VGGRPO: Towards World-Consistent Video Generation with 4D Latent Reward

Zhaochong An[1,2⋆], Orest Kupyn[1,3], Théo Uscidda[1,4], Andrea Colaco[1], Karan Ahuja[1], Serge Belongie[2], Mar Gonzalez-Franco[1], and Marta Tintore Gazulla[1]

[1] Google
[2] University of Copenhagen
[3] University of Oxford
[4] Institut Polytechnique de Paris

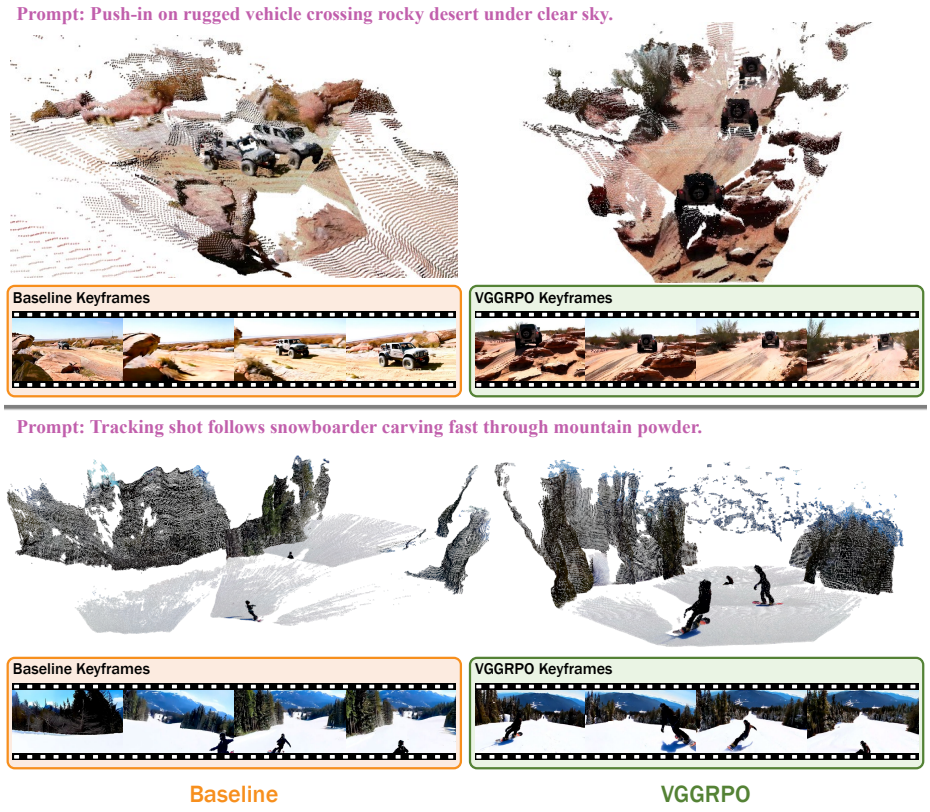

Fig. 1: World-consistent Video Generations with VGGRPO. We compare the baseline video diffusion model (*left*, orange) with the VGGRPO-aligned model (*right*, green). Each example depicts a challenging dynamic scene; we visualize representative keyframes from the generated video and reconstructed scene geometry from the inferred 4D scene representation. VGGRPO produces markedly more coherent scene structure and smoother camera motion over time, reducing geometric drift and structural artifacts in challenging dynamic settings. See our Project Page for more details.

⋆ This work was done during Zhaochong's internship at Google.

Abstract. Large-scale video diffusion models achieve impressive visual quality, yet often fail to preserve geometric consistency. Prior approaches improve consistency either by augmenting the generator with additional modules or applying geometry-aware alignment. However, architectural modifications can compromise the generalization of internet-scale pretrained models, while existing alignment methods are limited to static scenes and rely on RGB-space rewards that require repeated VAE decoding, incurring substantial compute overhead and failing to generalize to highly dynamic real-world scenes. To preserve the pretrained capacity while improving geometric consistency, we propose **VGGRPO** (**V**isual **G**eometry **GRPO**), a latent geometry-guided framework for geometry-aware video post-training. VGGRPO introduces a *Latent Geometry Model* (LGM) that stitches video diffusion latents to geometry foundation models, enabling direct decoding of scene geometry from the latent space. By constructing LGM from a geometry model with 4D reconstruction capability, VGGRPO naturally extends to dynamic scenes, overcoming the static-scene limitations of prior methods. Building on this, we perform latent-space Group Relative Policy Optimization with two complementary rewards: a *camera motion smoothness* reward that penalizes jittery trajectories, and a *geometry reprojection consistency* reward that enforces cross-view geometric coherence. Experiments on both static and dynamic benchmarks show that VGGRPO improves camera stability, geometry consistency, and overall quality while eliminating costly VAE decoding, making latent-space geometry-guided reinforcement an efficient and flexible approach to world-consistent video generation. Our project page is here.

1 Introduction

Recent video diffusion models [1, 43, 51, 69, 78] have achieved impressive visual fidelity and broad generalization by training on large volumes of diverse, high-quality data. However, they often lack 3D and motion consistency [6, 14, 48, 66, 67, 71], exhibiting geometric drift, unstable camera trajectory, and inconsistent scene structure. These issues are critical for downstream applications [2, 3, 13, 15, 25, 30, 31, 39, 54] such as embodied AI and physics-aware simulation, where stable camera motion and coherent 3D geometry are required.

To mitigate these issues, existing efforts largely follow two paradigms. The first paradigm *injects geometric structure into the generator* via additional conditioning modules [8, 53, 73] or extra loss components [10, 70]. For example, point cloud-conditioned diffusion models [8, 53] impose pixel-wise constraints from 3D inputs to improve static-scene consistency, while other approaches [4, 9, 21, 75] augment video diffusion with auxiliary geometry prediction to improve scene generation. While effective, these modifications often increase architectural and computational complexity and can constrain the model, weakening the broad generalization inherited from large-scale pretraining. The second paradigm performs *post-training alignment* inspired by reinforcement learning. Recent approaches adapt Direct Preference Optimization [41, 52] and compute rewards

from sparse epipolar constraints [29] or dense correspondences [11, 17] predicted by external geometry models [63]. However, these methods rely on offline preference data collection, yielding off-policy optimization. Moreover, rewards are typically evaluated in pixel space, requiring repeated VAE decoding, which significantly increases compute and memory overhead; RGB-based rewards are also sensitive to decoding noise and low-level pixel variations [16, 45], further weakening the optimization signals. Finally, these geometric reward formulations are limited to static scenes, as their underlying assumptions [29] and correspondence pipelines [11, 17] do not extend to complex dynamic videos.

In parallel, recent geometry foundation models [27, 63] have demonstrated that feed-forward networks can recover dense geometry and camera motion from static and dynamic image sequences, encoding strong geometric priors learned at scale. This raises a key question: *can we leverage these geometry priors while avoiding the cost and instability of RGB-based reward evaluation?*

We introduce **VGGRPO** (**V**isual **G**eometry **GRPO**), a latent geometry-guided, group-based reinforcement learning framework for video post-training. VGGRPO comprises two tightly coupled components. First, we construct a *Latent Geometry Model* (LGM) that connects video diffusion latents to a geometry foundation model via a lightweight stitching layer, thereby preserving its geometric priors. By operating directly in the VAE latent space, LGM further enables efficient scene-geometry extraction from latents without RGB decoding. Importantly, LGM is model-agnostic and can be instantiated with different geometry foundation models [27, 63], enabling VGGRPO to benefit from ongoing progress in this domain. When connected to models supporting dynamic 4D reconstruction [27], LGM allows VGGRPO to support dynamic videos, extending beyond the static-scene assumptions of prior geometry-consistency methods [11, 29].

Second, building on LGM, VGGRPO performs *latent-space* Group Relative Policy Optimization without repeated VAE decoding, substantially reducing the cost of group-based updates. To optimize for temporally smooth camera motion and coherent 3D structure across viewpoints, we design two complementary rewards: a *camera motion smoothness reward* that encourages stable camera trajectories, and a *geometry reprojection consistency reward* that enforces cross-view geometric coherence. Together, these rewards improve camera stability and 3D consistency, resulting in more realistic, world-consistent video generation, as shown in Fig. 1.

Extensive experiments show that VGGRPO yields consistent gains on both static- and dynamic-scene benchmarks across camera motion smoothness, geometric consistency, and overall video quality. Compared to RGB-based alignment strategies, VGGRPO latent rewards efficiently incorporate geometry priors and support dynamic scenes, providing a practical solution for world-consistent video post-training. Our contributions are summarized as follows:

- We show that reliable geometry-driven rewards can be computed directly in latent space, enabling efficient video post-training without repeated RGB decoding.

- We propose a **Latent Geometry Model** that stitches diffusion latents to geometry foundation models via a lightweight connector, enabling extraction of geometric predictions from latents without pixel-space inputs.
- We introduce **VGGRPO**, a latent-space, group-based reinforcement learning framework with complementary camera-motion and geometry rewards that jointly improve camera smoothness and 3D consistency for world-consistent video generation.

2 Related Work

Based on how existing methods improve geometric consistency, we divide the literature review into geometrically consistent video generation and diffusion model alignment methods.

2.1 Geometrically Consistent Video Generation

Large-scale diffusion models [18,22,33,44,46,49,56] trained with rectified flow [42] have significantly advanced video generation, yet often exhibit geometric drift that undermines scene realism. Geometrically and world-consistent video generation is crucial for downstream applications [5, 13, 36] requiring stable camera motion and coherent scene geometry. Existing approaches to improve geometric consistency broadly fall into two paradigms. *(i) Architecture-level geometry integration.* Point cloud-conditioned diffusion methods [8, 35, 53, 68, 73] introduces explicit 3D conditioning to improve static scene consistency. Other approaches augment diffusion models with auxiliary geometry prediction modules [20, 74]. World-consistent Video Diffusion [75] jointly models RGB and XYZ frames by treating 3D coordinates as an additional modality. GeoVideo [4] incorporates depth prediction with cross-frame consistency losses, while FantasyWorld [9] trains additional decoders to decode scene geometry alongside RGB frames. Although effective, these approaches increase architectural and computational complexity and are limited to static scenes. Furthermore, such modifications can restrict generative flexibility and weaken generalization. *(ii) Training-time regularization.* Another direction introduces extra supervision during training without adding new modules. Geometry Forcing [70] aligns diffusion features with a foundational geometry model [63], and ViCoDR [10] incorporates 3D correspondence losses into video diffusion training. However, these methods typically require full model fine-tuning or training from scratch, which can compromise the broad generalization from large-scale pretraining. In contrast, VGGRPO performs geometry-aware latent-space post-training with on-policy reward optimization, both improving geometric consistency and preserving generalization. Importantly, it naturally extends to dynamic scenes, which prior methods largely do not address.

2.2 Diffusion Model Alignment

Large-scale diffusion models [42] are trained to match broad data distributions, which may not align generations with task-specific objectives [47] (*e.g.*, aesthetics

or physical constraints). Post-training alignment tackles this gap. Early work on image generation [50,55] fine-tuned diffusion models on data filtered by aesthetic classifiers [57]. Later methods cast denoising as a sequential decision process: DDPO [7] and DPOK [12] apply policy-gradient updates under distributional constraints, while Diffusion-DPO [61] adapts Direct Preference Optimization (DPO) [52] to diffusion models using pairwise preferences data. Flow-DPO [41] further extends DPO to rectified-flow models. Some works target physical accuracy: PISA [32] improves physical stability via multi-component rewards, and PhysCorr [64] enhances physical realism through VLM-based rewards. More recently, alignment has been explored for geometry-aware video generation [65]. Epipolar-DPO [29] incorporates epipolar constraints, and VideoGPA [11] extends this direction with dense geometry rewards [63]; however, they assume static-scene consistency, limiting applicability to dynamic videos. Moreover, these approaches require computing rewards in pixel space and rely on offline preference data collection, incurring repeated RGB decoding and limiting optimization efficiency. In parallel, Group Relative Policy Optimization (GRPO) [58] offers an on-policy alternative by sampling from the current model during training, keeping reward signals on-policy without a fixed preference dataset. Flow-GRPO [40] and DanceGRPO [72] adapt this framework to flow-based generators [42], but still require RGB decoding to evaluate rewards; designing effective latent-space rewards remains challenging. Motivated by these limitations, we propose VGGRPO, which performs GRPO-based geometry alignment directly in latent space via a latent geometry model, removing the RGB decoding bottleneck and enabling flexible geometry-aware alignment for dynamic scenes.

3 Methodology

We now detail the formulation of VGGRPO, which couples a latent geometry model with group-based reinforcement optimization for geometry-aware video post-training. As shown in Fig. 2, our method comprises two components: (1) a *Latent Geometry Model* constructed via model stitching, which enables geometry extraction directly from diffusion latents without RGB decoding; and (2) *VGGRPO training* which performs latent-space GRPO using two complementary rewards: *camera motion smoothness* and *geometry reprojection consistency*. These rewards jointly encourage stable camera trajectories and cross-view geometric coherence. We first introduce preliminaries in Sec. 3.1, then describe the latent geometry model in Sec. 3.2 and VGGRPO training procedure in Sec. 3.3.

3.1 Preliminaries

Flow-Based Group Relative Policy Optimization formulates the denoising process of rectified flow models [42] as a multi-step MDP [7] and applies GRPO [34, 40, 58, 72, 76] with an ODE-to-SDE conversion for stochastic exploration. This framework operates on RGB video frames $\mathbf{x} = \{\mathbf{I}_i\}_{i=1}^{N}$. Let $\mathcal{P}$ be a set of text prompts and π_θ a policy parametrized by a velocity field $v_\theta(\mathbf{x}_t, t, p)$.

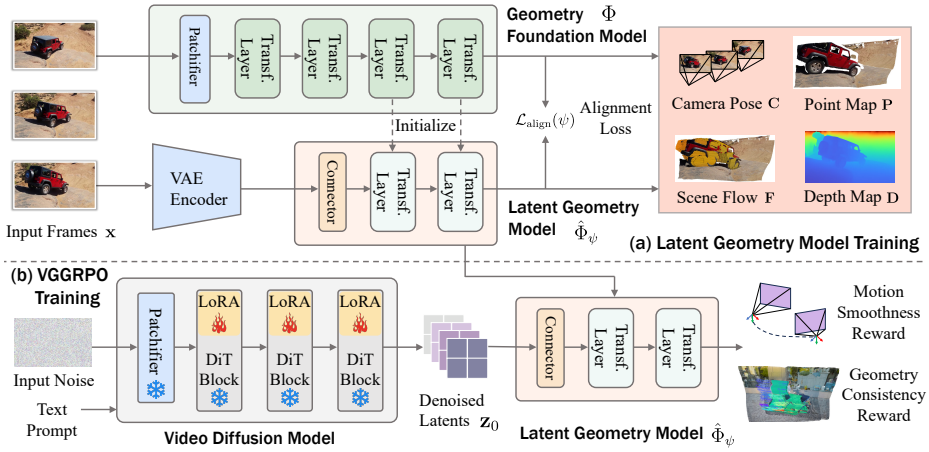

Fig. 2: Method Overview. (a) **Latent Geometry Model.** We connect latents from the diffusion VAE encoder to a geometry foundation model via a lightweight connector, yielding a Latent Geometry Model that predicts 4D scene geometry directly from video latents. (b) **VGGRPO training.** We perform latent-space GRPO using two complementary rewards, camera motion smoothness and geometry reprojection consistency, computed entirely in latent space with the latent geometry model. Together, these components align the video diffusion model toward 4D world-consistent generation on both static and dynamic scenes.

The goal is to maximize the expected reward with KL regularization toward a reference policy π_{ref}:

$$\max_{\theta} \mathbb{E}_{p\sim\mathcal{P},\,\mathbf{x}_0\sim\pi_\theta(\cdot|p)}[r(\mathbf{x}_0,p)] - D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \tag{1}$$

To optimize this objective, GRPO draws K samples from the current policy for each prompt, each consisting of a full denoising trajectory $(\mathbf{x}_T^k, \mathbf{x}_{T-1}^k, \ldots, \mathbf{x}_0^k)$ where $\mathbf{x}_t^k$ is the intermediate state at denoising step t. The per-step importance ratio and clipping operator are defined as:

$$\rho_t^k(\theta) = \frac{\pi_\theta(\mathbf{x}_{t-1}^k \mid \mathbf{x}_t^k, p)}{\pi_{\theta_{\text{old}}}(\mathbf{x}_{t-1}^k \mid \mathbf{x}_t^k, p)}, \quad \text{clip}_\varepsilon(\rho) = \text{clip}(\rho,\, 1{-}\varepsilon,\, 1{+}\varepsilon). \tag{2}$$

Each final clean sample $\mathbf{x}_0^k$ is scored by the reward function r, and the group-relative advantage is computed as:

$$A^k = \frac{r(\mathbf{x}_0^k,\, p) - \mu_r}{\sigma_r}, \tag{3}$$

with μ_r and σ_r the mean and standard deviation of $\{r(\mathbf{x}_0^k, p)\}_{k=1}^K$. The policy is then updated by maximizing the clipped surrogate objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{T}\sum_{t=0}^{T-1}\Big[\min\big(\rho_t^k(\theta)\,A^k,\ \text{clip}_\varepsilon(\rho_t^k(\theta))\,A^k\big) - \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Big]. \tag{4}$$

A full description of the framework, including the ODE-to-SDE conversion, closed-form importance ratio and KL divergence, as well as denoising reduction strategy, is provided in the Supplementary. In VGGRPO, we instantiate this framework in latent space with geometry-aware rewards computed directly from latents (Secs. 3.2 and 3.3).

Geometry Foundation Models are feed-forward transformers [63] that learn strong geometric priors from large-scale 3D-annotated data with minimal explicit 3D inductive biases. Given a sequence of N RGB frames $\{\mathbf{I}_i\}_{i=1}^N$ observing the same scene, a geometry model Φ predicts per-frame geometric representations

$$\{\mathbf{O}_i\}_{i=1}^N = \Phi\big(\{\mathbf{I}_i\}_{i=1}^N\big), \qquad \mathbf{O}_i = \{\mathbf{C}_i, \mathbf{D}_i, \mathbf{P}_i\}. \tag{5}$$

The per-frame outputs $\mathbf{O}_i$ typically include camera pose $\mathbf{C}_i$ (*e.g.*, rotation and translation), a depth map $\mathbf{D}_i$, and a 3D point map $\mathbf{P}_i$ expressed in a shared reference frame. Although these quantities are geometrically related, jointly predicting them during training has been shown to yield substantial performance gains. More recent models [24, 27, 59, 79] extend this formulation to dynamic 4D reconstruction by additionally predicting dynamic point maps or scene flow $\mathbf{F}_i$ within $\mathbf{O}_i$, enabling separation of static and dynamic components in the scene.

3.2 Latent Geometry Model

Geometry foundation models [27, 28, 37, 63] provide strong priors for scene geometry, but they operate in pixel space. Using them for reward computation therefore requires repeated VAE decoding of diffusion latents, resulting in substantial compute and memory overhead [45]. To eliminate this bottleneck, we construct a *Latent Geometry Model* by stitching video diffusion latents to a pretrained geometry foundation model, enabling geometry prediction directly from latents and allowing rewards to be computed in latent space.

Specifically, let $\mathcal{E}$ denote the encoder of a video VAE [62], which maps a video $\mathbf{x} = \{\mathbf{I}_i\}_{i=1}^N$ to latents $\mathbf{z} = \mathcal{E}(\mathbf{x})$. We denote by Φ a pretrained geometry model [27, 63] composed of L transformer layers $\{T_\ell\}_{\ell=1}^L$,

$$\Phi = T_L \circ T_{L-1} \circ \cdots \circ T_1, \tag{6}$$

which maps RGB sequences $\mathbf{x}$ to geometric predictions $\{\mathbf{O}_i\}_{i=1}^N$ as defined in Eq. (5). For convenience, we define the subnetwork spanning layers i to j as $\Phi_{i:j} = T_j \circ T_{j-1} \circ \cdots \circ T_i$.

To bypass the RGB input pathway, we replace the first $\hat{\ell}$ layers of Φ with a learned 3D convolutional connector S_ψ, parameterized by ψ, that maps VAE

latents directly into the intermediate feature space, giving the latent geometry model:

$$\hat{\Phi}_\psi = \Phi_{\hat{\ell}+1:L} \circ S_\psi. \tag{7}$$

The stitching layer $\hat{\ell}$ and the parameters ψ are found jointly by minimizing the feature alignment error over a calibration set of M videos $\mathbf{x}^1, \ldots, \mathbf{x}^M$:

$$\hat{\ell},\, \psi = \underset{\ell \in \{1,\ldots,L\},\, \psi}{\arg\min} \frac{1}{M} \sum_{m=1}^{M} \left\| S_\psi(\mathcal{E}(\mathbf{x}^m)) - \Phi_{1:\ell}(\mathbf{x}^m) \right\|_2^2. \tag{8}$$

We then fine-tune ψ together with the downstream layers $\Phi_{\hat{\ell}+1:L}$ to further reduce residual discrepancies between $\hat{\Phi}_\psi$ and the original geometry model Φ. Given RGB inputs $\mathbf{x}$, we minimize an alignment loss between their geometric predictions:

$$\mathcal{L}_{\text{align}}(\psi) = \sum_j \lambda_j \|\hat{\Phi}_{\psi,j}(\mathcal{E}(\mathbf{x})) - \Phi_j(\mathbf{x})\|_1, \tag{9}$$

where j indexes predicted geometry modalities (*e.g.*, pose $\mathbf{C}$, depth $\mathbf{D}$, point map $\mathbf{P}$, and scene flow $\mathbf{F}$), and λ_j are balancing weights. Afterwards, the resulting latent geometry model $\hat{\Phi}_\psi$ produces geometric predictions directly from latent representations:

$$\{\mathbf{C}_i, \mathbf{D}_i, \mathbf{P}_i, \mathbf{F}_i\}_{i=1}^N = \hat{\Phi}_\psi(\mathbf{z}), \tag{10}$$

where $\mathbf{F}_i$ is present when $\hat{\Phi}_\psi$ is constructed from geometry models with dynamic 4D capability [27]. We use these predictions to define geometry-aware rewards in Sec. 3.3 without RGB decoding, substantially improving efficiency during reinforcement optimization.

3.3 VGGRPO Training

With the latent geometry model $\hat{\Phi}_\psi$, we perform *latent-space GRPO* for geometry-aware video post-training, avoiding the compute and memory overhead of repeated VAE decoding in group-based updates. In practice, we observe that geometric inconsistency in generated videos is often driven by two factors: (i) jittery or unstable camera motion that induces temporal distortions and structural artifacts, and (ii) inconsistent 3D structure across views, where the same scene content is not geometrically aligned over time. Accordingly, our objective is *world-consistent video generation*, which requires both temporally smooth camera trajectories and cross-view coherent scene geometry. To this end, we define two complementary rewards from the geometry predicted by $\hat{\Phi}_\psi$ in Eq. (10): a *camera motion smoothness* reward and a *geometry reprojection consistency* reward. Together, they promote stable camera motion and cross-view geometric coherence, supporting world-consistent generation.

Camera Motion Smoothness Reward. To encourage stable and physically plausible camera motion, we define a smoothness reward based on the camera poses $\mathbf{C}_i$ predicted from the denoised video latents $\mathbf{z}_0$ by $\hat{\Phi}_\psi(\mathbf{z}_0)$. From $\mathbf{C}_i$ we

extract world-frame camera centers $\mathbf{c}_i$ and compute discrete velocities $\mathbf{v}_i = \mathbf{c}_{i+1} - \mathbf{c}_i$ and accelerations $\mathbf{a}_i = \mathbf{v}_i - \mathbf{v}_{i-1}$. Translational smoothness is then measured by the scale-normalized acceleration:

$$e_{\text{trans}}(\mathbf{z}_0) = \frac{1}{T-2} \sum_{i=2}^{T-1} \frac{\|\mathbf{a}_i\|_2}{\|\mathbf{v}_i\|_2 + \|\mathbf{v}_{i-1}\|_2}. \tag{11}$$

Smooth, near-constant-velocity trajectories yield $e_{\text{trans}}(\mathbf{z}_0) \approx 0$, while jittery motion produces larger values. Rotational smoothness is measured identically, replacing translational quantities with angular velocities $\boldsymbol{\omega}_i = \log_{SO(3)}(\mathbf{R}_i^\top \mathbf{R}_{i+1})$ and angular accelerations $\boldsymbol{\alpha}_i = \boldsymbol{\omega}_i - \boldsymbol{\omega}_{i-1}$:

$$e_{\text{rot}}(\mathbf{z}_0) = \frac{1}{T-2} \sum_{i=2}^{T-1} \frac{\|\boldsymbol{\alpha}_i\|_2}{\|\boldsymbol{\omega}_i\|_2 + \|\boldsymbol{\omega}_{i-1}\|_2}. \tag{12}$$

Similarly, $e_{\text{rot}}(\mathbf{z}_0) \approx 0$ for steady rotations and grows with abrupt orientation changes. The combined motion reward is:

$$r_{\text{motion}}(\mathbf{z}_0) = \frac{1}{2} \left(\frac{1}{1 + e_{\text{trans}}(\mathbf{z}_0)} + \frac{1}{1 + e_{\text{rot}}(\mathbf{z}_0)} \right). \tag{13}$$

Both error terms are mapped to $[0, 1]$ via $e \mapsto 1/(1+e)$, so r_{motion} is close to 1 for smooth trajectories and decreases toward 0 as jitter increases.

Geometry Reprojection Consistency Reward. We quantify cross-view geometric coherence using the point maps $\mathbf{P}_i$, depths $\mathbf{D}_i$, camera parameters $\mathbf{C}_i$, and scene flow $\mathbf{F}_i$ predicted by $\hat{\Phi}_\psi(\mathbf{z}_0)$, by reprojecting the predicted 3D structure into each view and comparing depths. We first construct a scene point cloud from the world-frame point maps $\{\mathbf{P}_i\}_{i=1}^N$. For static scenes, we aggregate all points across frames. For dynamic scenes, we use the predicted scene flow $\mathbf{F}_i$ to filter dynamic regions and aggregate only static points to obtain a stable scene representation. We project the resulting point cloud into view i using the predicted camera parameters $\mathbf{C}_i$, producing a rendered depth map $\hat{\mathbf{D}}_i$. We compare $\hat{\mathbf{D}}_i$ with the predicted depth $\mathbf{D}_i$ over valid projected pixels:

$$e_{\text{geo}}^{(i)}(\mathbf{z}_0) = \frac{1}{|\Omega_i|} \sum_{\mathbf{p} \in \Omega_i} \left| \hat{\mathbf{D}}_i(\mathbf{p}) - \mathbf{D}_i(\mathbf{p}) \right|, \tag{14}$$

where Ω_i is the set of valid projected pixels in view i. To focus on local failure cases, we define the geometry reward as the negated average error over the 3 worst views:

$$r_{\text{geo}}(\mathbf{z}_0) = -\frac{1}{3} \sum_{i \in \text{top-3}} e_{\text{geo}}^{(i)}(\mathbf{z}_0). \tag{15}$$

Alignment Policy Update. For each prompt, we sample K latent videos $\{\mathbf{z}_0^k\}_{k=1}^K$ from the current policy and score each with both rewards. Since r_{motion} and r_{geo} have different scales, we normalize each separately within the group

and form the advantage as their average, where $\mu_{\text{motion}}, \sigma_{\text{motion}}$ (resp. $\mu_{\text{geo}}, \sigma_{\text{geo}}$) denote the mean and standard deviation of each reward across the K samples:

$$A^k = \frac{1}{2}\left(\frac{r_{\text{motion}}(\mathbf{z}_0^k) - \mu_{\text{motion}}}{\sigma_{\text{motion}}} + \frac{r_{\text{geo}}(\mathbf{z}_0^k) - \mu_{\text{geo}}}{\sigma_{\text{geo}}}\right). \tag{16}$$

Substituting A^k into the GRPO objective (Eq. (4)) and computing importance ratios $\rho_t^k(\theta)$ over denoised latents $\mathbf{z}_t^k$ (in place of RGB frames $\mathbf{x}_t^k$), we maximize:

$$\mathcal{L}_{\text{VGGRPO}}(\theta) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{T}\sum_{t=0}^{T-1}\Big[\min\big(\rho_t^k(\theta)\,A^k,\ \text{clip}_\varepsilon(\rho_t^k(\theta))\,A^k\big) - \beta\, D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\Big]. \tag{17}$$

We stress that all rewards are computed from $\hat{\Phi}_\psi(\mathbf{z}_0)$ without decoding RGB frames, yielding an efficient geometry-aware post-training pipeline.

4 Experiments

We evaluate the aligned model across diverse scenarios, including static-scene, dynamic-scene, and standard benchmarks, demonstrating its effectiveness and generalization.

4.1 Experimental Setup

Implementation Details. We construct the latent geometry model by stitching to Any4D [27], a geometry foundation model that supports dynamic 4D reconstruction. The latent geometry model is trained for 20 epochs on a mixture of videos generated by the base diffusion model [62] and real videos from DL3DV [38], RealEstate10K [77], and MiraData [26]. For VGGRPO training, we fine-tune two text-to-video diffusion backbones at different scales, Wan2.1-1B and Wan2.2-5B [62], using LoRA [19] (rank $r = 32$, scaling factor $\alpha = 64$), with group size $G = 64$ and AdamW optimization (learning rate 1×10^{-4}, weight decay 1×10^{-4}). Training prompts are sampled from DL3DV, RealEstate10K, and MiraData, following prior geometry-alignment practice [29]. Additional details are provided in the supplementary material.

Baselines. We compare VGGRPO against representative baselines following prior work [11, 29]. For each baseline, we report the best-performing checkpoint to ensure fair comparison. We evaluate **Base Model**, the pretrained generator without post-training; **Supervised Fine-Tuning (SFT)**, flow-matching fine-tuning on real videos as a purely data-driven adaptation baseline; and two DPO alignment methods: **Epipolar-DPO** [29], which constructs preferences using classical epipolar geometry errors, and **VideoGPA** [11], which derives preferences from reprojection-based consistency scores using VGGT [63].

Method	Static			Dynamic		Sub. Cons.↑	Bg. Cons.↑	Aes. Qual.↑	Img. Qual.↑	Mot. Smooth.↑	Dyn. Deg.↑
	VQ↑	MQ↑	Epi.↓	VQ↑	MQ↑						
Base Model: Wan2.1-1B											
Base	-	-	0.133	-	-	0.7941	0.8930	0.5233	0.6178	0.9552	**0.9231**
SFT	45.26	46.84	0.137	40.00	39.00	0.8032	0.8896	0.5472	0.6256	0.9646	0.8795
Epipolar-DPO	54.21	55.79	**0.098**	45.50	43.00	0.8125	0.8916	0.5578	0.6461	0.9671	0.8816
VideoGPA	53.68	56.32	0.105	42.50	41.00	0.8068	0.8931	0.5562	0.6507	0.9650	0.8734
VGGRPO (Ours)	**59.47**	**66.84**	0.102	**57.00**	**63.00**	**0.8255**	**0.8974**	**0.5623**	**0.6585**	**0.9753**	0.9048
Base Model: Wan2.2-5B											
Base	-	-	0.142	-	-	0.8151	0.8958	0.4837	0.6402	0.9467	0.8692
SFT	46.32	52.63	0.129	33.00	51.00	0.8323	0.8925	0.4886	0.6159	0.9548	**0.9026**
Epipolar-DPO	52.11	58.95	0.101	38.00	54.50	0.8407	0.9054	0.4945	0.6275	0.9482	0.7603
VideoGPA	54.74	60.53	0.098	40.00	54.00	0.8511	0.9048	0.4920	0.6131	0.9518	0.7645
VGGRPO (Ours)	**62.63**	**68.42**	**0.093**	**56.50**	**66.00**	**0.8672**	**0.9056**	**0.5094**	**0.6843**	**0.9619**	0.8421

Table 1: Quantitative Comparison. The best and runner-up results are shown in **bold** and underlined, respectively. In both static- and dynamic-scene benchmarks, our method consistently improves geometric consistency and overall video quality over baselines across different base models, demonstrating the effectiveness of latent-space geometry-aware post-training.

4.2 Main Results

Quantitative Evaluation. To evaluate post-training across diverse scenarios, we construct two held-out benchmarks: 190 static-scene captions from DL3DV [38] and RealEstate10K [77], and 200 dynamic-scene captions from MiraData [26]. The dynamic split is curated to include challenging cases with significant non-rigid motion, making geometric drift and camera instability more pronounced. We evaluate from two perspectives: *general video quality* and *geometry-related quality*. For general video quality, we follow VBench [23] and report *subject consistency*, *background consistency*, *motion smoothness*, *aesthetic quality*, *imaging quality*, and *dynamic degree*. For geometry-related quality, following prior studies [11, 29], we report (i) *VideoReward* [41] win rates against the base model for *Visual Quality* (VQ) and *Motion Quality* (MQ), which reflect human preference, and (ii) *Sampson epipolar error* on the static split only (where epipolar assumptions hold), measuring point-to-epipolar-line distances as a classical geometric-precision metric. We report geometry-related metrics separately on the static and dynamic splits, and compute general video quality metrics over all 390 captions.

As shown in Tab. 1, VGGRPO consistently outperforms all baselines on geometry-related metrics. Across both static and dynamic splits, VGGRPO achieves higher motion quality and stronger geometric consistency, demonstrating the effectiveness of latent-space geometry-aware post-training. While prior geometry-aligned baselines [11, 29] can generalize to limited dynamic settings, their performance degrades noticeably on our dynamic benchmark with complex non-rigid motion, where maintaining coherent geometry is substantially harder. In contrast, VGGRPO remains robust under significant scene dynamics due to our 4D-aware latent geometry model and reward design. VGGRPO also improves VBench metrics overall, indicating that our improved geometric consistency does not come at the expense of perceptual quality.

Fig. 3: Qualitative Comparison on Static and Dynamic Scenes. We show the first, middle, and last frames of video generations for a static-scene prompt (*left*) and a dynamic-scene prompt (*right*), with a representative segment of each prompt shown at the top. All baselines exhibit inconsistent artifacts, including geometric drift, temporal flicker, and unstable camera motion. In contrast, VGGRPO produces more coherent scene structure with smoother camera trajectories across frames.

Qualitative Results. Fig. 3 presents qualitative comparisons on both static- and dynamic-scene settings, visualized using the first, middle, and last frames of each generated video. All baselines struggle to maintain world consistency, exhibiting unstable camera motion, geometric drift, and temporally inconsistent scene structure. In the static-scene example, the base model introduces spurious content (a second flower in the middle frame), while SFT shows signs of overfitting to biases in the limited real data, yielding less faithful structure. Epipolar-DPO and VideoGPA show noticeable temporal flicker in the early and late frames, respectively, indicating geometric instability. On the dynamic example, the gap widens: baselines degrade further in both visual quality and geometry, producing less stable frames and inconsistent structure, reflecting reduced robustness under complex dynamic scenes. In contrast, VGGRPO yields more coherent scene structure with smoother camera motion in both settings, highlighting the effectiveness of our geometry-guided latent-space post-training.

4.3 Additional Studies

Unless otherwise stated, we use Wan2.2-5B [62] as the base model and report *VideoReward* win rates for Visual Quality and Motion Quality, averaged over captions from both the static- and dynamic-scene splits, as well as the Sampson epipolar error (Epi.) on the static split to quantify geometric alignment.

Geo-FM	VQ↑	MQ↑	Epi.↓
VGGT	54.96	60.61	**0.090**
Any4D	**59.57**	**67.21**	0.093

(a) Impact of geometry FM.

r_{motion}	r_{geo}	VQ↑	MQ↑	Epi.↓
✓		55.60	63.40	0.104
✓	✓	**59.57**	**67.21**	**0.093**

(b) Impact of reward terms.

	VQ↑	MQ↑	Epi.↓	Time↓
Baseline	-	-	0.142	**44.35**
+ guidance	**52.63**	**52.37**	**0.136**	62.60

(c) Test-time reward guidance.

Model	Sub. Cons.↑	Bg. Cons.↑	Aes. Qual.↑	Img. Qual.↑	Mot. Smooth.↑	Dyn. Deg.↑
Baseline	0.9542	0.9528	0.5966	0.6733	0.9841	**0.4237**
Ours	**0.9644**	**0.9583**	**0.5991**	**0.6861**	**0.9895**	0.3962

(d) Generalization performance on standard VBench captions.

Reward	Time↓	Mem↓
RGB-based	54.73	76.80
Ours	**41.33**	**68.57**

(e) Efficiency study.

Table 2: Additional Studies. (a) **Impact of geometry foundation models:** VGGRPO generalizes across latent geometry models constructed from different geometry foundation models. (b) **Impact of reward components:** r_{motion} and r_{geo} provide complementary gains, and their combination yields the best overall performance. (c) **Test-time reward guidance:** Differentiable latent-space guidance through the latent geometry model improves geometric consistency at test time with modest runtime overhead and no training. (d) **Generalization:** Beyond improving world consistency, VGGRPO transfers to standard VBench captions, indicating improved generation quality and demonstrating robustness in diverse scenarios. (e) **Efficiency study:** Latent-space reward computation reduces both runtime and peak GPU memory compared to RGB-based rewarding. VQ/MQ denote *VideoReward* win rates for Visual and Motion Quality, Epi. denotes the Sampson epipolar error on static scenes. Time is measured in seconds, and Mem denotes peak GPU memory (GB). Best results are in **bold**.

Impact of geometry foundation models. Tab. 2a evaluates the effect of using different geometry foundation models [27, 63] to construct the latent geometry model. When stitching to VGGT [63], VGGRPO training uses only static-scene captions to respect VGGT's static-scene assumptions. As shown in Tab. 2a, the VGGT-based variant achieves slightly lower epipolar error on static scenes, while the Any4D-based variant attains higher VQ and MQ by benefiting from its dynamic-scene support. Both variants outperform prior post-training baselines [11, 29], indicating that VGGRPO is effective across different geometry foundation models.

Impact of reward components. As shown in Fig. 4, optimizing only r_{motion} substantially stabilizes the camera trajectory (red curve) compared to the shaky baseline, while geometric artifacts remain, *e.g.*, the inconsistent wall structure highlighted in the green circle. Adding r_{geo} effectively mitigates these reprojection inconsistencies while maintaining smooth camera motion, yielding more coherent scene geometry and improved perceptual quality. The quantitative results in Tab. 2b mirror these observations: r_{motion} alone improves alignment, and combining it with r_{geo} achieves the best results, confirming both the necessity and complementarity of the two rewards.

Test-time reward guidance. Our latent geometry model is differentiable, enabling gradient-based test-time guidance directly in latent space that steers the model toward improved geometric consistency without RGB decoding. More implementation details are provided in the Supplementary. Table 2c shows that

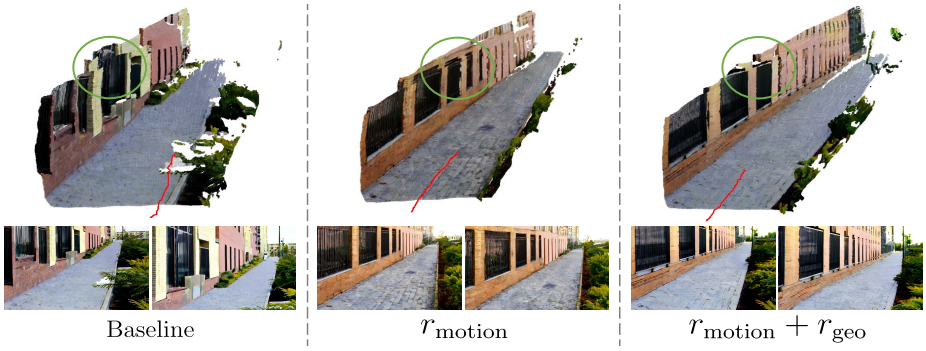

Fig. 4: Reward Components Ablation. The reconstructed scene visualizes the estimated camera trajectory (red curve), with the first and last frames shown below each reconstruction. Compared to the Baseline, optimizing the motion reward r_{motion} stabilizes camera motion, but geometric artifacts persist (green circle). Adding the reprojection consistency reward ($r_{\text{motion}}+r_{\text{geo}}$) further improves scene geometry while preserving smooth camera motion, proving the complementarity of both components.

applying reward guidance once every 20 denoising steps (out of 50 total) consistently improves geometric consistency with modest runtime overhead and no training, providing a training-free way to enhance geometry for a fixed video diffusion model. While the gains are smaller than those from full VGGRPO post-training, these results highlight the utility of our latent geometry model and reward formulation beyond training-time alignment.

Generalization. Beyond geometric consistency, we also verify that VGGRPO preserves the quality of general-purpose video generation. As shown in Tab. 2d, when evaluated on the standard VBench caption set [23], our model outperforms the baseline across most metrics, while slightly lower on Dynamic Degree. We attribute this to the definition of Dynamic Degree, which is computed based on the magnitude of RAFT optical flow [60]: by explicitly encouraging smoother camera trajectories, our motion reward reduces camera jitter and thus lowers the flow magnitude, even as perceptual motion quality improves. Importantly, these gains are achieved without specific training targeting general video quality, indicating that our geometry-aligned post-training preserves the model's original generalization. Overall, VGGRPO acts as a broadly applicable regularizer that improves perceptual quality and remains robust across diverse scenarios.

Efficiency study. Tab. 2e reports the wall-clock time (seconds) and peak GPU memory (GB) of the reward computation step (batch size 4) for RGB-based rewarding and our latent-based rewarding. Compared to RGB-based, our latent reward reduces peak memory from 76.80 GB to 68.57 GB and speeds up reward computation from 54.73 s to 41.33 s (13.40 s faster, 24.5% reduction), demonstrating the efficiency of computing geometry rewards directly in latent space.

5 Conclusion

We introduce VGGRPO, a geometry-aware post-training framework that aligns pretrained video diffusion models toward 4D world-consistent generation by optimizing directly in latent space. To this end, we develop a latent geometry model that connects video latents with geometry foundation models, leveraging their strong geometric priors to predict 4D scene geometry for reward computation. Building on this model, we perform latent-space GRPO, enabling efficient group-based policy updates without repeated VAE decoding. We further design two complementary rewards, camera motion smoothness and geometry reprojection consistency, which jointly promote temporally stable camera trajectories and cross-view coherent scene structure. This formulation applies to both static- and dynamic-scene videos, encouraging 4D world-consistent video generation. Across both static and dynamic benchmarks, VGGRPO substantially improves geometric consistency and camera smoothness while preserving strong visual fidelity, demonstrating that latent-space 4D geometry rewards provide an efficient and scalable approach for aligning video generation models.

Acknowledgements

We would like to thank Hyojun Go (ETH Zurich), Dominik Narnhofer (ETH Zurich), Nikolai Kalischek (Google), Mattia Segu (Google), and Fabian Manhardt (Google) for their valuable support and thoughtful discussion during the project. Zhaochong An and Serge Belongie are supported by funding from the Pioneer Centre for AI, DNRF grant number P1. Orest Kupyn is supported by a Google unrestricted gift.

References

1. An, Z., Jia, M., Qiu, H., Zhou, Z., Huang, X., Liu, Z., Ren, W., Kahatapitiya, K., Liu, D., He, S., et al.: Onestory: Coherent multi-shot video generation with adaptive memory. CVPR (2026)
2. An, Z., Li, Z., Ye, M., Qiao, F., Li, J., Wu, Z., Thengane, V., Li, C., Li, L., Van Gool, L., et al.: Video understanding: From geometry and semantics to unified models. Machine Intelligence Research (2026)
3. An, Z., Sun, G., Liu, Y., Li, R., Wu, M., Cheng, M.M., Konukoglu, E., Belongie, S.: Multimodality helps few-shot 3d point cloud semantic segmentation. In: ICLR (2025)
4. Bai, Y., Fang, S., Yu, C., Wang, F., Huang, Q.: Geovideo: Introducing geometric regularization into video generation model. NeurIPS (2025)
5. Ball, P.J., Bauer, J., Belletti, F., et al.: Genie 3: A new frontier for world models (2025)
6. Bhowmik, A., Korzhenkov, D., Snoek, C.G., Habibian, A., Ghafoorian, M.: Moalign: Motion-centric representation alignment for video diffusion models. ICLR (2026)

7. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. ICLR (2024)
8. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In: SIGGRAPH Asia. pp. 1–12 (2025)
9. Dai, Y., Jiang, F., Wang, C., Xu, M., Qi, Y.: Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. ICLR (2026)
10. Danier, D., Gao, G., McDonagh, S., Li, C., Bilen, H., Mac Aodha, O.: View-consistent diffusion representations for 3d-consistent video generation. arXiv preprint (2025)
11. Du, H., Ye, J., Cong, X., Li, R., Ni, J., Agarwal, A., Zhou, Z., Li, Z., Balestriero, R., Wang, Y.: Videogpa: Distilling geometry priors for 3d-consistent video generation. arXiv preprint (2026)
12. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. NeurIPS **36**, 79858–79885 (2023)
13. Gao, S., Liang, W., Zheng, K., Malik, A., Ye, S., Yu, S., Tseng, W.C., Dong, Y., Mo, K., Lin, C.H., et al.: Dreamdojo: A generalist robot world model from large-scale human videos. arXiv preprint (2026)
14. Gao, X., Wu, M., Yang, S., Yu, J., Taghavi, P., Lin, F., Tu, Z.: The pulse of motion: Measuring physical frame rate from visual dynamics. arXiv preprint (2026)
15. Ge, W., Shen, G., Feng, J., Wang, L., Lu, H., Tian, X., Tao, X., Chen, Y.C.: Campilot: Improving camera control in video diffusion model with efficient camera reward feedback. arXiv preprint (2026)
16. Go, H., Narnhofer, D., Bhat, G., Truong, P., Tombari, F., Schindler, K.: Text-to-3d by stitching a multi-view reconstruction network to a video generator. In: ICLR (2026)
17. Gu, L., Hur, J., Herrmann, C., Zhan, F., Zickler, T., Sun, D., Pfister, H.: Geco: A differentiable geometric consistency metric for video generation. arXiv preprint (2025)
18. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al.: Ltx-2: Efficient joint audio-visual foundation model. arXiv preprint (2026)
19. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
20. Hu, J., Liu, J., Yang, L., Zhang, X., Li, K., Zeng, S., Li, Y., Huang, H., Zhang, C., Lu, Y.: Geometry-as-context: Modulating explicit 3d in scene-consistent video generation to geometry context. In: CVPR (2026)
21. Huang, C.H.P., Mitra, N., Jeong, H., Yoon, J.S., Ceylan, D.: Jog3r: Towards 3d-consistent video generators. BMVC (2025)
22. Huang, X., Zhou, H., Yang, Q., Wen, S., Han, K.: Jova: Unified multimodal learning for joint video-audio generation. arXiv preprint (2025)
23. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR. pp. 21807–21818 (2024)
24. Jiang, Z., Zheng, C., Laina, I., Larlus, D., Vedaldi, A.: Geo4d: Leveraging video generators for geometric 4d scene reconstruction. In: ICCV. pp. 20658–20671 (2025)
25. Jiang, Z., Zhou, S., Jiang, Y., Huang, Z., Wei, M., Chen, Y., Zhou, T., Guo, Z., Lin, H., Zhang, Q., et al.: Wovr: World models as reliable simulators for post-training vla policies with rl. arXiv preprint (2026)

26. Ju, X., Gao, Y., Zhang, Z., Yuan, Z., Wang, X., Zeng, A., Xiong, Y., Xu, Q., Shan, Y.: Miradata: A large-scale video dataset with long durations and structured captions. NeurIPS **37**, 48955–48970 (2024)
27. Karhade, J., Keetha, N., Zhang, Y., Gupta, T., Sharma, A., Scherer, S., Ramanan, D.: Any4d: Unified feed-forward metric 4d reconstruction. arXiv preprint (2025)
28. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. 3DV (2026)
29. Kupyn, O., Manhardt, F., Tombari, F., Rupprecht, C.: Epipolar geometry improves video generation models. arXiv preprint (2025)
30. Le, M.Q., Zhu, Y., Kalogeiton, V., Samaras, D.: What about gravity in video generation? post-training newton's laws with verifiable rewards. arXiv preprint (2025)
31. Li, C., Liu, H., Liu, Y., Feng, B.Y., Li, W., Liu, X., Chen, Z., Shao, J., Yuan, Y.: Endora: Video generation models as endoscopy simulators. In: International conference on medical image computing and computer-assisted intervention. pp. 230–240. Springer (2024)
32. Li, C., Michel, O., Pan, X., Liu, S., Roberts, M., Xie, S.: Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. ICML (2025)
33. Li, D., Fei, Z., Li, T., Dou, Y., Chen, Z., Yang, J., Fan, M., Xu, J., Wang, J., Gu, B., et al.: Skyreels-v3 technique report. arXiv preprint (2026)
34. Li, R., Liang, Y., Ni, Z., Huang, H., Zhang, C., Li, X.: Growing with the generator: Self-paced grpo for video generation. arXiv preprint (2025)
35. Li, R., Torr, P., Vedaldi, A., Jakab, T.: Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In: ICCV. pp. 25690–25699 (2025)
36. Liang, H., Cao, J., Goel, V., Qian, G., Korolev, S., Terzopoulos, D., Plataniotis, K.N., Tulyakov, S., Ren, J.: Wonderland: Navigating 3d scenes from a single image. In: CVPR. pp. 798–810 (2025)
37. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. ICLR (2026)
38. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: CVPR. pp. 22160–22169 (2024)
39. Liu, H., Li, C., Li, Z., Wu, Y., Li, W., Yang, Z., Zhang, Z., Lin, Y., Han, S., Feng, B.: Ir3d-bench: Evaluating vision-language model scene understanding as agentic inverse rendering. NeurIPS **38** (2026)
40. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. NeurIPS (2025)
41. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Xia, M., Wang, X., et al.: Improving video generation with human feedback. In: NeurIPS (2025)
42. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. ICLR (2023)
43. Liu, Z., Ren, W., Liu, H., Zhou, Z., Chen, S., Qiu, H., Huang, X., An, Z., Yang, F., Patel, A., et al.: Tuna: Taming unified visual representations for native unified multimodal models. CVPR (2026)
44. LumaLabs: Dream machine (06 2024), `https://lumalabs.ai/dream-machine`, accessed: 2024

45. Mi, X., Yu, W., Lian, J., Jie, S., Zhong, R., Liu, Z., Zhang, G., Zhou, Z., Xu, Z., Zhou, Y., et al.: Video generation models are good latent reward models. arXiv preprint (2025)
46. OpenAI: Video generation models as world simulators (2024), https://openai.com/index/video-generation-models-as-world-simulators/, accessed: 2024
47. Pan, P., Zhao, J., Lin, Y., Lin, C., Li, C., Liu, H., Shen, T., Mu, Y.: Id-crafter: Vlm-grounded online rl for compositional multi-subject video generation. In: CVPR. pp. 36627–36637 (2026)
48. Park, B., Go, H., Nam, H., Kim, B.H., Chung, H., Kim, C.: Steerx: Creating any camera-free 3d and 4d scenes with geometric steering. In: ICCV. pp. 27326–27337 (2025)
49. PikaLabs: Pika 1.5 (10 2024), https://pika.art/, accessed: 2024
50. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint (2023)
51. Qiu, H., Liu, S., Zhou, Z., An, Z., Ren, W., Liu, Z., Schult, J., He, S., Chen, S., Cong, Y., et al.: Histream: Efficient high-resolution video generation via redundancy-eliminated streaming. arXiv preprint (2025)
52. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. NeurIPS **36**, 53728–53741 (2023)
53. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: CVPR. pp. 6121–6132 (2025)
54. Ren, Z., Wei, Y., Yu, X., Luo, G., Zhao, Y., Kang, B., Feng, J., Jin, X.: Videoworld 2: Learning transferable knowledge from real-world videos. arXiv preprint (2026)
55. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
56. Runway: Gen-3 (06 2024), https://runwayml.com/, accessed: 2024
57. Schuhmann, C.: Laion-aesthetics (2022), https://laion.ai/blog/laion-aesthetics/, accessed: 2023-11-10
58. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint (2024)
59. Sucar, E., Insafutdinov, E., Lai, Z., Vedaldi, A.: V-dpm: 4d video reconstruction with dynamic point maps. arXiv preprint (2026)
60. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. pp. 402–419. Springer (2020)
61. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: CVPR. pp. 8228–8238 (2024)
62. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint (2025)
63. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR. pp. 5294–5306 (2025)
64. Wang, P., Wang, W., Li, Q.: Physcorr: Dual-reward dpo for physics-constrained text-to-video generation with automated preference selection. arXiv preprint (2025)

65. Wang, W., He, X., Gu, Y., Yang, Y., Zhang, Z., He, Y., Ding, Y., Hu, X., Chen, D.Y., He, Z., et al.: World-r1: Reinforcing 3d constraints for text-to-video generation. ICML (2026)
66. Wang, Z., Wang, T., Zhang, H., Zuo, X., Wu, J., Wang, H., Sun, W., Wang, Z., Cao, C., Zhao, H., et al.: Worldcompass: Reinforcement learning for long-horizon world models. arXiv preprint (2026)
67. Wang, Z., Hu, Y., Wang, H., Chen, F., Liu, Y., Li, W., Lei, Y.: Chain of event-centric causal thought for physically plausible video generation. CVPR (2026)
68. Wang, Z., Lin, H., Yoon, J., Cho, J., Zhang, Y., Bansal, M.: Anchorweave: World-consistent video generation with retrieved local spatial memories. arXiv preprint (2025)
69. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint (2025)
70. Wu, H., Wu, D., He, T., Guo, J., Ye, Y., Duan, Y., Bian, J.: Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. ICLR (2026)
71. Xue, H., Chen, Q., Wang, Z., Huang, X., Shechtman, E., Xie, J., Chen, Y.: Mogan: Improving motion quality in video diffusion via few-step motion adversarial post-training. arXiv preprint (2025)
72. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint (2025)
73. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. TPAMI (2025)
74. Zhang, H., Chen, K., Zhang, Z., Chen, H.H., Lyu, Y., Zhang, Y., Yang, S., Zhou, K., Chen, Y.: Dualcamctrl: Dual-branch diffusion model for geometry-aware camera-controlled video generation. arXiv preprint (2025)
75. Zhang, Q., Zhai, S., Martin, M.A.B., Miao, K., Toshev, A., Susskind, J., Gu, J.: World-consistent video diffusion with explicit 3d modeling. In: CVPR. pp. 21685–21695 (2025)
76. Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., Liu, M.Y.: Diffusionnft: Online diffusion reinforcement with forward process. ICLR (2026)
77. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. SIGGRAPH (2018)
78. Zhou, Z., Liu, S., Liu, H., Qiu, H., An, Z., Ren, W., Liu, Z., Huang, X., Ng, K.W., Xie, T., et al.: Scaling zero-shot reference-to-video generation. CVPR (2026)
79. Zhu, R., Lu, J., Hu, W., Han, X., Cai, J., Shan, Y., Zheng, C.: Motioncrafter: Dense geometry and motion reconstruction with a 4d vae. CVPR (2026)