LoGo: Local-Global Rewards for Consistent Long-Horizon Video Generation

Anonymous authors
Paper under double-blind review

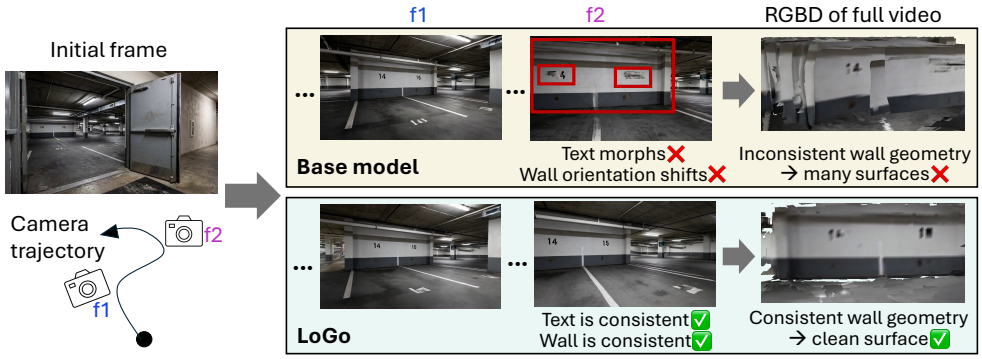


Figure 1: Camera-controlled video models often break 3D consistency during generation (top row). Here, a video generated by an existing video model shows both local inconsistencies (the numbers on the wall are distorted) and global inconsistencies (the wall's orientation shifts). We introduce LoGo, a post-training approach that combines global and spatially localized rewards to reliably improve consistency in generation (bottom row).

Abstract

Camera-controlled video models are rapidly advancing toward long generation horizons and complex camera control. A key failure mode is 3D inconsistency: as the camera moves, objects lose permanence and scene structures shift. Existing post-training techniques, which assign a single scalar reward to the entire generation, are poorly suited to correcting these inconsistencies over long horizons. We introduce LoGo, which blends global and spatially localized rewards for camera-controlled video models. The local reward provides fine-grained credit assignment, which substantially improves 3D consistency, while the global reward preserves camera following and video quality. Across three base models, LoGo shows a clear advantage on DL3DV and TrajectoryBench, a new benchmark for long-horizon, complex-camera-control generation that current evaluations lack. LoGo effectively reduces local object shifts, artifacts, and global scene changes, illustrating the importance of credit assignment in post-training video models.

1 Introduction

The rapid development in video generation has enabled us to generate *worlds* from image and text prompts. Growing camera control capabilities enable users to freely navigate the generated space. But, what qualities are necessary to transform these generates frames into "worlds" that users can immerse themselves in and embodied applications can leverage?

A fundamental requirement is consistency: scenes should have persistent structure, and objects should have permanence. However, current models struggle; Kupyn et al. (2025) highlights classical geometric errors, Rakheja et al. (2025) studies object and relation permanence, and Ye et al. (2026) focuses on revisit consistency. These works highlight the difficulty of consistent video generation. In this work, we focus on 3D consistency: **(1)** Globally, the spatial structure and layout should persist as the user navigates the scene; **(2)** Locally, objects should stay consistent—they should not suddenly appear or disappear, become corrupted by artifacts, or otherwise shift in geometry or appearance.

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.



Figure 2: A scalar reward cannot capture localized failures: the door inconsistency is "washed out" by the overall good structure of the rollout. LoGo catches the localized inconsistency (right).

Current approaches improve video generation consistency through post-training. They apply DPO (Rafailov et al., 2023) or Flow-GRPO (Liu et al., 2026a) with geometry rewards such as epipolar constraints (Kupyn et al., 2025), depth reprojection (An et al., 2026; Du et al., 2026), or Gaussian reconstruction (Wang et al., 2026a). These approaches all collapse the reward to a single value for the entire generated video, without fine-grained credit assignment. This may suffice for short-horizon generation (80 frames), but new challenges emerge as horizons grow. Chief among them is 3D consistency failures, which tend to be localized and are thus difficult to capture with a single global reward. Such failures surface in our experiments generating up to 400 frames. For example, in Fig. 2, the door inconsistency in the rollout is local, yet the clip earns a high score for its otherwise sound geometry, and the error goes unpenalized.

We propose LoGo, which enhances the global reward with a spatially localized reward. LoGo directly performs credit assignment in 3D space. As shown by the reward heatmap on the right of Fig. 2, LoGo pinpoints the door inconsistency, promoting local as well as global consistency. This local reward improves 3D consistency, and when combined with camera and aesthetic rewards it preserves or even improves video quality and camera following.

Current benchmarks for 3D consistency (Duan et al., 2025) focus on short-horizon (80-frame) generations with simple camera commands like "push forward." To address the lack of 3D consistency evaluation for long-horizon generation, we introduce TrajectoryBench, a new benchmark centered on complex camera trajectories. TrajectoryBench challenges models to generate space far beyond the initial view and to revisit the same regions from new angles, exposing failure modes that arise under these demanding conditions and giving the community a testbed to study them.

Across three base video models, LoGo shows a strong advantage compared to baseline methods using global reward structure. LoGo's advantage persists across three different post-training algorithms, demonstrating the robustness of our method. We make the following contributions: **(1)** We propose LoGo, a 3D-based reward structure that improves geometric consistency in camera-controlled video generation via combined global and spatially localized rewards, achieving state-of-the-art on two challenging benchmarks; **(2)** Our reward framework is general, yielding consistent gains across a range of base models – Lingbot2 (Gao et al., 2026), Lyra2 (Shen et al., 2026), and UniWorld (Zhou et al., 2026) – and post-training techniques – DiffusionNFT (Zheng et al., 2026), Flow-GRPO (Liu et al., 2026a), and DRaFT (Clark et al., 2024); **(3)** We release TrajectoryBench, a benchmark for long-horizon video generation with complex camera trajectories.

2 Related Work

Post-training for 3D consistency of video models. While some prior works improve 3D consistency via architecture (Xu et al., 2024), memory (Ren et al., 2025), or training objective (Xiang et al., 2026), a large number of works perform post-training. They propose geometry-based reward models, such as epipolar constraint in Epipolar-DPO (Kupyn et al., 2025), depth reprojection in VG-GRPO (An et al., 2026), VideoGPA (Du et al., 2026), VIGOR (Yin et al., 2026) and GeoVideo (Bai et al., 2026), or Gaussian splatting (Kerbl et al., 2023) in World-R1(Wang et al., 2026a). Methods like CamVerse (Wang et al., 2026d), CamPilot (Ge et al., 2026) use similar rewards to post-train for camera control. While prior works show the efficacy of geometry rewards, they focus on the short-horizon, 80-frame WAN model (Wan et al., 2025), and are limited to simple camera trajectories.

Credit assignment in RL. Credit assignment in RL has proven impactful in LLM reasoning (Uesato et al., 2022; Tao et al., 2026). In computer vision, reward localization has been explored for certain tasks, as in SpatialFlow-GRPO (Yang et al., 2026) for editing, LongCat (Team et al., 2026b) for omni model post-training, and CreFlow (Ni et al., 2026) for robotic video generation. While they

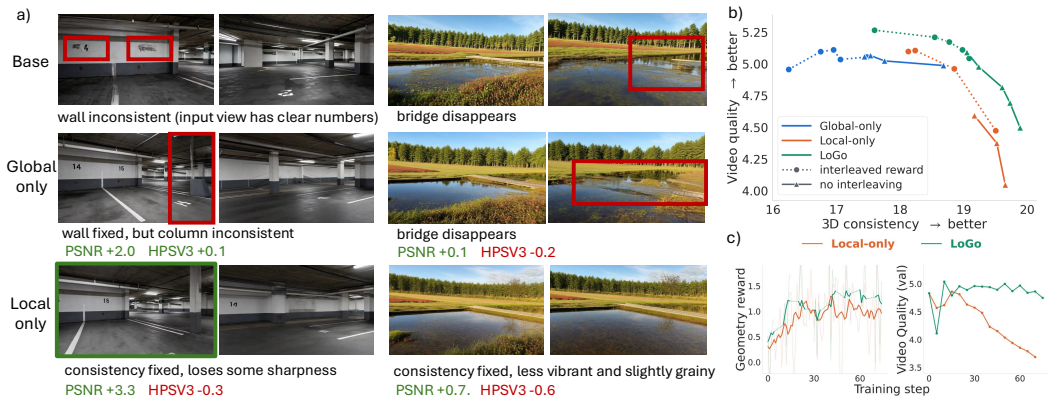


Figure 3: Motivation of LoGo. **(a)** A global reward fails to improve 3D consistency; local reward improves consistency, measured in PSNR, but degrades video quality, measured in HPSv3. This motivates combining the two in LoGo. **(b)** LoGo (green) extends the Pareto frontier in the consistency-quality tradeoff. **(c)** Reward curves and quality evaluation during training: local reward degrades video quality over time; LoGo improves consistency without compromising quality.

adopt some version of localized reward like loss masking (Ni et al., 2026), they do not comprehensively explore credit assignment. In this work, we focus on a reward design that incorporates proper credit assignment, which is central to improving 3D consistency over long horizons.

Long-horizon video generation post-training. Recent efforts enable long-horizon generation via autoregressive rollout of flow-matching chunks (Shen et al., 2026; Sun et al., 2025; Gao et al., 2026). While these models have been post-trained, existing methods are limited to specific aspects. (Wang et al., 2026c) and DreamX-World 1.0 (Team et al., 2026a) post-train only for camera following, a "local" property that can be evaluated per-frame without considering the full video. Wang et al. (2026c) train on individual sampled autoregressive chunks, without considering credit assignment across the video. Gu et al. (2026) use post-training to make reversible actions "cancel" each other over long horizons, which differ from our focus on 3D consistency and credit assignment.

3 Method

The multifaceted challenge of improving 3D consistency in long-horizon video generation motivates each component of the LoGo recipe. We explain each component below: **(1)** spatially-localized credit assignment in §3.1; **(2)** local-global mixing in §3.2; **(3)** reward interleaving in §3.3.

3.1 Reward localization in 3D

Given a base model and a 3D consistency objective, the naive approach is to use the objective as a scalar reward for post-training. However, as shown in Fig. 2, such a reward focuses on the overall scene structure, "washing out" local inconsistencies. As a result, training only with a global reward is ineffective: as shown in Fig. 3, a global reward fails to fix existing errors (in the lake example, the bridge still disappears) or introduces new ones (in the garage example, it morphs the column).

Modern video models fail in three, largely local, ways: **(1)** inconsistent object appearance, **(2)** hallucinated or disappearing objects, and **(3)** local artifacts. A good reward should penalize these failures directly rather than merely ranking full-length videos. Motivated by this, we design a reward localization method that operates directly in 3D space for accurate credit assignment during post-training.

Given a video, we first construct a scene point cloud by passing keyframes to VGGT-Ω (Wang et al., 2025) and unproject each pixel using the predicted depth into a shared coordinate system. As shown in Fig. 4a, we voxelize this shared 3D space into a voxel grid, and calculate a spatially localized error per voxel: each voxel's error is defined as the average reprojection error of every pixel unprojected into the voxel. For a pixel, the reprojection error is defined as a weighted combination of RGB error, *i.e.,* RGB projection of the scene point cloud into that pixel compared with the pixel's generated RGB, and depth error, *i.e.,* depth projection of the scene point cloud into that pixel compared to the pixel's VGGT-predicted depth.

Let $\mathcal{P}$ be the scene point cloud, c_j the camera pose of frame j, and $I_j(x)$ the pixel of the j-th frame at location x. For each voxel i, $\mathcal{S}_i$ denotes the set of pixels that unproject into the voxel. If Proj_{RGB} and Proj_{D} is the RGB and depth projection function, respectively, the local voxel error is defined as:

$$e_i = \frac{1}{|\mathcal{S}_i|} \sum_{(x,j)\in\mathcal{S}_i} \alpha \left\| \text{Proj}_{\text{RGB}}[\mathcal{P}, c_j, x] - I_j^{\text{RGB}}(x) \right\|_2^2 + (1-\alpha) \left| \text{Proj}_{\text{D}}[\mathcal{P}, c_j, x] - I_j^{\text{D}}(x) \right| \quad (1)$$

For small group sizes during post-training, a cross-voxel normalization per rollout can be added to the score to highlight the difference between systematically good or bad rollouts. In our experiments, we set the voxel size to $0.1\times$ the 90-percentile (P90) scene depth. For a 240-frame video, this yields a median of roughly 3,000 voxels per scene, with a median of 200 points per voxel.

We visualize our local reward via projected heatmaps in Figure 4, with voxel advantages projected onto keyframe patches. Our reward localization scheme can correctly capture each category of local inconsistency that we observe, including appearance inconsistency of the plant in the first example, hallucinated objects in the second example of the suddenly-appearing bookshelf, as well as local artifacts, as seen in the last example as a yellow blob on the left side of keyframe 2.



Figure 4: **a)**: illustration of voxel reward aggregation. **b)**: visualization of local reward projected to video frames.

3.2 LoGo: Blending local and global reward

While the local reward is highly effective at improving 3D consistency, using it in isolation causes video quality to degrade over the course of training (Fig. 3c). We attribute this to reward hacking: as in LLMs, where dense rewards can induce such behavior (Zhuang et al., 2026), the local reward fixes inconsistencies at the expense of overall quality, reducing sharpness and vibrancy (Fig. 3a). The pattern is consistent with reward hacking, since blurrier, smoother edges lower depth errors and less vibrant (closer-to-grey) colors lower RGB errors.

In contrast, a global reward, defined as the average reprojection error across all pixels and frames, proves more robust to this degradation. The Pareto curves in Fig. 3b make the trade-off clear: the local reward (orange) sacrifices video quality (measured by HPSv3 (Ma et al., 2025)) for geometric consistency (measured by reprojection PSNR), whereas the global reward (blue) largely preserves quality. To capture both the geometric consistency gains of the local reward and the video quality of the global reward, we blend them as a weighted sum of the normalized reward – the optimality probability in DiffusionNFT (Zheng et al., 2026).

$$r_{i,\text{LoGo}} = \tfrac{1}{2} + \tfrac{1}{4}\,\text{clip}\left(\frac{\bar{e}_{i,\text{local}} - e_{i,\text{local}}}{Z_{e_{i,\text{local}}}}, -1, 1\right) + \tfrac{1}{4}\,\text{clip}\left(\frac{\bar{e}_{\text{global}} - e_{\text{global}}}{Z_{e_{\text{global}}}}, -1, 1\right) \quad (2)$$

This blended local-global reward expands the Pareto frontier in the consistency-quality tradeoff, shown with green in Fig. 3b. Over the course of training, LoGo improves the geometry reward, *i.e.,* the weighted sum of global reward and average local reward across voxels, without degrading aesthetics, as show in Fig. 3c.

3.3 Reward interleaving

To ensure our post-trained models are robust across multiple axes, such as video quality and camera control, we interleave post-training with rewards targeting these properties. We adopt a cyclic schedule: k steps with r_{LoGo}, then m steps of aesthetic reward – HPSv3 (Ma et al., 2025) – and n steps of camera reward – equally weighted sum of rotation and translation errors, which preserves or improves other metrics while still improving consistency. In Fig. 3b, the dotted half of each curve uses reward interleaving, and the solid half does not. Our main results use DiffusionNFT (Zheng et al., 2026); in the Appendix we further demonstrate efficacy with Flow-GRPO (Liu et al., 2026a) and DRaFT (Clark et al., 2024).

Efficiency. LoGo adds minimal overhead – only 2.1% of the total reward calculation time and 0.5% of the total per-step training time. LoGo's overhead comes only from computing the pixel-voxel correspondence (9 seconds per step out of 1920 seconds of total per-step training time on 64 H100s). Training time is dominated by the video rollout, VGGT invocation and backpropagation, all independent of LoGo.

4 EXPERIMENTS

We now turn to experiments to demonstrate the efficacy of LoGo. We first introduce TrajectoryBench, which addresses the evaluation gap for long-horizon video generation and complex camera control. Then, we compare LoGo against post-training baselines across three base video models, showing a consistent advantage. Finally, we ablate LoGo's design choices, isolating the impact of a spatially localized reward over a global reward on long-horizon generation.

4.1 BENCHMARKS

TrajectoryBench. We introduce TrajectoryBench, a new benchmark for long-horizon video generation with complex camera trajectories. Compared to existing benchmarks (Duan et al., 2025), it contains challenging scenarios along three axes: **(1)** trajectories that move far from the initial view, testing whether a model can generate unseen content; **(2)** trajectories that revisit the same area from multiple viewing angles, testing 3D consistency; and **(3)** trajectories with space "transitions," such as passing through a door, testing whether a model can synthesize entirely new scenes.

TrajectoryBench contains 2,000 examples across three difficulty levels (*easy, medium, hard*) and three space types (indoor, outdoor, transition), each pairing an input image with a camera trajectory. The easy split is subsampled from WorldScore (80 frames, one camera motion) for compatibility with existing benchmarks; medium includes transition scenes that go through a door into a new space, and contains concatenated three single-camera-motion commands from WorldScore for indoor and outdoor scens; and hard split packs seven to eight camera motions into 240 frames, using our own sourced initial images. Fig. 5 contrasts a WorldScore trajectory (easy) with a hard TrajectoryBench one in expansion and revisit modes: WorldScore trajectories are simple, whereas TrajectoryBench tests generation from views far from the initial frame or repeated revisits. TrajectoryBench also provides comprehensive metrics spanning 3D consistency, camera-following, and video quality.

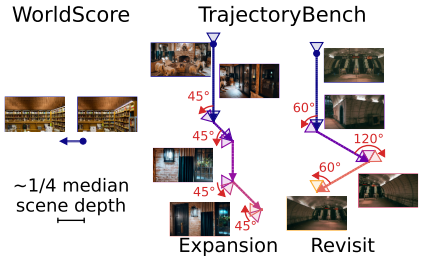


Figure 5: Camera trajectories of TrajectoryBench vs. WorldScore.

DL3DV. We evaluate on the DL3DV (Ling et al., 2024) 100-scene holdout set used by prior works (Du et al., 2026). However, Du et al. (2026) evaluate with only the initial image and text prompt, and omit camera control. Since our focus is on camera-controlled video models, we additionally condition on DL3DV's original camera trajectories and evaluate camera control accuracy.

4.2 TRAINING AND EVALUATION SETTINGS

Base models and training setting. We post-train three state-of-the-art camera-controlled video generation base models: (1) Lingbot-world-v2 (Gao et al., 2026), a 14B autoregressive model that generates 12 frames at a time with KV cache, (2) Lyra-2 (Shen et al., 2026), a 14B model that generates 80 frames per chunk, with a sparse 3D cache that supports camera-based retrieval; and (3) UniWorld (Zhou et al., 2026), a 14B bidirectional model trained on 80-frame clips. Because UniWorld's quality degrades greatly beyond 80 frames, we subsample longer trajectories into 80-frame generations. We train all models on 1700 DL3DV scenes until convergence, which takes 16-36 hours on 64 H100s. More details can be found in §A.4 in the Appendix.

Metrics. We evaluate 3D consistency, camera following, and video quality. For **3D consistency**, we measure RGBD reprojection PSNR (Bai et al., 2026; Du et al., 2026), its depth-only variant MVCS (Bai et al., 2026), Gaussian reconstruction error (Wang et al., 2026b), and epipolar er-

	Model	3D consistency					Camera		Video quality
		PSNR-V ↑	PSNR-D ↑	PSNR-GS ↑	MVCS ↑	Epipolar ↓	RPE° ↓	RPE trans ↓	VR-VQ ↑
Lingbot2	Base	16.9	16.0	16.9	0.842	1.507	13.2	0.108	0.21
	VideoGPA	17.1	16.1	17.0	0.847	1.486	13.1	0.107	0.22
	World-R1	17.0	16.0	16.9	0.839	1.515	13.1	0.108	0.21
	Global-only (ours)	18.0	17.1	17.8	0.872	1.306	11.2	0.102	0.31
	LoGo (ours)	**19.0**	**18.2**	**18.4**	**0.894**	**1.111**	8.9	0.103	0.37
UniWorld	Base	17.6	16.8	17.5	0.881	1.260	13.8	0.132	0.39
	VideoGPA	17.7	16.9	17.5	0.882	1.275	13.8	0.132	0.39
	World-R1	17.9	17.1	18.2	0.869	1.401	13.1	0.129	0.22
	Global-only (ours)	18.3	17.5	18.0	0.892	1.195	13.3	0.126	0.37
	LoGo (ours)	**18.8**	**18.0**	**18.6**	**0.901**	**1.108**	14.0	0.130	0.41
Lyra2	Base	18.4	17.7	19.0	0.897	1.550	7.9	0.100	-0.23
	VideoGPA	18.0	17.4	18.6	0.896	1.511	7.7	0.099	-0.21
	World-R1	18.8	18.1	19.4	0.899	1.550	7.7	0.101	-0.20
	Global-only (ours)	19.4	18.6	19.7	0.900	1.396	7.6	0.099	-0.21
	LoGo (ours)	**19.7**	**19.0**	**20.0**	**0.905**	**1.352**	7.5	0.101	-0.19

Table 1: **Results on TrajectoryBench.** PSNR-V and PSNR-D measure RGBD reprojection PSNR (using VGGT and DA3), PSNR-GS is based on Gaussian reconstruction. LoGo achieves strong gains in all consistency metrics across three base models. It increases PSNR-D by 2.2dB from Lingbot2 base model, and reduces epipolar error by 26%. Meanwhile, LoGo preserves, and often improves, camera following and video quality. Standard errors are in Table 6 in the Appendix.

	Model	3D consistency					Camera		Video quality
		PSNR-V ↑	PSNR-D ↑	PSNR-GS ↑	MVCS ↑	Epipolar ↓	RPE° ↓	RPE trans ↓	VR-VQ ↑
Lingbot2	Base	13.3	13.1	15.6	0.719	2.161	20.52	0.082	-0.422
	VideoGPA	13.3	13.2	15.6	0.728	2.429	20.15	0.081	-0.409
	World-R1	13.6	13.3	15.7	0.717	2.271	20.13	0.070	-0.429
	Global-only (ours)	14.0	13.8	16.3	0.772	2.107	22.68	0.086	-0.305
	LoGo (ours)	**16.1**	**15.8**	**17.3**	**0.795**	1.365	18.33	0.077	-0.283
UniWorld	Base	14.7	14.5	16.6	0.748	1.510	12.94	0.057	-0.128
	VideoGPA	14.7	14.6	16.7	0.752	1.623	13.31	0.058	-0.129
	World-R1	14.9	14.7	17.3	0.747	1.926	15.45	0.064	-0.234
	Global-only (ours)	14.9	14.7	16.7	0.778	1.557	12.18	0.058	-0.134
	LoGo (ours)	**15.4**	**15.1**	**17.4**	**0.807**	1.429	12.64	0.055	-0.121
Lyra2	Base	15.2	14.8	16.8	0.819	2.448	22.72	0.148	-0.557
	VideoGPA	15.0	14.7	16.6	0.813	2.595	21.95	0.143	-0.508
	World-R1	15.3	15.2	16.9	0.832	2.213	21.72	0.140	-0.523
	Global-only (ours)	15.6	15.3	17.1	0.826	2.244	22.17	0.145	-0.525
	LoGo (ours)	**15.9**	**15.6**	**17.3**	**0.844**	1.863	22.38	0.145	-0.513

Table 2: **Results on DL3DV.** LoGo yields larger gains than baseline methods across three base models for all 3D consistency metrics, while maintaining camera control and video quality, with up to 37% decrease in epipolar error. Standard errors are shown in Table 7 in the Appendix.

ror (Kupyn et al., 2025). We report depth metrics using VGGT (Wang et al., 2025) and Depth Anything V3 (Lin et al., 2025) to avoid model bias. For **camera control**, we report relative pose error (RPE): the VGGT-estimated camera translation and rotation between frames versus the ground truth. For **video quality**, we measure video quality (VQ) score from VideoReward (Liu et al., 2026b), a Qwen-based (Yang et al., 2024) reward model trained on human preference.

Baselines. We compare with state-of-the-art 3D consistency post-training methods. Prior methods are based on WAN (Wan et al., 2025), a 80-frame image/text-to-video model without camera control. Since there has been no prior work on post-training camera-controlled models for 3D consistency, we extend these methods to this setting. We compare to the following: **(1)** VideoGPA (Du et al., 2026), a DPO (Rafailov et al., 2023) method using a single scalar VGGT (Wang et al., 2025) reprojection reward for a whole video; **(2)** World-R1 (Zhou et al., 2026) performs Flow-GRPO (Liu et al., 2026a) using VLM feedback on Gaussian reconstruction, yielding a scalar reward per rollout.

4.3 Results

LoGo post-training improves consistency, while maintaining or improving video quality and camera control. On both TrajectoryBench (Table 1) and DL3DV (Table 2), LoGo significantly improves 3D consistency from the base models, with up to +2.7dB PSNR and 37% decrease of epipolar error. For the best model, Lingbot2, LoGo also improves camera control and video quality. LoGo's effectiveness is shown qualitatively in Fig. 6, addressing both global and local inconsistencies, in-

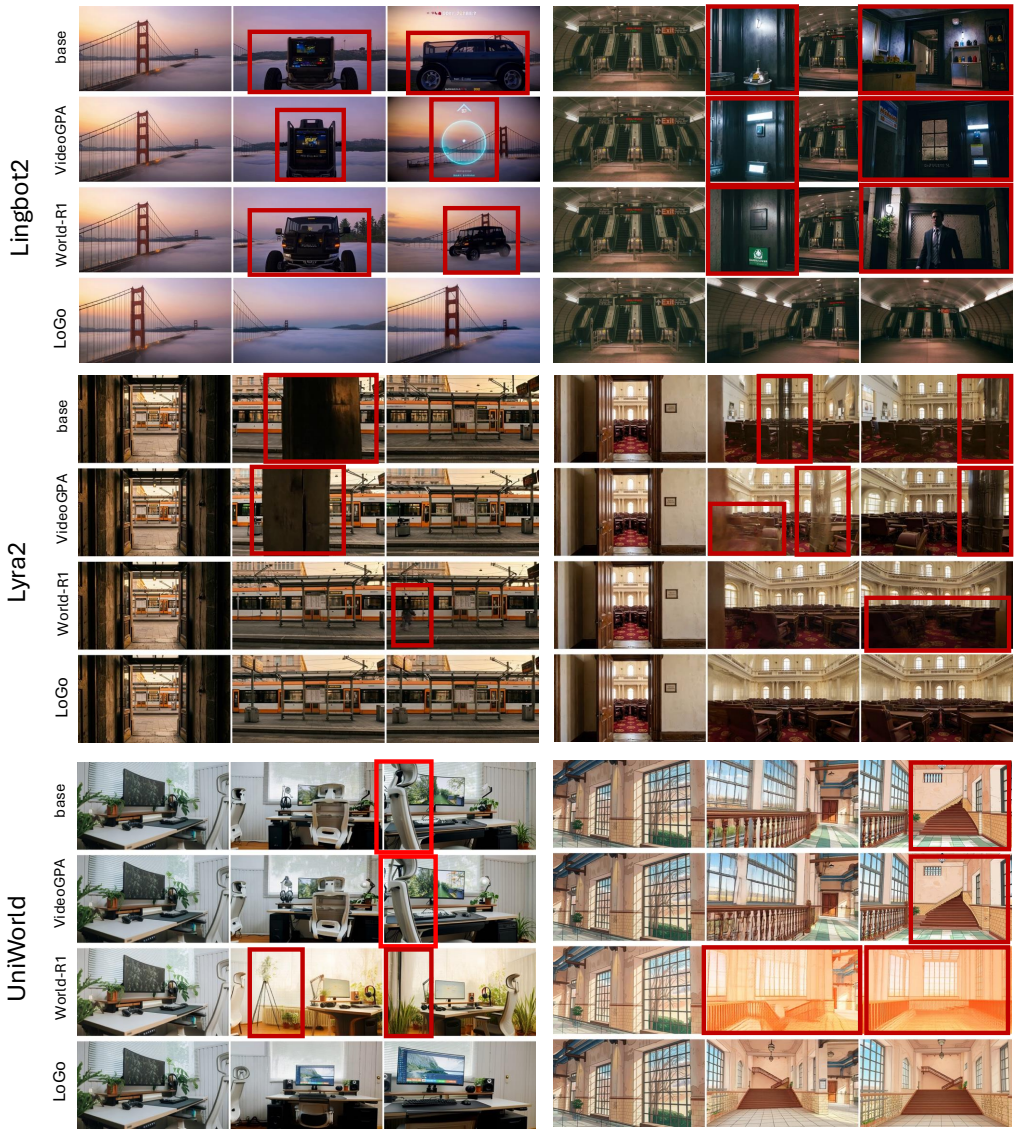


Figure 6: **Qualitative results of LoGo vs. baselines**. Red boxes indicate inconsistent regions. LoGo effectively fixes both global and local inconsistencies, including scene change (top right), hallucinated objects (top left), floater artifacts (middle), and inconsistent object appearance (bottom). Baseline post-training methods are not capable of resolving these issues, while LoGo is.

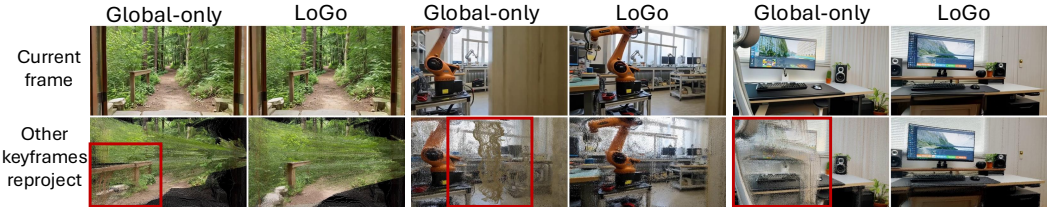


Figure 7: **Global-only reward vs. LoGo.** A 3D-consistent video should closely match its reprojections, though not perfectly, since keyframes in long videos overlap only sparsely. Global reward incurs inconsistencies, highlighted in red, such as hallucinated object (double railing), artifact (floater in the robot scene), and object appearance change (chair). LoGo, by combining global reward with local reward, effectively removes these inconsistencies.

	Model	PSNR-V ↑			PSNR-D ↑			Epipolar ↓		
		E	M	H	E	M	H	E	M	H
Lingbot2	Base	22.6	18.2	15.3	20.6	17.6	14.5	0.524	0.681	2.034
	VideoGPA	22.7	18.3	15.4	20.7	17.7	14.6	0.533	0.684	1.997
	World-R1	22.7	18.1	15.4	20.6	17.4	14.5	0.526	0.681	2.046
	LoGo	**23.8**	**20.2**	**17.6**	**21.6**	**19.7**	**17.0**	**0.490**	**0.601**	**1.439**
UniWorld	Base	23.0	17.2	16.7	21.2	16.4	16.1	0.450	**0.760**	1.623
	VideoGPA	23.1	17.2	16.8	21.2	16.4	16.2	0.450	0.773	1.642
	World-R1	22.4	17.3	17.2	20.5	16.5	16.6	0.462	0.860	1.805
	LoGo	**23.2**	**18.2**	**18.1**	**21.3**	**17.4**	**17.5**	**0.446**	**0.760**	**1.379**
Lyra2	Base	22.2	17.9	17.8	21.1	17.2	17.3	0.519	0.874	2.027
	VideoGPA	22.1	17.7	17.3	21.1	17.1	16.7	0.519	0.878	1.963
	World-R1	22.2	18.1	18.4	21.0	17.5	17.8	0.529	0.939	1.999
	LoGo	**22.5**	**18.8**	**19.4**	**21.2**	**18.1**	**18.8**	**0.510**	**0.847**	**1.723**

Table 3: **Difficulty breakdown on TrajectoryBench.** We show results on easy, medium, and hard categories of TrajectoryBench. PSNR-V denotes RGBD reprojection error with VGGT, and PSNR-D with DA3. LoGo's margin over the baselines widens as difficulty increases across all three base models. This trend also holds for other metrics found in Table 6 in the Appendix.

cluding scene change (subway changes into a room, top-right), hallucinated objects (car appears in golden gate bridge scene, top left), artifacts (column-like floaters for Lyra2), and object appearance consistency (chair orientation change, bottom left). All these inconsistencies are fixed by LoGo. We provide the full videos in the supplementary materials.

LoGo shows stronger gains than prior post-training methods, partly due to credit assignment. Prior methods, VideoGPA and World-R1, show moderate gains from base models. For Lingbot2, on both benchmarks the prior methods improve by 0.1dB for PSNR-GS and by 0.2dB for PSNR-D. We hypothesize LoGo's advantage comes partly from credit assignment via the local reward. In Table 2, our global-only variant's advantage over baselines is moderate, whereas LoGo shows stronger gains. This comparison showcases the importance of credit assignment via the local reward.

Fig. 6 shows qualitatively that baseline methods fail to fix the base models' inconsistencies: in the golden gate bridge example, both VideoGPA and World-R1 change the hallucinated object but do not remove it. World-R1, while reducing floaters (middle left), generates other artifacts like a morphing human-shadow-like objects in the train station and morphs the chairs in the auditorium (middle right). In the cartoon stylized example, UniWorld base model and all baselines transform the hallway with door into a staircase scene, while LoGo is able to generate a coherent stair scene.

Fig. 7 further compares LoGo with our global-only variant qualitatively, visualizing the point cloud reprojection from other keyframes versus the generated frame. A global-only reward fails to fix local inconsistencies like double railing, artifacts, and the altered chair orientation. This corroborates the quantitative finding that credit assignment is an important contributor to LoGo's advantage.

LoGo's advantage widens as the trajectories become more difficult. TrajectoryBench's increasing difficulty tier allows us to further understand LoGo's advantage. While all metrics degrade as the difficulty increases from easy to difficult, as expected, Table 3 shows that LoGo's advantage grows with difficulty. While LoGo's advantage is moderate on easy trajectories (+1dB PSNR-D for Lingbot2), the gain is much higher on hard trajectories (+2.5dB). The easy/medium/hard breakdown for all metrics with standard errors are shown in Table 6 in the Appendix.

5 Ablations and Analyses

We analyze the performance of LoGo over increasing video lengths, compare different credit assignment methods for LoGo's local reward, and ablate each component of LoGo's recipe. We further provide an ablation of voxel granularity and show LoGo's robustness on other RL algorithms.

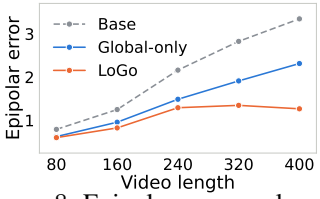


Figure 8: Epipolar error vs. length.

LoGo's gain strengthens as video length increases. We evaluate videos ranging from 80 to 400 frames in length. As shown in Fig. 8, longer-horizon generation is more challenging for the base model, Lingbot2, yielding

	3D Consistency					Camera		Video Quality
	PSNR-V ↑	PSNR-D ↑	PSNR-GS ↑	MVCS ↑	Epipolar ↓	RPE° ↓	RPE trans ↓	VR-VQ ↑
Base	15.3	14.5	15.7	0.812	2.034	13.2	0.108	0.21
Global-only	16.5	15.7	16.7	0.848	1.726	11.2	0.102	0.31
Window	17.0	16.3	17.1	0.856	1.666	9.5	**0.094**	0.32
Framewise	17.0	16.4	16.9	0.857	1.531	10.7	0.101	0.24
Patchwise	16.9	16.2	17.0	0.865	1.660	**8.4**	0.100	0.24
Voxel	**17.8**	**17.2**	**17.7**	**0.874**	**1.381**	8.9	0.107	**0.38**

Table 4: **Comparison of different local rewards.** We compare voxel (used in LoGo) with framewise, 2D patchwise, and 40-frame window credit assignment. Voxel-based yields the best consistency gains. Easy/medium/hard breakdown and standard errors are in Table 8 in the Appendix.

	3D Consistency					Camera		Video Quality	
	PSNR-V ↑	PSNR-D ↑	PSNR-GS ↑	MVCS ↑	Epipolar ↓	RPE° ↓	RPE trans ↓	VR-VQ ↑	HPSv3 ↑
Base	16.9	16.0	16.9	0.842	1.507	13.2	0.108	0.21	5.09
Local-only	**19.2**	18.2	**18.7**	0.893	1.078	11.2	0.116	0.32	4.47
+ interleave	**19.2**	**18.3**	18.6	**0.897**	**1.072**	**8.9**	0.107	**0.38**	4.94
+ global (final)	19.0	18.2	18.4	0.894	1.111	**8.9**	**0.103**	0.37	**5.18**

Table 5: **Ablating the LoGo recipe.** Using local-only reward yields best 3D consistency but drops camera and aesthetics metrics. Adding in reward interleaving improves other metrics, but still degrades HPSv3 compared to the base model. LoGo, which includes both reward interleaving and local-global blending, yields the best overall result. Standard errors are in Table 11 in the Appendix.

higher epipolar errors. LoGo leads at all lengths, and its advantage widens with the growing frame length – from 24% epipolar error reduction at 80 frames to 62% at 400 frames.

Voxel credit assignment in 3D wins over frame-or-patch-level assignment. There are multiple methods of credit assignment, such as assiging credit to windows of frames, each frame, or 2D patches, in addition to 3D voxels. Table 4 compares these credit assignment methods, and shows voxels to be the strongest. This is intuitive since voxels occupy the 3D space where inconsistencies occur. Localizing per-frame or per-patch can be inaccurate: since camera following is imperfect, a fixed frame or patch usually does not correspond to the same region across various rollouts.

Ablating the LoGo recipe. As explained in §3.2 and §3.3, local-global blending and reward interleaving improve video quality and camera following. We ablate the effect of each by adding them in one at a time in Table 5. Only using the local voxel reward yields the highest 3D consistency metrics at the cost of camera and aesthetics. Applying reward interleaving without local-global blending improves camera and video quality, but still degrades HPSv3. Blending with global reward further improves camera and aesthetics, surpassing the base model in every metric.

LoGo's advantage persists with other post-training algorithms. While our main results use DiffusionNFT (Zheng et al., 2026), we also show LoGo's advantage with other post-training methods, including FlowGRPO (Liu et al., 2026a) and DRaFT (Clark et al., 2024) in Table 9 in the Appendix.

Voxel granularity. Across voxel sizes from 0.05× to 0.7× 90-percentile (P90) scene depth, finer granularity up to 0.1× P90 scene depth improves both 3D consistency and camera control (§A.7.2).

6 Conclusion and Limitations

We propose LoGo, a reward design which enables camera-controlled video models to generate 3D-consistent worlds by post-training with a local-global reward blending. LoGo, via effective 3D credit assignment, outperforms prior methods. Motivated by the lack of 3D consistency evaluation for long-context camera-controlled video generation, we propose TrajectoryBench, which we hope will serve as a testbed for future development. We release code, checkpoints, and benchmark.

LoGo has two main limitations: **(1)** Very long-horizon generation (beyond 400 frames) remains challenging and might require new memory design, not just better reward design; **(2)** LoGo focuses on static scenes, and extending it to dynamic scenes is an exciting direction for future work. There, 4D reconstruction techniques such as Zhang et al. (2026) could be used to compute rewards similar in spirit to LoGo.

7 AI USE STATEMENT

In this paper, we used LLMs to polish language (asking ChatGPT to help with awkward sentences) and for visualization of schematic illustrations. We also used AI coding tools to help speed up standard code implementation. We have not used generative AI tools for the core research idea or to write the content of the paper.

REFERENCES

Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge Belongie, Mar Gonzalez-Franco, and Marta Tintore Gazulla. Vggrpo: Towards world-consistent video generation with 4d latent reward. *arXiv preprint arXiv:2603.26599*, 2026.

Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regularization into video generation model. *Advances in Neural Information Processing Systems*, 38:57602–57622, 2026.

Kevin Clark, Paul Vicol, Kevin Swersky, and David Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, volume 2024, pp. 4793–4822, 2024.

Hongyang Du, Junjie Ye, Xiaoyan Cong, Runhao Li, Jingcheng Ni, Aman Agarwal, Zeqi Zhou, Zekun Li, Randall Balestriero, and Yue Wang. Videogpa: Distilling geometry priors for 3d-consistent video generation. *arXiv preprint arXiv:2601.23286*, 2026.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 27713–27724, 2025.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

Wenhang Ge, Guibao Shen, Jiawei Feng, Luozhou Wang, Hao Lu, Xingye Tian, Xin Tao, and Ying-Cong Chen. Campilot: Improving camera control in video diffusion model with efficient camera reward feedback. *arXiv preprint arXiv:2601.16214*, 2026.

Google. Nano banana, 2026. URL https://aistudio.google.com/docs/image-generation.

Bohai Gu, Yueyang Yuan, Taiyi Wu, Dazhao Du, Jian Liu, Xiaoyi Pang, Jie Zhang, Xiaocheng Lu, Haobin Zhong, Xiaotong Zhao, et al. Worldcycle: Self-verifiable reinforcement learning for long-horizon video world models. *arXiv preprint arXiv:2608.04964*, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Orest Kupyn, Théo Uscidda, Marta Tintore Gazulla, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models. *arXiv preprint arXiv:2510.21615*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023.

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.

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, 2026a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2026b.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025.

Zhenyang Ni, Yijiang Li, Ruochen Jiao, Simon Sinong Zhan, Sipeng Chen, Zhenfei Yin, Minshuo Chen, Philip Torr, Zhaoran Wang, and Qi Zhu. Creflow: Corrective reflow for sparse-reward embodied video diffusion rl. *arXiv preprint arXiv:2605.14274*, 2026.

Pexels. Pexels: Free stock photos, 2026. URL https://www.pexels.com. Accessed: 2026-09-17.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Akshat Rakheja, Aarsh Ashdhir, Aryan Bhattacharjee, and Vanshika Sharma. World consistency score: a unified metric for video generation quality. *arXiv preprint arXiv:2508.00144*, 2025.

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.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

Leitian Tao, Baolin Peng, Wenlin Yao, Tao Ge, Hao Cheng, Mike Hang Wang, Jianfeng Gao, and Sharon Li. Trace: Turn-level reward assignment via credit estimation for long-horizon agents. *arXiv preprint arXiv:2607.13988*, 2026.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026a.

Meituan LongCat Team, Xunliang Cai, Meng Cheng, Feng Gao, Zhe Kong, Jiamu Li, Le Li, Weiheng Li, Hongyu Liu, Shuai Tan, et al. Longcat-video-avatar 1.5 technical report. *arXiv preprint arXiv:2605.26486*, 2026b.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

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.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

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*, 2026a.

Weijie Wang, Xiaoxuan He, Youping Gu, Zeyu Zhang, Yefei He, Yanbo Ding, Xirui Hu, Donny Y. Chen, Zhiyuan He, Yuqing Yang, Yifan Yang, and Bohan Zhuang. World-r1: Reinforcing 3d constraints for text-to-video generation. In *ICML*, 2026b.

Zehan Wang, Tengfei Wang, Haiyu Zhang, Xuhui Zuo, Junta Wu, Haoyuan Wang, Wenqiang Sun, Zhenwei Wang, Chenjie Cao, Hengshuang Zhao, et al. Worldcompass: Reinforcement learning for long-horizon world models. *arXiv preprint arXiv:2602.09022*, 2026c.

Zhaoqing Wang, Xiaobo Xia, Zhuolin Bie, Jinlin Liu, Dongdong Yu, Jia-Wang Bian, and Changhu Wang. Taming camera-controlled video generation with verifiable geometry reward. In *European Conference on Computer Vision*, pp. 625–642. Springer, 2026d.

Xunzhi Xiang, Zixuan Duan, Yabo Chen, Zhengxuan Wei, Guiyu Zhang, Zixiao Gu, Zhe Gao, Haibin Huang, Chi Zhang, Qi Fan, et al. Videoweave: Unlocking geometric consistency in video generation via joint geometry-video modeling. *arXiv preprint arXiv:2606.14162*, 2026.

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.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Yankai Yang, Yancheng Long, Wei Chen, Xingyu Lu, Hongyang Wei, Bin Wen, Fan Yang, Tingting Gao, Han Li, and Shuo Yang. Spatialflow-grpo: Where spatial credit drives image editing. *arXiv preprint arXiv:2606.26872*, 2026.

Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. Mind: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026.

Tengjiao Yin, Jinglei Shi, Heng Guo, and Xi Wang. Vigor: Video geometry-oriented reward for temporal generative alignment. *arXiv preprint arXiv:2603.16271*, 2026.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently reconstructing dynamic scenes one d4rt at a time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7382–7392, June 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.

Haiyang Zhou, Wangbo Yu, Chaoran Feng, Xunyu Zhou, Yonghong Tian, and Li Yuan. Uniworldview: Large-baseline view synthesis via video diffusion models, 2026. URL https://arxiv.org/abs/2608.04701.

Yuchen Zhuang, Di Jin, Jiaao Chen, Wenqi Shi, Hanrui Wang, and Chao Zhang. Workforceagent-r1: Incentivizing reasoning capability in llm-based web agents via reinforcement learning. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 34–49, 2026.

A APPENDIX

In the Appendix, we first describe the qualitative video examples uploaded in the supplementary material. We also include additional explanation of the localized reward calculation, additional experimental result tables, additional training, evaluation, and benchmark construction details, as well as additional ablations and analyses. We also upload the code of LoGo in the supplementary materials.

A.1 QUALITATIVE VIDEO EXAMPLES

We provide video examples in the supplementary materials in `videos/index.html`. The videos include both a comparison of LoGo with baselines, and a comparison of LoGo with global-only reward. We categorize the examples by the base models' failure modes: (1) Base model changes the scene upon revisit. (2) Base model generates artifacts or lighting shifts. (3) Base model hallucinates objects. (4) Base model makes existing objects disappear, or change unreasonably.

In order to fit in the 100MB limit of supplementary materials, we downsample the videos by $0.5\times$ from 832×480 to 416×240, which slightly degrades the video quality.

A.2 RL REWARD FORMULATION WITH LOGO

We instantiate voxelized credit assignment by making the advantage calculation non-uniform across a sample. For example, for DiffusionNFT, the canonical implementation is using a scalar r with

$$\mathcal{L}_{\text{NFT}} = \mathbb{E}\Big[r\|v_\theta^+ - v\|^2 + (1-r)\|v_\theta^- - v\|^2\Big]$$

With LoGo, we assign a different r per latent spatio-temporal patch by averaging the optimality probability (normalized reward) for all voxels that project to this patch:

$$\mathcal{L}_{\text{NFT},j}^{(k)} = \frac{1}{|A|}\mathbb{E}\Big[\sum_{i\in A=\{i:S_i\cap I_j(k)\neq\emptyset\}} r_i\|v_{\theta^+,j}^{(k)} - v_j^{(k)}\|^2 + (1-r_i)\|v_{\theta^-,j}^{(k)} - v_j^{(k)}\|^2\Big] \tag{3}$$

where S_i denotes all pixels that project to voxel i, and $I_j(k)$ is the set of pixels in frame j patch k, and A denotes all voxels that project onto the current patch. The FlowGRPO (Liu et al., 2026a) objective can be decomposed similarly.

A.3 ADDITIONAL EXPERIMENT RESULTS

We provide full results referenced in the main text below. We present the full easy/medium/hard breakdown for TrajectoryBench evaluation for all metrics. We also show standard errors to demonstrate the statistical significance of metric improvements.

Table 6 compares LoGo with baselines, with full breakdown of easy/medium/hard split of all 3D metrics and standard errors on TrajectoryBench. Table 7 shows LoGo's comparison with baselines on DL3DV with standard errors. Table 8 shows full difficulty breakdown for the ablation of different localization methods, demonstrating that voxel localization (used in LoGo) is the strongest, with growing advantage as the difficulty increases. Table 9 shows full difficulty breakdown of the LoGo evaluation with various post-training algorithms, including DiffusionNFT, Flow-GRPO, and DRaFT, with standard errors. Table 10 shows various metrics' improvement under different voxel granularities for voxelization. Table 11 shows standard errors for the LoGo recipe ablation.

A.4 ADDITIONAL TRAINING DETAILS

We perform post-training with LoGo on Lyra2, Lingbot2, and UniWorld on 1700 DL3DV scenes. For training, we use DL3DV's given camera trajectories based on COLMAP (Schönberger & Frahm, 2016). Since each model's translation unit is different from the COLMAP unit, we perform scaling of $0.4\times$ to convert to Lyra2 unit, and $1\times$ for UniWorld. Lingbot2 does not need scaling since it self-normalizes by each scene's peak step. We subsample the first half of the scene trajectories into 240 frames to ensure appropriate scene coverage and camera motion speed. We train all models with a batch size of 8 and group size of 16. For Lingbot2 and UniWorld, we use a group size of 16

	Model	3D consistency														Camera		Video quality
		PSNR-V ↑			PSNR-D ↑			PSNR-GS ↑			MVCS ↑			Epipolar ↓			RPE° ↓	RPE trans ↓	VR-VQ ↑
		E	M	H	E	M	H	E	M	H	E	M	H	E	M	H			
Lingbot2	Base	22.6	18.2	15.3	20.6	17.6	14.5	21.0	18.0	15.7	0.918	0.878	0.812	0.524	0.681	2.034	13.2	0.108	0.21
	VideoGPA	22.7±0.03	18.3±0.08	15.4±0.04	20.7±0.05	17.7±0.08	14.6±0.05	21.1±0.05	18.1±0.07	15.8±0.04	0.916±0.002	0.880±0.004	0.820±0.005	0.533±0.011	0.684±0.013	1.997±0.052	13.1±0.14	0.107±0.001	0.22±0.004
	World-R1	22.7±0.06	18.1±0.08	15.4±0.04	20.6±0.08	17.4±0.08	14.5±0.05	21.1±0.06	18.0±0.07	15.7±0.04	0.917±0.002	0.873±0.005	0.809±0.005	0.526±0.012	0.681±0.012	2.046±0.054	13.1±0.14	0.108±0.001	0.21±0.004
	Global (ours)	22.9±0.14	19.3±0.11	16.5±0.05	20.8±0.13	18.8±0.11	15.7±0.06	21.4±0.11	18.9±0.09	16.7±0.05	0.916±0.005	0.908±0.006	0.848±0.005	0.521±0.013	0.649±0.015	1.726±0.057	11.2±0.17	0.102±0.001	0.31±0.006
	LoGo (ours)	**23.8±0.16**	**20.2±0.11**	**17.6±0.05**	**21.6±0.15**	**19.7±0.11**	**17.0±0.06**	**21.7±0.11**	**19.4±0.09**	**17.4±0.05**	**0.942±0.006**	**0.929±0.006**	**0.871±0.006**	**0.490±0.010**	**0.601±0.012**	**1.439±0.052**	8.9±0.18	0.103±0.001	0.37±0.006
UniWorld	Base	23.0	17.2	16.7	21.2	16.4	16.1	20.6	17.5	16.9	0.900	0.898	0.871	0.450	**0.760**	1.623	13.8	0.132	0.39
	VideoGPA	23.1±0.02	17.2±0.02	16.8±0.02	21.2±0.05	16.4±0.04	16.2±0.03	20.6±0.05	17.6±0.03	16.9±0.02	0.900±0.001	0.898±0.001	0.871±0.002	0.450±0.002	0.773±0.010	1.642±0.036	13.8±0.06	0.132±0.000	0.39±0.002
	World-R1	22.4±0.07	17.3±0.08	17.2±0.06	20.5±0.08	16.5±0.08	16.6±0.06	20.1±0.09	17.9±0.08	18.0±0.05	0.894±0.003	0.883±0.006	0.858±0.005	0.462±0.003	0.860±0.015	1.805±0.042	13.1±0.19	0.129±0.001	0.22±0.006
	Global (ours)	22.8±0.05	17.6±0.05	17.6±0.05	20.8±0.06	16.9±0.07	17.1±0.06	20.5±0.07	17.9±0.06	17.6±0.04	0.899±0.001	0.901±0.005	0.887±0.003	0.447±0.003	0.773±0.017	1.514±0.039	13.3±0.14	0.126±0.001	0.37±0.004
	LoGo (ours)	**23.2±0.05**	**18.2±0.06**	**18.1±0.06**	**21.3±0.07**	**17.4±0.07**	**17.5±0.06**	**21.0±0.09**	**18.5±0.06**	**18.2±0.05**	**0.904±0.002**	**0.913±0.004**	**0.895±0.003**	**0.446±0.002**	**0.760±0.012**	**1.379±0.036**	14.0±0.18	0.130±0.001	0.41±0.004
Lyra2	Base	22.2	17.9	17.8	21.1	17.2	17.3	21.6	18.9	18.5	**0.924**	0.922	0.881	0.519	0.874	2.027	7.9	0.100	-0.23
	VideoGPA	22.1±0.04	17.7±0.06	17.3±0.04	21.1±0.05	17.1±0.06	16.7±0.04	21.6±0.06	18.7±0.06	18.0±0.04	0.920±0.003	0.926±0.003	0.880±0.003	0.519±0.003	0.878±0.033	1.963±0.050	7.7±0.11	0.099±0.001	-0.21±0.003
	World-R1	22.2±0.05	18.1±0.06	18.4±0.04	21.0±0.07	17.5±0.06	17.8±0.04	21.7±0.07	19.3±0.08	19.0±0.04	0.920±0.003	0.929±0.003	0.883±0.003	0.529±0.007	0.939±0.034	1.999±0.051	7.7±0.12	0.101±0.001	-0.20±0.003
	Global (ours)	**22.5±0.05**	18.6±0.06	19.0±0.04	**21.3±0.07**	17.9±0.06	18.4±0.04	**21.8±0.07**	19.6±0.07	19.4±0.04	0.921±0.003	0.929±0.003	0.884±0.003	0.513±0.003	0.869±0.023	1.784±0.050	7.6±0.11	0.099±0.001	-0.21±0.003
	LoGo (ours)	**22.5±0.04**	**18.8±0.06**	**19.4±0.04**	21.2±0.06	**18.1±0.07**	**18.8±0.05**	**21.8±0.08**	**19.8±0.07**	**19.7±0.04**	0.922±0.003	**0.930±0.003**	**0.892±0.003**	**0.510±0.003**	**0.847±0.025**	**1.723±0.051**	7.5±0.10	0.101±0.001	-0.19±0.003

Table 6: TrajectoryBench evaluation difficulty breakdown. Geometry columns are split by difficulty tier E / M / H for easy, medium, hard splits. LoGo achieves better 3D consistency than baseline methods in all twelve metrics across three splits for all three models, while maintaining or improving camera control and video quality, with the strongest advantage on the hard split. ± denotes paired standard error.

	Model	3D consistency					Camera		Video quality
		PSNR-V ↑	PSNR-D ↑	PSNR-GS ↑	MVCS ↑	Epipolar ↓	RPE° ↓	RPE trans ↓	VR-VQ ↑
Lingbot2	Base	13.3	13.1	15.6	0.719	2.161	20.52	0.082	-0.422
	VideoGPA	13.3±0.10	13.2±0.11	15.6±0.09	0.728±0.014	2.429±0.158	20.15±1.215	0.081±0.005	-0.409±0.010
	World-R1	13.6±0.11	13.3±0.10	15.7±0.11	0.717±0.017	2.271±0.154	20.13±1.681	0.070±0.004	-0.429±0.012
	Global (ours)	14.0±0.15	13.8±0.14	16.3±0.14	0.772±0.016	2.107±0.152	22.68±1.745	0.086±0.006	-0.305±0.017
	LoGo (ours)	**16.1±0.22**	**15.8±0.20**	**17.3±0.21**	**0.795±0.021**	1.365±0.198	18.33±1.504	0.077±0.006	-0.283±0.018
UniWorld	Base	14.7	14.5	16.6	0.748	1.510	12.94	0.057	-0.128
	VideoGPA	14.7±0.03	14.6±0.07	16.7±0.05	0.752±0.005	1.623±0.082	13.31±0.458	0.058±0.001	-0.129±0.008
	World-R1	14.9±0.12	14.7±0.12	17.3±0.14	0.747±0.018	1.926±0.102	15.45±1.432	0.064±0.002	-0.234±0.015
	Global (ours)	14.9±0.10	14.7±0.11	16.7±0.07	0.778±0.013	1.557±0.079	12.18±0.514	0.058±0.002	-0.134±0.011
	LoGo (ours)	**15.4±0.11**	**15.1±0.12**	**17.4±0.09**	**0.807±0.014**	1.429±0.072	12.64±0.949	0.055±0.002	-0.121±0.012
Lyra2	Base	15.2	14.8	16.8	0.819	2.448	22.72	0.148	-0.557
	VideoGPA	15.0±0.09	14.7±0.08	16.6±0.09	0.813±0.011	2.595±0.216	21.95±0.549	0.143±0.003	-0.508±0.011
	World-R1	15.3±0.11	15.2±0.13	16.9±0.12	0.832±0.010	2.213±0.200	21.72±0.768	0.140±0.003	-0.523±0.011
	Global (ours)	15.6±0.11	15.3±0.13	17.1±0.10	0.826±0.013	2.244±0.253	22.17±0.619	0.145±0.003	-0.525±0.010
	LoGo (ours)	**15.9±0.12**	**15.6±0.13**	**17.3±0.10**	**0.844±0.012**	1.863±0.225	22.38±0.690	0.145±0.002	-0.513±0.012

Table 7: DL3DV evaluation with standard errors. ± denotes paired standard errors. LoGo yields more improvement in 3D consistency compared to baseline post-training methods on three different base models for all four metrics, while maintaining camera control and video quality. The improvement is significant, with up to 37% decrease in epipolar error.

Reward readout	3D consistency															Camera		Video quality
	PSNR-V ↑			PSNR-D ↑			PSNR-GS ↑			MVCS ↑			Epipolar ↓			RPE° ↓	RPE trans ↓	VR-VQ ↑
	E	M	H	E	M	H	E	M	H	E	M	H	E	M	H			
Base	22.6	18.2	15.3	20.6	17.6	14.5	21.0	18.0	15.7	0.918	0.878	0.812	0.524	0.681	2.034	13.2	0.108	0.21
Global	22.9±0.14	19.3±0.11	16.5±0.05	20.8±0.13	18.8±0.11	15.7±0.06	21.4±0.11	18.9±0.09	16.7±0.05	0.916±0.005	0.908±0.006	0.848±0.005	0.521±0.013	0.649±0.015	1.726±0.057	11.2±0.17	0.102±0.001	0.31±0.006
Window	23.3±0.17	19.9±0.11	17.0±0.05	21.2±0.15	19.3±0.11	16.3±0.06	21.6±0.12	**19.4±0.09**	17.1±0.05	0.918±0.005	0.914±0.006	0.856±0.005	0.539±0.020	0.635±0.014	1.666±0.053	9.5±0.17	**0.094±0.001**	0.32±0.006
Framewise	23.5±0.17	19.7±0.11	17.0±0.05	21.3±0.16	19.0±0.11	16.4±0.06	21.5±0.11	19.0±0.09	16.9±0.05	0.929±0.005	0.918±0.006	0.857±0.005	0.531±0.021	0.617±0.013	1.531±0.053	10.7±0.17	0.101±0.001	0.24±0.006
Patchwise	23.4±0.16	19.6±0.11	16.9±0.06	21.2±0.15	18.8±0.11	16.2±0.06	21.5±0.12	19.0±0.09	17.0±0.05	0.930±0.005	0.918±0.006	0.865±0.005	0.509±0.010	0.621±0.013	1.660±0.054	**8.4±0.18**	0.100±0.001	0.24±0.006
Voxel	**23.9±0.14**	**20.3±0.11**	**17.8±0.05**	**21.6±0.14**	**19.7±0.11**	**17.2±0.06**	**21.7±0.11**	**19.4±0.09**	**17.7±0.05**	**0.937±0.005**	**0.934±0.006**	**0.874±0.005**	**0.499±0.029**	**0.587±0.013**	**1.381±0.053**	8.9±0.18	0.107±0.001	**0.38±0.006**

Table 8: Full ablation table of different ways to localize the reward. We compare voxel reward (used in LoGo as local reward), and other localization schemes such as framewise, 2D patchwise, frame windows (40f windows). Voxel localization yields the best improvement because it normalizes per-group rollouts in the native 3D space where 3D consistency is measured, without being affected by shifts in region represented per frame/patch caused by slightly different camera following in each rollout. This advantage persists across difficulty splits, and is most significant on the "hard" split, which has longer generation horizon and more complex camera motions.

and learning rate of 3e-4. For Lyra2, we train with a learning rate 3.33e-5 for more stability. For Lingbot2 and UniWorld, we use an equal weighting (0.5-0.5) of RGB and depth error in the reward. For Lyra2, we use a 0.3-RGB, 0.7-depth weighting to reduce reward hacking. For UniWorld, since it uses BLIP-2 (Li et al., 2023) captioner for inference, we also train with BLIP-2 captions. We do not use captions for Lingbot2 and Lyra2. Models are trained till the geometry reward converges, around 16-36 hours on 64 H100s for the three base models.

	RL algorithm	3D consistency															Camera		Video quality
		PSNR-V ↑			PSNR-D ↑			PSNR-GS ↑			MVCS ↑			Epipolar ↓			RPE° ↓	RPE trans ↓	VR-VQ ↑
		E	M	H	E	M	H	E	M	H	E	M	H	E	M	H			
	Base	23.0	17.2	16.7	21.2	16.4	16.1	20.6	17.5	16.9	0.900	0.898	0.871	0.450	0.760	1.623	13.8	0.132	0.39
DiffusionNFT	Ours-global	22.8±0.05	17.6±0.05	17.6±0.05	20.8±0.06	16.9±0.07	17.1±0.06	20.5±0.07	17.9±0.06	17.6±0.04	0.899±0.001	0.901±0.005	0.887±0.003	0.447±0.003	**0.773±0.017**	1.514±0.039	**13.3±0.14**	**0.126±0.001**	0.37±0.004
	LoGlo	**23.1±0.05**	**17.9±0.06**	**17.9±0.06**	**21.2±0.07**	**17.2±0.07**	**17.4±0.06**	**20.9±0.08**	**18.4±0.07**	**18.0±0.05**	**0.903±0.002**	**0.912±0.004**	**0.895±0.003**	**0.446±0.003**	0.781±0.017	**1.436±0.037**	13.9±0.18	0.130±0.001	**0.40±0.004**
FlowGRPO	Ours-global	23.2±0.03	17.7±0.03	17.5±0.03	21.3±0.06	16.8±0.04	16.8±0.04	20.9±0.10	18.0±0.04	17.5±0.03	0.901±0.001	**0.902±0.002**	0.880±0.002	0.456±0.001	0.788±0.014	**1.500±0.035**	13.8±0.07	**0.132±0.000**	**0.37±0.003**
	LoGlo	**23.3±0.03**	**17.8±0.03**	**17.6±0.03**	**21.5±0.05**	**17.0±0.04**	**16.9±0.04**	**21.0±0.08**	**18.1±0.04**	**17.6±0.03**	**0.902±0.001**	**0.902±0.002**	**0.883±0.002**	**0.454±0.002**	**0.764±0.008**	1.553±0.039	**13.7±0.10**	**0.132±0.000**	0.36±0.003
DRaFT	Ours-global	**23.0±0.02**	17.1±0.02	16.6±0.02	**21.1±0.05**	16.2±0.04	16.0±0.03	**20.6±0.07**	17.4±0.03	16.7±0.02	**0.899±0.000**	**0.894±0.001**	0.869±0.002	0.453±0.004	0.768±0.011	1.726±0.037	**14.0±0.07**	**0.132±0.000**	**0.39±0.003**
	LoGlo	**23.0±0.02**	**17.2±0.02**	**16.8±0.02**	**21.1±0.04**	**16.3±0.04**	**16.2±0.03**	**20.6±0.06**	**17.5±0.03**	**16.8±0.02**	**0.899±0.001**	0.893±0.002	**0.871±0.002**	**0.449±0.002**	**0.765±0.013**	**1.670±0.038**	**14.0±0.08**	**0.132±0.000**	**0.39±0.003**

Table 9: Full difficulty breakdown of TrajectoryBench evaluation of LoGo with three different post-training algorithms: DiffusionNFT, Flow-GRPO, and DRaFT. LoGo's advantage persists regardless of post-training algorithms. Paired standard error is displayed in each cell.

Voxel size	3D consistency															Camera		Video quality
	PSNR-V ↑			PSNR-D ↑			PSNR-GS ↑			MVCS ↑			Epipolar ↓			RPE° ↓	RPE trans ↓	VR-VQ ↑
	E	M	H	E	M	H	E	M	H	E	M	H	E	M	H			
Base	22.6	18.2	15.3	20.6	17.6	14.5	21.0	18.0	15.7	0.918	0.878	0.812	0.524	0.681	2.034	13.2	0.108	0.21
0.05	24.1±0.16	**20.7±0.11**	18.6±0.06	21.8±0.16	20.1±0.11	17.8±0.06	21.7±0.13	**19.7±0.09**	18.2±0.06	0.932±0.005	0.935±0.007	0.880±0.005	**0.466±0.010**	0.579±0.014	**1.229±0.051**	9.2±0.17	0.108±0.001	0.42±0.006
0.1	24.3±0.15	20.5±0.11	**18.7±0.06**	21.8±0.15	19.9±0.11	**17.9±0.07**	**21.8±0.10**	19.6±0.09	**18.4±0.06**	0.941±0.005	0.939±0.007	0.885±0.006	0.486±0.014	**0.577±0.014**	1.235±0.051	**9.0±0.19**	0.111±0.001	0.42±0.006
0.3	**24.6±0.16**	**20.7±0.11**	18.4±0.06	**22.2±0.16**	**20.2±0.11**	17.6±0.07	21.7±0.11	19.6±0.09	18.0±0.06	**0.947±0.006**	**0.940±0.007**	**0.893±0.006**	0.467±0.012	0.581±0.013	1.246±0.051	10.5±0.18	0.117±0.001	0.42±0.006
0.5	23.6±0.12	19.9±0.11	17.5±0.06	21.4±0.12	19.3±0.11	16.6±0.06	21.6±0.10	19.2±0.09	17.5±0.05	0.933±0.004	0.923±0.006	0.865±0.005	0.484±0.016	0.595±0.013	1.450±0.053	14.5±0.18	0.119±0.001	0.39±0.006
0.7	24.2±0.17	20.3±0.11	17.7±0.06	21.8±0.16	19.7±0.11	17.0±0.06	21.8±0.11	19.4±0.09	17.6±0.05	0.935±0.005	0.929±0.006	0.882±0.005	0.489±0.014	0.588±0.013	1.321±0.052	10.3±0.19	**0.107±0.001**	**0.43±0.006**

Table 10: Ablation of localization voxel size on TrajectoryBench, with difficulty breakdown and paired standard errors. Decrease of voxel size generally increases 3D consistency, up to 0.1 × median scene depth. We choose voxel size of 0.1 based on this ablation.

	3D consistency					Camera		Video quality	
	PSNR-V ↑	PSNR-D ↑	PSNR-GS ↑	MVCS ↑	Epipolar ↓	RPE° ↓	RPE trans ↓	VR-VQ ↑	HPSv3 ↑
Base	16.9	16.0	16.9	0.842	1.507	13.2	0.108	0.21	5.09
Local reward	**19.2±0.05**	18.2±0.05	**18.7±0.04**	0.893±0.004	1.078±0.034	11.2±0.12	0.116±0.001	0.32±0.004	4.47±0.019
+ interleave	**19.2±0.05**	**18.3±0.05**	18.6±0.04	**0.897±0.004**	**1.072±0.034**	**8.9±0.13**	0.107±0.001	**0.38±0.004**	4.94±0.013
+ global (final)	19.0±0.05	18.2±0.05	18.4±0.04	0.894±0.004	1.111±0.033	**8.9±0.13**	**0.103±0.001**	0.37±0.004	**5.18±0.010**

Table 11: Ablation on LoGo recipe with paired standard errors. Using local reward only yields best 3D consistency but worst camera and aesthetics metrics. Adding in reward interleaving improves other metrics, but still degrades HPSv3 compared to the base model. LoGo, which includes both reward interleaving and local-global blending, yields the best overall result.

A.5 Additional evaluation details

For depth PSNR evaluation, we use leave-one-out reprojection: we fuse keyframes (every 20th frame) from the video, leaving out the current frame, and project the point cloud using the given frame's camera pose. Leave-one-out is used because for long trajectories, keyframes can be spatially sparse and overall point cloud reprojection could be dominated by each frame's self-reprojection if not excluded.

A.6 Additional details on TrajectoryBench construction

TrajectoryBench contains 2000 tasks, each composed of an initial image, a camera trajectory, and an optional text prompt. We divide the benchmark into easy, medium, and hard splits. The scene types include indoor, outdoor, and "extension": one space going through a door into another space.

The easy split is a subsample of WorldScore of 80 frames and 1 camera action, sourced from a subset of WorldScore (Duan et al., 2025) to be comparable with existing benchmarks, thus putting the more difficult tasks in context. While WorldScore has a small split of "composite camera action" evaluation, it tests I2V models via three separate 80-frame, single-camera-action generations. Our medium split includes indoor, outdoor, and "extension" scene types. For indoor and outdoor, tests 3-action composition with a single 240-frame generation using the same initial images for comparability. Additionally, the medium split also includes "extension" scenes which go through a door into a new space. For the hard splits (up to 240 frames and 8 camera motions), we source initial images for the indoor and outdoor scenes from Pexels (Pexels, 2026). Indoor covers residential, commercial, institutional, industrial, recreational, cultural and transit spaces. Outdoor covers civic gathering, transportation, agricultural (farms, vineyards), commercial, recreational, and natural spaces. For the "extension" scene type, since we require the door to be of specific size and location, and want to maintain scene diversity, we leverage Nano Banana (Google, 2026) to generate spaces



Figure 9: Visualization of LoGo's local reward. The red-blue overlay is the normalized score per voxel projected back to image patches - blue means good and red means bad. In this example, a tree that did not exist in earlier is hallucinated in the middle of the video and splits into two. This is caught by LoGo's local reward component, which concentrates on the tree region, and subsequently fixed by post-training with LoGo.

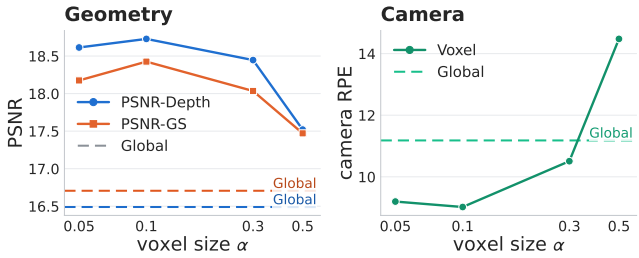


Figure 10: Effect of voxel size on geometry metric and camera following metrics. Both improve with finer voxel size up till 0.1× P90 scene depth.

covering diverse scene types, such as ballroom, indoor pool, aircraft hangar, and courtroom. The camera trajectories are human-designed using language primitives (go back, go left, etc.), mapped into camera paths, and evaluated on base models to adjust the translation and rotation amount to ensure the viewer 1) does not crash into walls or objects, and 2) cover a substantial sweep of spaces.

We also include stylized scenes: 20% of the benchmark is stylized via Nano Banana into eight styles, including Japanese anime, classical oil painting, Minecraft, watercolor, geometric low poly 3D, 3D studio cartoon (Pixar style), cyberpunk, and Victorian steampunk. The diverse coverage of TrajectoryBench and the inclusion of various difficulty levels of trajectories make enable a comprehensive evaluation of models.

A.7 Additional Analyses and Ablations

A.7.1 Additional visualization of LoGo's local reward.

We visualize voxel-localized credit assignment by back-projecting the score to image patches. The two keyframes show a tree non-existent in the start of the video hallucinated in the middle of the frame and splits into two parts. This inconsistency is caught by the local reward of LoGo– the red (penalty) area concentrates around the tree. Training with LoGo removes the hallucinated tree.

A.7.2 Voxel granularity ablation.

We further ablate voxel granularity to understand what leads to the best credit assignment performance. Figure 10 shows the effect of voxel granularity on geometry and camera control metrics. Since our scenes comprise of indoor and outdoor scenes of diverse scales, we define voxel size as a multiple of 90-percentile (P90) scene depth. Comparing voxel sizes of 0.05× to 0.7× P90 scene depth, we find that up to 0.1× P90 scene depth, finer granularity improves both 3D consistency geometry metrics (depth reprojection PSNR) and camera control. Finer-grained credit assignment

intuitively helps improve the 3D objective better. Regarding camera, we hypothesize that penalizing a large voxel can incentivize the model to just turn the camera less to "see" less of the penalized voxel, whereas fine-grained penalty does not provide this incentive: turning camera less could avoid the penalized small voxel, but also potentially other good voxels. Voxels that are too fine-grained no longer brings much improvement, since most inconsistencies can already be captured by the 0.1 granularity.

A.8 CODE

We provide code in `code/` of the supplementary materials.