GEO-ALIGN: VIDEO GENERATION ALIGNMENT VIA METRIC GEOMETRY REWARD

Anonymous authors
Paper under double-blind review

ABSTRACT

Camera-controlled video generation has achieved remarkable progress in recent years. However, existing video-to-video re-rendering methods primarily rely on Supervised Fine-Tuning using synthetic datasets. At present, there is an extreme scarcity of synchronized, multi-view real-world video data. Consequently, the prevailing paradigm often exhibits limited generalization when processing out-of-distribution real-world videos, with models struggling to accurately adhere to physical scales and camera trajectories. To bridge this gap, we propose Geo-Align, the first Reinforcement Learning framework specifically designed for camera-controlled video re-rendering. Built upon a pretrained model, we optimize the model through a scale-aware perceptual reward mechanism. Specifically, we introduce a metric 3D estimator to extract precise camera trajectories from generated videos, explicitly penalizing deviations in rotation and translation. Furthermore, we meticulously designed a data pipeline strategy based on real-world conditioning videos and target camera trajectories derived from synthetic data, eliminating the reliance on paired data. Extensive experiments demonstrate that Geo-Align consistently outperforms existing supervised learning baselines in both precise camera controllability and visual fidelity, indicating the effectiveness of our method.

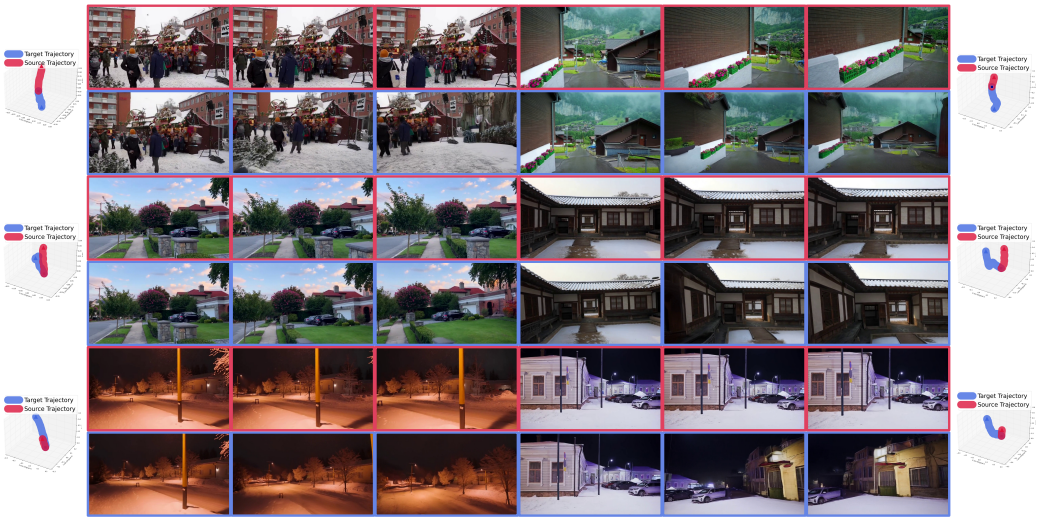

Figure 1: Given a conditioning video, Geo-Align synthesizes a novel view video according to the target camera trajectory.

1 INTRODUCTION

Camera controllability plays a vital role in video generation, particularly in fields such as film production and game engine rendering. In this paper, we focus on the video retake task. Formulated as a video-to-video generation problem, this task requires a model to synthesize a novel-view video

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.

along a target camera trajectory, given a conditioning video and the target trajectory as inputs. Recent methods such as ReCamMaster (Bai et al., 2025) and ReDirector (Park et al., 2025b), have successfully re-rendered dynamic scenes from input videos along new camera trajectories by training on synthetic datasets generated via engines like Unreal Engine. While TrajectoryCrafter (Yu et al., 2025) and CogNVS (Chen et al., 2025a) are methods based on reconstruction, warping, and subsequent completion. However, the current supervised learning paradigm for generating videos with novel camera trajectories faces two core bottlenecks:

Data Scarcity: Unlike camera-controlled video generation conditioned on a single initial frame, video retake requires multi-view video data for supervised training. Given the scarcity of such real-world data, implicit condition methods (Bai et al., 2025; Park et al., 2025b) predominantly rely on synthetic datasets, while warping-based methods (Yu et al., 2025; Van Hoorick et al., 2024; Chen et al., 2025a) rely on point cloud renderings to synthesize target videos, constructing such data is highly non-trivial. While fine-tuning on synthetic data yields impressive results, these models often exhibit significant domain shift when performing inference on real-world scenes.

Metric Ambiguity: Camera pose annotations for existing real-world videos are often scale-less. Even the MultiCam-Video data constructed by ReCamMaster (Bai et al., 2025) only provides metric information for synthetic data. Standard SFT loss functions focus on pixel-level or feature-level reconstruction rather than explicitly optimizing for physically meaningful, metric-level camera alignment, frequently leading to scale drift in generated trajectories.

To address these challenges, we propose Geo-Align, a framework that introduces Reinforcement Learning (RL) to directly optimize the physical alignment and visual quality of camera movements. Unlike previous SFT paradigms (Bai et al., 2025; Park et al., 2025b) that rely on time-synchronized ground-truth videos from multiple camera angles, reinforcement learning methods do not require video data corresponding to the target camera trajectory. Since real-world conditioning videos are easily obtainable, we can post-train the model via RL as long as we have the target camera trajectory. We adopt a fusion strategy combining real and synthetic data. During RL training, the conditioning videos are real-world captures. For the target camera trajectories, we sample from OmniWorld (Zhou et al., 2025) gaming data, which provides a rich variety of natural camera movements. Since gaming trajectories are typically non-metric, we perform rescaling using Truncated Gaussian Sampling. Specifically, we sample the maximum values for rotation and translation between adjacent frames within defined thresholds and rescale the camera trajectories to reasonable scales accordingly.

We utilize a Verifiable Geometry Reward to train our model, which compares the camera trajectories estimated from the generated video (via MapAnything (Keetha et al., 2025)) against the target trajectories. A metric evaluator is introduced to mitigate metric-related reward hacking during the reinforcement learning process. This effectively penalizes degenerate solutions—such as the model producing a shape-preserving but slow-moving trajectory in response to a rapid target trajectory. To prevent visual degradation during geometric optimization and preserve the model's priors, we also incorporate aesthetic rewards, utilizing VideoAlign (Liu et al., 2025b) and HPSv3 (Ma et al., 2025) as the reward models. We freeze the majority of the model's parameters, training only the self-attention layers.

We evaluate our model on the DAVIS (Pont-Tuset et al., 2017) datasets across the ten target camera trajectory categories defined by ReCamMaster (Bai et al., 2025). Results demonstrate that our RL-trained model not only improves accuracy in following target trajectories on real-world data but also outperforms the original model across various aesthetic evaluation metrics. Our core contributions are as follows:

- **Reinforcement Learning for Video Retake**: We utilize metric geometry model to extract rotation and translation errors. This enables our model to better align with geometric constraints and achieve more accurate metric scaling in real-world conditioning videos. Furthermore, we incorporate aesthetic rewards to enhance the overall quality of the generated videos.

- **Fusion Data Strategy**: We leverage MapAnything (Keetha et al., 2025) to extract camera poses from Citywalk (Li et al., 2025b) dataset as real-world conditioning priors. By combining this with Truncated Gaussian Sampling to rescale target trajectories from gaming data, we enhance training diversity and bridge the scale gap between source videos and target trajectories. Furthermore, it circumvents the necessity of paired multi-view video data.

- **State-of-the-Art (SOTA) Performance**: Our RL-trained model achieves SOTA performance on the DAVIS (Pont-Tuset et al., 2017) dataset across ReCamMaster's (Bai et al., 2025) 10 trajectory types, consistently improving both camera trajectory fidelity and overall visual aesthetics. Qualitative comparisons further demonstrate a noticeable improvement in the quality of the generated videos.

2 Related Work

2.1 Camera-Controlled Video Retake

Camera-controlled video retake (Bahmani et al., 2025; He et al., 2024; Wang et al., 2024b; Go et al., 2025) aims to synthesize novel views from existing footage by redirecting camera trajectories through generative models. Early approaches predominantly rely on explicit geometric transformations, utilizing external depth estimators (Hu et al., 2025; Chen et al., 2025b) and point trackers (Karaev et al., 2024; Xiao et al., 2024) to warp input frames before refining them with video diffusion models (Wan et al., 2025b; Kong et al., 2024; Yang et al., 2024), as seen in methods like TrajectoryCrafter (Yu et al., 2025) and CogNVS (Chen et al., 2025a). However, these explicit methods frequently suffer from warping artifacts that propagate directly into the synthesized output, particularly under dynamic camera motions or complex scene structures. To bypass explicit warping, implicit methods (Chen et al., 2025a; Jeong et al., 2025; Lu et al., 2025; Zhang et al., 2025a) such as Generative Camera Dolly (GCD) (Van Hoorick et al., 2024) and ReCamMaster (Bai et al., 2025) condition models directly on camera extrinsic parameters, internalizing multi-view geometry through synthetic datasets. While recent advancements like ReDirector (Park et al., 2025b) extend this implicit paradigm to handle variable-length inputs and dynamic motions via Rotary Camera Encoding (RoCE). ll these frameworks fundamentally rely on supervised fine-tuning, where the primary bottleneck is the severe scarcity of time-synchronized multi-view video data. Since constructing such datasets from real-world footage is exceedingly difficult, existing SFT methods (Bai et al., 2025; Park et al., 2025b) are forced to rely heavily on synthetic data.

2.2 Feed-Forward 3D Reconstruction

Recent feed-forward models directly predict scene geometry without traditional SfM optimization (Triggs et al., 1999; Snavely et al., 2006; Wu, 2013; Schönberger & Frahm, 2016). DUSt3R (Wang et al., 2024a) pioneered this by regressing dense point maps from unconstrained images. To handle continuous visual streams, methods (Lan et al., 2025; Wang & Agapito, 2024; Zhuo et al., 2026; Liu et al., 2025c) such as CUT3R Wang et al. (2025b) and WinT3R (Li et al., 2025c) introduced stateful memory and sliding-window mechanisms for efficient online perception. Concurrently, models (Zhang et al., 2026; Yang et al., 2025; Deng et al., 2026; Leroy et al., 2024; Zhang et al., 2025b) like VGGT (Wang et al., 2025a), π^3 (Wang et al., 2025c), and Depth Anything 3 (Lin et al., 2025) have scaled into unified foundational architectures capable of jointly inferring multi-view geometry, cameras, and depth. Despite these advances, achieving accurate metric-scale reconstruction remains challenging. To address this, MapAnything (Keetha et al., 2025) introduces a universal framework specifically for metric 3D reconstruction. By employing a factored representation that decouples camera poses and depth into scale-invariant components and explicit global scales, MapAnything (Keetha et al., 2025) robustly maps local geometry into a unified metric space without test-time optimization.

2.3 Group Relative Policy Optimization in Generative Models

Group Relative Policy Optimization (GRPO) (Guo et al., 2025) has recently emerged as a powerful online reinforcement learning framework for aligning generative models (Fei et al., 2025; Geng et al., 2025; Xue et al., 2025). In flow-matching (Lipman et al., 2022) domains, Flow-GRPO (Liu et al., 2025a) enables online RL via ODE-to-SDE conversion, while MixGRPO (Li et al., 2025a) further improves optimization efficiency by introducing a mixed ODE-SDE sliding window sampling mechanism. This paradigm has similarly advanced video generation: GrndCtrl (He et al., 2025) utilizes GRPO for physically grounded world modeling, and recent frameworks (Wang et al., 2025d; Ge et al., 2026; Wang et al., 2026) adopt verifiable geometry rewards to optimize precise camera-controlled video generation. Another line of work enhance synthesis quality by incorporating explicit

(Wu et al., 2025a; Kupyn et al., 2025; Yin et al., 2026) or implicit (An et al., 2026; Yan et al., 2025; Du et al., 2026; Wu et al., 2025b) geometric constraints as reward signals to enforce multi-view consistency. Furthermore, LongCat-Video (Team et al., 2025) demonstrates robust multi-reward RLHF in foundational video models by introducing crucial stabilization techniques—specifically, employing max group standard deviation to bound reward variances within groups, and utilizing policy and KL loss reweighting to dynamically balance optimization and prevent reward hacking. Building upon these advancements, our method synergistically integrates the efficient mixed sampling framework of MixGRPO (Li et al., 2025a) with the max group standard deviation and policy/KL loss reweighting strategies from LongCat-Video (Team et al., 2025), achieving highly stable and computationally efficient policy optimization.

3 METHODOLOGY

3.1 OVERVIEW

Given an input conditioning video and a user-specified, unseen camera trajectory, we aim to re-render and generate a novel view video sequence. Formally, let $\mathbf{x}_{1:N}$ denote the conditioning video of length N, and c be the corresponding text prompt. To guide the generation process along a designated path, the model is additionally conditioned on a target camera trajectory $\mathbf{P}^{tgt}_{1:N}$, including target camera intrinsic parameters $\mathbf{K}^{tgt}_{1:N}$ and extrinsic parameters $\mathbf{E}^{tgt}_{1:N}$.

Our framework is built upon a pretrained video world model, denoted as $\mathcal{W}_\theta$. During the iterative generation process (e.g., diffusion or flow matching), the model predicts the denoised representation (or velocity vector) given a noisy latent $\mathbf{z}_t$ at timestep t. The conditional generation process can be formulated as:

$$\hat{\mathbf{v}}_\theta = \mathcal{W}_\theta(\mathbf{z}_t, t, \mathcal{C}), \tag{1}$$

where $\mathcal{C} = \{\mathbf{x}_{1:N}, c, \mathbf{P}^{tgt}_{1:N}\}$ encapsulates all the multimodal conditioning signals. Although fine-tuning video foundation models multi-view videos offers a viable solution to this task, the inherent scarcity of such data remains a significant bottleneck. Relying solely on supervised fine-tuning often leads to geometric inconsistencies and suboptimal camera control. Therefore, the adoption of RL frees us from multi-view data dependencies, unlocking the potential to train on vastly larger and more diverse data. Our goal is to optimize the model parameters θ to maximize a composite reward function $\mathcal{R}$, which comprehensively evaluates the alignment between the generated video $\mathbf{y}_{1:N}$ and the target trajectory $\mathbf{P}^{tgt}_{1:N}$, as well as the overall video quality. The RL objective is defined as:

$$\max_\theta \mathbb{E}_{\mathbf{y}_{1:N}\sim\mathcal{W}_\theta(\cdot|\mathcal{C})}\left[\mathcal{R}(\mathbf{y}_{1:N}, \mathbf{P}^{tgt}_{1:N})\right]. \tag{2}$$

By directly optimizing this reward, the model is guided to strictly adhere to the prescribed target trajectory while maintaining superior spatiotemporal fidelity.

3.2 MULTI-DIMENSIONAL REWARD DESIGN

Verifiable Geometry Reward. To enforce rigorous spatial alignment between the generated video $\mathbf{y}_{1:N}$ and the designated target trajectory, we introduce a verifiable Geometry Reward. We construct our 3D evaluator upon MapAnything (Keetha et al., 2025), a metric feed-forward 3D reconstruction model. By feeding the generated video into the 3D evaluator, we extract the predicted camera trajectory, comprising translations $\hat{\mathbf{t}}_{1:N}$ and rotations $\hat{\mathbf{R}}_{1:N}$. The geometric discrepancy is then quantified against the input target trajectory $(\mathbf{t}_{1:N}, \mathbf{R}_{1:N})$ across two dimensions. Specifically, we compute the weighted Euclidean deviation for translation and the angular deviation for rotation:

$$D_{trans} = \sum_{i=1}^{N} w_i \|\mathbf{t}_i - \hat{\mathbf{t}}_i\|_2, \tag{3}$$

$$D_{rot} = \sum_{i=1}^{N} w_i \arccos\left(\frac{\mathrm{Tr}(\mathbf{R}_i^\top \hat{\mathbf{R}}_i) - 1}{2}\right), \tag{4}$$

where w_i represents the temporal weight for the i-th frame. A key empirical observation motivates this weighting scheme: pretrained video generative models typically exhibit strong adherence to the

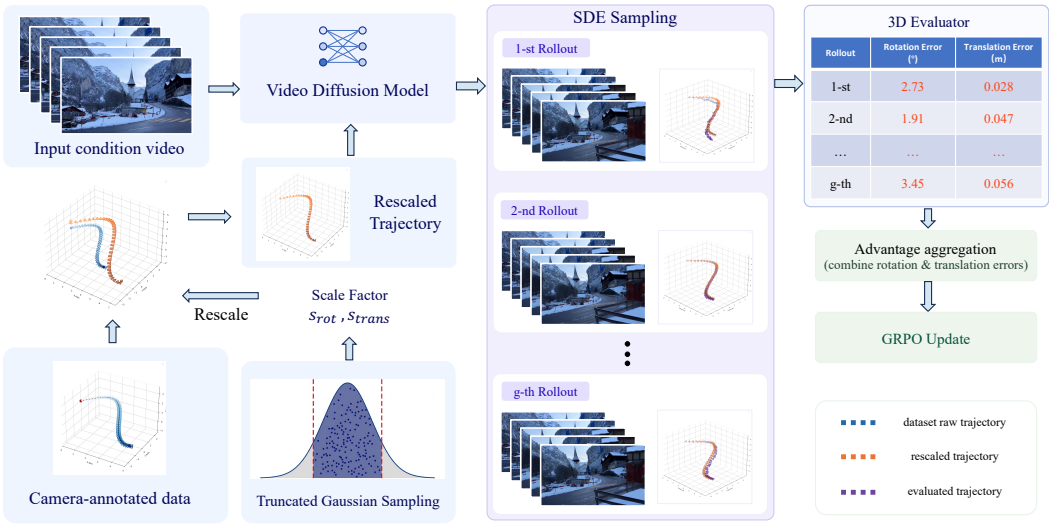

Figure 2: **Geo-Align pipeline.** Given a conditioning video, we sample a camera trajectory from other camera-annotated data and scale it to a plausible range, with the scaling factor drawn from a truncated Gaussian distribution. After the model generates a set of rollout videos, a metric 3D evaluator assesses the camera trajectory of each sample to compute geometry rewards. Finally, the model is optimized via Group Relative Policy Optimization (Guo et al., 2025).

conditioning trajectory in the initial frames, but suffer from severe error accumulation and spatial drift in the latter frames. Because the latter frames more accurately reflect the model's true predictive capability and are the primary bottleneck in trajectory control, we design w_i as a monotonically increasing function of time i (e.g., $w_1 < w_2 < \cdots < w_N$). This temporally progressive weighting mechanism explicitly penalizes long-term drift and forces the RL process to prioritize the optimization of challenging latter frames.

Perceptual and Aesthetic Reward. Optimizing solely for geometric alignment can inadvertently lead to reward hacking, resulting in perceptual degradation or unnatural artifacts. To preserve and enhance the visual fidelity of the synthesized video, we incorporate multidimensional aesthetic and quality rewards. First, we leverage the VideoAlign (Liu et al., 2025b) evaluator to assess sequence-level dynamics, yielding a visual quality score (s_{vis}) and a motion quality score (s_{mot}). Furthermore, to guarantee superior single-frame visual aesthetics and high-frequency details, we utilize HPSv3 (Ma et al., 2025) to evaluate the perceptual quality of each individual frame.

3.3 Flow Matching Optimization via GRPO

To efficiently optimize the pretrained flow matching model $\mathcal{W}_\theta$ for trajectory-controlled generation, we employ Group Relative Policy Optimization (Guo et al., 2025). Traditional PPO (Schulman et al., 2017) relies on a memory-intensive value model for baseline estimation. GRPO resolves this memory constraint by removing the value model and leveraging the relative scores within a group of outputs to compute the advantage. Given the prohibitively long group sampling time of video generation models, we adopt the sliding-window sampling strategy from MixGRPO (Li et al., 2025a). This mechanism restricts stochastic sampling and gradient updates strictly to an active temporal window, significantly accelerating convergence. Furthermore, since directly summing multi-dimensional rewards is mathematically unstable, we aggregate the feedback in the advantage space. Following the max group standard deviation strategy (as in LongCat Video (Team et al., 2025)), we robustly normalize each reward dimension $k \in \{rot, trans, vis, mot, hps\}$ within a group of G sampled rollouts to prevent the over-amplification of low-variance noise:

$$\hat{A}_k^{(j)} = \frac{r_k^{(j)} - \mu_k}{\max(\sigma_k, \epsilon)}, \tag{5}$$

where μ_k and σ_k are the group mean and standard deviation. The total advantage $A_{total}^{(j)}$ is then formulated as:

$$A_{total}^{(j)} = \sum_{i \in k} \lambda_i \hat{A}_i^{(j)}. \tag{6}$$

Standard GRPO incorporates a KL-divergence penalty to anchor the policy to the pretrained model. However, to maximize the model's exploratory capability on entirely novel, out-of-distribution target camera trajectories, we remove this KL penalty. Incorporating a timestep-aware policy loss weight w_t to balance gradients across diffusion stages as in LongCat Video (Team et al., 2025), our final objective function is:

$$\mathcal{J}(\theta) = \mathbb{E}_{t,\mathbf{z}_t,\mathcal{C}} \left[\frac{1}{G} \sum_{j=1}^{G} w_t \min \left(\rho_t^{(j)} A_{total}^{(j)}, \text{clip} \left(\rho_t^{(j)}, 1-\epsilon_c, 1+\epsilon_c \right) A_{total}^{(j)} \right) \right]. \tag{7}$$

where $\rho_t^{(j)}$ denotes the policy probability ratio, and ϵ_c is the clipping hyperparameter.

3.4 Metric-Aware Data Sampling Pipeline

RL removes the need for paired ground-truth target videos. We use in-the-wild CityWalk (Li et al., 2025b) videos as conditioning inputs and estimate their source camera trajectories with MapAnything (Keetha et al., 2025). Target trajectories are sampled independently from the OmniWorld (Zhou et al., 2025) gaming dataset. These trajectories provide diverse camera motions but lack an absolute physical scale and can contain severe rotations.

We rescale each trajectory using its maximum frame-to-frame translation and rotation speeds, v_{trans}^{max} and v_{rot}^{max}. Desired maximum speeds τ_{trans} and τ_{rot} are drawn from truncated Gaussian distributions with bounded supports, retaining motion diversity while controlling scale:

$$s_{trans} = \frac{\tau_{trans}}{v_{trans}^{max} + \epsilon}, \qquad s_{rot} = \frac{\tau_{rot}}{v_{rot}^{max} + \epsilon}, \tag{8}$$

$$\tilde{\mathbf{t}}_i = s_{trans} \mathbf{t}_i, \qquad \tilde{\mathbf{R}}_i = \exp(s_{rot} \log(\mathbf{R}_i)). \tag{9}$$

Here ϵ prevents division by zero. This decoupled pipeline combines real-world appearance with rescaled synthetic camera motion without requiring synchronized multi-view videos. Appendix A.1 gives the full speed definitions, sampling distributions, and rescaling procedure.

4 Experiments

4.1 Implementation Details

We adopt ReDirector (Park et al., 2025b) which is based on Wan2.1 (Wan et al., 2025a) 1.3B as our foundational pretrained video generation model. Following our proposed metric-aware data sampling pipeline, we continuously draw conditioning videos from the CityWalk (Li et al., 2025b) dataset and physically rescaled target trajectories from the OmniWorld (Zhou et al., 2025) dataset. For the verifiable geometric reward, MapAnything (Keetha et al., 2025) is employed as the frozen 3D evaluator. To preserve the strong spatiotemporal prior of the pretrained base model while enabling precise spatial control, we employ a parameter-efficient fine-tuning strategy: during the RL optimization, we solely update the weights of the self-attention layers, keeping all other network components strictly frozen. The model is configured to generate video sequences of $N = 81$ frames at a spatial resolution of 480×832. During inference and RL sampling, the continuous flow matching generation process is discretized into $T = 25$ denoising timesteps. For the GRPO (Guo et al., 2025) reinforcement learning framework, we follow the efficient mixed sampling framework of MixGRPO (Li et al., 2025a) and set the group size to $G = 12$ video rollouts per condition to compute the robust group-normalized advantages. The network is optimized for a total of 140 RL iterations using a constant learning rate of $\eta = 1 \times 10^{-4}$. The post-training process is distributed across 64 NVIDIA A800 GPUs, consuming about 130 hours.

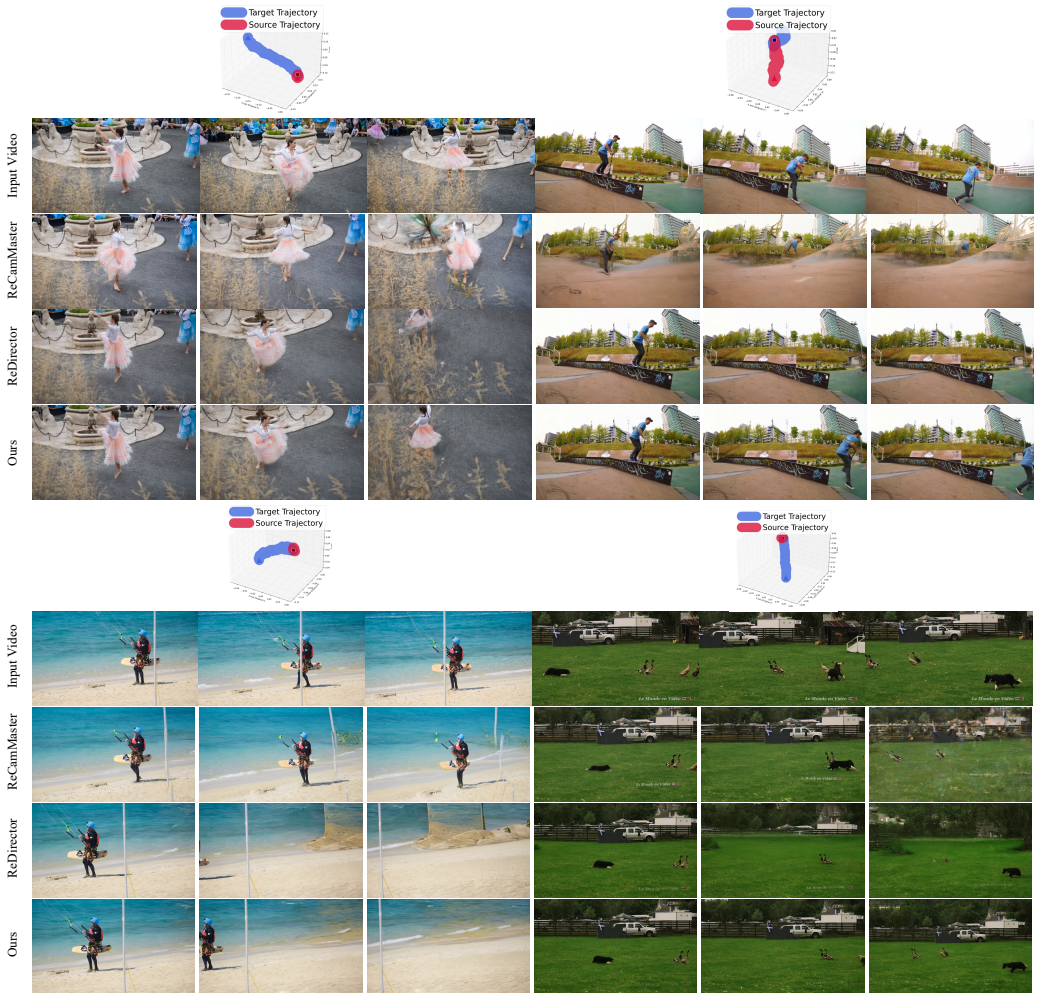

Figure 3: **Qualitative results on the DAVIS (Pont-Tuset et al., 2017) dataset.** Geo-Align demonstrates superior capabilities in maintaining geometric consistency between the foreground subject and the background, whereas other methods suffer from varying degrees of distortion.

4.2 Baselines

We compare our method against two categories of state-of-the-art baselines. The first category comprises explicit warping-based methods, specifically TrajectoryCrafter (Yu et al., 2025) and CogNVS (Chen et al., 2025a). As these models are limited to generating fewer than 49 frames during inference. The second category consists of models conditioned on implicit camera extrinsics, including ReCamMaster (Bai et al., 2025) and ReDirector (Park et al., 2025b), which are capable of generating 81 or more frames.

4.3 Evaluation protocol

We follow the evaluation protocol of ReDirector (Park et al., 2025b), using 50 videos from the DAVIS dataset. By applying 10 ReCamMaster (Bai et al., 2025) camera trajectories per video, we construct 500 test cases with lengths varying from tens to nearly a hundred frames. We restrict TrajectoryCrafter (Yu et al., 2025) and CogNVS (Chen et al., 2025a) to a maximum of 49 frames to prevent performance degradation; for the other methods, the evaluated frame length matches the dataset defaults.

Table 1: **Quantitative comparison on DAVIS.** Explicit-warping methods use at most 49 frames; implicit-conditioning methods use dataset-default lengths. Bold marks the best result within each group; underlining marks the second-best implicit-conditioning result.

	Explicit warping		Implicit conditioning		
Metric	CogNVS	Trajectory Crafter	ReCam Master	ReDirector	Ours
Subject consistency ↑	0.9134	**0.9145**	0.9050	<u>0.9098</u>	**0.9151**
Background consistency ↑	**0.9342**	0.9331	**0.9190**	0.9150	<u>0.9179</u>
Aesthetic quality ↑	0.4830	**0.5252**	<u>0.5146</u>	0.5141	**0.5168**
Imaging quality ↑	**0.6398**	0.6394	0.6611	<u>0.6821</u>	**0.6842**
Temporal flickering ↑	**0.9460**	0.9432	**0.9644**	0.9537	<u>0.9548</u>
Motion smoothness ↑	0.9768	**0.9800**	**0.9867**	0.9856	<u>0.9862</u>
Dyn-MEt3R ↑	0.8037	**0.8244**	0.7971	<u>0.8497</u>	**0.8573**
MEt3R ↓	0.2736	**0.2136**	0.3485	<u>0.3130</u>	**0.3077**
TransErr ↓	0.0367	**0.0293**	0.0245	<u>0.0149</u>	**0.0129**
RotErr ↓	**6.9499**	10.4340	2.3175	<u>1.4635</u>	**1.3645**

Table 2: **Quantitative reward ablation on DAVIS.** Beyond improving geometric consistency and camera accuracy, the geometry reward also improves visual quality. Bold denotes the best value and underlining the second best.

Metric	ReDirector	Video quality reward	Full reward
Subject consistency ↑	0.9098	<u>0.9130</u>	**0.9147**
Background consistency ↑	0.9150	<u>0.9174</u>	**0.9181**
Aesthetic quality ↑	0.5141	<u>0.5145</u>	**0.5163**
Imaging quality ↑	<u>0.6821</u>	0.6818	**0.6828**
Temporal flickering ↑	0.9537	**0.9549**	<u>0.9548</u>
Motion smoothness ↑	0.9856	<u>0.9861</u>	**0.9863**
Dyn-MEt3R ↑	0.8497	<u>0.8531</u>	**0.8550**
MEt3R ↓	0.3130	<u>0.3117</u>	**0.3082**
TransErr ↓	0.0149	<u>0.0147</u>	**0.0140**
RotErr ↓	<u>1.4635</u>	1.6082	**1.3895**

For our metrics, we use ViPE (Huang et al., 2025) to extract camera parameters to compute TransErr and RotErr. We also apply MEt3R (Asim et al., 2025) for input video consistency, Dyn-MEt3R (Park et al., 2025a) for geometric consistency, and VBench (Huang et al., 2024) for comprehensive aesthetic evaluation. Moreover, to evaluate complex trajectories, we compare our method with our base model (ReDirector (Park et al., 2025b)) under different camera speeds. Since large camera movements can produce consecutive featureless frames (e.g., sky, water or non-textured ground) that lead to ViPE (Huang et al., 2025) estimation failures, we perform this speed-wise comparison on a reliable subset of 40 DAVIS (Pont-Tuset et al., 2017) videos to ensure fairness, applying 10 ReCamMaster (Bai et al., 2025) camera trajectories per video.

4.4 Main Results

Table 1 demonstrates the effectiveness of our reinforcement learning framework. Compared to the baseline (Park et al., 2025b), our RL-trained model exhibits notable improvements in Geometric Consistency, Camera Accuracy, and overall video quality. The gains in Geometric Consistency and Camera Accuracy are particularly substantial, validating that the incorporation of the geometry reward successfully guides the model to follow target trajectories with higher precision. Furthermore, Table 3 presents the evaluation results across varying camera speeds. Our model consistently outperforms the baseline under more complex target trajectories, clearly illustrating the enhanced robustness across diverse trajectory conditions achieved through our RL training.

Table 3: **Quantitative comparison across camera speeds.** Each speed uses the same 40-video DAVIS subset and ten trajectories per video. Bold indicates the better value in each pair.

Camera speed	1.0		1.5		2.0	
Metric	ReDirector	Ours	ReDirector	Ours	ReDirector	Ours
Subject consistency ↑	0.9130	**0.9178**	0.8922	**0.8983**	0.8726	**0.8783**
Background consistency ↑	0.9143	**0.9174**	0.9004	**0.9035**	0.8881	**0.8911**
Aesthetic quality ↑	0.5168	**0.5191**	0.5085	**0.5111**	0.4984	**0.5017**
Imaging quality ↑	0.6755	**0.6782**	0.6713	**0.6741**	0.6624	**0.6668**
Temporal flickering ↑	0.9540	**0.9552**	0.9486	**0.9495**	0.9457	**0.9469**
Motion smoothness ↑	0.9857	**0.9862**	0.9838	**0.9844**	0.9820	**0.9828**
Dyn-MEt3R ↑	0.8555	**0.8638**	0.8368	**0.8458**	0.8249	**0.8338**
MEt3R ↓	0.3206	**0.3150**	0.3354	**0.3307**	0.3496	**0.3444**
TransErr ↓	0.0140	**0.0123**	0.0157	**0.0129**	0.0161	**0.0153**
RotErr ↓	1.3689	**1.2156**	**1.5476**	1.5491	1.9246	**1.8821**

Qualitatively, Figure 3 visualizes our results on the DAVIS (Pont-Tuset et al., 2017) dataset. Our method demonstrates superior capabilities in maintaining geometric consistency between the foreground subject and the background. Notably, under large camera motions, baseline methods such as ReCamMaster (Bai et al., 2025) and ReDirector (Park et al., 2025b) frequently suffer from severe degradation, including subject disappearance and background blurring. In contrast, our approach significantly mitigates these collapse scenarios, robustly preserving both subject and background details. We attribute this enhanced stability to our RL training paradigm, which effectively exposes the model to a broader and more complex distribution of camera trajectories.

4.5 Ablation Study

To validate the effectiveness of our proposed mechanisms, we conduct an ablation study by retraining the model under different reward configurations. All models are trained for 140 steps on 16 A800 GPUs and evaluated against the baseline (ReDirector (Park et al., 2025b)) using the identical evaluation protocol and dataset as in Table 1. Specifically, we compare the full reward formulation (incorporating both aesthetic and geometry rewards) against an ablated variant trained exclusively with the aesthetic reward. As shown in Table 2, relying solely on the aesthetic reward yields marginal overall performance improvements and even leads to a noticeable degradation in rotation accuracy. In contrast, integrating the geometry reward not only substantially enhances camera accuracy but also contributes to better visual quality and geometric consistency.

Additional CityWalk visualizations and DAVIS comparisons are provided in Appendix A.4.

5 Conclusion

In this work, we introduce Geo-Align, a reinforcement learning framework designed for camera-controlled video retake. First, we design a reward mechanism using a metric 3D evaluator to explicitly optimize how accurately the generated video follows the target camera trajectory. Second and crucially, we propose a hybrid data sampling strategy combining real videos and scaled camera trajectories from synthetic data. This effectively mitigates the field's reliance on scarce time-synchronized multi-view video data, allowing the model to train on far more diverse scenes and complex trajectories. Empirical results validate that our approach consistently enhances camera control precision, consistency, and visual quality over existing baselines.

AI USE STATEMENT

Generative AI tools were used to assist with language translation, English expression, and grammar correction. They were not used to develop the research ideas, implement the research methods, conduct experiments, derive mathematical formulations, or create scientific figures. The authors are responsible for the final content of this work, including its text, claims, and artifacts.

ETHICS STATEMENT

Our framework can support creators, animators, and filmmakers by providing more precise camera control over generative video tools. However, high-fidelity video generation can also be misused to create deepfakes or misleading visual content. We encourage the development of synthetic-media detection and watermarking mechanisms alongside these generative capabilities. The third-party assets and their licenses are listed in Appendix A.3.

REPRODUCIBILITY STATEMENT

The reward formulation and trajectory-sampling pipeline are described in Section 3. Section 4 specifies the base model, trainable components, rollout configuration, training resources, evaluation protocol, baselines, and ablation settings. Appendix A.3 lists the datasets, models, and evaluation tools used in this work, together with their sources and licenses.

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, 2026. URL https://arxiv.org/abs/2603.26599.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6034–6044, 2025.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22875–22889, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025.

Kaihua Chen, Tarasha Khurana, and Deva Ramanan. Reconstruct, inpaint, test-time finetune: Dynamic novel-view synthesis from monocular videos. *arXiv preprint arXiv:2507.12646*, 2025a.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22831–22840, 2025b.

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it – pushing vggt's limits on kilometer-scale long rgb sequences, 2026. URL https://arxiv.org/abs/2507.16443.

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, 2026. URL https://arxiv.org/abs/2601.23286.

Senyu Fei, Siyin Wang, Li Ji, Ao Li, Shiduo Zhang, Liming Liu, Jinlong Hou, Jingjing Gong, Xianzhong Zhao, and Xipeng Qiu. Srpo: Self-referential policy optimization for vision-language-action models, 2025. URL https://arxiv.org/abs/2511.15605.

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, 2026. URL https://arxiv.org/abs/2601.16214.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, Linus, Di Wang, and Jie Jiang. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again, 2025. URL https://arxiv.org/abs/2507.22058.

Hyojun Go, Byeongjun Park, Jiho Jang, Jin-Young Kim, Soonwoo Kwon, and Changick Kim. Splatflow: Multi-view rectified flow model for 3d gaussian splatting synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21524–21536, 2025.

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.

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.

Haoyang He, Jay Patrikar, Dong-Ki Kim, Max Smith, Daniel McGann, Ali-akbar Agha-mohammadi, Shayegan Omidshafiei, and Sebastian Scherer. Grndctrl: Grounding world models via self-supervised reward alignment. *arXiv preprint arXiv:2512.01952*, 2025.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2005–2015, 2025.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

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

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pp. 18–35. Springer, 2024.

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

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Orest Kupyn, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models, 2025. URL https://arxiv.org/abs/2510.21615.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer, 2025. URL https://arxiv.org/abs/2508.10893.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r, 2024. URL https://arxiv.org/abs/2406.09756.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Yiming Cheng, Miles Yang, Zhao Zhong, and Liefeng Bo. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025a.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675*, 2025b.

Zizun Li, Jianjun Zhou, Yifan Wang, Haoyu Guo, Wenzheng Chang, Yang Zhou, Haoyi Zhu, Junyi Chen, Chunhua Shen, and Tong He. Wint3r: Window-based streaming reconstruction with camera token pool. *arXiv preprint arXiv:2509.05296*, 2025c.

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.

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

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. *arXiv preprint arXiv:2505.05470*, 2025a.

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. *arXiv preprint arXiv:2501.13918*, 2025b.

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos, 2025c. URL https://arxiv.org/abs/2412.09401.

Dongyue Lu, Ao Liang, Tianxin Huang, Xiao Fu, Yuyang Zhao, Baorui Ma, Liang Pan, Wei Yin, Lingdong Kong, Wei Tsang Ooi, et al. See4d: Pose-free 4d generation via auto-regressive video inpainting. *arXiv preprint arXiv:2510.26796*, 2025.

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

Byeongjun Park, Hyojun Go, Hyelin Nam, Byung-Hoon Kim, Hyungjin Chung, and Changick Kim. Steerx: Creating any camera-free 3d and 4d scenes with geometric steering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27326–27337, 2025a.

Byeongjun Park, Byung-Hoon Kim, Hyungjin Chung, and Jong Chul Ye. Redirector: Creating any-length video retakes with rotary camera encoding. *arXiv preprint arXiv:2511.19827*, 2025b.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM SIGGRAPH 2006 Papers*, pp. 835–846, 2006.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025.

Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pp. 298–372. Springer, 1999.

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, 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, 2025a. URL https://arxiv.org/abs/2503.20314.

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*, 2025b.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory, 2024. URL https://arxiv.org/abs/2408.16061.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024a.

Weijie Wang, Xiaoxuan He, Youping Gu, Yifan Yang, Zeyu Zhang, Yefei He, Yanbo Ding, Xirui Hu, Donny Y. Chen, Zhiyuan He, Yuqing Yang, and Bohan Zhuang. World-r1: Reinforcing 3d constraints for text-to-video generation, 2026. URL https://arxiv.org/abs/2604.24764.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-Equivariant Visual Geometry Learning. *arXiv preprint arXiv:2507.13347*, 2025c.

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. *arXiv preprint arXiv:2512.02870*, 2025d.

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, 2024b.

Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, pp. 127–134. IEEE, 2013.

Fan Wu, Jiacheng Wei, Ruibo Li, Yi Xu, Junyou Li, Deheng Ye, and Guosheng Lin. Ic-world: In-context generation for shared world modeling, 2025a. URL https://arxiv.org/abs/2512.02793.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling, 2025b. URL https://arxiv.org/abs/2507.07982.

Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20406–20417, 2024.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing grpo on visual generation, 2025. URL https://arxiv.org/abs/2505.07818.

Tianyi Yan, Wencheng Han, Xia Zhou, Xueyang Zhang, Kun Zhan, Cheng zhong Xu, and Jianbing Shen. Rlgf: Reinforcement learning with geometric feedback for autonomous driving video generation, 2025. URL https://arxiv.org/abs/2509.16500.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass, 2025. URL https://arxiv.org/abs/2501.13928.

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

Tengjiao Yin, Jinglei Shi, Heng Guo, and Xi Wang. Vigor: Video geometry-oriented reward for temporal generative alignment, 2026. URL https://arxiv.org/abs/2603.16271.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 100–111, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2050–2062, 2025a.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion, 2025b. URL https://arxiv.org/abs/2410.03825.

Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views, 2026. URL https://arxiv.org/abs/2502.12138.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer, 2026. URL https://arxiv.org/abs/2507.11539.

A Additional Details and Results

A.1 Metric-Aware Trajectory Sampling Details

Specifically, for the conditioning inputs, we utilize in-the-wild CityWalk (Li et al., 2025b) videos, which encompass a diverse array of static and dynamic scenes across both indoor and outdoor environments. The source camera trajectories for these uncalibrated conditioning videos are estimated using MapAnything (Keetha et al., 2025). Conversely, to inject a rich and complex repertoire of camera motions into the model, we sample the target trajectories from the OmniWorld (Zhou et al., 2025) gaming dataset. However, drawing target trajectories directly from gaming data introduces critical optimization bottlenecks: these trajectories lack an absolute physical metric scale and frequently exhibit severe rotation.

To guarantee the physical plausibility and kinematic stability of the target camera poses during RL training, we introduce a rescaling mechanism. First, we calculate the maximum frame-to-frame translation speed v_{trans}^{max} and rotation speed v_{rot}^{max} of the raw target trajectory $\mathbf{P}_{1:N}^{tgt}$:

$$v_{trans}^{max} = \max_{i \in [1, N-1]} \|\mathbf{t}_{i+1} - \mathbf{t}_i\|_2, \tag{10}$$

$$v_{rot}^{max} = \max_{i \in [1, N-1]} \|\log(\mathbf{R}_i^\top \mathbf{R}_{i+1})^\vee\|_2, \tag{11}$$

where $\mathbf{t}_i$ and $\mathbf{R}_i$ denote the translation vector and rotation matrix at frame i, respectively, and $(\cdot)^\vee$ maps the skew-symmetric matrix in the Lie algebra $\mathfrak{so}(3)$ to its corresponding rotation vector. To ensure the trajectory speeds fall within a reasonable physical bound while maintaining data diversity, we sample target maximum speeds, τ_{trans} and τ_{rot}, from Truncated Gaussian Distributions:

$$\tau_{trans} \sim \mathcal{N}_{trunc}(\mu_t, \sigma_t^2, [a_t, b_t]), \tag{12}$$

$$\tau_{rot} \sim \mathcal{N}_{trunc}(\mu_r, \sigma_r^2, [a_r, b_r]), \tag{13}$$

where $[a_t, b_t]$ and $[a_r, b_r]$ define the strict physical bounds for translation and rotation speeds, concentrating the sampling probability around natural human-walking or steady-cam speeds. Finally, we compute the rescaling factors for translation and rotation, denoted as s_{trans} and s_{rot} respectively:

$$s_{trans} = \frac{\tau_{trans}}{v_{trans}^{max} + \epsilon}, \quad s_{rot} = \frac{\tau_{rot}}{v_{rot}^{max} + \epsilon}, \tag{14}$$

where ϵ is a small constant to prevent division by zero. The target trajectory is then uniformly rescaled to yield the modified physical-aware trajectory $\tilde{\mathbf{P}}_{1:N}^{tgt}$:

$$\tilde{\mathbf{t}}_i = s_{trans} \mathbf{t}_i, \tag{15}$$

$$\tilde{\mathbf{R}}_i = \exp\left(s_{rot} \log(\mathbf{R}_i)\right). \tag{16}$$

This rescaling protocol effectively eliminates unnatural camera jumps and aligns the synthetic gaming trajectories with real-world metric scales, significantly stabilizing the RL optimization landscape.

A.2 Limitations

Our reinforcement learning approach improves the accuracy of camera trajectory adherence in generated videos while simultaneously enhancing overall video quality. However, the model remains susceptible to failure when faced with excessively fast rotations, large translations, or large foreground objects close to the camera. Furthermore, inputs dominated by dynamic objects frequently lead to artifacts, such as the flickering or vanishing of those dynamic elements. Thirdly, the RL training process is highly time-consuming because it requires sampling the model to generate multiple complete videos for each batch, and the video generation process itself is inherently slow. Therefore, accelerating the RL training process remains a compelling direction for future exploration.

A.3 Assets and Licenses

We summarize the assets used in our research, including their licenses and accessibility, in Table 4. All assets are used in accordance with their respective terms.

Table 4: Summary of used assets (datasets, models, and code).

Asset	Citation	License	Link/Source
ReCamMaster	(Bai et al., 2025)	MIT License	Repository
ReDirector	(Park et al., 2025b)	Apache License 2.0	Repository
TrajectoryCrafter	(Yu et al., 2025)	Apache License 2.0	Repository
MixGRPO	(Li et al., 2025a)	Non-Commercial License v1.1.1	Repository
MapAnything	(Keetha et al., 2025)	Apache License 2.0	Repository
OmniWorld	(Zhou et al., 2025)	BY-NC-SA 4.0	Repository
CityWalk	(Li et al., 2025b)	No commercial license	Repository
ViPE	(Huang et al., 2025)	Apache License 2.0	Repository
MEt3R	(Asim et al., 2025)	MIT License	Repository
Dyn-MEt3R	(Park et al., 2025a)	CC BY 4.0	Repository
VBench	(Huang et al., 2024)	Apache License 2.0	Repository

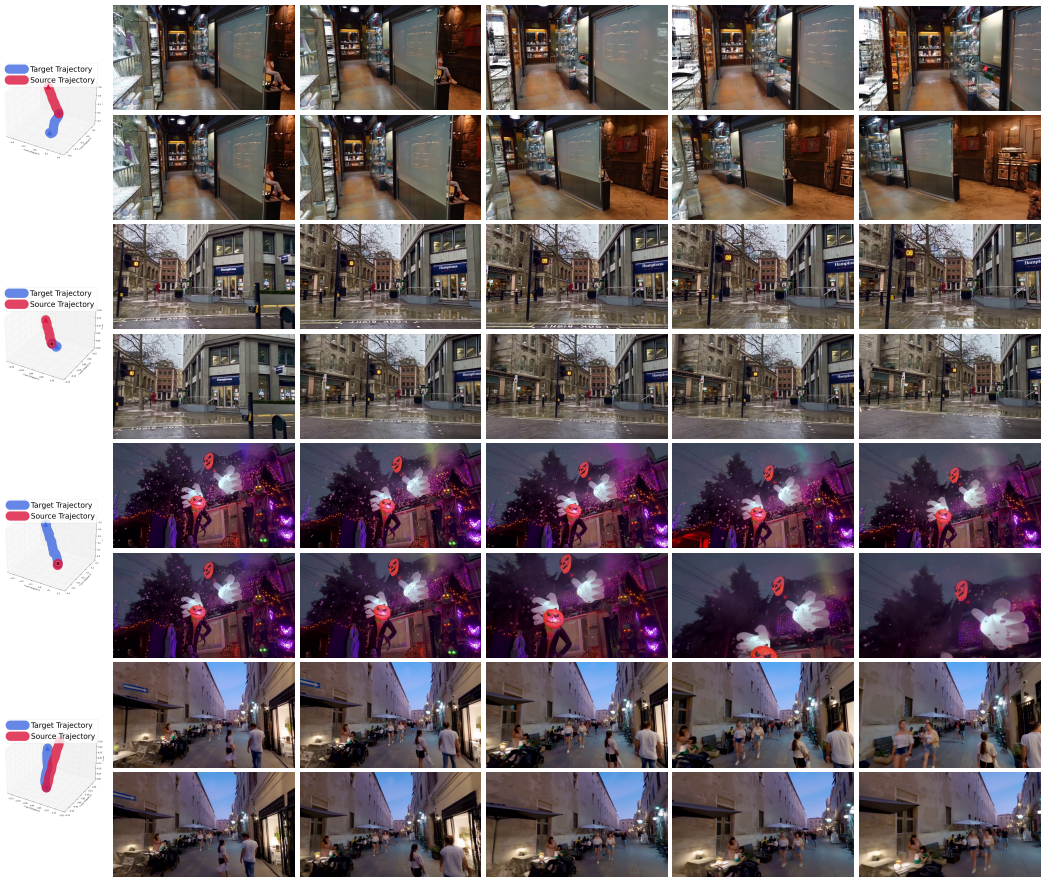

Figure 4: **More visualization results on CityWalk (Li et al., 2025b) dataset.** For each example, the top row illustrates the input video, whereas the bottom row visualizes our results following the target trajectory.

A.4 Additional Qualitative Results

Figure 4 shows additional CityWalk examples. Figure 5 provides further DAVIS comparisons, and Figure 6 illustrates failure cases.



Figure 5: **More qualitative comparisons on the DAVIS (Pont-Tuset et al., 2017) dataset.** For each example, the top row illustrates the input video, while the second and third rows present the results of ReDirector (Park et al., 2025b) and our model, respectively.



Figure 6: **Failure cases.** The model remains susceptible to failure when faced with excessively fast rotations, large translations, or large foreground objects close to the camera.