **Additional qualitative 4D reconstruction examples.** Input and generated keyframes accompany target camera trajectories (blue: source; orange: target). Pointmaps at the middle keyframe compare input-only and joint reconstruction using Pi3X (top two rows) and OmniX (bottom two). Each pair retains only source-view points under the same inspection camera and crop, with static merging off. Corners mark corresponding regions.

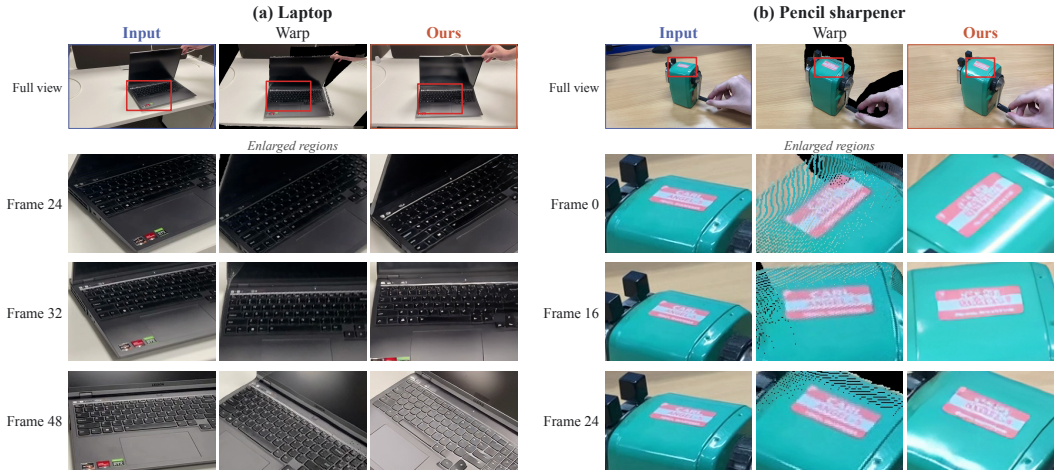

Figure 25: **Failure cases in preserving fine visual details.** Input frames, geometric warps, and generated views are shown for a laptop and a pencil sharpener. Enlarged regions reveal distorted keyboard details and inconsistent label text across frames. Red boxes indicate the enlarged regions.