REVIEW: NOVEL-VIEW VIDEO GENERATION WITH SPATIOTEMPORAL CONSISTENCY

Anonymous authors
Paper under double-blind review

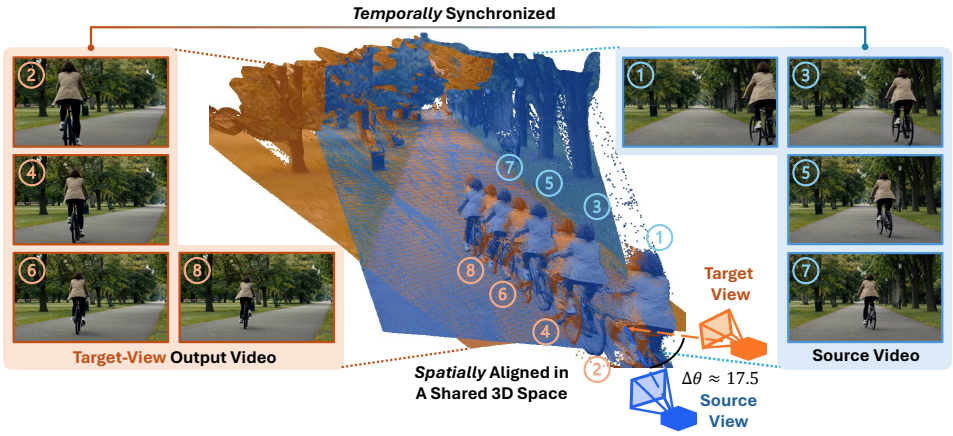

Figure 1: **ReView: replaying the same event from a new view.** We interleave source-video frames (blue) and generated target-view frames (orange) at corresponding time indices. Observed from different viewpoints, the two sequences depict the same action unfolding and remain spatially aligned in a shared 3D space, demonstrating faithful event replay with spatiotemporal consistency.

ABSTRACT

We study novel-view video generation: given an input video and a target camera viewpoint, the goal is to synthesize a high-quality video that faithfully replays the same dynamic event from the target view while preserving *spatiotemporal consistency* with the source video. We introduce *ReView*, a two-stage framework built on a pretrained video generation model. First, we adapt the model to novel-view generation through supervised training on a mixture of synthesized monocular-video pairs, simulated multi-view videos, and real-world multi-view images. We then apply reinforcement learning (RL) as a post-training stage to further improve spatiotemporal consistency. Specifically, we introduce a *Cross-View Reprojection Reward* that promotes cross-view geometric consistency through 3D reprojection and visual matching. To improve the efficiency of video RL, we further propose *Progressive-Horizon Reinforcement Learning*, which progressively shifts updates from single-frame to joint-video rollouts through single-frame, mixed, and full-video stages. This strategy attains a geometric consistency performance comparable to full-video RL using less than one third as many training compute. Experiments show improved viewpoint accuracy and cross-view consistency, with competitive novel-view reconstruction and video quality. These results suggest that pretrained video priors, together with multi-view supervision and geometry-aware post-training, offer a practical route toward more scalable 4D generation.

1 INTRODUCTION

Recent progress in video generation (Wan et al., 2025; Wu et al., 2025; Team et al., 2025; HaCohen et al., 2026; Seedance et al., 2026) has enabled realistic motion and appearance in complex dynamic scenes, while advances in 3D generation (Gao et al., 2024; Shi et al., 2024; Yu et al., 2025a; Xiang et al., 2026; Han et al., 2026) have made it possible to synthesize geometrically coherent content

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.

that can be observed from different viewpoints. Bridging these two capabilities naturally raises a challenging question: how can a video model reproduce an observed dynamic event from a new viewpoint while preserving its spatial structure and temporal evolution?

In this work, we study *direct novel-view video generation*: given a source video and a prescribed target viewpoint, we aim to replay the source event from that viewpoint throughout the output sequence, without a gradual camera transition from the source view. Beyond visual fidelity, this requires **spatiotemporal consistency** with the source (Figure 1). Spatially, corresponding content should agree with a shared 3D scene and the requested camera transformation. Temporally, the output should preserve the source event's motion, ordering, and synchronization. The goal is a faithful alternative observation of the same event, rather than a visually similar but independently generated sequence. Camera-trajectory-controllable methods (Bai et al., 2025a; Ren et al., 2025; Zhao et al., 2026; Xu et al., 2026; Fu et al., 2026) address generation under continuously moving viewpoints. We focus instead on direct redirection with a potentially large viewpoint change imposed from the outset, rather than a gradual camera transition toward the target view. Without smooth, high-overlap views, the model must immediately relocate source content, complete newly revealed regions, and preserve the same event throughout the target video. This setting places particular demands on geometric grounding and synchronized cross-view dynamics, and requires the model to preserve the same event across two spatially distinct observations.

We introduce **ReView**, a two-stage framework built on a pretrained video model (Wan et al., 2025). Supervised adaptation first equips the model with target-aligned geometric control and source-reference fusion: projected source observations specify target-view locations, while reference attention supplies complementary appearance and motion context. To address scarce synchronized multi-view supervision, we combine reprojection-based pseudo pairs from monocular videos (Yu et al., 2025b), calibrated multi-view videos collected through our game-engine and robot-simulation pipelines, and real-world multi-view images (Ling et al., 2024). Monocular reprojection provides scalable training pairs but cannot supply independent observations of content hidden from the source camera. Our data engine pipelines address this limitation by capturing the same dynamic event from multiple calibrated cameras, providing direct supervision for newly visible content and evolving occlusions. Real-world multi-view images further broaden the appearance diversity and generalization. Training on this mixture yields our supervised model, *ReView-Base*.

Supervised adaptation provides a strong initialization, but its generative objective does not directly measure source–output geometric agreement, leaving visually convincing outputs susceptible to geometric drift and temporal deviations. We therefore apply reinforcement learning (RL) (Liu et al., 2026; Zheng et al., 2026) to shift ReView-Base toward geometrically and temporally consistent generations. Our *Cross-View Reprojection Reward (CVRR)* unifies visual matching and geometric reprojection. For temporally aligned source–target frames, visual matching identifies where source content appears in the generated view, while geometric reprojection predicts where it should appear under the prescribed camera transformation. Closer agreement between these locations yields a higher reward. During video RL, this reward encourages source-content preservation and viewpoint adherence. Video RL is nevertheless expensive: every sample requires multiple full-video rollouts involving repeated denoising of spatiotemporal tokens (Liu et al., 2026). Fewer policy-update transitions or reward-evaluated frames reduce part of this cost without eliminating full-video generation costs. We introduce *Progressive-Horizon Reinforcement Learning (PHRL)*, which progressively shifts updates from single-frame to joint-video generation. The curriculum begins with inexpensive single-frame rollouts for cross-view alignment, proceeds to mixed single-frame and video updates, and concludes with full-video refinement. Our contributions are summarized as follows:

- **Direct, spatiotemporally consistent novel-view generation.** We introduce ReView with viewpoint-aligned geometric control and source-reference fusion to preserve source appearance and dynamics under prescribed viewpoint. We also develop game-engine and robot-simulation pipelines that collect synchronized multi-view videos for supervised adaptation.
- **Efficient geometry-aware video RL.** We develop a video RL pipeline (PHRL) that couples a cross-view reprojection reward with a progressive frame-to-video curriculum. Low-cost single-frame updates establish cross-view alignment, while subsequent mixed and video-only updates refine the shared policy under joint generation.
- **Empirical validation of consistency and efficiency.** Experiments demonstrate improved viewpoint accuracy, cross-view consistency, and motion reconstruction with competitive video quality.

PHRL matches full-video-only RL across geometric consistency, motion accuracy, and perceptual quality, with more than $3\times$ higher training efficiency.

2 Related Works

Camera-controlled video generation and video reshooting. MotionCtrl (Wang et al., 2024), CameraCtrl (He et al., 2024; 2025), and follow-up work (Bahmani et al., 2025; Wang et al., 2026b) condition video generation on camera motion or pose sequences, whereas video reshooting requires preserving a source video's appearance and synchronized dynamics. One line of work (Bai et al., 2025a; Zhang et al., 2025; Van Hoorick et al., 2026) addresses this task by injecting source-video as well as camera conditions and training on synchronized videos rendered from multiple cameras. Another grounds target-view generation in explicit geometry (Van Hoorick et al., 2024; Chen et al., 2026; Yu et al., 2025b; Ren et al., 2025; Hu et al., 2025; Lin et al., 2026; Cai et al., 2026), using video-depth estimates as conditioning or maintaining temporally persistent 4D point clouds to provide target-view geometric guidance. Compared with trajectory-based camera control, directly replaying the complete source event from a spatially separated viewpoint remains less explored. This motivates our focus on target-view geometric grounding and synchronized cross-view dynamics.

Reinforcement learning for diffusion models. RL supports language-model alignment and reasoning (Ouyang et al., 2022; Guo et al., 2025; Shao et al., 2024) and has been adapted to diffusion post-training (Black et al., 2024). Building on these early formulations, Flow-GRPO (Liu et al., 2026) converts deterministic flow sampling into a marginal-preserving SDE, enabling online exploration and group-relative policy optimization while reducing the denoising steps used for training. Subsequent methods address complementary costs: DanceGRPO (Xue et al., 2025), MixGRPO (Li et al., 2025), and Flash-GRPO (He et al., 2026) sparsify denoising-trajectory updates, while Branch-GRPO (Li et al., 2026) shares computation across branched rollouts. At the sample and objective levels, replay-based methods reuse historical trajectories (Zhang et al., 2026a; Wang et al., 2026a), whereas DiffusionNFT (Zheng et al., 2026) replaces reverse-trajectory likelihood optimization with a forward-process objective. For video models, sparse-frame rewards reduce decoding and evaluation costs but generally retain full latent rollouts (Prabhudesai et al., 2024; Yuan et al., 2024); ContentV and Astrolabe instead use fixed shorter rollouts or architecture-specific local-clip optimization (Lin et al., 2025b; Zhang et al., 2026b). Full-video rollouts nevertheless remain substantially more expensive than single-frame generation. While these strategies improve update efficiency, reuse computation, or restrict temporal extent, the allocation between single-frame and joint-video training remains less explored. Our progressive curriculum addresses this allocation, learning alignment through inexpensive single-frame updates before shifting toward joint-video refinement.

3 Method

3.1 Problem Formulation

Let $\mathbf{X}^s = \{\mathbf{I}_i^s\}_{i=1}^F$ be a monocular source video and $\Delta\mathbf{C} = [\mathbf{R} \mid \mathbf{t}]$ the requested source-to-target camera transformation. We generate a target video $\mathbf{X}^t = \{\mathbf{I}_i^t\}_{i=1}^F$ that depicts the same event from the requested viewpoint throughout the sequence, preserving source appearance, dynamics, and temporal alignment. The target viewpoint applies from the first output frame, rather than being reached through a gradual camera transition. ReView first adapts a pretrained video model through geometric control, source-reference fusion, and cross-view supervision, then applies geometry-aware RL with a progressive frame-to-video curriculum.

3.2 Geometry-Grounded Conditional Video Transformer

Explicit target-view grounding. We estimate temporally consistent source depths $\widehat{\mathbf{D}} = \{\widehat{\mathbf{D}}_i\}_{i=1}^F$ using DepthCrafter. Given source and target intrinsics $\mathbf{K}^s$ and $\mathbf{K}^t$, we back-project a homogeneous source pixel $\mathbf{u} = (u, v, 1)^\top$ into 3D, transform it by $\Delta\mathbf{C}$, and project it into the target image:

$$\widetilde{\mathbf{u}} \sim \mathbf{K}^t \left(\mathbf{R} \left[\widehat{\mathbf{D}}_i(\mathbf{u})(\mathbf{K}^s)^{-1}\mathbf{u} \right] + \mathbf{t} \right). \tag{1}$$

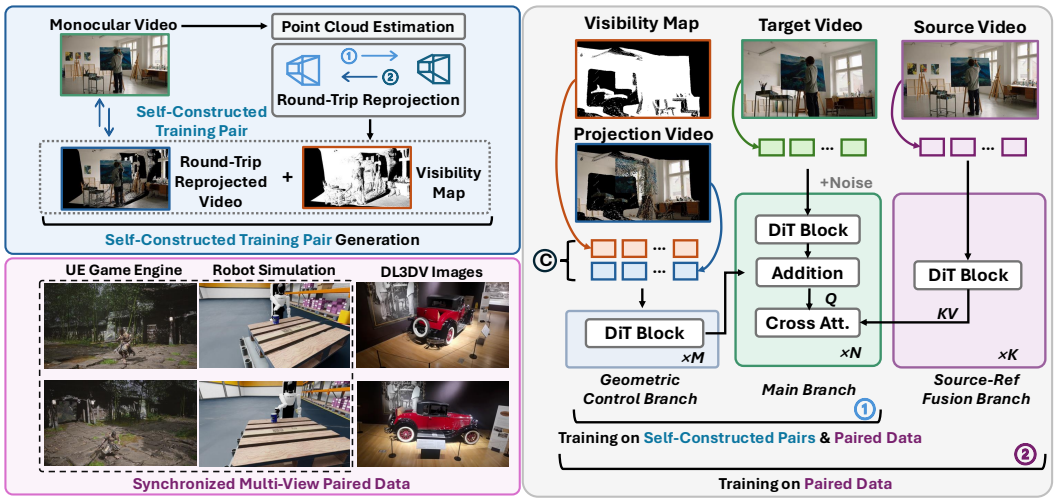

Figure 2: **Left:** We use three data sources: monocular videos, synchronized synthetic multi-view videos, and real-world multi-view images. **Right:** Our two-stage supervised adaptation injects geometric and source-reference conditions through the two branches.

Depth and camera translation use compatible coordinate and scale conventions. We retain projections with positive target-camera depth within the target image and resolve collisions with a z-buffer, producing a partial target-view render $\mathbf{P} = \{\mathbf{P}_i\}_{i=1}^{F}$ and a binary visibility mask $\mathbf{A} = \{\mathbf{A}_i\}_{i=1}^{F}$. The mask marks target pixels supported by projected source geometry; pixels outside this support require completion by the generative model. This representation makes the camera condition spatially explicit, placing observed source content directly at its predicted target-view locations.

Spatiotemporally aligned control. We build on the pretrained Wan2.2-5B video diffusion transformer and encode the render and mask as $\mathbf{z}^p = \mathcal{E}_p(\mathbf{P})$ and $\mathbf{z}^a = \mathcal{E}_a(\mathbf{A})$ on the target latent grid, where $\mathcal{E}_p$ and $\mathcal{E}_a$ map the conditioning inputs to the spatial and temporal layout of target tokens. Their channel-wise concatenation, $\mathbf{c}^g = [\mathbf{z}^p; \mathbf{z}^a]$, conditions a ControlNet-style (Zhang et al., 2023) branch initialized from backbone blocks. At its first block, projected control tokens are added to the noisy target tokens; the resulting state propagates through the branch, producing a residual at each control block that is injected after its corresponding backbone block. Zero-initialized output projections make the branch initially contribute no residual, while the aligned feature grids maintain target-view grounding throughout denoising.

Source-reference fusion. Reprojection provides explicit geometric guidance but is limited by source visibility and can be unreliable near depth discontinuities, thin structures, and moving-object boundaries. We therefore complement geometric guidance with source-event appearance and motion context through source-reference attention. Specifically, we encode the complete source video as $\mathbf{z}^s = \mathcal{E}(\mathbf{X}^s)$, where $\mathcal{E}$ is the backbone video encoder, and patchify the resulting latent into reference tokens. Within each fusion module, self-attention first updates the reference tokens. Cross-attention then uses target denoising tokens as queries and source-reference tokens as keys and values. This lets target tokens retrieve source appearance and motion context beyond the partial render, complementing the explicit target-view locations supplied by geometric control (Figure 2). Block configurations, initialization, and conditioning perturbations are detailed in Appendix B.1.

3.3 Cross-View Supervision

Multi-View Synchronized Dataset. Reprojection-based pseudo pairs cannot provide independent target-view observations of content hidden from the source camera. This is particularly limiting for direct viewpoint changes, where newly revealed regions may remain unobserved throughout the source sequence. Synchronized collections such as SynCamMaster (Bai et al., 2025b) instead use rendered multi-view videos. To broaden scene and interaction diversity, we collect calibrated, time-synchronized videos through game-engine and robot-simulation pipelines (Wang et al., 2026c). The game collection contains 1,000 sequences with 25 views each, covering visually rich environments and character dynamics; the robot collection contains 2,000 manipulation episodes with 12 views each, covering robot movements and object interactions. Together, they provide direct supervision

for newly visible content and evolving occlusions across viewpoints. Each sequence contains videos and associated camera parameters, with corresponding frame indices observing the same scene state. Sampling two cameras yields a training tuple $(\mathbf{X}^s, \mathbf{X}^t, \Delta\mathbf{C})$, with the relative transformation computed from calibrated poses. These synchronized observations preserve event identity and timing while varying viewpoint, enabling supervision of both appearance and motion beyond source-view warping. Our collection complements monocular pseudo pairs and real-world multi-view images during supervised adaptation. Appendix C describes the capture pipelines, camera layouts, pair construction, and dataset visualizations; training data recipes are specified in Appendix B.2.

Supervised flow matching. Given a clean target latent $\mathbf{z}_0$, Gaussian noise $\boldsymbol{\epsilon}$, and $\sigma \in [0, 1]$, define $\mathbf{z}_\sigma = (1-\sigma)\mathbf{z}_0 + \sigma\boldsymbol{\epsilon}$ with target velocity $\mathbf{v}^\star = \boldsymbol{\epsilon} - \mathbf{z}_0$. The conditional vector field v_Θ minimizes

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\mathbf{z}_0,\boldsymbol{\epsilon},\sigma}\left[w(\sigma)\,\|v_\Theta(\mathbf{z}_\sigma; \mathbf{z}^p, \mathbf{z}^a, \mathbf{z}^s) - \mathbf{v}^\star\|_2^2\right], \tag{2}$$

where $w(\sigma)$ is the prescribed loss weight. We train the conditioning pathways sequentially. First, we optimize only the geometric control branch on pseudo and paired data, keeping the pretrained backbone frozen. We then freeze the learned control branch as well and train only the source-reference fusion modules on paired data (Figure 2). Establishing target-view grounding before learning reference fusion reduces the risk that source conditioning bypasses the geometric condition.

3.4 Cross-view Consistency Reinforcement

3.4.1 Cross-view Consistency Reward

To complement flow-matching supervision, our *Cross-View Reprojection Reward (CVRR)* measures agreement between visual matching and geometric reprojection. Visual matching locates source content in the generated view, while reprojection predicts its location under the requested camera transformation. For each temporally corresponding source–target frame pair, SuperPoint and LightGlue identify correspondences $\{(\mathbf{q}_j^s, \mathbf{q}_j^t)\}_{j=1}^J$. Joint two-view geometry inference provides source depth $\widehat{\mathbf{D}}^s$ and intrinsics $(\widehat{\mathbf{K}}^s, \widehat{\mathbf{K}}^t)$. We project each matched source point to $\widehat{\mathbf{q}}_j^t$ via Equation 1, combining the estimated depth and intrinsics with the prescribed transformation $\Delta\mathbf{C}$, all expressed in compatible coordinate and scale conventions. RL holds this input pose fixed, whereas evaluation-time RE@SG uses estimated poses. Over valid matches $\mathcal{V}$, we compute the median pixel error $e_i = \mathrm{median}_{j\in\mathcal{V}}\,\|\widehat{\mathbf{q}}_j^t - \mathbf{q}_j^t\|_2$ at the original image resolution and define

$$r_i = r_{\mathrm{geo}}(\mathbf{I}_i^s, \widehat{\mathbf{I}}_i^t) = \frac{1}{s}\,\mathrm{clip}(r_{\max} - e_i, r_{\min}, r_{\max}), \tag{3}$$

where $r_{\min}$ and $r_{\max}$ bound the clipped value and $s > 0$ scales the reward; the fixed camera argument is omitted for brevity. Insufficient valid matches often indicate poor preservation of source content in the generated view; such pairs receive the reward floor $r_{\min}/s$. Video rewards average corresponding frame pairs as $R_F(\mathbf{X}^s, \widehat{\mathbf{X}}^t) = \frac{1}{F}\sum_{i=1}^F r_i$.

Visual matching and geometric reprojection serve complementary roles. Photometric comparisons with source-derived references can penalize legitimate view-dependent shading and appearance changes or favor source-copying shortcuts. Feature matching instead identifies recognizable source content while tolerating moderate appearance variation, without requiring identical RGB values across viewpoints. Reprojection additionally assesses whether that content occupies locations consistent with the requested camera; visual similarity alone is insufficient. Together, they encourage source-content preservation at geometrically appropriate locations rather than source-view copying. The comparison also follows the source timeline: as the event evolves, each generated frame is evaluated against the corresponding source observation. Motion captured by reliable correspondences therefore contributes to the reward, discouraging frozen or temporally misaligned reconstructions. During joint-video RL, these spatial and time-aligned signals encourage preservation of the source event's appearance and motion from the requested viewpoint. Appendix D illustrates the correspondence patterns associated with content distortion, temporal mismatch, and camera errors.

3.4.2 Frame-to-video policy optimization

Full-video RL repeatedly denoises multiple spatiotemporal candidates. Evaluating fewer reward frames or updating fewer denoising transitions saves computation but leaves full-video generation

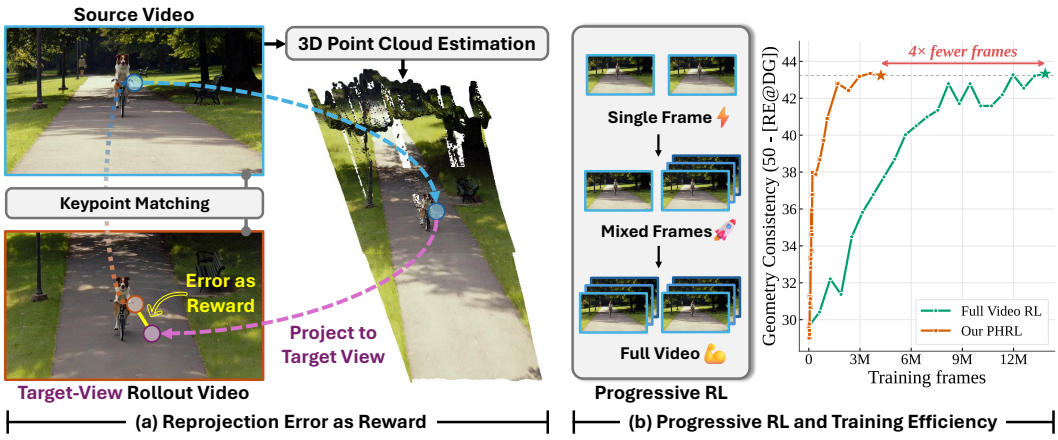

Figure 3: **Cross-view reward and progressive video RL.** (a) CVRR compares visually matched target keypoints with source points reprojected under the prescribed camera transformation; closer agreement yields higher reward. (b) PHRL progresses from single-frame initialization to mixed single-frame/video updates and full-video refinement. All stages optimize the same policy with CVRR, increasing the share of joint-video updates while keeping video rollouts at a fixed length.

expensive. Our *Progressive-Horizon Reinforcement Learning* (PHRL) learns cross-view alignment through low-cost single-frame rollouts before shifting toward joint-video refinement (Figure 3).

Let $J_1(\theta)$ be the expected reward for single-frame generation at a uniformly sampled source-frame index, and $J_F(\theta)$ the expected mean frame reward under joint F-frame generation. Both modes share policy parameters θ and the same cross-view reward function. Their allocation is described by

$$J_\alpha^{\text{mix}}(\theta) = (1-\alpha)J_1(\theta) + \alpha J_F(\theta), \qquad \alpha \in [0,1], \tag{4}$$

where we increase the video-update probability α, forming the three-stage schedule in PHRL:

1. **Single-frame initialization** ($\alpha = 0$)**.** Learn cross-view alignment using single-frame rollouts.
2. **Mixed frame-to-video transfer** ($\alpha = 1/3$)**.** Sample single-frame and video updates with expected ratio of 2:1, retaining inexpensive updates while adapting the policy to joint generation.
3. **Full-video refinement** ($\alpha = 1$)**.** Refine the policy exclusively with video rollouts.

Detailed stage lengths and rollout settings are given in Appendix B.4.

Across all three stages, we optimize the shared LoRA parameters using Flow-GRPO (Liu et al., 2026). All candidates within an update use the same execution mode. For each condition, G candidates share the selected frame indices and receive either the single-frame reward or video reward R_F. Stochastic SDE transitions enable exploration, and group-relative advantages guide clipped policy updates. The complete optimization objective and normalization rules are given in Appendix B.5.

The relationship between the two execution modes can be characterized through their induced output distributions. For bounded frame rewards with range $B = \overline{r} - \underline{r}$, the objectives satisfy

$$|J_F(\theta) - J_1(\theta)| \le B\,\overline{\delta}_F(\theta), \tag{5}$$

where $\overline{\delta}_F(\theta)$ is the average total-variation distance between frame marginals from joint-video generation and outputs of single-frame generation, over conditions and frame indices. When these distributions are close, the single-frame objective approximates the video objective, motivating its use for low-cost initialization. Formal definitions, proofs, and mixed-objective analysis are in Appendix A.

4 Experiments

4.1 Implementation and Evaluation

Implementation. We initialize ReView from Wan2.2-5B (Wan et al., 2025), sequentially train the geometry-control and source-reference modules with cross-view supervision, and apply geometry-aware Flow-GRPO (Liu et al., 2026) with LoRA and three-stage PHRL. Supervised training and evaluation use 61-frame 720p videos unless otherwise specified; RL uses single-frame or 41-frame rollouts. Appendix B details the training data, optimization, hardware, and RL schedule.

Table 1: **Viewpoint-control accuracy and cross-view 3D consistency.** Throughout this paper, **Bold** and underlined values denote best and second-best results, respectively.

Method	Translation Numerical error ⇓	Translation Direction error ⇓	Rotation error ⇓	RE@SG ⇓
ReCamMaster	0.322	8.72	9.12	28.64
TrajectoryCrafter	0.225	5.69	6.75	21.29
EX-4D	0.198	5.01	6.15	15.86
GEN3C	0.195	4.83	6.36	13.79
Vista4D	0.187	4.27	5.07	8.51
Ours	**0.175**	**3.95**	**4.56**	**6.73**

Table 2: **Novel-view reconstruction and motion accuracy on N3DV and iPhone.** The prefix "m" denotes covisibility masking. EPE compares inter-frame flow estimated from generated and synchronized ground-truth target videos.

Method	N3DV							iPhone						
	mPSNR ⇑	mSSIM ⇑	mLPIPS ⇓	PSNR ⇑	SSIM ⇑	LPIPS ⇓	EPE ⇓	mPSNR ⇑	mSSIM ⇑	mLPIPS ⇓	PSNR ⇑	SSIM ⇑	LPIPS ⇓	EPE ⇓
ReCamMaster	14.08	0.433	0.422	13.82	0.424	0.540	5.092	10.93	0.405	0.619	10.97	**0.395**	0.676	5.075
TrajectoryCrafter	18.60	**0.713**	0.288	15.98	0.530	0.366	4.002	13.89	**0.493**	0.530	12.99	0.325	0.586	2.833
EX-4D	18.55	0.635	**0.240**	16.92	0.552	0.358	4.678	13.05	0.450	0.583	12.87	0.308	0.599	4.357
GEN3C	18.15	0.621	0.302	16.54	0.547	0.389	4.808	12.99	0.439	0.579	12.70	0.294	0.575	3.562
Vista4D	18.68	0.604	0.275	16.69	0.587	**0.315**	2.356	14.25	0.475	**0.430**	14.02	0.317	**0.492**	1.175
Ours	**19.56**	0.672	0.256	**17.35**	**0.589**	0.338	**2.078**	**14.39**	0.482	0.463	**14.11**	0.331	0.531	**1.093**

Evaluation benchmarks. We assess open-world generalization and controlled target-view reconstruction on complementary reference-free and reference-based benchmarks. The reference-free benchmark contains 280 in-the-wild real-world videos and 120 synthesized by state-of-the-art video diffusion models, spanning diverse scenes, events, and motion patterns to evaluate viewpoint control, cross-view consistency, and perceptual quality without synchronized target references. Reference-based evaluation uses synchronized real-world multi-view data from iPhone (Gao et al., 2022) and N3DV (Li et al., 2022) to measure pixel-aligned novel-view fidelity. For N3DV, we randomly select three source–target camera pairs per scene, yielding 18 synchronized pairs; for iPhone, we use five valid sequences following prior video-reshooting evaluations (Yu et al., 2025b; Lin et al., 2026).

4.2 Main Results

Viewpoint control and cross-view consistency. On the reference-free benchmark, we compare estimated generated-video poses with requested target poses using translation numerical error, translation direction error, and rotation error. Cross-view 3D consistency is measured by RE@SG (Kant et al., 2025), a per-frame reprojection error computed from SuperPoint source–target correspondences (DeTone et al., 2018; Lindenberger et al., 2023) and DA3 geometry (Lin et al., 2025a). Evaluation uses estimated camera poses consistent with previous works, unlike the fixed input pose in our RL reward. ReView ranks first on all four metrics (Table 1), lowering the second-best RE@SG from 8.51 to 6.73 and achieving the lowest translation and rotation errors. These results establish the geometric-control performance of the complete system.

Novel-view reconstruction and motion accuracy. On iPhone and N3DV, we measure agreement with synchronized target video ground truth using PSNR, SSIM, and LPIPS (Zhang et al., 2018) and their covisibility-masked variants (mPSNR, mSSIM, mLPIPS), which assess source-content preservation in target regions observable from the source. Following Vista4D (Lin et al., 2026), we also report optical-flow endpoint error (EPE): the Euclidean difference between corresponding inter-frame flow vectors estimated from generated and ground-truth target videos. EPE measures reference-motion agreement beyond framewise reconstruction (formalized in Appendix B.7). ReView ranks first or second on all 14 metrics (Table 2) and achieves the lowest EPE on both datasets: 2.078 versus Vista4D's 2.356 on N3DV (11.8% lower) and 1.093 versus 1.175 on iPhone (7.0% lower). Together with leading mPSNR and competitive perceptual scores, these gains demonstrate faithful co-visible content preservation and improved spatial-temporal reconstruction, adding temporal evidence beyond the framewise geometric-consistency results in Table 1.

Video fidelity. We compute FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018) between generated and corresponding source videos. We also report five VBench dimensions (Huang et al.,

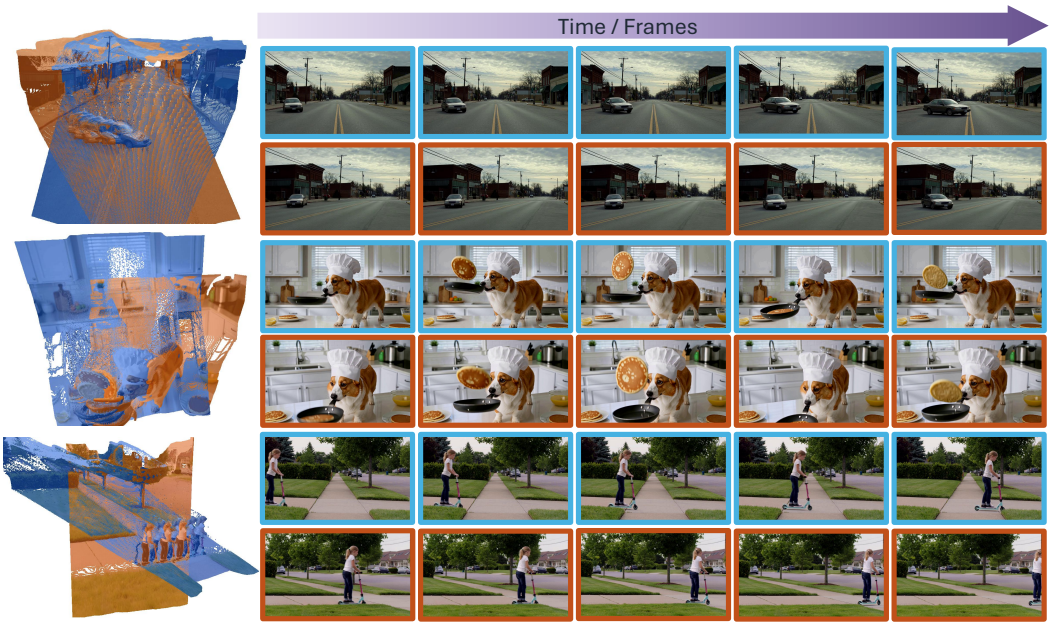

Figure 4: **Qualitative examples of direct novel-view video generation.** Source and generated frames are shown over time alongside their interleaved 3D point clouds, jointly illustrating temporal alignment and spatial consistency across viewpoints.

Table 3: **Video fidelity and temporal quality** on the reference-free benchmark.

Method	Base Model Size	FID ⇓	FVD ⇓	CLIP-T ⇑	VBench				
					Aesthetic quality ⇑	Imaging quality ⇑	Subject consistency ⇑	Background consistency ⇑	Temporal style ⇑
ReCamMaster	1.3B	88.7736	**233.8969**	0.3245	0.5777	0.6535	0.9680	0.9658	0.2788
TrajectoryCrafter	5B	128.0080	398.8139	0.3192	0.5704	0.6830	0.9627	0.9728	0.2680
EX-4D	14B	109.5105	298.6227	0.3189	0.5505	0.6355	0.9666	0.9702	0.2705
GEN3C	7B	105.4698	367.1137	0.3198	0.5616	0.6758	0.9437	0.9631	0.2741
Vista4D	14B	90.4365	246.7042	**0.3278**	**0.6025**	**0.6852**	0.9678	**0.9788**	**0.2811**
Ours	5B	**88.2768**	240.9733	0.3202	0.5815	0.6843	**0.9682**	0.9739	0.2799

2024), aesthetic quality, imaging quality, subject consistency, background consistency, and temporal style, and CLIP-T text alignment (Radford et al., 2021). ReView leads FID and subject consistency and ranks second in FVD and the remaining four VBench dimensions (Table 3), complementing geometric and motion-accuracy gains with competitive perceptual quality. ReCamMaster has the lowest FVD but larger pose errors and RE@SG; Vista4D leads several quality metrics with a 14B backbone versus ReView's 5B. Under our 720p evaluation setting, Vista4D requires over two hours of inference per video, compared with approximately five minutes for ReView. Besides, Figure 4 presents temporally ordered examples of direct novel-view video generation, complementing the reconstruction and flow-based measurements with a sequence-level view of appearance and motion.

4.3 Ablation Studies

Architecture and conditioning design. Table 4 ablates conditioning designs and post-training: *Camera Param.* replaces target-aligned geometric control with camera-parameter embeddings; *w/o Ref. Fusion* removes source-reference fusion while retaining geometric control; *Token Concat.* replaces both pathways with temporal concatenation of patchified projected-video, source-video, and noisy target-video tokens, following Vista4D-style in-context conditioning (Lin et al., 2026). *ReView-Base* and *ReView-RL* denote the supervised and PHRL-trained models, respectively. Relative to ReView-Base, using camera embeddings leads to increased RE@SG and reduced mPSNR/mSSIM, supporting dense target-aligned guidance over implicit camera conditioning. Removing reference fusion worsens RE@SG and lowers PSNR from 16.85 to 14.75, supporting complete source appearance as a complement to partial reprojections. Token concatenation remains competitive with ReView-Base but triples backbone sequence length for three equally sized streams, in-

Table 4: **Ablation of the conditioning architecture and geometry-aware post-training.** All architecture variants are trained with the supervised objective; ReView-RL further applies PHRL to ReView-Base. The prefix "m" denotes covisibility-masked evaluation.

Variant	Viewpoint Control and Cross-View Consistency				Novel-View Synthesis on N3DV					
	Translation Numerical error ⇓	Translation Direction error ⇓	Rotation error ⇓	RE@SG ⇓	mPSNR ⇑	mSSIM ⇑	mLPIPS ⇓	PSNR ⇑	SSIM ⇑	LPIPS ⇓
Camera Param.	0.297	7.50	9.20	25.89	14.15	0.465	0.386	14.11	0.427	0.415
w/o Ref. Fusion	0.211	4.82	5.80	15.22	17.69	0.589	0.303	14.75	0.505	0.396
Token Concat.	0.205	4.77	5.49	14.55	18.20	0.597	0.310	17.01	0.595	0.320
ReView-Base	0.208	4.93	5.89	14.36	18.02	0.610	0.297	16.85	0.592	0.324
ReView-RL	0.175	3.95	4.56	6.73	19.56	0.672	0.256	17.35	0.589	0.338

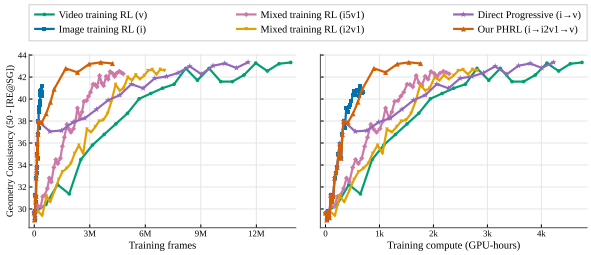

Figure 5: **RL schedule efficiency.** Evaluation score $50 - \mathrm{RE@SG}$ (↑) versus generated training frames (left) and GPU-hours (right). PHRL approaches video-only RL at lower cost.

	N3DV			VBench	
Variant	mPSNR ↑	mSSIM ↑	EPE ↓	Aes. ↑	Sty. ↑
Base	18.02	0.610	3.99	0.5915	0.2808
i	18.53	0.615	2.68	0.5633	0.2701
i5v1	18.92	0.628	2.25	0.5802	0.2773
v	19.53	0.681	1.99	0.5893	0.2805
PHRL	19.56	0.672	2.08	0.5815	0.2799

Table 5: **RL schedule quality.** Base is ReView-Base; *i* and *v* denote single-frame and video RL. Aes./Sty. denote aesthetic quality and temporal style.

creasing pairwise self-attention interactions nine-fold. Our design preserves target-token sequence length through separate control and reference computations. Geometry-aware post-training reduces RE@SG by 53.1% (14.36 to 6.73) and increases mPSNR by 1.54 dB. ReView-RL leads eight of ten metrics, with small regressions in full-frame SSIM and LPIPS.

Progressive-horizon reinforcement learning. Starting from ReView-Base with the same training reward, we compare PHRL with single-frame-only RL (*i*), full-video-only RL (*v*), fixed single-frame-to-video update ratios of 5:1 and 2:1 (*i5v1*, *i2v1*), and a direct *i*→*v* schedule without the mixed stage. Figure 5 plots the evaluation score $50 - \mathrm{RE@SG}$ (higher is better) against cumulative generated training frames and GPU-hours; the score uses evaluation-time estimated poses rather than CVRR's fixed input pose. Training-frame counts include all generated candidates (Appendix B.4). Single-frame RL improves rapidly but plateaus below full-video RL. Fixed mixtures narrow the gap, with *i2v1* finishing above *i5v1*. Direct *i*→*v* and full-video RL reach similar final scores but require substantially more computation than PHRL. With mixed updates before video-only refinement, PHRL achieves comparable geometric consistency using significantly fewer generated frames and GPU-hours. Table 5 complements geometric scores with reconstruction, motion accuracy, and video quality. PHRL lowers EPE to 2.08 from 3.99 for ReView-Base and 2.68 for single-frame RL, while improving both VBench scores over single-frame RL. Compared with full-video RL, PHRL has similar mPSNR (19.56 vs. 19.53), slightly lower mSSIM (0.672 vs. 0.681), and slightly higher EPE (2.08 vs. 1.99). Aesthetic quality (0.5815 vs. 0.5893) and temporal style (0.2799 vs. 0.2805) also remain close. Together, the curve and table demonstrate efficient geometric improvement with motion reconstruction and perceptual quality close to full-video-only RL.

5 Conclusion

We introduced ReView for direct novel-view video generation, preserving the source event's appearance and dynamics from a prescribed viewpoint. ReView combines target-aligned geometric control and source-reference fusion with synchronized multi-view supervision and a cross-view reprojection reward. Our PHRL curriculum progressively shifts training from single-frame to joint-video rollouts. Experiments demonstrate improved viewpoint accuracy, cross-view consistency, and motion reconstruction with competitive video quality. PHRL achieves performance comparable to full-video-only RL using less than one third as many training compute, demonstrating the effectiveness of progressive frame-to-video optimization for efficient video RL.

AI USE STATEMENT

The use of large language models (LLMs) in this work was limited to language editing and literature discovery. For language editing, LLMs were used to improve the clarity and readability of the manuscript, refine wording and sentence structure, and correct grammatical errors. For literature discovery, LLMs assisted in identifying potentially relevant prior work and locating related references. All references included in the manuscript were reviewed and verified by the authors. LLMs were not used to generate research ideas, develop the methodology, design or conduct experiments, implement the proposed method, analyze data, interpret experimental results, or formulate scientific claims or conclusions. These aspects of the work were undertaken by the authors. All AI-assisted revisions were carefully reviewed for technical accuracy and fidelity to the intended meaning. The authors take full responsibility for the accuracy and integrity of the final manuscript, including all AI-assisted content.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the proposed method, experimental setup, model configurations, training hyperparameters, datasets, and evaluation protocols in the main paper and appendix. Upon publication, we will publicly release the source code for training and evaluation to further facilitate reproduction of these results.

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, 2025.

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.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Hongrui Cai, Junjie Luo, Zhihong Fu, Shengnan Zhu, Jiawei Wen, Wanquan Feng, Songtao Zhao, and Qian He. Scaling4d: Pushing the frontier of video novel view synthesis through large-scale monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11174–11184, 2026.

Kaihua Chen, Tarasha Khurana, and Deva Ramanan. Reconstruct, inpaint, test-time finetune: Dynamic novel-view synthesis from monocular videos. *Advances in Neural Information Processing Systems*, 38:47723–47753, 2026.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 337–33712. IEEE, 2018.

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. *Advances in Neural Information Processing Systems*, 35: 33768–33780, 2022.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Yue Han, Chong Li, Zhening Liu, Cong Huang, Fang Deng, Yong Liu, Fangyun Wei, and Yan Lu. Ink3d: Sculpting 3d assets with extremely complex textures via video generative models. *arXiv preprint arXiv:2607.01222*, 2026.

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.

Xiaoxuan He, Siming Fu, Zeyue Xue, Weijie Wang, Ruizhe He, Yuming Li, Dacheng Yin, Shuai Dong, Haoyang Huang, Hongfa Wang, et al. Flash-grpo: Efficient alignment for video diffusion via one-step policy optimization. *arXiv preprint arXiv:2605.15980*, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Yash Kant, Ethan Weber, Jin Kyu Kim, Rawal Khirodkar, Su Zhaoen, Julieta Martinez, Igor Gilitschenski, Shunsuke Saito, and Timur Bagautdinov. Pippo: High-resolution multi-view humans from a single image. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16418–16429. IEEE, 2025.

Junzhe Li, Yutao Cui, Tao Huang, Weijie Kong, Yiming Cheng, Chuxuan Zeng, Yinping Ma, Chun Fan, Miles Yang, Zhao Zhong, et al. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025.

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5511–5521. IEEE, 2022.

Yuming Li, Yikai Wang, Zhongyu Zhao, Ming Lu, Qi She, Shanghang Zhang, et al. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 20839–20866, 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*, 2025a.

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, 2026.

Wenfeng Lin, Renjie Chen, Boyuan Liu, Shiyue Yan, Ruoyu Feng, Jiangchuan Wei, Yichen Zhang, Yimeng Zhou, Chao Feng, Jiao Ran, et al. Contentv: Efficient training of video generation models with limited compute. *arXiv preprint arXiv:2506.05343*, 2025b.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 17581–17592. IEEE, 2023.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International conference on learning representations*, volume 2025, pp. 1045–1064, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Bowen Ping, Chengyou Jia, Minnan Luo, Hangwei Qian, and Ivor Tsang. Flow-factory: A unified framework for reinforcement learning in flow-matching models. *arXiv preprint arXiv:2602.12529*, 2026.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 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.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *International conference on learning representations*, volume 2024, pp. 39838–39859, 2024.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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.

Basile Van Hoorick, Dian Chen, Shun Iwase, Pavel Tokmakov, Muhammad Zubair Irshad, Igor Vasiljevic, Swati Gupta, Fangzhou Cheng, Sergey Zakharov, and Vitor Campagnolo Guizilini. Anyview: Synthesizing any novel view in dynamic scenes. *arXiv preprint arXiv:2601.16982*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jin Wang, Jianxiang Lu, Guangzheng Xu, Comi Chen, Haoyu Yang, Linqing Wang, Peng Chen, Mingtao Chen, Zhichao Hu, Longhuang Wu, et al. Tagrpo: Boosting grpo on image-to-video generation with direct trajectory alignment. *arXiv preprint arXiv:2601.05729*, 2026a.

Weijie Wang, Xiaoxuan He, Youping Gu, Yifan Yang, Zeyu Zhang, Yefei He, Yanbo Ding, Xirui Hu, Donny Y Chen, Zhiyuan He, et al. World-r1: Reinforcing 3d constraints for text-to-video generation. *arXiv preprint arXiv:2604.24764*, 2026b.

Wenbo Wang, Fangyun Wei, Qixiu Li, Xi Chen, Yaobo Liang, Chang Xu, Jiaolong Yang, and Baining Guo. Mobilemanibench: Simplifying model verification for mobile manipulation. In *European Conference on Computer Vision*, pp. 242–259. Springer, 2026c.

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.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14419–14429, 2026.

Tian-Xing Xu, Zi-Xuan Wang, Guangyuan Wang, Li Hu, Zhongyi Zhang, Peng Zhang, Bang Zhang, and Song-Hai Zhang. Ucm: Unified modeling of camera control and memory with time-aware positional encoding warping for world models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5916–5926. IEEE, 2025a.

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, 2025b.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6463–6474. IEEE, 2024.

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, 2025.

Liyu Zhang, Kehan Li, Tingrui Han, Tao Zhao, Yuxuan Sheng, Shibo He, and Chao Li. Op-grpo: Efficient off-policy grpo for flow-matching models. *arXiv preprint arXiv:2604.04142*, 2026a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Songchun Zhang, Zeyue Xue, Siming Fu, Jie Huang, Xianghao Kong, Y Ma, Haoyang Huang, Nan Duan, and Anyi Rao. Astrolabe: Steering forward-process reinforcement learning for distilled autoregressive video models. *arXiv preprint arXiv:2603.17051*, 2026b.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4245–4257, 2026.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026.

Appendix

This supplement presents the conditional analysis of PHRL (Appendix A), implementation and evaluation details (Appendix B), synchronized data collection (Appendix C), and qualitative analysis of CVRR (Appendix D). The four full-page dataset visualizations are collected at the end.

A Conditional Analysis of the Frame-to-Video Proxy

This appendix formalizes the relationship between the single-frame proxy and the video reward objective in Section 3.4.2. The analysis bounds their expected-reward gap through the mismatch between induced frame distributions, then characterizes mixed-objective optimization under an additional near-optimality assumption.

A.1 Objectives and frame marginals

Let $\mathbf{c} \sim \mathcal{D}$ denote a conditioning example containing the source video and requested target-view controls, and let $\mathbf{c}_i$ denote the information available in single-frame execution at frame index i. An F-frame rollout induces a joint distribution $p_\theta^F(\widehat{\mathbf{X}}^t \mid \mathbf{c})$. Its expected mean frame reward is

$$J_F(\theta) = \mathbb{E}_{\mathbf{c}\sim\mathcal{D}}\mathbb{E}_{\widehat{\mathbf{X}}^t\sim p_\theta^F(\cdot|\mathbf{c})}[R_F(\mathbf{X}^s, \widehat{\mathbf{X}}^t)]. \tag{6}$$

Here the requested camera is part of the conditioning information used to evaluate R_F. Let $p_{\theta,i}^F(\cdot \mid \mathbf{c})$ denote the marginal distribution of frame i in the jointly generated video. By contrast, $p_{\theta,i}^1(\cdot \mid \mathbf{c}_i)$ denotes the output distribution when the same policy generates a single frame directly. Uniform frame-index sampling defines

$$J_1(\theta) = \mathbb{E}_{\mathbf{c}\sim\mathcal{D}}\mathbb{E}_{i\sim\mathrm{Unif}\{1,\dots,F\}}\mathbb{E}_{\widehat{\mathbf{I}}_i^t\sim p_{\theta,i}^1(\cdot|\mathbf{c}_i)}[r_i(\widehat{\mathbf{I}}_i^t; \mathbf{c}_i)]. \tag{7}$$

In both objectives, $r_i(\widehat{\mathbf{I}}_i^t; \mathbf{c}_i)$ is the same frame reward from Equation 3, evaluated against the corresponding source frame and requested camera. The comparison uses the same conditioning distribution and uniformly weighted frame indices. The two execution modes share parameters, but their output distributions need not coincide because temporal interactions and available conditioning context may differ.

By linearity of expectation, the frame-averaged video reward R_F gives

$$J_F(\theta) = \frac{1}{F}\sum_{i=1}^{F}\mathbb{E}_{\mathbf{c}\sim\mathcal{D}}\mathbb{E}_{\widehat{\mathbf{I}}_i^t\sim p_{\theta,i}^F(\cdot|\mathbf{c})}[r_i(\widehat{\mathbf{I}}_i^t; \mathbf{c}_i)]. \tag{8}$$

This identity holds even when the jointly generated frames are temporally dependent. It exposes the shared frame-level reward, but does not equate direct single-frame generation with the marginals of a video rollout.

A.2 A bounded objective gap

Lemma 1 (Bounded single-frame proxy gap)**.** *Suppose $r_i(\cdot; \mathbf{c}_i) \in [\underline{r}, \overline{r}]$ for every frame and condition, and let $B = \overline{r} - \underline{r}$. Define*

$$\overline{\delta}_F(\theta) = \mathbb{E}_{\mathbf{c}\sim\mathcal{D}}\left[\frac{1}{F}\sum_{i=1}^{F}\mathrm{TV}\big(p_{\theta,i}^F(\cdot \mid \mathbf{c}), p_{\theta,i}^1(\cdot \mid \mathbf{c}_i)\big)\right], \tag{9}$$

where $\mathrm{TV}(p,q) = \sup_{\mathcal{A}} |p(\mathcal{A}) - q(\mathcal{A})|$ over measurable image sets. Then $|J_F(\theta) - J_1(\theta)| \le B\,\overline{\delta}_F(\theta)$, as stated in Equation 5.

Proof. For any reward in $[\underline{r}, \overline{r}]$, centering at $m = (\underline{r} + \overline{r})/2$ gives

$$|\mathbb{E}_p[r] - \mathbb{E}_q[r]| \le 2\|r - m\|_\infty\,\mathrm{TV}(p,q) \le B\,\mathrm{TV}(p,q).$$

Apply this inequality to each pair of conditional frame distributions in Equation 8 and average over conditions and frame indices. □

For CVRR, the reward lies in $[r_{\min}/s, r_{\max}/s] = [0.2, 5]$, including the reward floor assigned to invalid matches, so $B = (r_{\max} - r_{\min})/s = 4.8$. The bound is informative when the induced frame distributions are sufficiently close; bounded rewards alone do not establish such closeness.

A conditional implication for proxy improvements. For parameters θ_a before and θ_b after single-frame training, applying Lemma 1 at both endpoints yields

$$\begin{aligned} J_F(\theta_b) - J_F(\theta_a) \geq J_1(\theta_b) - J_1(\theta_a) \\ - B\left[\overline{\delta}_F(\theta_a) + \overline{\delta}_F(\theta_b)\right]. \end{aligned} \tag{10}$$

This follows by writing the video-objective difference as the proxy improvement plus the two endpoint objective gaps and bounding each gap. A proxy improvement larger than the endpoint mismatch penalty is therefore sufficient for a video-objective improvement. The condition is not automatic: the mismatch may be large or may increase during single-frame training. Without controlling or estimating this mismatch, the inequality provides a conditional interpretation of transfer rather than a guarantee that proxy training improves the video policy.

A.3 Mixed-objective comparison

For the mixed reward objective in Equation 4, Lemma 1 implies

$$|J_F(\theta) - J_\alpha^{\text{mix}}(\theta)| = (1-\alpha)|J_F(\theta) - J_1(\theta)| \leq B(1-\alpha)\overline{\delta}_F(\theta). \tag{11}$$

The next result compares an approximately optimal mixed-objective policy with an optimal video-objective policy in the same parameter class, under an additional uniform mismatch assumption.

Corollary 1 (Conditional video-objective suboptimality bound)**.** *Let Θ be the admissible set of LoRA parameters, assume $\sup_{\theta\in\Theta}\overline{\delta}_F(\theta) \leq \delta$, and suppose a maximizer $\theta_F^\star \in \arg\max_{\theta\in\Theta} J_F(\theta)$ exists. For $\eta_\alpha \geq 0$, if $\widehat{\theta}_\alpha \in \Theta$ satisfies $J_\alpha^{\text{mix}}(\widehat{\theta}_\alpha) \geq \sup_{\theta\in\Theta} J_\alpha^{\text{mix}}(\theta) - \eta_\alpha$, then*

$$J_F(\theta_F^\star) - J_F(\widehat{\theta}_\alpha) \leq \eta_\alpha + 2B(1-\alpha)\delta. \tag{12}$$

Proof. Equation 11 and the uniform mismatch assumption give $\sup_{\theta\in\Theta}|J_F(\theta) - J_\alpha^{\text{mix}}(\theta)| \leq B(1-\alpha)\delta$. Adding and subtracting the mixed-objective values at $\theta_F^\star$ and $\widehat{\theta}_\alpha$ yields

$$\begin{aligned} J_F(\theta_F^\star) - J_F(\widehat{\theta}_\alpha) &= [J_F(\theta_F^\star) - J_\alpha^{\text{mix}}(\theta_F^\star)] \\ &\quad + [J_\alpha^{\text{mix}}(\theta_F^\star) - J_\alpha^{\text{mix}}(\widehat{\theta}_\alpha)] \\ &\quad + [J_\alpha^{\text{mix}}(\widehat{\theta}_\alpha) - J_F(\widehat{\theta}_\alpha)] \\ &\leq 2B(1-\alpha)\delta + \eta_\alpha. \end{aligned}$$

The two outer terms are bounded by the uniform objective gap; the middle term is at most η_α by the assumed mixed-objective optimality tolerance. □

At $\alpha = 0$, the bound becomes $\eta_0 + 2B\delta$; at $\alpha = 1$, the proxy term vanishes because the mixed objective equals the video objective. For fixed mismatch and optimization tolerances, increasing α reduces the proxy contribution to the bound. This does not establish that the achieved video reward improves monotonically with α: the optimization residual and the actual distributional mismatch may change across training stages.

Scope of the analysis. These results characterize expected reward objectives rather than the normalized, clipped Flow-GRPO surrogate in Appendix B.5. Their interpretation depends on the stated marginal-mismatch assumptions and, for Corollary 1, near-optimality of the mixed objective. They motivate single-frame initialization and increasing video allocation without prescribing the number or cost of policy updates. Training efficiency and motion accuracy are assessed empirically in Figure 5 and Table 5.

B Implementation Details

We expand the architecture, supervised adaptation, reward computation, and RL settings summarized in Section 4.1. Dataset capture and examples are grouped separately in Appendix C.

B.1 Architecture and conditioning

The repetition labels M, N, and K in Figure 2 denote the numbers of control blocks, backbone blocks, and reference-fusion modules, respectively.

Geometric control branch. ReView uses Wan2.2-5B as its pretrained video backbone. The render and visibility-mask encoders map their inputs to the target latent grid; concatenating their features along the channel dimension yields a 96-channel control tensor. The ControlNet-style branch contains eight DiT blocks initialized from backbone blocks $\{0, 4, 8, 12, 16, 20, 24, 28\}$. At the branch input, projected control tokens are added to the noisy target tokens. The control state then passes through successive branch blocks, each producing a residual injected after its corresponding backbone block, forming the Geometric Control Branch in Figure 2.

The residual output projections are initialized to zero, so the branch initially leaves the pretrained backbone output unchanged. Its patchify kernel is initialized by duplicating the pretrained target-latent kernel along the input-channel dimension. During supervised training, we perturb the render latents to discourage dependence on idealized reprojections and improve robustness to depth artifacts.

Source-reference fusion. The source video is encoded and patchified into reference tokens. Fusion modules, initialized from their corresponding backbone blocks, are inserted after backbone blocks $\{0, 8, 16, 24\}$, forming the Source-Ref Fusion Branch in Figure 2. Each module first updates the reference tokens with self-attention, then uses them as keys and values for cross-attention from the target denoising tokens. This provides access to source appearance and motion information where projected evidence is absent or unreliable, including thin structures and moving-object boundaries.

B.2 Supervised adaptation

We initialize ReView from the pretrained Wan2.2-5B video model (Wan et al., 2025). Training follows the two-stage procedure in Figure 2.

Geometry-control adaptation. We first optimize the geometry-control branch while keeping the pretrained backbone frozen. This stage uses 127K videos filtered from OpenVid-1M (Nan et al., 2025), from which pseudo source–target pairs are constructed through double reprojection (Yu et al., 2025b), together with 20K synchronized video pairs from our robot-manipulation collection, 10K pairs from our Unreal Engine game collection, and 5K multi-view image pairs from DL3DV (Ling et al., 2024). We train the geometry-control branch for 12,000 optimization steps.

Source-reference adaptation. We then freeze both the backbone and the geometry-control branch and optimize the source-reference fusion modules for 6,000 steps, using 30K robot-manipulation video pairs, 10K Unreal Engine video pairs, and 17K DL3DV multi-view image pairs.

Shared settings. Both stages use AdamW with a learning rate of 1×10^{-5} and a global batch size of 64 on 64 AMD MI300 GPUs. Unless otherwise specified, supervised training and evaluation use 720p videos of 61 frames.

B.3 Cross-view reward computation

For each temporally corresponding source–target frame pair, SuperPoint and LightGlue provide visual correspondences, while joint two-view geometry inference supplies source depth and source/target intrinsics. We bilinearly sample source depth at each matched keypoint and apply the back-projection and target-view projection in Equation 1. The estimated geometry and prescribed

camera transformation are expressed in compatible coordinate and scale conventions. During RL, reprojection uses the fixed input source-to-target pose; this differs from evaluation-time RE@SG, which uses estimated poses.

We discard matches with invalid source depth, negative target-camera depth, or projected locations outside the target image. Coordinates are rescaled to the original image resolution before computing distances. For the valid set $\mathcal{V}$, the median error used in Equation 3 is

$$e_i = e_{\text{med}}(\mathbf{I}_i^s, \widehat{\mathbf{I}}_i^t) = \text{median}_{j\in\mathcal{V}} \left\| \widehat{\mathbf{q}}_j^t - \mathbf{q}_j^t \right\|_2 . \tag{13}$$

We detect at most 2,048 keypoints and require at least 50 valid matches. The reward parameters are $(r_{\min}, r_{\max}, s) = (2, 50, 10)$, giving a reward range of $[0.2, 5]$. A pair with fewer than 50 valid matches receives the floor $r_{\min}/s = 0.2$ rather than a reward based on its remaining correspondences. For a video rollout, all temporally corresponding frame-pair rewards, including floor values, contribute to the mean R_F.

Figure 6 visualizes the matching and projection signals; Appendix D discusses their roles in content, motion, and viewpoint consistency.

B.4 Progressive-horizon reinforcement learning

Starting from the supervised checkpoint, we perform geometry-aware post-training with Flow-GRPO (Liu et al., 2026), using the Flow-Factory implementation (Ping et al., 2026). Following the mixed ODE–SDE sampling strategy (Li et al., 2025), each rollout contains two stochastic SDE transitions and 18 deterministic ODE transitions. We optimize rank-128 LoRA parameters while keeping the remaining model parameters frozen. Each update uses 24 conditioning samples with a group size of $G = 32$. The RL training pool contains 7K conditioning examples.

Our PHRL schedule begins with 280 single-frame updates, followed by 60 mixed updates that select single-frame and video execution with probabilities $2/3$ and $1/3$, respectively, and concludes with 80 full-video updates. For a single-frame update, we randomly sample the target frame index for each conditioning example. Video rollouts in the mixed and final stages contain $F = 41$ frames.

Training-frame accounting. For a sequence of rollout-collection updates indexed by u, let B_u be the number of conditioning examples, G the candidates per example, and $H_u \in \{1, F\}$ the selected rollout length. The cumulative generated-frame count is $N_{\text{frames}} = G \sum_u B_u H_u$. This counts every generated candidate frame, rather than only frames decoded for reward evaluation or denoising transitions used for policy updates. For mixed stages, the realized choices of H_u determine the count; the sampling probability specifies only its expectation. Figure 5 reports this count alongside training compute in GPU-hours, using evaluation score $50 - \text{RE@SG}$ rather than training CVRR on the vertical axis.

B.5 Flow-GRPO optimization

We optimize the execution mode selected by PHRL using Flow-GRPO (Liu et al., 2026), with stochastic SDE transitions inserted into the flow sampling trajectory. For each condition, we generate G candidates at a common temporal support. Candidate g receives reward R_g, given by the frame reward for single-frame execution or the frame-averaged reward R_F for video execution. Its normalized advantage is

$$A_g = \frac{R_g - \frac{1}{G}\sum_{j=1}^{G} R_j}{\text{std}_{\text{global}}(R) + \epsilon_{\text{num}}}, \tag{14}$$

where the denominator uses the reward standard deviation across the rollout batch and $\epsilon_{\text{num}} > 0$ stabilizes normalization. We clip the normalized advantage as $\widetilde{A}_g = \text{clip}(A_g, -a_{\max}, a_{\max})$.

Let $H \in \{1, F\}$ denote the rollout length and K the number of stochastic denoising transitions retained per rollout for policy optimization. For retained transition k, write $\mathbf{z}_{g,k}^{\text{in}}$ and $\mathbf{z}_{g,k}^{\text{out}}$ for its input and output latents, and let $q_{\theta,k}^H$ denote its conditional transition density. Using $\mathbf{c}_H$ for the corresponding video or single-frame condition, the likelihood ratio is

$$\rho_{g,k}(\theta) = \frac{q_{\theta,k}^H(\mathbf{z}_{g,k}^{\text{out}} \mid \mathbf{z}_{g,k}^{\text{in}}, \mathbf{c}_H)}{q_{\theta_{\text{old}},k}^H(\mathbf{z}_{g,k}^{\text{out}} \mid \mathbf{z}_{g,k}^{\text{in}}, \mathbf{c}_H)}, \tag{15}$$

where θ_{old} is the policy used to collect the rollouts. We minimize the clipped surrogate

$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{GK}\sum_{g=1}^{G}\sum_{k=1}^{K}\min\Big(\rho_{g,k}(\theta)\widetilde{A}_g, \text{clip}(\rho_{g,k}(\theta), 1-\varepsilon, 1+\varepsilon)\widetilde{A}_g\Big). \tag{16}$$

Equation 4 describes the allocation at the reward-objective level; the implemented updates use the normalized, clipped surrogate above.

B.6 Reference-free camera sampling

For the reference-free benchmark, we construct requested target viewpoints by randomly combining camera offsets relative to the source view. Yaw and pitch magnitudes are sampled from $[10^\circ, 25^\circ]$ and $[8^\circ, 15^\circ]$, respectively, with either sign. We sample radial offsets $r \in [0, 0.1]$ and horizontal and vertical offsets $x, y \in [0.02, 0.12]$. These combinations specify the target-camera transformations used for novel-view generation.

B.7 Optical-flow evaluation of motion accuracy

Following the flow-based motion comparison in Vista4D (Lin et al., 2026), EPE compares temporal motion in the generated video with that in the synchronized ground-truth target video. The reference is therefore another observation from the requested target viewpoint, not the source video or its partial reprojection. This makes the comparison sensitive to motion reconstruction in addition to per-frame appearance.

Matched temporal pairs. Let Φ denote the optical-flow estimator and (i, j) an evaluated pair of frame indices. Applying the same estimator to generated and ground-truth target frames gives

$$\widehat{\mathbf{f}}_{i\to j} = \Phi(\widehat{\mathbf{I}}_i^t, \widehat{\mathbf{I}}_j^t), \qquad \mathbf{f}_{i\to j}^{\text{ref}} = \Phi(\mathbf{I}_i^t, \mathbf{I}_j^t). \tag{17}$$

Each two-dimensional vector represents the displacement of a pixel from the first frame to the second. Both videos must use the same temporal indices and frame interval, so that corresponding vectors describe motion over the same portion of the event. The reference flow is estimated from ground-truth RGB frames; it is not a ground-truth flow annotation.

Endpoint error. For an evaluated pixel set Ω_{ij} on a common target-image grid, the mean endpoint error for this frame pair is

$$e_{ij}^{\text{flow}} = \frac{1}{|\Omega_{ij}|}\sum_{\mathbf{u}\in\Omega_{ij}}\left\|\widehat{\mathbf{f}}_{i\to j}(\mathbf{u}) - \mathbf{f}_{i\to j}^{\text{ref}}(\mathbf{u})\right\|_2. \tag{18}$$

Here Ω_{ij} specifies the pixels included in the evaluation average; it need not be the source–target covisibility mask used by the masked reconstruction metrics. The vector difference penalizes errors in motion magnitude and direction. Lower EPE indicates closer agreement with the reference event's motion, complementing PSNR, SSIM, and LPIPS, which compare individual frames. Unlike a smoothness score, EPE compares against reference motion and therefore penalizes missing motion where the reference flow is nonzero.

C Synchronized Multi-View Data and Visualization

Our game-engine and robot-simulation pipelines capture calibrated, time-synchronized observations of the same dynamic event. They complement monocular reprojection pairs with independently observed target content: game environments provide character-centered motion in visually rich scenes, while robot simulations provide object interactions and state changes. The two captured collections are described below; stage-specific training-pair counts, including additional DL3DV image pairs, appear in Appendix B.2.

C.1 GAME-ENGINE CAPTURE

We collect videos from high-fidelity AAA games built with Unreal Engine. Our pipeline controls camera position and orientation at runtime independently of the gameplay camera and character inputs. This separation lets us observe scripted actions from configured viewpoints without tying capture to the player's viewing direction. We arrange 25 cameras around the character, spanning a full 360° azimuthal range and two pitch levels, and record synchronized views of the evolving scene.

Automated keyboard and mouse inputs drive combinations of character actions and traversal. Varying environments, action sequences, and character appearances yields 1,000 synchronized multi-view sequences. Within each sequence, the cameras observe the same event and temporal progression, allowing training pairs to differ in viewpoint while retaining a shared scene state.

C.2 ROBOT-SIMULATION CAPTURE

Following the simulation-based data-generation paradigm of MobileManiBench, we place 12 cameras around a robot interaction workspace and record synchronized manipulation episodes. Tasks include grasping, placing, opening, and closing, capturing coordinated robot and object motion as well as object-state transitions. We obtain 2,000 distinct action–scene–object combinations by varying at least one component between episodes. Recording each episode from all 12 cameras provides complementary observations of the same interaction and its changing visibility.

C.3 SYNCHRONIZATION AND TRAINING-PAIR CONSTRUCTION

Each sequence is represented as $\mathcal{S} = \{(\mathbf{X}^{(v)}, \mathbf{C}^{(v)})\}_{v=1}^{V}$, where $\mathbf{X}^{(v)} = \{\mathbf{I}_i^{(v)}\}_{i=1}^{F}$ and $\mathbf{C}^{(v)}$ contains camera intrinsics and extrinsics. Here $V = 25$ for the game collection and $V = 12$ for the robot collection. Corresponding frame indices observe the same underlying scene state. Selecting two distinct cameras yields $(\mathbf{X}^s, \mathbf{X}^t, \Delta\mathbf{C})$, with the relative transformation computed from calibrated poses. Unlike source-view warping, these pairs provide independent target observations of newly visible content, view-dependent appearance, and evolving occlusions. The collection sizes above count captured sequences; adaptation-stage totals count sampled training pairs.

C.4 READING THE MULTI-VIEW VISUALIZATIONS

Figures 7 and 8 show game-engine examples; Figures 9 and 10 show robot-simulation examples. These four full-page figures are collected at the end of the supplement. In each grid, columns vary the camera viewpoint and rows advance through time. Reading across a row compares synchronized observations of one scene state; reading down a column follows the event from a single camera. This layout highlights the supervision provided by the collection: paired views preserve event identity and timing while changing visible surfaces, spatial layout, and occlusion.

D QUALITATIVE ANALYSIS OF THE CROSS-VIEW REWARD

Figure 6 illustrates the effectiveness of CVRR. Visual matching identifies target-image locations corresponding to source keypoints, while geometric reprojection predicts their locations under the prescribed camera transformation. Their separation is the reprojection error used in Equation 3. Figure 6 illustrates how CVRR captures three representative generation errors: content distortion, temporal mismatch and lag, and camera shift and inaccuracies. These failures affect different aspects of the output, but all produce substantial disagreement between visually matched keypoints and their geometrically projected locations. Content distortion disrupts spatial correspondences, frozen frames fail to follow the evolving source event, and inaccurate viewpoints place content inconsistently with the requested camera transformation. The resulting large reprojection errors are penalized by CVRR, whereas the spatiotemporally consistent example shows closer agreement and receives a higher reward. These examples illustrate how a unified reward captures errors in content, temporal alignment, and camera adherence.

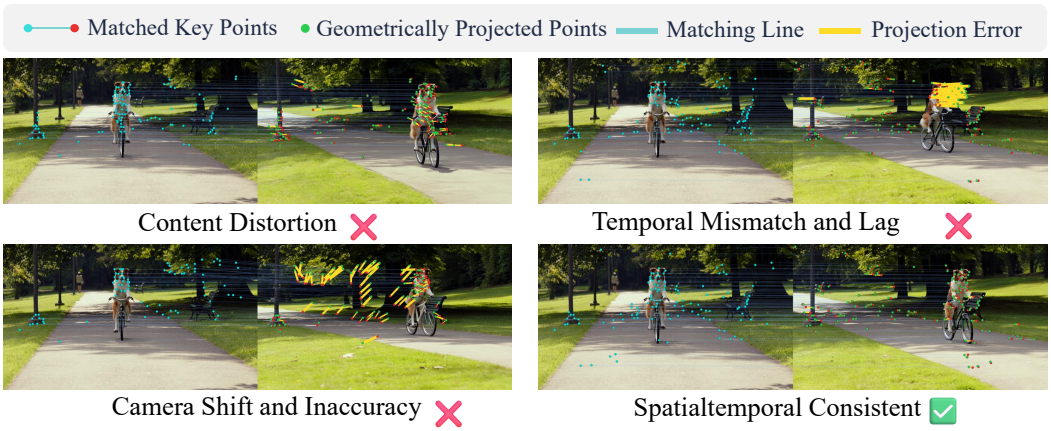

Figure 6: **CVRR captures diverse generation errors.** Content distortion, temporal mismatch and lag, and inaccurate viewpoints all exhibit large discrepancies between visually matched and geometrically projected locations, leading to lower rewards. The spatiotemporally consistent result exhibits closer agreement between the two signals.

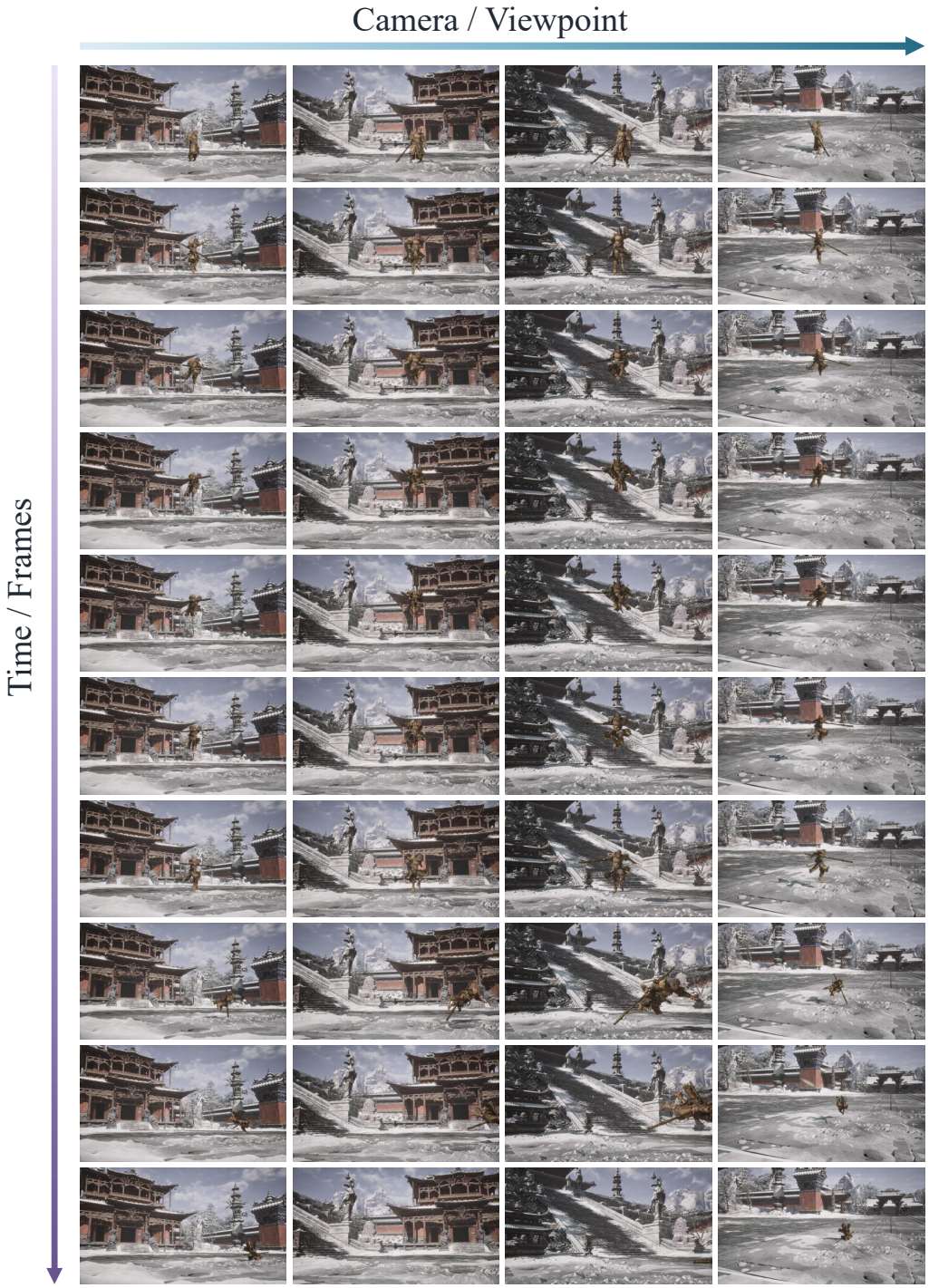

Figure 7: **Synchronized game-engine videos: example 1.** Columns show different camera viewpoints and rows show successive sampled times. Each row observes the same scene state across cameras, while each column follows the event from one viewpoint. These observations provide paired appearance and motion supervision.



Figure 8: **Synchronized game-engine videos: example 2.** Columns show different camera viewpoints and rows show successive sampled times. Each row observes the same scene state across cameras, while each column follows the event from one viewpoint. These observations provide paired appearance and motion supervision.

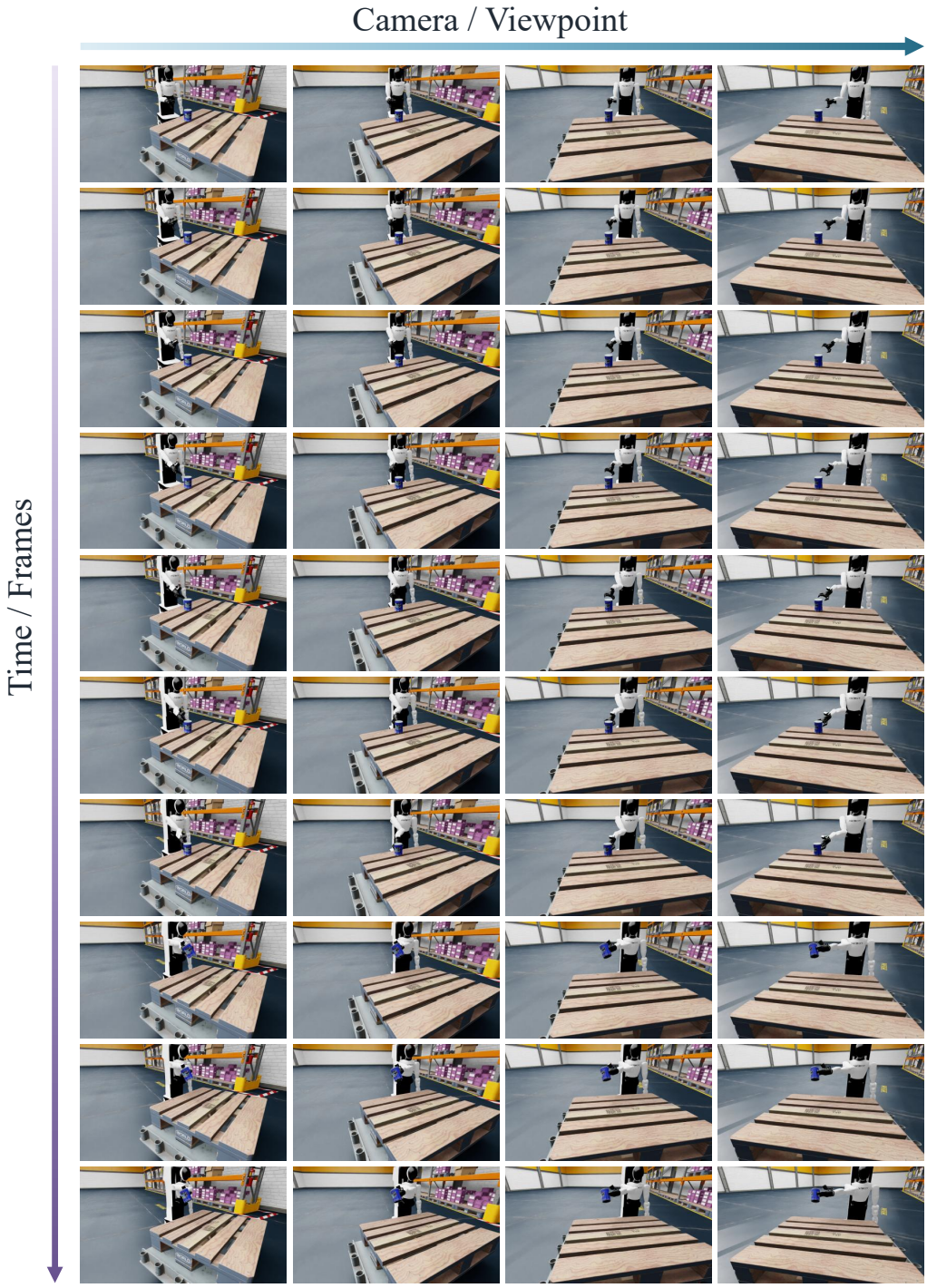

Figure 9: **Synchronized robot-simulation videos: example 1.** Columns show different camera viewpoints and rows show successive sampled times. Each row observes the same scene state across cameras, while each column follows the event from one viewpoint. These observations provide paired appearance and motion supervision.

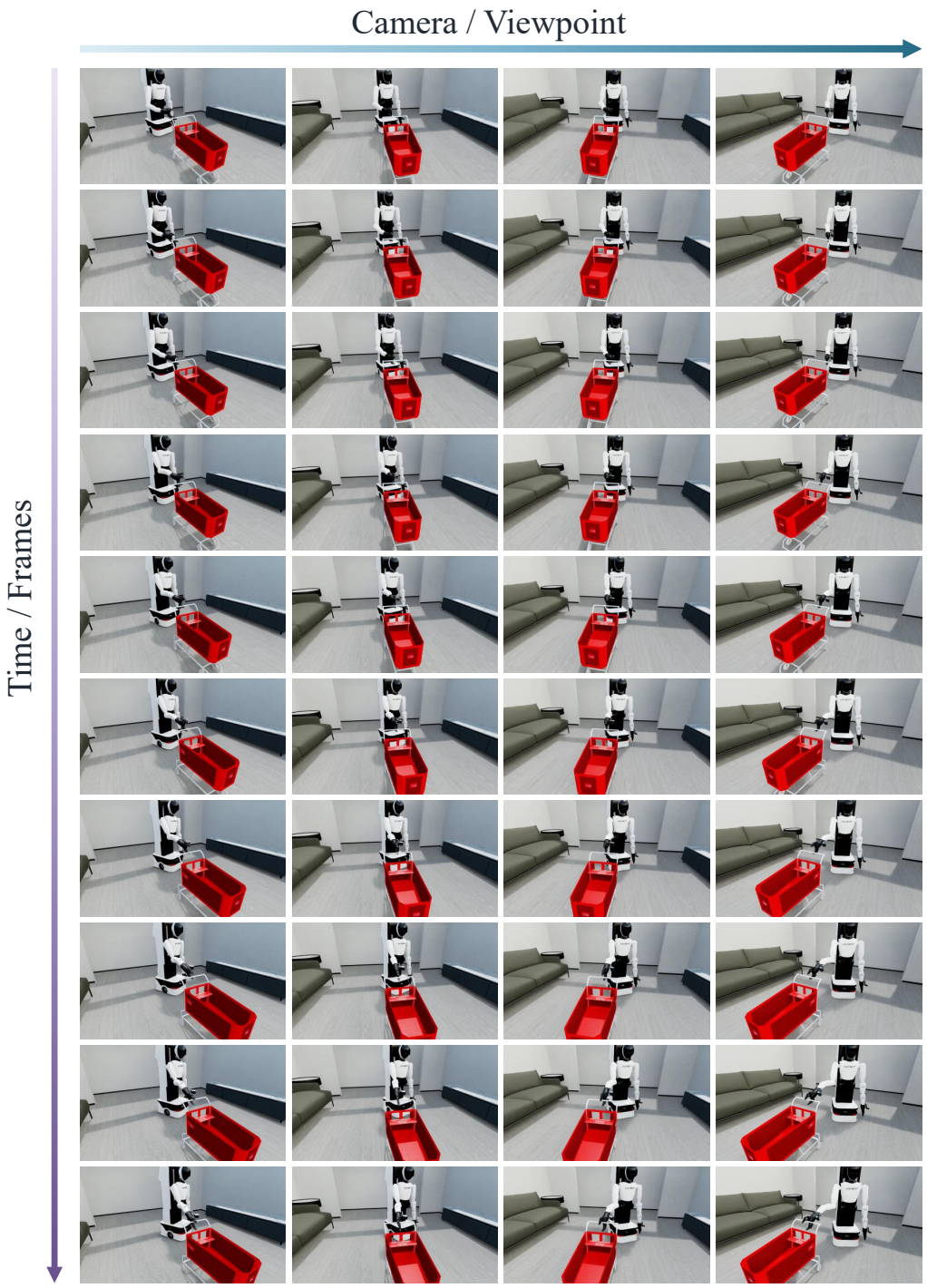

Figure 10: **Synchronized robot-simulation videos: example 2.** Columns show different camera viewpoints and rows show successive sampled times. Each row observes the same scene state across cameras, while each column follows the event from one viewpoint. These observations provide paired appearance and motion supervision.