TEST-TIME GEOMETRIC ALIGNMENT FOR CAMERA-CONTROLLED VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Camera-controlled video generative models produce visually compelling results yet routinely violate the geometry of the scenes they depict. Some existing remedies inject 3D priors at training time, requiring generator-specific optimization. We instead treat geometric consistency as an inference-time alignment problem: scene observations and generated videos should depict the same underlying geometry, and embeddings from geometric foundation models provide a proxy for this agreement. We introduce **GeoReward**, a training-free, differentiable reward defined by geometry-embedding similarity between scene references and generated videos. The reward supports candidate search and gradient-based guidance of denoising trajectories, enabling generation quality to improve as test-time compute scales. Empirically, GeoReward improves geometric consistency in both image-conditioned and video-conditioned generation across different base generators and diverse static and dynamic scenes. Video comparisons are available at `https://georeward-review.github.io`.

1 INTRODUCTION

Video generation is increasingly asked to behave like a renderer of a coherent world rather than a producer of plausible pixels. World models, embodied simulation, and controllable content creation all require that a generated video depict a single, persistent 3D scene, and camera-controlled generation makes this requirement explicit: given observations of a scene and a target camera trajectory, models such as ReCamMaster (Bai et al., 2025), GEN3C (Ren et al., 2025), and TrajectoryCrafter (Yu et al., 2025) must re-render *the same scene* from a different viewpoint—geometric consistency is a core correctness criterion of the task.

Current video generative models do not consistently meet this criterion. Analyses of video diffusion representations identify limitations in their geometric structure (Wu et al., 2026); dedicated benchmarks measure 3D-consistency failures across generators (Duan et al., 2025); and GeCo (Gu et al., 2025) detects depth-reprojection and deformation errors in generated videos. The failure is not marginal in our own measurements. We evaluate camera-motion types represented during training (*e.g.*, pans, tilts, and arcs) as well as out-of-distribution (OOD) loop trajectories (*e.g.*, lemniscate and arc out-and-back), where OOD trajectories test both generalization beyond the standard camera presets and the ability to preserve scene structure when returning to previously visited viewpoints. On natural videos, the baseline's mean fused GeCo error on these loops is 4.65 times that on standard trajectories (Table 3).

Training-time approaches address geometric inconsistency by updating the generator. Geometry Forcing aligns diffusion features with those of a geometric foundation model during training (Wu et al., 2026); World-R1 post-trains generators with reinforcement learning against 3D-aware rewards (Wang et al., 2026a). Both require resource-intensive generator-specific optimization. Yet geometric consistency need not be equally desirable for every request: faithful scene rendering benefits from it, whereas artistic videos may intentionally depict surreal spatial transitions or object morphing. Inference-time alignment offers the flexibility to apply a geometric objective when needed and omit it otherwise, without changing the generator's weights. Other preferences, such as aesthetic style, motion intensity, and reference-subject fidelity, similarly motivate separating a generative model from application-specific inference-time objectives.

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.

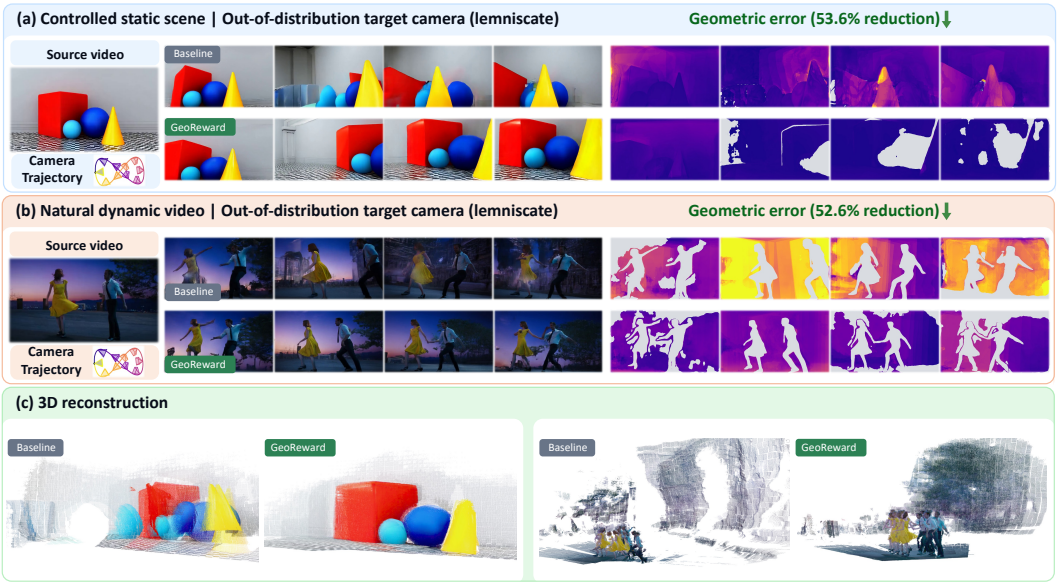


Figure 1: GeoReward enables **training-free** improvements in 3D awareness for camera-controlled video generation. Compared with baseline (Bai et al., 2025), GeoReward produces more visually and geometrically consistent generations across diverse scenes and subjects (a, b), even under out-of-distribution target camera trajectories. GeCo error maps (Gu et al., 2025) reveal reduced depth inconsistencies and structural distortions, supporting more accurate 3D reconstruction in (c).

We therefore treat geometric consistency as an inference-time alignment problem. A trained generator defines a distribution over many plausible denoisings of the same input, with different levels of geometric fidelity. External rewards can help recover preferred outputs by searching or steering its sampling process (Ma et al., 2025; Singhal et al., 2025; Yuan et al., 2026). The central question is how to obtain a useful geometric reward without training a task-specific reward model.

Our premise is that features from pretrained geometric foundation models can provide this feedback. The source video and the camera-redirected target video should depict the same underlying scene despite their different viewpoints. Trained to recover 3D structure from images, these models offer representations for comparing source video frames and generated frames across viewpoints without requiring pixel alignment. We use their embedding similarity as a differentiable proxy for scene consistency. This formulation requires neither ground-truth 3D nor an additional learned reward head; its effectiveness as a geometric signal is evaluated through consistency measurements and comparisons with generic visual embeddings.

We instantiate this idea as **GeoReward**, a generator-agnostic framework that uses geometry-derived rewards for candidate search, gradient guidance, and their combination. We evaluate it on a DFoT-based image-to-video model (Song et al., 2025) and ReCamMaster (Bai et al., 2025), a Wan2.1-based video-to-video model (Wan Team, 2025), keeping both generators frozen. For image-to-video generation, we deliberately use a Geometry Forcing checkpoint that already incorporates geometric feature alignment during training (Wu et al., 2026). Improvements on this checkpoint demonstrate that inference-time alignment can provide additional value even after geometry-aware training.

Our contributions are as follows:

- We show that geometry models are useful for improving video generation. To this end, we introduce an effective, training-free, and differentiable geometric consistency reward, **GeoReward**, using geometric foundation-model features as the critic.
- We demonstrate improvements on two frozen generators. On RealEstate10K (Zhou et al., 2018), GeoReward obtains additional fidelity gains (PSNR, LPIPS, SSIM) on a Geometry Forcing checkpoint (Wu et al., 2026) (Table 1). On ReCamMaster (Bai et al., 2025) video-to-video generation, GeoReward search reduces mean per-trajectory fused GeCo error by 13.2–16.4% on controlled scenes and 6.2–10.3% on natural videos (Tables 2 and 3).

- We highlight the scaling behavior of GeoReward: performance improves as the candidate pool grows. We further show that search composes with efficiency techniques that reduce its cost.
- We ablate three geometry models as critics—π^3 (Wang et al., 2026b), VGGT (Wang et al., 2025), and Depth Anything 3 (Lin et al., 2025)—and show that each improves geometric consistency on RealEstate10K (Zhou et al., 2018), with π^3 performing best.

2 Related Work

Camera-controlled video generation. A growing line of work conditions video generators on camera trajectories, either injecting pose embeddings into pretrained text-to-video models (He et al., 2025; Wang et al., 2024) or using source-video conditioning (Bai et al., 2025; Park et al., 2026), explicit 3D caches (Ren et al., 2025), decoupled view transformation (Yu et al., 2025), infinite-homography warps (Kim et al., 2026), or completed 4D reconstructions (Cao et al., 2026). These methods develop the *conditioning mechanism*. Our method is complementary: it retains the evaluated generators' conditioning mechanisms and improves their outputs through inference-time sampling and selection.

Geometric consistency of generated video. The geometric deficiencies of video diffusion models are documented by representation analyses (Wu et al., 2026), unified benchmarks (Duan et al., 2025), and consistency metrics that expose depth-reprojection and deformation errors (Gu et al., 2025). Training-time approaches include Geometry Forcing, which aligns intermediate diffusion features with VGGT features (Wu et al., 2026), and World-R1, which post-trains generators against 3D-aware rewards (Wang et al., 2026a); geometry foundation models have also been used to enhance downstream perception (Gao et al., 2026). Inference-time approaches include SteerX, which uses pose-free reconstruction models to supply geometric steering rewards (Park et al., 2025). VG-GRPO learns a latent 4D reward for GRPO post-training and also evaluates test-time guidance (An et al., 2026). VIGOR (Yin et al., 2026) builds a geometric reward from VGGT, scoring a generation's *internal* multi-view consistency through point tracking and reprojection, and applies it to preference post-training, best-of-N selection for bidirectional generators, and streaming search. GeoReward anchors inference-time feedback to external observations of the intended scene through frozen reconstruction-model embeddings, without explicit tracking or reprojection for scoring. In source-referenced camera control, comparing time-aligned observations across viewpoints avoids imposing cross-frame scene rigidity, making the formulation applicable to dynamic source videos.

Inference-time alignment of generative models. Steering a frozen diffusion model with an external reward is well studied for images: best-of-N selection and search over seeds (Ma et al., 2025), particle-based methods such as Feynman–Kac steering (Singhal et al., 2025) and value-based resampling (Li et al., 2025), and gradient-based guidance that perturbs the score with the reward gradient (Dhariwal & Nichol, 2021; Bansal et al., 2023). For video, Video-T1 uses a vision–language verifier for temporal branching (Liu et al., 2025), and WMReward re-purposes the predictive error of a latent world model (V-JEPA 2) as a physics-plausibility reward for search and guidance (Yuan et al., 2026; Assran et al., 2025). We adopt the same alignment machinery but target a different failure mode with a different class of reward model: WMReward uses a *predictive* model of dynamics, while GeoReward compares generation and scene references through *reconstruction*-model embeddings. The feature comparison replaces prediction error as the reward, with viewpoint tolerance assessed empirically.

3 Method: Test-Time Geometric Scaling

We consider conditional video generation. A frozen generator defines a distribution $p_\theta(\mathbf{y} \mid \mathbf{s}, c)$ over videos $\mathbf{y} = (y_1, \ldots, y_T)$, where the conditioning $\mathbf{s}$ is a context image or a source video, and c is a target camera trajectory. We denote the scene references available to the reward by $\mathbf{q}$, separately from the generator's conditioning. Because the sampler is stochastic, repeated sampling yields plausible outputs of varying geometric quality; our goal is to spend test-time compute to obtain more consistent ones. Following the inference-time alignment framing (Yuan et al., 2026; Singhal et al., 2025), we target a reward-tilted distribution

$$p^*(\mathbf{y} \mid \mathbf{s}, c, \mathbf{q}) \;\propto\; w(r(\mathbf{y}; \mathbf{q}))\; p_\theta(\mathbf{y} \mid \mathbf{s}, c), \tag{1}$$

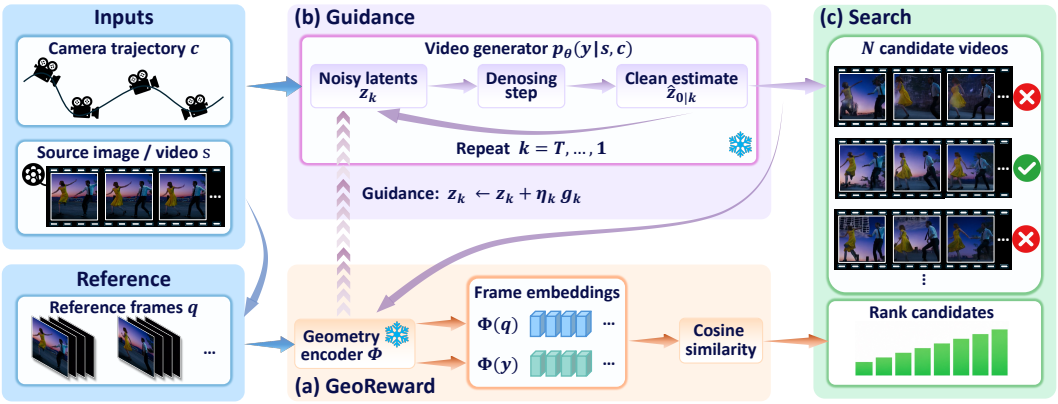


Figure 2: **Method overview.** GeoReward leverages a frozen geometry foundation model to improve geometric consistency in camera-controlled video generation. (a) GeoReward: We encode source-video reference frames and generated frames with a shared frozen geometry encoder and compute their cosine similarity to obtain a scene-referenced reward. (b) Guidance: Reward gradients steer the denoising trajectory by updating the noisy latents based on clean-sample estimates. (c) Search: We rank N candidate videos by their reward and select the highest-scoring output. Search can also rank guided candidates, combining both mechanisms.

where $r(\mathbf{y}; \mathbf{q})$ is a proxy for consistency with the scene references and $w(\cdot)$ is a nonnegative weighting function. Selection requires a stochastic sampler and reference frames; gradient guidance additionally requires a differentiable path from the sampling state to the reward. Neither scheme updates generator or critic weights.

3.1 GeoReward: A Geometric-Consistency Reward

Why reconstruction-model embeddings. The reference and the generation should depict the same underlying scene, but may differ in viewpoint or time. Direct pixel comparisons can penalize legitimate viewpoint changes, while reprojection-based comparisons require estimated geometry and correspondences. We seek a representation that tolerates viewpoint changes while remaining sensitive to scene identity and structure. Feed-forward geometry reconstruction models—π^3 (Wang et al., 2026b), VGGT (Wang et al., 2025), and Depth Anything 3 (Lin et al., 2025)—provide pretrained features for this comparison. Their extraction paths are differentiable and require no ground-truth 3D.

Reward definition. Let $\Phi(\mathbf{v})_f \in \mathbb{R}^D$ denote the embedding of frame f in video $\mathbf{v}$, obtained by mean-pooling patch tokens at a fixed model layer and applying ℓ_2 normalization. π^3 uses its image encoder independently on each frame; VGGT and DA3 use features that also depend on other frames within the processed chunk. The extraction layers and preprocessing are specified in Appendix A.5. Given corresponding reference and generated frame indices $\mathcal{F}$ at aligned timesteps, the base reward is their mean cosine similarity:

$$r(\mathbf{y}; \mathbf{q}) \;=\; \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \langle \Phi(\mathbf{q})_f,\, \Phi(\mathbf{y})_f \rangle \,. \tag{2}$$

Equation 2 rewards feature agreement; it is not a direct reconstruction error or a guarantee of geometric consistency.

Score aggregation. One design choice matters in practice—*aggregation*: besides the time-aligned form of Eq. 2 we consider segment-level variants that compare each reference frame against a window of generated frames and aggregate by max, top-k mean, or softmax weighting. An optional motion term compares directions of feature change between consecutive reference segments (Appendix A.1). Section 4.1 specifies the reference protocols; Appendix A.2 ablates aggregation.

3.2 SAMPLING SCHEMES

We use three schemes for the alignment objective in Eq. 1, mirroring gradient-free, gradient-based, and combined inference-time alignment (Yuan et al., 2026).

Best-of-N search. We draw N candidates $\{\mathbf{y}^{(i)}\}_{i=1}^{N}$ from the frozen generator with independent noise seeds and keep

$$i^* = \arg\max_{i\in\{1,\dots,N\}} r\big(\mathbf{y}^{(i)};\mathbf{q}\big), \qquad \mathbf{y}^* = \mathbf{y}^{(i^*)}. \tag{3}$$

For independent candidates with a continuous reward distribution, selection induces the weight $w(r) = F(r)^{N-1}$ in Eq. 1, where F is the cumulative distribution function of the reward under the generator (Beirami et al., 2025; Gui et al., 2024). Our selectors retain the first candidate in evaluation order when scores tie. Search permits parallel candidate generation, requires only forward evaluations of the critic, and caches scores so that every prefix budget $N' \le N$ can be evaluated without further generation.

Gradient guidance. Search cannot fix a candidate pool in which every sample is geometrically poor; guidance instead steers trajectories during sampling toward higher reward, motivated by the exponential weighting $w(r) = \exp(\lambda r)$ and classifier guidance (Dhariwal & Nichol, 2021). At denoising step k with sampling state $\mathbf{z}_k$, we form the clean-sample estimate $\hat{\mathbf{z}}_{0|k}$, decode the frames touched by the reward, and compute its gradient:

$$\mathbf{g}_k = \nabla_{\mathbf{z}_k} r\big(\mathcal{D}(\hat{\mathbf{z}}_{0|k});\mathbf{q}\big), \tag{4}$$

where $\mathcal{D}$ is the latent decoder, or the identity for a pixel-space model. The sampling state then takes an ascent step on the reward,

$$\mathbf{z}_k \;\leftarrow\; \mathbf{z}_k + \eta_k\,\mathbf{g}_k, \tag{5}$$

where η_k is a per-step scale. The clean estimate is $(\mathbf{z}_k - \sqrt{1-\bar{\alpha}_k}\,\epsilon_\theta)/\sqrt{\bar{\alpha}_k}$ for DDIM, using the denoiser form of Tweedie's identity (Efron, 2011), and $\mathbf{z}_k - \sigma_k \mathbf{v}_\theta$ for the flow-matching sampler. The gradient is incorporated through the sampler's update, with sampler-specific scaling and normalization. The evaluated implementations use either full-denoiser backpropagation or a detached-denoiser approximation; their exact gradient paths, update parameterizations, and schedules are given in Table 8. We restrict the reward to a sparse set of guidance frames to bound memory. The guidance implementation builds on the frame-level guidance of Jang et al. (2026).

Search + guidance. The two compose naturally: generate all N candidates with guidance, then select the best by Eq. 3. Guidance steers individual candidates toward higher reward, and selection retains the highest-scoring completed video. Their relative performance depends on the budget and setting (Sections 4.2 and 4.3).

Adaptive compute and tree search for reducing cost. Both schemes compose with standard test-time efficiency techniques. Progressive search draws candidates sequentially and stops once the observed reward exceeds a threshold, so compute concentrates on inputs whose early candidates score poorly. Denoising-time tree search exploits the fact that the reward can score the clean estimate $\hat{\mathbf{z}}_{0|k}$ of Eq. 4 without gradients at intermediate steps: candidates are pruned by successive halving during sampling, so a larger candidate pool is explored at the denoiser budget of plain best-of-N. Section 4.5 evaluates both extensions.

4 EXPERIMENTS

We evaluate GeoReward on static scenes (Section 4.2) and then on more challenging dynamic videos (Section 4.3). We next examine the choice of reward features (Section 4.4). Lastly, we evaluate compute cost and integration with efficient sampling and search (Section 4.5).

4.1 EXPERIMENTAL SETUP

Implementation details. Our pipeline is training-free and uses off-the-shelf pretrained generators and critics, with all model weights frozen. For static scene evaluation, we use the Geometry Forcing

checkpoint (Wu et al., 2026) of the Diffusion Forcing Transformer (DFoT) (Song et al., 2025), generating 256 frames at 256×256 from one context image, pretrained on the RealEstate10K dataset. For dynamic scene evaluation, we adopt ReCamMaster (Bai et al., 2025), based on Wan2.1-T2V-1.3B (Wan Team, 2025), which generates 81 frames at 832×480 conditioned on an 81-frame source video. ReCamMaster works on both static and dynamic scenes. ReCamMaster uses ten standard camera presets from motion families represented during training: pan right/left, tilt up/down, zoom in/out, translate up/down with rotation, and arc left/right with rotation. We also evaluate two harder out-of-distribution (OOD) loop-closure trajectories: lemniscate and arc out-and-back.

The main comparisons use $N{=}8$ candidates against an unguided $N{=}1$ baseline. Geometry Forcing and both ReCamMaster datasets use the same two disjoint seed groups, with results averaged across groups to assess whether gains persist across different random samples. Appendices A.1 and B.1 specify the reference frames, samplers, and scoring configurations for Geometry Forcing and ReCamMaster, respectively. All search and guidance experiments for both models fit within a single NVIDIA A100 with 80 GB memory.

Evaluation datasets. We evaluate Geometry Forcing on 100 validation clips from RealEstate10K (Zhou et al., 2018), a dataset of static real-estate videos with estimated camera trajectories. We evaluate ReCamMaster on both a synthetic, controlled static set and a natural, dynamic set. The natural-video set contains 120 source–trajectory cases: ten source videos selected from TV and movie clips, each re-rendered along twelve target trajectories. These videos include moving subjects, occlusions, cinematic camera motion, and view-dependent appearance changes, making preservation of scene structure challenging. The controlled set also contains 120 source–trajectory cases, combining ten Wan2.1-generated source videos with the same twelve trajectories. These sources depict approximately static scenes with geometric primitives, textured floors, and architectural structures, reducing the contribution of independent object motion to the consistency evaluation (Appendix B.3).

Evaluation metrics. On RealEstate10K, we report FID (Heusel et al., 2017), PSNR, LPIPS (Zhang et al., 2018), and SSIM for visual fidelity. We also report the recorded revisit error (RVE) (Xiao et al., 2025). For geometric consistency, we report GeCo depth, motion, and fused errors (GeCo_d, GeCo_m, and GeCo_f) (Gu et al., 2025) on RealEstate10K and both ReCamMaster sets. GeCo evaluates videos of static scenes using depth-reprojection errors and discrepancies between estimated optical flow and camera-induced rigid flow (Gu et al., 2025; Zhang et al., 2025); lower depth and motion errors indicate fewer depth inconsistencies and non-rigid distortions, respectively, while the fused error summarizes both. We apply SAM 2 (Ravi et al., 2025) static-region masks to natural videos and evaluate full frames on controlled scenes. To assess visual quality, we report frame–source CLIP similarity (CLIP-V) and VBench (Huang et al., 2024) subject consistency (SC), background consistency (BC), aesthetic quality (AQ), and imaging quality (IQ). We also report ViPE-estimated rotational absolute trajectory error (ATE) (Huang et al., 2025) to assess camera following (Appendix B.4). Unless specified otherwise, GeCo gains are mean per-trajectory percentage reductions relative to the unguided $N{=}1$ baseline, after averaging sources and seed groups.



Figure 3: **RealEstate10K: selection after GeoReward.** Two examples, each showing ground truth (top), the single-sample baseline Geometry Forcing (middle), and π^3 best-of-8 selection (bottom).

4.2 Results on Static Scenes

Table 1: **Quantitative comparison on the RealEstate10K dataset: additional gains after Geometry Forcing are geometry-model agnostic.** π^3 and VGGT use eight references with top-k+motion aggregation; DA3 uses four with segment-max. BoN: best-of-8 search; ∇: guidance. GeCo errors $\times 10^{-3}$. Bold/underlining denotes the best/second-best displayed values.

Method	FID↓	PSNR↑	LPIPS↓	SSIM↑	GeCo_f ↓	GeCo_d ↓	GeCo_m ↓	RVE↓
Geometry Forcing	49.8	11.85	0.513	0.388	4.45	5.35	3.48	230.9
+GeoReward-VGGT (BoN)	50.9	11.89	0.512	0.389	4.36	<u>5.18</u>	3.46	234.4
+GeoReward-DA3 (BoN)	49.3	11.92	<u>0.511</u>	0.391	4.31	<u>5.18</u>	<u>3.41</u>	230.9
+GeoReward-π^3 (BoN)	**47.5**	<u>12.05</u>	**0.501**	**0.399**	**4.26**	**5.10**	<u>3.41</u>	**228.0**
+GeoReward-π^3 (BoN+∇)	<u>48.0</u>	**12.07**	**0.501**	<u>0.398</u>	<u>4.30</u>	5.21	**3.40**	<u>228.9</u>

Table 2: **(a) ReCamMaster on controlled static scenes.** BoN: best-of-8; ∇: guidance. GeoReward with both BoN and BoN+∇ improves the performance. BoN+∇ strengthens gains on harder OOD trajectories. GeCo errors $\times 10^{-3}$. Bold/underlining denotes the best/second-best displayed values.

Trajectory	Method	Geometric consistency				Visual quality				
		GeCo_d ↓	GeCo_m ↓	GeCo_f ↓	Gain (%)	CLIP-V↑	SC↑	BC↑	AQ↑	IQ↑
Standard (10)	ReCamMaster	3.32	2.14	2.70	—	96.1	96.8	95.8	45.5	66.7
	+GeoReward-π^3 (BoN)	**2.83**	**1.84**	**2.33**	**+13.2**	**97.3**	**98.1**	**96.6**	<u>46.2</u>	**67.3**
	+GeoReward-π^3 (BoN+∇)	<u>3.25</u>	<u>1.93</u>	<u>2.53</u>	<u>+5.3</u>	<u>96.6</u>	<u>97.4</u>	<u>96.3</u>	**46.4**	<u>66.8</u>
OOD (2 loops)	ReCamMaster	9.59	16.54	15.68	—	92.9	91.0	93.5	<u>43.3</u>	64.6
	+GeoReward-π^3 (BoN)	<u>8.23</u>	<u>13.19</u>	<u>12.74</u>	<u>+16.4</u>	<u>94.5</u>	<u>93.5</u>	<u>94.1</u>	**44.7**	<u>66.2</u>
	+GeoReward-π^3 (BoN+∇)	**7.78**	**12.02**	**11.27**	**+20.9**	**94.9**	**93.8**	**94.4**	**44.7**	**66.7**

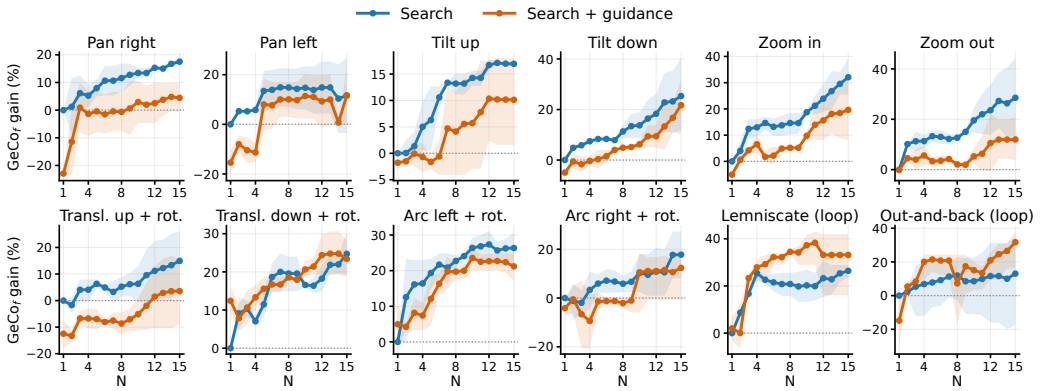


Figure 4: **Per-trajectory scaling on controlled scenes.** Fused GeCo error reduction relative to unguided baseline $N{=}1$. Both search and guidance demonstrate scaling trends as the search budget increases. Search improves all 12 trajectories, and **guidance further improves both OOD loops**.

Additional gains on RealEstate10K after geometry-aware training. Geometry Forcing (Wu et al., 2026) already aligns intermediate diffusion features with geometry-model features during training. GeoReward further improves this fixed checkpoint: when used as selection critics, VGGT, DA3, and π^3 all improve average PSNR, LPIPS, and SSIM (Table 1). Among the three critics, π^3 provides the strongest overall fidelity, improving FID from 49.8 to 47.5, PSNR from 11.85 to 12.05, LPIPS from 0.513 to 0.501, and SSIM from 0.388 to 0.399. Figure 3 shows qualitative comparisons on RealEstate10K, where GeoReward produces higher-fidelity results than the baseline.

Search consistently improves static-scene generation. Best-of-N search provides a simple default across the two generators. In addition to the fidelity gains on RealEstate10K, search reduces the mean per-trajectory fused GeCo error by 13.2% on standard controlled trajectories and 16.4% on controlled OOD loops (Table 2(a)). Both depth and motion errors decrease, while the reported visual-quality scores improve over ReCamMaster.

Guidance strengthens gains on harder camera trajectories. Controlled scenes make geometry under camera motion easier to assess by reducing independent object motion, although unintended motion and appearance changes remain possible confounds. On the OOD loops, guidance+search increases the mean per-trajectory GeCo reduction from 16.4% to 20.9%. Figure 4 shows that larger candidate pools provide further gains: as we increase the sampling candidate number from $N = 1$ to N=15, search keeps improving all twelve trajectories, and guidance further increases the gain on both OOD loop trajectories.

4.3 Results on Dynamic Scenes

Table 3: **(b) ReCamMaster on natural dynamic videos.** BoN: best-of-8; ∇: guidance. GeoReward with both BoN and BoN+∇ improves the performance. GeCo errors $\times 10^{-3}$. Bold/underlining denotes the best/second-best displayed values.

Trajectory	Method	Geometric consistency				Visual quality				
		$\text{GeCo}_d \downarrow$	$\text{GeCo}_m \downarrow$	$\text{GeCo}_f \downarrow$	Gain (%)	CLIP-V↑	SC↑	BC↑	AQ↑	IQ↑
Standard (10)	ReCamMaster	6.40	2.64	4.24	—	96.4	91.9	<u>92.8</u>	57.6	52.9
	+GeoReward-π^3 (BoN)	<u>6.22</u>	<u>2.27</u>	<u>3.93</u>	<u>+6.2</u>	**96.9**	**92.5**	**93.0**	**58.2**	<u>53.4</u>
	+GeoReward-π^3 (BoN+∇)	**6.14**	**2.22**	**3.88**	**+6.9**	<u>96.8</u>	<u>92.4</u>	**93.0**	<u>58.0</u>	**54.1**
OOD (2 loops)	ReCamMaster	10.46	20.41	19.71	—	<u>94.7</u>	86.3	**90.5**	<u>55.3</u>	50.5
	+GeoReward-π^3 (BoN)	**10.02**	**18.11**	**17.65**	**+10.3**	**95.0**	**87.0**	<u>90.4</u>	**55.4**	<u>50.9</u>
	+GeoReward-π^3 (BoN+∇)	<u>10.22</u>	<u>19.84</u>	<u>19.21</u>	<u>+3.6</u>	<u>94.7</u>	<u>86.7</u>	<u>90.4</u>	55.0	**51.1**

Geometry improvements extend to natural videos. GeoReward also improves geometric consistency on the more challenging natural-video set (Table 3(b)). Search reduces mean per-trajectory fused GeCo error by 6.2% on standard trajectories and 10.3% on OOD loops; guidance+search also improves both groups. Figure 1 includes a natural-video comparison under an OOD camera trajectory. Independent subject motion, occlusions, and changes in visibility and appearance make these videos harder to evaluate and align. Guidance gains are less uniform here than on controlled scenes, motivating future rewards that better distinguish legitimate scene dynamics from geometric distortions.

Geometry gains accompany improvements in visual quality. Table 2(a) and Table 3(b) show that both search and guidance+search improve subject consistency and imaging quality in the standard and OOD groups. GeoReward gain is not obtained by degrading visual quality.

4.4 Ablation Study

Geometry-aware features provide a stronger selection signal. We compare π^3 with DINOv2 (Oquab et al., 2024) and CLIP (Radford et al., 2021) to test whether geometry-aware features improve selection over generic visual similarity. All selectors rank the same eight candidates using the same 81 source-frame correspondences and are evaluated with GeCo. π^3 achieves the best pooled mean-error reduction of 8.9%, compared with 6.0% for DINOv2 and 7.5% for CLIP (Table 4). Generic visual embeddings thus also provide useful selection signals, while π^3 yields the strongest result among the tested configurations. Implementation details are given in Appendix B.2.

Table 4: **Matched reward-feature ablation.** Fused GeCo error ($\times 10^{-3}$, $\downarrow$) on the natural set.

Selector	$\text{GeCo}_f \downarrow$
ReCamMaster	6.82
π^3	**6.21** (**+8.9%**)
DINOv2 ViT-B/14	6.41 (+6.0%)
CLIP ViT-B/16	6.30 (+7.5%)

Table 5: **Denoising-time tree search at comparable denoiser work (≈ same budget).** Fused GeCo error ($\times 10^{-3}$, $\downarrow$) on controlled scenes 1–5, 60 videos.

Camera	N=1	Best-of-8	Tree-20
Standard	2.57	2.22 (+13.5%)	**1.89 (+26.6%)**
Lemniscate	27.22	**20.84 (+23.4%)**	21.04 (+22.7%)
Arc out-and-back	7.12	6.48 (+9.0%)	**5.82 (+18.3%)**
All twelve	5.00	4.13 (+17.5%)	**3.81** (**+23.9%**)

4.5 Compute Cost and Integration with Efficient Search

Table 6: **Computation cost.** For each model, the first row shows absolute values; subsequent rows show overhead multipliers. All rows use one A100.

Method	Time	Memory	N
Geometry Forcing	633.1 ± 1.1 s	33.0 GB	1
+ GeoReward (BoN)	$\times N$	$\times 1.00$	8
+ GeoReward (∇)	$\times 1.10$	$\times 0.77$	1
+ GeoReward (BoN+∇)	$\times 1.10N$	$\times 0.77$	8
ReCamMaster	596.2 ± 1.6 s	6.3 GB	1
+ GeoReward (BoN)	$\times N$	$\times 1.00$	8
+ GeoReward (∇)	$\times 2.76$	$\times 3.28$	1
+ GeoReward (BoN+∇)	$\times 2.76N$	$\times 3.28$	8

Table 7: **Progressive-search threshold sweep.** Geometry Forcing on RealEstate10K. "all" denotes improvement in all metrics over $N{=}1$.

Policy	avg N	FID↓	PSNR↑	LPIPS↓	SSIM↑	all
Fixed $N{=}1$	1.0	49.2	11.85	0.514	0.387	
$\tau_{\mathrm{r}}{=}0.9950$	3.1	48.5	11.88	0.511	0.389	✓
$\tau_{\mathrm{r}}{=}0.9960$	3.5	48.3	11.86	0.511	0.390	✓
$\tau_{\mathrm{r}}{=}0.9965$	3.8	48.0	11.87	0.509	0.391	✓
$\tau_{\mathrm{r}}{=}0.9970$	4.3	48.0	11.90	0.509	0.392	✓
$\tau_{\mathrm{r}}{=}0.9975$	4.9	47.8	11.91	0.508	0.393	✓
$\tau_{\mathrm{r}}{=}0.9980$	5.9	47.7	11.92	0.507	0.393	✓
Fixed $N{=}8$	8.0	47.6	11.98	0.505	0.395	✓

Runtime and memory. Inference-time alignment trades additional computation for improved performance. Table 6 reports the overhead induced by GeoReward on both generators. GeoReward (BoN) parallelizes as N vanilla sampling runs under the per-GPU memory bound of the baseline, so its aggregate footprint scales linearly with the number of candidates while wall-clock time can remain that of a single run; reward scoring adds under 0.5%. Guidance introduces extra computation from gradient backpropagation, proportional to the number of guided denoising steps and decoded guidance frames: the detached gradient path on Geometry Forcing keeps both overheads small, whereas on ReCamMaster the reward backpropagates through the full DiT. In general, GeoReward can be adapted to the available computation budget with appropriate sampling methods—stopping early when a candidate suffices, or reallocating a fixed denoiser budget over a larger pool—offering a spectrum of compute–performance tradeoffs.

Efficient sampling improves the cost–performance trade-off. GeoReward can be combined with efficient search to retain fidelity gains with fewer generated candidates or obtain larger geometric gains at comparable denoiser work. On the evaluated Geometry Forcing pool, progressive search stops once the reward exceeds a threshold. At $\tau_{\mathrm{r}}{=}0.9975$, it uses 4.9 candidates per video on average—approximately 39% fewer full generations than best-of-8—while improving FID, PSNR, LPIPS, and SSIM over $N{=}1$ and approaching best-of-8 performance (Table 7). For ReCamMaster, tree search starts from twenty candidates and prunes them during denoising. Both methods use the same 50-step sampler and classifier-free guidance, with the pruning schedule chosen to approximately match their denoiser budgets: 820 forward evaluations for tree search versus 800 for best-of-8, a difference of 2.5%. Table 5 shows that tree search nearly doubles the reduction in pooled mean fused GeCo error on standard trajectories, from 13.5% to 26.6%. Across all twelve trajectories, the reduction increases from 17.5% to 23.9%. Appendices A.3 and B.5 provide the implementation details and limitations.

5 Conclusion

GeoReward uses off-the-shelf geometric foundation-model features as a training-free reward for searching and steering camera-controlled video generation. Experiments across two base generators and diverse static and dynamic scenes demonstrate its applicability to image-conditioned and video-conditioned generation. On Geometry Forcing, geometric critics improve per-frame fidelity, adding value after geometry-aware training. On ReCamMaster, search achieves a 9.8% mean per-trajectory reduction in fused GeCo error across natural and controlled videos. Progressive and tree search further improve the compute–performance trade-off.

Limitations and future work. On natural standard trajectories, BoN search has no statistically detectable mean change in rotational ATE ($+0.05^\circ$), whereas guidance+search increases it by 1.04° relative to the unguided baseline (Appendix B.4). Future work should better balance geometric consistency with target-camera adherence. Broader evaluation across generators, scenes, and camera trajectories is also needed, as computational constraints limit the present transfer study.

AI USE STATEMENT

Generative AI tools (Claude and Codex) assisted with literature review, drafting and editing. The existing research workflow also used AI assistance for code implementation and job orchestration. Wan2.1 generated the controlled static-scene source videos. The numerical revisions here use recorded experimental results. The authors are responsible for the final submission.

REPRODUCIBILITY STATEMENT

Appendix A.1 specifies the Geometry Forcing sampler, reference frames, selection score and guidance hyperparameters (Table 8); Section 4.1 and Appendix B.3 specify the ReCamMaster generator, trajectories, seed groups, and the controlled scene set; and Appendix B.4 gives the camera-following and visual-quality protocols. Appendix C documents exact model checkpoints and frozen/unfrozen weights. Code, data, and camera trajectories will be released.

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. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *arXiv preprint arXiv:2506.09985*, 2025.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-Controlled Generative Rendering from a Single Video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal Guidance for Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 843–852, 2023.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. In *International Conference on Machine Learning (ICML)*, 2025.

Wei Cao, Hao Zhang, Fengrui Tian, Yulun Wu, Yingying Li, Shenlong Wang, Ning Yu, and Yaoyao Liu. FreeOrbit4D: Training-Free Arbitrary Camera Redirection for Monocular Videos via Foreground-Complete 4D Reconstruction. In *ACM SIGGRAPH 2026 Conference Papers*, 2026.

Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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 (ICCV)*, 2025.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Yulu Gao, Bohao Zhang, Zongheng Tang, Jitong Liao, Wenjun Wu, and Si Liu. VGGT-Segmentor: Geometry-enhanced cross-view segmentation. *arXiv preprint arXiv:2604.13596*, 2026. URL https://arxiv.org/abs/2604.13596.

Leslie Gu, Junhwa Hur, Charles Herrmann, Fangneng Zhan, Todd Zickler, Deqing Sun, and Hanspeter Pfister. GeCo: Evaluating Geometric Consistency for Video Generation via Motion and Structure. *arXiv preprint arXiv:2512.22274*, 2025.

Lin Gui, Cristina Gârbacea, and Victor Veitch. BoNBoN alignment for large language models and the sweetness of best-of-n sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling Camera Control for Video Diffusion Models. In *International Conference on Learning Representations (ICLR)*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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 (CVPR)*, 2024.

Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Jaehong Yoon, Soo Ye Kim, Zhe Lin, and Sung Ju Hwang. Frame Guidance: Training-Free Guidance for Frame-Level Control in Video Diffusion Models. In *International Conference on Learning Representations (ICLR)*, 2026.

Min-Jung Kim, Jeongho Kim, Hoiyeong Jin, Junha Hyung, and Jaegul Choo. Infinite-Homography as Robust Conditioning for Camera-Controlled Video Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gökçen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-Free Guidance in Continuous and Discrete Diffusion Models with Soft Value-Based Decoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Haotong Lin, Sili Chen, et al. Depth Anything 3: Recovering the Visual Space from Any Views. *arXiv preprint arXiv:2511.10647*, 2025.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-T1: Test-Time Scaling for Video Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps. *arXiv preprint arXiv:2501.09732*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024.

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. *arXiv preprint arXiv:2503.12024*, 2025.

Byeongjun Park, Byung-Hoon Kim, Hyungjin Chung, and Jong Chul Ye. ReDirector: Creating Any-Length Video Retakes with Rotary Camera Encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations (ICLR)*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A General Framework for Inference-Time Scaling and Steering of Diffusion Models. In *International Conference on Machine Learning (ICML)*, 2025.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-Guided Video Diffusion. In *International Conference on Machine Learning (ICML)*, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards Accurate Generative Models of Video: A New Metric & Challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Wan Team. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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. In *International Conference on Machine Learning (ICML)*, 2026a.

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. In *International Conference on Learning Representations (ICLR)*, 2026b.

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*, 2024.

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. In *International Conference on Learning Representations (ICLR)*, 2026.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term Consistent World Simulation with Memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Tengjiao Yin, Jinglei Shi, Heng Guo, and Xi Wang. VIGOR: Video Geometry-Oriented Reward for Temporal Generative Alignment. *arXiv preprint arXiv:2603.16271*, 2026.

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 (ICCV)*, 2025.

Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdzal, and Adriana Romero-Soriano. Inference-Time Physics Alignment of Video Generative Models with Latent World Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16118–16129, June 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yuchen Zhang, Nikhil Keetha, et al. UFM: A Simple Path towards Unified Dense Correspondence with Flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo Magnification: Learning View Synthesis Using Multiplane Images. *ACM Transactions on Graphics*, 37(4), 2018.

Table 8: **Guidance configuration per generator.** $\mathbf{g}$ denotes the reward gradient with respect to the guided variable.

	Geometry Forcing (DFoT, pixel-space DDIM, 50 steps)	ReCamMaster (latent flow matching, 50 steps)
Clean-sample estimate	$\hat{\mathbf{z}}_{0\|k} = (\mathbf{z}_k - \sqrt{1-\bar{\alpha}_k}\,\epsilon_\theta)/\sqrt{\bar{\alpha}_k}$	$\hat{\mathbf{z}}_{0\|k} = \mathbf{z}_k - \sigma_k \mathbf{v}_\theta$
Gradient path	reward $\to \hat{\mathbf{z}}_{0\|k}$ only; $\partial\hat{\mathbf{z}}_{0\|k}/\partial\mathbf{z}_k \approx 1/\sqrt{\bar{\alpha}_k}$	through the sparse VAE decoder and the full DiT (sequence-parallel, activation checkpointing)
Update	$\tilde{\epsilon} = \epsilon_\theta - \sqrt{1-\bar{\alpha}_k}\,\lambda\,\mathbf{g}/\sqrt{\bar{\alpha}_k}$, λ=1	$\mathbf{z}_k \leftarrow \mathbf{z}_k + \eta\,\mathbf{g}/\|\mathbf{g}\|_2$, η=3
Guided steps	all 50 DDIM steps of the keyframe pass (references are context frames in the interpolation passes)	steps 3–50 of 50
Guided frames $\mathcal{F}_{\text{guid}}$	the eight spread references $\{17, 51, \ldots, 255\}$	every fifth frame $\{5, 10, \ldots, 80\}$ (16 frames)
Decoding for the reward	none (pixel space)	latents downsampled $4\times$, guided frames only
Classifier-free guidance	history guidance 4.0	5.0 on the sampler step; positive branch only on the guided pass

A Geometry Forcing: Implementation and Evaluation

A.1 Protocol and Reward Configurations

Geometry Forcing rewards use held-out ground-truth frames, excluding the context image in the main configuration. Sampling is hierarchical: keyframe prediction at density $1/16$, followed by interpolation, with 50 DDIM steps and history guidance scale 4.0. We evaluate two disjoint seed groups, denoted seed group 1 and seed group 2. In Table 1, π^3 and VGGT use spread-8 references with segment top-k + motion aggregation; DA3 uses spread-4 references with segment-max aggregation. Appendix A.2 defines these reference sets and scores.

Guidance hyperparameters. Table 8 lists the guidance configuration of each generator. Two clarifications qualify the table. First, the ReCamMaster clean-sample estimate uses the flow-matching convention $\hat{\mathbf{z}}_{0|k} = \mathbf{z}_k - \sigma_k \mathbf{v}_\theta$, where σ_k is the flow time of the shift-5 sampling schedule and $\mathbf{v}_\theta$ predicts the noise-minus-data velocity; the guided forward pass uses the conditional branch only. Second, both updates are heuristic ascent steps rather than exact gradients of the tilted distribution in Eq. 1: the ReCamMaster update retains only the gradient direction (global ℓ_2 normalization) with a fixed per-step displacement η, and the Geometry Forcing update uses the detached-denoiser approximation $\partial\hat{\mathbf{z}}_{0|k}/\partial\mathbf{z}_k \approx 1/\sqrt{\bar{\alpha}_k}$.

A.2 Reference-Frame and Aggregation Ablations

We examine two choices in the Geometry Forcing reward: which reference frames to use and how to aggregate their similarities with generated frames. Both ablations select from the same eight-candidate pool per video.

Reference-frame placement. The suffix in ref-4/ref-8 and spread-4/spread-8 denotes the number of reference frames. The *ref* sets sample the full 256-frame clip approximately uniformly, including frame 0 (the context image). The *spread* sets use ground-truth frames at generated keyframe indices, starting at frame 17 and excluding the context image. Table 9 gives the exact zero-based indices and compares the four sets on 20 clips with top-k + motion aggregation. Ref-8 has the lowest FVD (Unterthiner et al., 2018) for each critic on this subset; these estimates do not establish a preferred reference set across the full evaluation.

Aggregation modes. Let c_{ij} denote the cosine similarity between reference embedding q_i and the embedding of generated frame j. *Time-aligned* scoring averages similarities at matching frame indices. The segmented modes instead divide the generated clip at midpoints between consecutive

Table 9: **Reference-frame ablation.** FVD↓ on the first 20 Geometry Forcing clips, with $N{=}8$ and top-k + motion aggregation. Frame 0 is the context image. Best per critic in **bold**.

References	Frame indices	π^3	VGGT	DA3
ref-4	$0, 85, 170, 255$	424.5	469.1	447.3
ref-8	$0, 36, 73, 109, 146, 182, 219, 255$	**414.8**	**460.3**	**426.5**
spread-4	$17, 102, 170, 255$	440.8	465.5	469.8
spread-8	$17, 51, 85, 119, 153, 187, 221, 255$	437.3	468.8	451.5
$N{=}1$ baseline	—		472.3	

reference indices and compare each reference with all generated frames in its segment S_i. *Segment-max* takes the highest similarity in each segment; *segment-top-k* averages the four highest similarities ($k{=}4$). *Segment-softmax* takes a weighted average, with weights $w_{ij} = \mathrm{softmax}_{j \in S_i}(c_{ij}/0.07)$. The segment scores are then averaged across references.

Top-k + motion augments the segment-top-k score a with a feature-change term. Let p_i be the ℓ_2-normalized mean embedding of the top-k generated frames in S_i. The final score is $a + 0.25m$, where $m = \mathrm{mean}_i \cos(q_{i+1} - q_i,\, p_{i+1} - p_i)$ compares successive reference and generated feature changes. Guidance and progressive search use a alone.

Table 10 holds the spread-8 references fixed and evaluates all five modes on 100 videos. The FVD ranking depends on the critic: segment-top-k gives the lowest FVD for π^3, top-k + motion for VGGT, and segment-max for DA3.

Table 10: **Aggregation ablation.** FVD↓ on 100 Geometry Forcing videos, with $N{=}8$ and spread-8 references. Best per critic in **bold**.

Aggregation	π^3	VGGT	DA3
time-aligned	242.8	242.3	246.1
segment-max	242.4	250.3	**242.0**
segment-top-k	**235.7**	247.3	246.2
segment-softmax	236.6	242.9	245.6
top-k + motion	241.4	**239.3**	243.2

A.3 Progressive Search: Stopping Rule and Threshold Sweep

The progressive-search policy of Section 4.5 generates up to eight candidates in a fixed order, stopping when the best segment-top-k π^3 score exceeds a threshold τ_{r}. It returns the highest-scoring candidate generated so far. Table 7 reports the threshold sweep. The main-text operating point, $\tau_{\mathrm{r}}{=}0.9975$, uses 4.9 candidates on average; $\tau_{\mathrm{r}}{=}0.9980$ uses 5.9 and approaches the fixed-$N{=}8$ results. Thresholds were selected on this evaluation pool and have not been independently validated; the policy does not adaptively invoke guidance.

A.4 Runtime and Memory

Table 6 reports computation cost per method. For Geometry Forcing (256 frames at 256×256, hierarchical DDIM-50 sampling on an A100-80GB), π^3 scoring takes 16 s for all eight candidates, approximately 0.3% of their generation time. Single-frame critic forwards take 8, 44, and 15 ms for π^3, VGGT, and DA3, respectively. Detached-gradient guidance adds 10.4% runtime per candidate and uses 25.3 GB peak allocated memory, compared with 33.0 GB for the baseline sampler.

A.5 Embedding Extraction

Each critic maps RGB frames in $[0, 1]$ to per-frame ℓ_2-normalized embeddings by resizing the inputs, running the frozen model, and mean-pooling each frame's patch tokens. The same differentiable extraction supports candidate scoring and gradient guidance; Table 11 summarizes the model-specific settings.

Table 11: **Critic embedding extraction.** All produce per-frame, ℓ_2-normalized, mean-pooled patch-token embeddings; they differ in which network computes the tokens and whether frames interact.

	π^3	VGGT-1B	DA3 ViT-L
Tokens from	encoder only	full aggregator	DINOv2 backbone
Cross-frame attention	none	frame + global layers	within chunk
Input resolution	adaptive[a]	518^2	504^2
Feature layer	encoder output	last aggregator	layer 23 (last)
Embedding dim	1024	2048[b]	1024

[a] Aspect-preserving under a 255k-pixel budget (patch size 14). [b] Camera and register tokens are excluded before pooling. VGGT and DA3 process chunks of 16 frames in this implementation.

The π^3 loader instantiates `Pi3X`, the π^3 authors' extended checkpoint release (Wang et al., 2026b), with a DINOv2 ViT-L/14 image encoder with register tokens. Inputs are bilinearly resized to patch-14-compatible dimensions and use the model's channel normalization. The encoder weights are initialized from the public DINOv2 ViT-L/14 with-registers release and were updated by π^3's geometry training (relative weight deviation 7–11% across early, middle, and late transformer blocks), so the compared features are geometry-trained rather than generic DINOv2 features. Checkpoint revisions for all models are listed in Appendix C.

B ReCamMaster: Implementation and Evaluation

B.1 Sampling and Reference Frames

ReCamMaster generates 81 frames at 832×480 with 50 flow-matching steps. The reward uses the source video without target ground truth ($\mathbf{q} = \mathbf{s}$), pairing frames at the same time along different camera paths. Selection scores all 81 time-aligned frame pairs; guidance steers on every fifth frame at 48 of the 50 sampling steps (Table 8). Both the natural and controlled sets use the same two disjoint seed groups as the Geometry Forcing main comparison. ReCamMaster search and guidance fit on one A100-80GB GPU.

B.2 Matched Appearance-Control Protocol

The natural-video control uses the same eight candidates per video, reference-frame pairs, and GeCo evaluation as π^3. DINOv2 uses the 768-dimensional CLS output of `dinov2_vitb14`; CLIP uses the 512-dimensional projected image embedding of `openai/clip-vit-base-patch16`. RGB frames are bilinearly resized to 224×224, with ImageNet normalization for DINOv2 and model-specific normalization for CLIP. Encoders run in FP16; embeddings are converted to FP32 and ℓ_2-normalized. Selection maximizes the mean cosine similarity over all 81 time-aligned frame pairs.

π^3 preprocessing first resizes and center-crops source frames to the generation dimensions (832×480), then applies its adaptive patch-grid resize. The appearance scorer resizes the original full source frames directly to 224×224. Consequently, candidate pools, frame pairing, and evaluation are matched, while source preprocessing, backbone size, and feature pooling differ. The control does not isolate geometry supervision from those choices.

B.3 Controlled Static-Scene Protocol

The ten sources are generated with Wan2.1-T2V-1.3B (81 frames at 832×480, 50 steps) using prompts for static scenes and a fixed camera. They include geometric primitives on checkerboard or textured floors, stacked wooden blocks, objects on a table, and a stone-column corridor. Textured surfaces provide correspondences for geometric evaluation. Each source is re-rendered along ten standard presets and two loop-closure trajectories, and evaluation covers the full frame. The prompts encourage static content but do not eliminate unintended source motion or appearance changes.

Table 12: **Camera-following measurements on natural standard trajectories.** ViPE-estimated errors on ten videos and ten presets. $N{=}1$ vs. best-of-8 within each candidate pool; the last column counts rotational ATE/RPE wins out of 100 pairs. Rotation is in degrees; translation is in meters and uses translating presets only. Table 14 reports means across both seed groups and paired uncertainty.

Method, seed group	ATE rot. $N{=}1{\to}8$	RPE rot. $N{=}1{\to}8$	ATE transl. $N{=}1{\to}8$	$N{=}8$ wins
Search, group 1	3.87 → 3.64	0.93 → 0.88	0.41 → 0.40	53 / 49
Search, group 2	3.45 → 3.78	0.84 → 0.91	0.39 → 0.46	48 / 48
Search+guid., group 1	5.07 → 5.01	1.28 → 1.24	0.55 → 0.57	45 / 44
Search+guid., group 2	3.79 → 4.38	0.92 → 1.05	0.52 → 0.58	46 / 42

Table 13: **Camera-following error on loop closures** (two trajectories × ten natural videos per seed group; ATE/RPE rotation in deg, translation in m).

Method, seed group	N	ATE rot.	ATE transl.	RPE rot.	RPE transl.
Search, group 1	1 / 8	50.70 / 52.44	1.66 / 1.93	15.49 / 15.65	0.50 / 0.54
Search, group 2	1 / 8	53.21 / 52.76	1.65 / 1.83	16.01 / 15.64	0.48 / 0.52
Search+guid., group 1	1 / 8	50.20 / 53.49	1.83 / 1.97	15.35 / 15.71	0.54 / 0.57
Search+guid., group 2	1 / 8	52.50 / 52.83	1.95 / 1.81	15.88 / 15.75	0.56 / 0.50

B.4 Camera-Following Controls

Table 14: **Camera following on natural standard trajectories.** Mean rotational ATE (degrees) over ten sources, ten presets, and two seed groups. Δ is the change from ReCamMaster; positive values indicate degradation. Paired 95% CIs resample the ten source videos.

Method	Rotational ATE↓	ΔATE	95% CI of ΔATE
ReCamMaster ($N{=}1$)	3.66	—	—
+GeoReward-π^3 (BoN)	3.71	$+0.05$	$[-0.18, +0.29]$
+GeoReward-π^3 (BoN+∇)	4.70	$+1.04$	$[+0.52, +1.57]$

Camera-following protocol. Tables 12 and 13 report errors on standard and loop-closure trajectories, respectively. We prepend the source video's first frame as a pose anchor, estimate camera poses with ViPE (Huang et al., 2025), and compare them with the target extrinsics using the ATE/RPE pipeline of Kim et al. (2026). Poses are expressed relative to the first frame and aligned by per-trajectory median-scale normalization. Translation errors use only presets 5–10, since presets 1–4 are pure rotations. On loop closures, mean rotational ATE is approximately 50–54° across configurations; these measurements cannot separate generation error from pose-estimation failure.

Camera-following results on natural videos. Table 14 compares BoN search and guidance+search with the unguided ReCamMaster baseline. We compute paired percentile bootstrap intervals using 5,000 resamples of the ten source videos, retaining all ten presets and both seed groups together within each sampled source. The evaluation contains 200 paired cases but only ten source-video clusters. Search has no statistically detectable mean change on this evaluation; this does not establish equivalence. Guidance+search increases mean rotational ATE, with a confidence interval excluding zero.

Table 12 separates selection effects by seed group. Guided $N{=}1$ already includes guidance, so its comparison with guided $N{=}8$ measures only selection, whereas Table 14 compares the full guidance+search pipeline with the unguided baseline. Selection changes mean rotational ATE in opposite directions across seed groups. Within each candidate pool, rotational ATE/RPE win rates are 42–53%, with absolute median changes below 0.13°. Paired translation changes also have small medians, although some pool means increase.

Table 15: **Tree search versus best-of-8: denoiser work and measured end-to-end cost.** GPU-seconds count both GPUs of the sequence-parallel tree run.

	Best-of-8	Tree-20	Overhead
Denoiser forwards	800	820	$+2.5\%$
End-to-end GPU-seconds per scene	4770	5470	$+15\%$

B.5 Denoising-Time Tree Search

Result at comparable denoiser work. Table 5 reports controlled scenes 1–5, averaged over two independent seed groups of 60 videos each, comparing 820 denoiser forwards for tree search with 800 for best-of-8. Tree search wins 70/100 per-video comparisons on standard trajectories (38/50 and 32/50 in the two groups) but trails best-of-8 marginally on the lemniscate. The comparison changes the initial candidate count and pruning schedule together, so it does not isolate their individual effects. On natural videos, the tested dynamic-pruning variant did not improve on plain search.

Matching denoiser work. Tree search starts from twenty candidates and retains 10/5/3 after steps 10/20/30 of a 50-step flow-matching sampler. Classifier-free guidance requires two DiT forwards per step. The schedule therefore uses

$$C_{\text{tree}} = 2(20 \cdot 10 + 10 \cdot 10 + 5 \cdot 10 + 3 \cdot 20) = 820, \qquad C_{\text{BoN}} = 2 \cdot 8 \cdot 50 = 800.$$

This increases the initial pool by $2.5\times$ at 2.5% more denoiser forwards. At each checkpoint, we reuse the current velocity prediction to form $\hat{\mathbf{z}}_{0|k} = \mathbf{z}_k - \sigma_k \mathbf{v}_\theta$; obtaining this estimate requires no additional DiT evaluation.

Sparse, low-resolution scoring. Intermediate ranking decodes only sixteen frames, at indices $\{5, 10, \ldots, 80\}$, from two-latent-frame windows. Before VAE decoding, each window is bilinearly downsampled by four along each spatial dimension. The decoded frames are thus 208×120, compared with the final 832×480 output, reducing the spatial area processed by the decoder to $1/16$. The critic applies its standard preprocessing to these frames and compares their features with cached source features. Both decoding and scoring run without gradients. Only the three final survivors undergo full-resolution, 81-frame decoding and final reward evaluation.

Runtime accounting. The schedule makes $20+10+5 = 35$ intermediate scoring calls. Recorded component timings are approximately 1–2 s per sparse decode-and-score call and 10 s per denoising step with classifier-free guidance on A100 hardware. These component estimates imply roughly 35–70 s of probe overhead, in addition to 410 denoising steps across candidates, compared with 400 for best-of-8. Runtime also benefits from encoding the source and prompts once per source–trajectory case and fully decoding only three finalists instead of eight. At matched per-step speed, the sampling costs can be decomposed as

$$T_{\text{BoN}} \approx 400 t_{\text{step}} + 8 t_{\text{final}} + T_{\text{other}}^{\text{BoN}},$$
$$T_{\text{tree}} \approx 410 t_{\text{step}} + 35 t_{\text{probe}} + 3 t_{\text{final}} + T_{\text{other}}^{\text{tree}},$$

where t_{step} includes both classifier-free guidance forwards, t_{probe} includes sparse decoding and reward scoring, t_{final} includes full-video decoding and scoring, and T_{other} covers setup, transfers, and output overhead. Table 15 reports the measured end-to-end comparison on the recorded runs.

Probe validation. In two validation cases without pruning, the Spearman correlation between probe and final full-resolution reward ranges from 0.55 to 0.90 across steps 5–40. This checks agreement with the selection reward, not necessarily with GeCo, and is insufficient to establish reliability across scenes.

C Checkpoints and Reproducibility

All model weights are frozen throughout; no generator, critic, or auxiliary model is fine-tuned. Generators: the public Geometry Forcing DFoT checkpoint

(`geometry_forcing_state_dict.ckpt` from `Haoyuwu/GeometryForcing`) and the official ReCamMaster checkpoint (`step20000.ckpt`) over Wan2.1-T2V-1.3B (DiT, UMT5 text encoder, and VAE weights from the official release). Critics and evaluators, with Hugging Face revisions: the π^3 critic uses the Pi3X release (`yyfz233/Pi3X` @ `bb1deea4`; image encoder with the multimodal head disabled) on both generators, VGGT (`facebook/VGGT-1B` @ `860abec7`), DA3 (`depth-anything/DA3-LARGE-1.1` @ `0e109ae3`), the appearance-control CLIP (`openai/clip-vit-base-patch16` @ `57c21647`), SAM 2 (`facebook/sam2.1-hiera-large` @ `665f8e2a`), and DINOv2 via `torch.hub` (`facebookresearch/dinov2`; ViT-B/14 for the appearance control, ViT-L/14 inside GeCo).

We retain candidate seeds, generation configurations, and cached outputs for re-evaluation.