Epipolar Geometry Improves Video Generation Models

Anonymous authors
Paper under double-blind review

Abstract

Video generation models have advanced significantly through the latent diffusion transformers trained with rectified flow techniques. Yet these models still struggle with geometric inconsistencies, unstable motion, and visual artifacts that break the illusion of realistic 3D scenes. 3D-consistent video generation could significantly impact numerous downstream applications in generation and reconstruction tasks. We explore how epipolar geometry constraints improve modern video diffusion models. Despite using massive training data, these models fail to capture fundamental geometric principles. We align diffusion models using pairwise epipolar geometry constraints via preference-based optimization, directly addressing unstable trajectories and geometric artifacts through mathematically principled geometric enforcement. Our approach efficiently enforces geometric principles without requiring end-to-end differentiability. Evaluation demonstrates that classical geometric constraints provide more stable optimization signals than modern learned metrics. Training on static scenes with dynamic cameras ensures metric quality while the model generalizes to various dynamic scenes. By bridging data-driven learning with classical computer vision, we reduce epipolar error by 31% and improve human-rated consistency from 54% to 72% without compromising visual quality.

1 Introduction

Video generation has witnessed remarkable progress, with recent models (OpenAI, 2024; Wiedemer et al., 2025; Polyak et al., 2025; Wang et al., 2025a; Kong et al., 2024) producing increasingly realistic content from text and image conditions. This advancement has spurred researchers to repurpose these powerful video models for broader applications, including animation (Yang et al., 2024), virtual worlds generation (He et al., 2025), and novel view synthesis (Zhou et al., 2025). Video diffusion models are trained on vast volumes of data, developing a strong understanding of object appearance, motion patterns, and scene composition. Many recent works aim to utilize these priors in various downstream tasks (Jiang et al., 2025; Voleti et al., 2024; Chen et al., 2024). Despite this, these models still struggle to maintain perfect 3D consistency, often producing content with imperfect geometry, unstable motion, and perspective flaws, even though almost all of the training data is 3D-consistent. Some approaches for enhancing 3D consistency rely on noise optimization (Liu & Vahdat, 2025), explicit guidance through point clouds (Zhang et al., 2024; Hou et al., 2024), or camera parameters (Zheng et al., 2024). Nevertheless, noisy control signals can constrain the model's generation capabilities, and latent-space optimization makes it difficult to compute direct geometric losses.

With the rising popularity of reinforcement learning for model alignment (Rafailov et al., 2023; Shao et al., 2024; Ouyang et al., 2022), post-training alignment has gained attention in diffusion model research. Methods such as VideoReward (Liu et al., 2025) finetune vision-language models on human preference data, enabling direct supervision through reward models. However, human-annotated quality scores introduce noisy signals and are expensive to collect. Human judgments are inherently subjective and may not capture geometric principles ensuring 3D consistency. The gap between subjective human evaluations and objective geometric requirements creates an opportunity for alignment methods that leverage more mathematically grounded metrics for video quality assessment.

We propose a simple approach that bridges modern video diffusion models with classical computer vision algorithms. Rather than incorporating explicit 3D guidance during generation, we use well-established non-

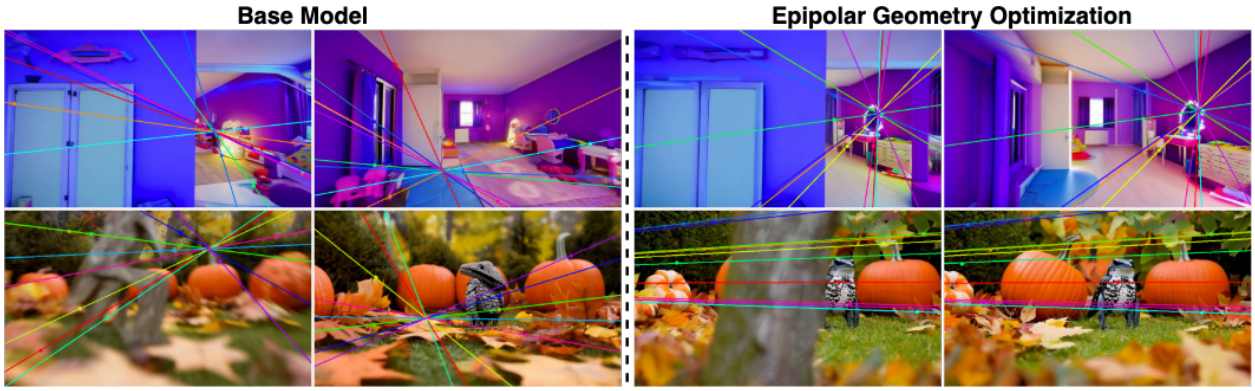


Figure 1: First and middle frame from videos. The baseline model produces geometrically inconsistent outputs with artifacts and unnatural motion. Aligned model generates visibly improved results with smoother camera trajectories, reduced artifacts, and enhanced 3D consistency.

differentiable geometric constraints as reward signals in a preference-based finetuning framework. Specifically, we leverage epipolar geometry constraints to assess 3D consistency between frames. By sampling multiple videos conditioned on the same prompt, we generate diverse camera trajectories that vary in geometric coherence. Epipolar geometry metrics provide reliable signals for identifying which generations better adhere to the principles of projective geometry, enabling us to rank videos and to create training pairs that guide the model toward improved geometric consistency.

Our method implements this through Direct Preference Optimization (DPO) (Rafailov et al., 2023), requiring only relative rankings rather than absolute reward values. This bypasses the difficulties of directly using non-differentiable computer vision algorithms in the training loop. By finetuning the model to prioritize generations that satisfy principled geometric constraints, we guide it towards generating inherently more 3D-consistent videos without restricting creative capabilities or requiring explicit 3D supervision. The aligned model shows enhanced 3D consistency, smoother camera trajectories, and fewer artifacts compared.

While simple in nature, this paper shows that a basic geometric constraint, described in 1982 (Sampson, 1982), can recover what video models fail to do, even after large-scale training on billion-scale data: 3D consistency.

In summary, the key contributions are as follows:

Epipolar Geometry Optimization: We introduce a method for finetuning video diffusion models using epipolar geometry constraints as reward signals, particularly leveraging Sampson distance to enhance 3D video consistency without needing differentiability. The models finetuned with reliable signals from classical vision algorithms achieve superior consistency and quality, significantly reducing artifacts and unstable motion in generated content. Our approach demonstrates that aligning models with fundamental geometric principles enables the generalizable shaping of the model's sample distribution, which is independent of the video content.

Analysis of Alignment Challenges: We analyze two important challenges in optimizing 3D scene consistency: motion-consistency tradeoff and generalization to dynamic scenes, and propose simple yet efficient ways to address both. Finetuning with static motion penalty prevents the degenerate solution of improving consistency by reducing motion, while a combination of implicit KL-divergence, compact trainable adapter, and adaptive timestamp-weighting allows the model to generalize well to dynamic content even when trained on static scenes with dynamic cameras.

Comprehensive Evaluation Framework: We develop an extensive evaluation protocol measuring perceptual quality, 3D consistency, motion stability, visual fidelity, and generalization across diverse scenarios. We compare multiple alignment approaches, demonstrating that classical geometric constraints provide more stable optimization signals than modern learned metrics. In addition, we release a dataset of 162,000

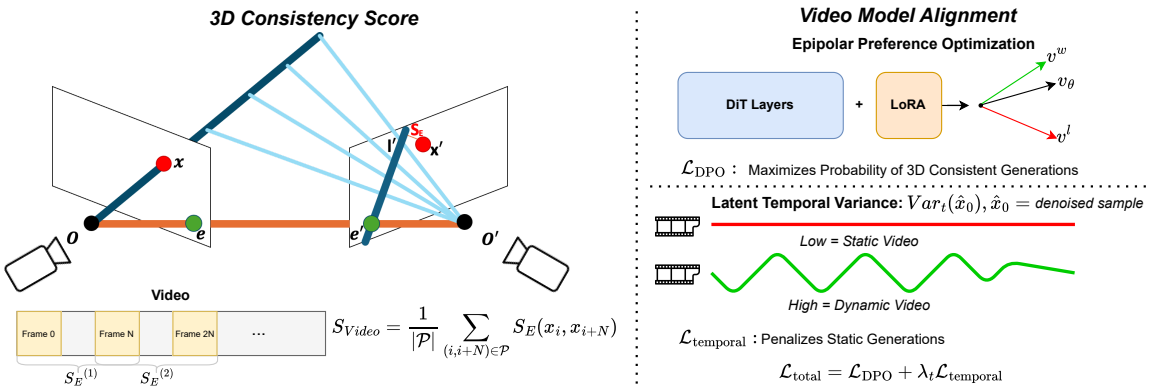


Figure 2: **Epipolar Geometry Optimization.** Left: In 3D-consistent videos, corresponding points must lie on epipolar lines; we compute the distance from matched points to their expected epipolar lines in sampled frame pairs. Right: Flow-DPO training with a static penalty aligns the model toward geometrically consistent outputs.

generated videos with geometric consistency annotations. The data and code enable further research in geometry-aware video generation.

2 Background and Related Work

Flow Matching: Flow Matching (Lipman et al., 2022) learns a time-dependent vector field that transforms samples from a simple prior (e.g., standard Gaussian) into samples from a target distribution. Given samples from an unknown data distribution $q(x_0)$, the goal is to learn a vector field $v_\theta(x_t, t)$ that generates a probability path $p_t(x)$ interpolating between the prior and data distributions. The Flow Matching objective regresses a neural network against a target vector field $u_t(x)$ that generates this path:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,x\sim p_t} \|v_\theta(x,t) - u_t(x)\|^2 . \tag{1}$$

Modern video diffusion models (Wang et al., 2025a) adopt Rectified Flow (Liu et al., 2022b), defining a linear interpolation between data $x_0 \sim q(x_0)$ and noise $x_1 \sim \mathcal{N}(0, I)$:

$$x_t = (1-t)x_0 + tx_1, \quad t \in [0,1] . \tag{2}$$

This yields a constant target velocity $v = x_1 - x_0$, simplifying the objective to:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,x_0,x_1} \|v - v_\theta(x_t, t)\|^2 \tag{3}$$

At inference, samples are generated by integrating the learned velocity field from $t = 1$ (noise) to $t = 0$ (data).

Epipolar Geometry: Epipolar geometry describes the projective relationship between two views of the same scene, depending only on camera parameters and relative pose. When a 3D point $\mathbf{X} \in \mathbb{R}^3$ is observed by two cameras, its projections $\mathbf{x} \in \mathbb{R}^2$ and $\mathbf{x}' \in \mathbb{R}^2$ in the respective image planes are geometrically constrained. The 3D point, camera centers, and both image points must lie on a common plane called the *epipolar plane*. This plane intersects each image along a line called the *epipolar line*: a point at location $\mathbf{x}$ in the first image has its correspondence $\mathbf{x}'$ in the second image lying on the epipolar line $\mathbf{l}' = \mathbf{F}\mathbf{x}$. The fundamental matrix $\mathbf{F} \in \mathbb{R}^{3\times 3}$ encodes this relationship algebraically using homogeneous coordinates. For corresponding points observing the same 3D point:

$$\mathbf{x}'^\top \mathbf{F} \mathbf{x} = 0 . \tag{4}$$

The Sampson error (Sampson, 1982) quantifies how well a pair of corresponding points satisfies the epipolar constraint:

$$S_E = \frac{(\mathbf{x}'^\top \mathbf{F} \mathbf{x})^2}{\|\nabla_{\mathbf{x}}(\mathbf{x}'^\top \mathbf{F}\mathbf{x})\|^2 + \|\nabla_{\mathbf{x}'}(\mathbf{x}'^\top \mathbf{F}\mathbf{x})\|^2} . \tag{5}$$

The numerator measures algebraic violation, while the denominator normalizes by the gradient magnitude, providing a first-order approximation to the geometric distance between each point and its epipolar line.

Flow Models Alignment: Since image and video latent flow models are trained on internet-scale noisy data, efficient finetuning and alignment strategies have emerged. A common approach frames alignment as reinforcement learning, maximizing a reward function while keeping the model close to a reference distribution:

$$\max_{\theta} \mathbb{E}_{x_0 \sim p_\theta}\left[r(x_0)\right] - \beta\, D_{\mathrm{KL}}\left[p_\theta \| p_{\mathrm{ref}}\right], \tag{6}$$

where $r(x_0)$ is a reward function and β controls the KL-divergence regularization strength. Early methods such as DRAFT (Clark et al., 2023) and AlignProp (Prabhudesai et al., 2023) backpropagate through differentiable reward functions, while DPOK (Fan et al., 2023) and DDPO (Black et al., 2023a) employ policy gradient methods to handle non-differentiable rewards.

Diffusion-DPO (Wallace et al., 2024) solves Equation (6) analytically, eliminating explicit reward model access. Given preference pairs (x_0^w, x_0^l) where x_0^w is preferred over x_0^l, define the noise prediction error $\mathcal{E}_\theta(\mathbf{x}_t, t) = \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2$. We write $\mathcal{E}_\theta^w$ for the error on the preferred sample and $\mathcal{E}_\theta^l$ for the less preferred; similarly, $\mathcal{E}_{\mathrm{ref}}^{w,l}$ denotes errors under the reference model. The Diffusion-DPO objective is:

$$\mathcal{L}_{\mathrm{DPO}} = -\mathbb{E}\Big[\log \sigma\Big(-\tfrac{\beta}{2}\big(\mathcal{E}_\theta^w - \mathcal{E}_{\mathrm{ref}}^w - \mathcal{E}_\theta^l + \mathcal{E}_{\mathrm{ref}}^l\big)\Big)\Big], \tag{7}$$

where $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. This requires only pairwise preferences and enables direct finetuning in latent space.

For rectified flow models (Lipman et al., 2022; Liu et al., 2022a; Albergo & Vanden-Eijnden, 2022), the noise prediction relates to velocity prediction via $\mathcal{E} = (1-t)^2\mathcal{V}$. Define the velocity error $\mathcal{V}_\theta(\mathbf{x}_t, t) = \|v - v_\theta(\mathbf{x}_t, t)\|^2$ with $v = \boldsymbol{\epsilon} - \mathbf{x}_0$. Using analogous notation $\mathcal{V}_\theta^{w,l}$ and $\mathcal{V}_{\mathrm{ref}}^{w,l}$, and $\beta_t = \beta(1-t^2)$, the Flow-DPO loss (Liu et al., 2025) is:

$$\mathcal{L}_{\text{F-DPO}} = -\mathbb{E}\Big[\log \sigma\Big(-\tfrac{\beta_t}{2}\big(\mathcal{V}_\theta^w - \mathcal{V}_{\mathrm{ref}}^w - \mathcal{V}_\theta^l + \mathcal{V}_{\mathrm{ref}}^l\big)\Big)\Big]. \tag{8}$$

In practice, alignment methods rely on learned reward models to generate preference pairs. Latent image diffusion models (Podell et al., 2023; Rombach et al., 2022) finetune on data ranked by aesthetics classifiers (Schuhmann, 2022), while VideoReward (Liu et al., 2025) trains VLMs on human annotations. However, learned rewards require expensive annotations, capture subjective preferences that may not reflect geometric correctness, and can produce inconsistent signals across different scene types. Recent work explores structured reward signals: DSO (Li et al., 2025b) employs DPO to align 3D generators with physical soundness, and PISA (Li et al., 2025a) improves the physical stability of video generators using multi-component reward functions. Our method leverages classical computer vision algorithms to provide objective, mathematically grounded preference signals based on epipolar geometry, resulting in more reliable alignment with 3D consistency principles. Since the initial public release of this work, several subsequent papers have expanded the direction of geometry-aware video alignment. VideoGPA Du et al. (2026) and VIGOR Yin et al. (2026) apply DPO with a scoring based on foundational geometry models Wang et al. (2025b) ; VIGOR [link] develops a geometry-oriented reward for temporal generative alignment; GeoFlow Ackermann et al. (2026) and VGGRPO An et al. (2026) explicitly compute reward for dynamic scenes and extend the alignment to GRPO Shao et al. (2024). Geo-Align Li et al. (2026) combines metric-scale reward from MapAnything Keetha et al. (2025) with GRPO to improve camera-controlled video generation. Taken together, they provide additional evidence that enforcing 3D consistency is a useful and increasingly important direction for video generation research. While these works share the goal of enforcing stronger geometric or 3D consistency in generated videos, they differ from our method in the supervision signal, reward construction, and optimization procedure. Our contribution is to demonstrate the effectiveness of simple, reliable epipolar geometric consistency as a basis for preference optimization in video generation.

Video Generation Models: Recent advances in video generation have been dominated by closed-source models trained on proprietary datasets, including Sora (OpenAI, 2024), Runway Gen-3 (Runway, 2024), Luma AI (LumaLabs, 2024), Pika (PikaLabs, 2024), and Veo (Google DeepMind, 2024). Yet their closed nature limits opportunities for fine-tuning. Open-source alternatives have emerged: Stable Video Diffusion (Blattmann

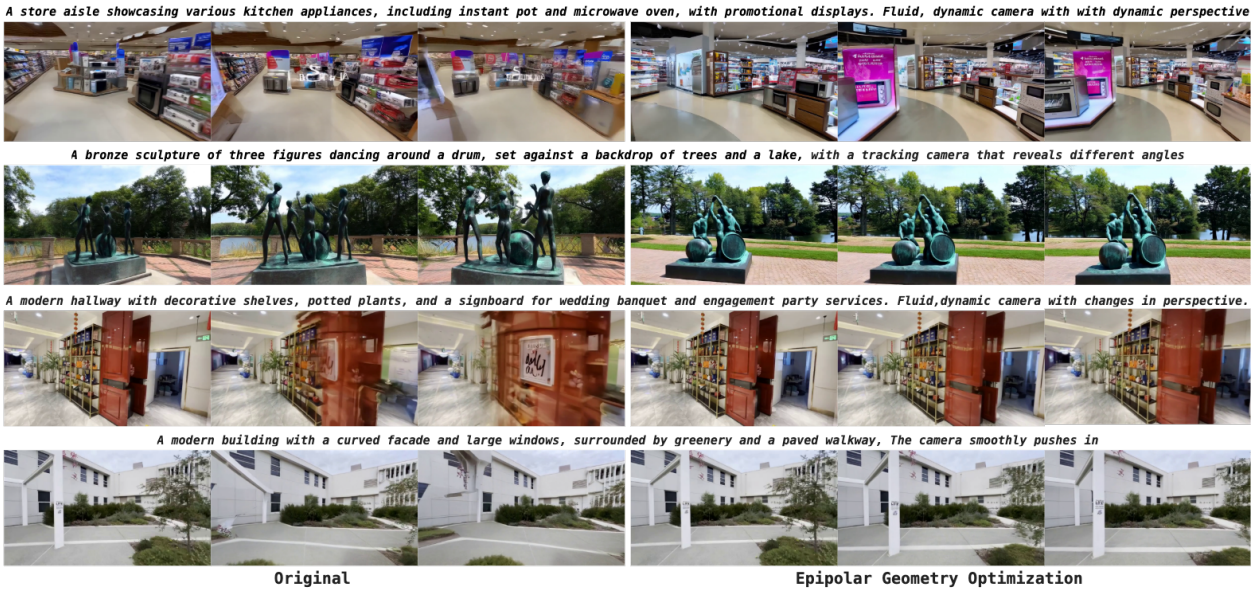


Figure 3: **Qualitative Evaluation:** Visual comparison between the videos generated by the base and finetuned model. First two rows: Wan-2.1-T2V (Wang et al., 2025a), Last two: Wan-2.1-I2V. Our finetuning significantly reduces artifacts and enhances motion smoothness, resulting in more geometrically consistent 3D scenes. Best seen in the supplementary video.

et al., 2023), Hunyuan-Video (Kong et al., 2024), LTX-Video (HaCohen et al., 2024), and Wan-2.1 (Wang et al., 2025a). We use Wan-2.1 (1.3B parameters) for its efficient 3D VAE architecture, which is well-suited for adaptation. These models are trained on enormous data volumes exceeding specific application needs, making domain-aware alignment valuable. V3D (Chen et al., 2024) finetunes for 3D reconstruction, while VideoReward (Liu et al., 2025) introduced RL-based alignment. However, prior methods rely on subjective human preferences or VLMs mimicking them. Our approach optimizes against mathematical rules, aligning models with fundamental geometric principles.

3 Method

We aim to align pretrained video diffusion models to generate geometrically consistent 3D scenes from text or image prompts. We propose an alignment strategy leveraging classical epipolar geometry within a preference-based optimization framework. Traditional reinforcement learning methods (Black et al., 2023b; Shao et al., 2024) require explicit reward functions and access to final samples, impractical for video models due to non-differentiable rewards and high denoising costs. Our key observation is that while epipolar geometry constraints do not produce smooth, globally comparable loss surfaces across scene types, the relative intra-prompt error remains consistent. When generating multiple videos with fixed conditioning, diffusion sampling's stochastic nature yields outputs with varying geometric consistency. Epipolar error effectively quantifies relative 3D consistency, with higher values indicating lower consistency. This finding aligns with the direct preference optimization (DPO) paradigm, which requires only relative metrics to determine preference between output pairs rather than absolute reward values. DPO's pairwise comparison nature eliminates the need for globally normalized reward functions, instead leveraging reliable local ranking provided by epipolar geometry measurements to guide model alignment toward more geometrically consistent video generation.

3.1 Objective Function

Given the pretrained video generator p_{ref} taking text prompt T and optional first frame I, generating $x_0 \sim p_{\text{ref}}(x_0|T, I^*)$ where $I^* \in \{I, \emptyset\}$, we want to learn p_θ optimized for 3D-consistent video sequences. One

choice is to optimize:

$$\max_{\theta} \mathbb{E}_{c\sim\mathcal{D}_c, x_0\sim p_\theta}\big[r(x_0)\big] - \beta\, D_{\mathrm{KL}}\big[p_\theta \| p_{\mathrm{ref}}\big]\,, \tag{9}$$

where $r(x_0)$ outputs 3D consistency scores, $c{=}(T, I^*)$ is the conditioning, and β weights the KL-divergence keeping p_θ close to p_{ref}. However, $r(x_0)$ relies on non-differentiable algorithms, and evaluating geometric consistency requires complete video generation, which is expensive.

This motivates Flow-DPO (Equation (8)), requiring only pairwise preferences. Given dataset $\mathcal{D}{=}\{(c, x_0^w, x_0^l)\}$ where x_0^w has better consistency than x_0^l (lower Sampson error), let $\mathcal{V}_\theta^*{=}\|v^*{-}v_\theta(\mathbf{x}_t^*, t)\|^2$ denote the velocity prediction error. We optimize:

$$\mathcal{L} = -\mathbb{E}\Big[\log\sigma\Big(-\tfrac{\beta_t}{2}\big(\mathcal{V}_\theta^w - \mathcal{V}_{\mathrm{ref}}^w - \mathcal{V}_\theta^l + \mathcal{V}_{\mathrm{ref}}^l\big)\Big)\Big], \tag{10}$$

where $\mathbf{x}_t^* = (1{-}t)\mathbf{x}_0^* + t\boldsymbol{\epsilon}^*$, $v^* = \boldsymbol{\epsilon}^* - \mathbf{x}_0^*$, $\beta_t = \beta(1{-}t^2)$.

Motion-Consistency Tradeoff: Static bias is one of the key challenges in optimizing model consistency, significantly increasing the complexity of the task. In theory, a perfectly static video (single image repeated) achieves perfect consistency yet fails as a generative video task. To prevent degenerate solutions where the model reduces motion to achieve 3D consistency, we add a temporal variation penalty:

$$\mathcal{L}_{\mathrm{temp}} = -\lambda\, \mathbb{E}\big[\mathrm{Var}_t(\hat{x}_0)\big]\,, \tag{11}$$

where $\hat{x}_0 = x_t + (1{-}t)v_\theta(x_t, t)$ is the predicted clean sample, variance is computed across the temporal dimension, and $\lambda{=}0.001$. Our final objective is $\mathcal{L}_{\mathrm{total}} = \mathcal{L} + \mathcal{L}_{\mathrm{temp}}$.

Minimizing this loss encourages the model to improve denoising performance on preferred samples $\mathbf{x}_t^w$ relative to less preferred samples $\mathbf{x}_t^l$, guiding the predicted velocity field v_θ to align with videos exhibiting better 3D consistency while preserving motion quality.

3.2 Preference Score Computation

We evaluate the 3D consistency of generated videos by measuring how well they satisfy the epipolar constraints. Given a pair of frames $\mathbf{x}_i$ and $\mathbf{x}_j$ from a generated video, we first compute a set of point correspondences using SIFT Lowe (1999) feature matching. While we validate the method with a simple, robust handcrafted descriptor, the pipeline can also leverage more recent learned descriptors Lindenberger et al. (2023); Sun et al. (2021); Potje et al. (2024). These correspondences provide a robust set of matching points between the different viewpoints. We then estimate the fundamental matrix using the normalized 8-point algorithm within a RANSAC Fischler framework to handle outliers. The Sampson error (Equation (5)) provides a first-order approximation to the geometric distance between a point and its epipolar line. Lower Sampson error values indicate better adherence to projective geometry constraints and, thus, more consistent 3D structure in the generated videos.

To obtain a single consistency score for ranking videos, we uniformly sample every N-th frame from a video of length L, yielding frames $\{\mathbf{x}_0, \mathbf{x}_N, \mathbf{x}_{2N}, \ldots\}$. We compute the mean Sampson error across consecutive frame pairs:

$$S_V = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} S_E(\mathbf{x}_i, \mathbf{x}_j)\,, \tag{12}$$

where $\mathcal{P} = \{(0, N), (N, 2N), \ldots\}$ denotes consecutive frame pairs and $S_E(\mathbf{x}_i, \mathbf{x}_j)$ is the mean Sampson error over all point correspondences between frames i and j. Lower scores indicate better geometric consistency. For preference pair construction, given videos generated from the same prompt, we assign x_0^w to the video with lower S_V and x_0^l to the one with higher S_V.

While we validate our method with SIFT, a simple and robust handcrafted descriptor, the pipeline can leverage more recent learned descriptors (Lindenberger et al., 2023; Sun et al., 2021; Potje et al., 2024). We analyze the impact of descriptor choice in our ablation studies (Figure 6a).

3.3 Generalization to Dynamic Scenes

Epipolar constraints assume the relationship between frames can be described by a single fundamental matrix, valid only for static scenes with camera motion. Dynamic objects violate this assumption, corrupting the preference signal. This motivates a key design choice: we train exclusively on static scenes with dynamic cameras, ensuring high geometric consistency metric accuracy. Despite this, our method generalizes to dynamic content through three mechanisms. First, finetuning only a low-rank adapter (LoRA) Hu et al. (2022) preserves the base model's capacity for diverse dynamic content. Second, implicit KL-divergence regularization in the Flow-DPO objective keeps the finetuned model close to the reference distribution, preventing overfitting to static scene characteristics. Third, we align the model with a timestep-dependent loss weighting that assigns lower weights on early steps $w(t) = \exp\big(-2(t-0.5)^2\big)$. This weighting is motivated by the observation that general scene motion is encoded in the first few timestamps Wu et al. (2024), whereas geometric consistency is often violated later, when fine spatial details are generated. Together, these choices allow for the encoding of generalizable modifications to the model's sample distribution that are independent of the video content, as shown in Figure 4b.

3.4 Implementation Details

We conduct experiments with a state-of-the-art open-source video diffusion model Wan2.1 (Wang et al., 2025a), with 1.3 billion parameters.

Offline Dataset Generation: Since epipolar constraints require static scenes with camera motion, we extract text prompts from Zheng et al. (2025), which provides diverse indoor and outdoor scene descriptions with dynamic camera movements. To enhance training data quality, we employ Gemma-3 VLM to expand original prompts with more challenging camera motion descriptions, increasing geometric complexity and ensuring sufficient variance in 3D consistency. We generate three videos per caption to increase variation in 3D consistency quality, as preliminary experiments showed pairs from just two samples often lacked meaningful geometric differences. We implement rigorous data filtering: in addition to removing near-static videos, we only sample pairs where $(\text{metric}(x_{\text{win}}) - \text{metric}(x_{\text{lose}}) > \tau) \wedge (\text{metric}(x_{\text{win}}) > \epsilon)$, eliminating pairs where both videos have similar consistency and ensuring we only learn from meaningful gaps. In total, we generate 24,000 triplets for text-to-video and 30,000 triplets for image-to-video training, requiring approximately 1,980 GPU hours on NVIDIA A6000s.

Training Configuration: We implement our approach using Low-Rank Adaptation (LoRA) (Hu et al., 2022) with rank r=64 and α=128. This strategy eliminates the need to store the reference model in memory, since the base model with the adapter disabled naturally serves as p_{ref} during training. We train with a batch size of 32 for 10,000 iterations using the AdamW (Loshchilov & Hutter, 2017) optimizer with a learning rate of 5×10^{-6} and 500 warmup steps. The finetuning takes 2 days on 4 A6000 GPUs.

4 Experiments

We evaluate our epipolar-aligned model across three dimensions: 3D consistency, motion stability, and generalization to dynamic scenes, demonstrating that classical geometric constraints provide more reliable optimization signals than learned metrics while improving video generation quality.

4.1 Evaluation Setup

We evaluate on 400 videos from DL3DV (Ling et al., 2024) and RealEstate10K (Zhou et al., 2018) test sets, using Gemma-3 VLM (Team et al., 2025) to generate challenging camera motion descriptions. For generalization, we test on VBench 2.0 (Huang et al., 2024), MiraData (Ju et al., 2024), and VideoReward (Liu et al., 2025) benchmarks extending beyond static scenes. We measure performance using: (1) VideoReward VLM for motion quality assessment, (2) VBench protocol (Huang et al., 2024) for standardized motion and visual quality metrics, (3) classical geometric consistency via Sampson epipolar error, and (4) 3D reconstruction quality via Gaussian Splatting to validate downstream task impact.

Human Evaluation: We conduct two-stage human evaluation. First, annotators label videos as geometrically consistent based on visible artifacts and motion stability. This reveals that the baseline produces consistent videos only 54.1% of the time, confirming significant room for improvement. Second, annotators perform pairwise comparisons between baseline and aligned versions, demonstrating that our approach preserves quality for already-consistent content while dramatically improving inconsistent cases (60.4% vs 7.5% win rate).

4.2 3D Consistency

Table 1: **3D Consistency Evaluation:** Epipolar aligned model improves 3D Scene Reconstruction and is preferred by human evaluators.

	3D Consistency Metrics		**3D Scene Reconstruction**			**Human Eval**
Method	Sampson Error ↓	Perspective Realism ↑	PSNR ↑	SSIM ↑	LPIPS ↓	Consistency Rate
Baseline	0.190	0.426	22.32	0.706	0.343	54.1%
Ours	**0.131**	**0.428**	**23.13**	**0.729**	**0.315**	**71.8%**

We validate that epipolar geometry alignment improves 3D consistency using three approaches that test different aspects of geometric quality.

3D Scene Reconstruction: We test whether generated videos support accurate 3D scene reconstruction using VGGT (Wang et al., 2025b) to extract scene parameters and camera trajectories. We initialize 3D Gaussian Splatting from extracted scene structure, run 7000 optimization iterations using Splatfacto (Tancik et al., 2023) on 80% of frames, and evaluate reconstruction fidelity on the remaining 20%. Our model demonstrates substantial improvements: PSNR increases from 22.32 to 23.13 (+3.6%), SSIM improves from 0.706 to 0.729 (+3.2%), and LPIPS decreases from 0.343 to 0.315 (−8.2%). These gains demonstrate that epipolar alignment produces videos with genuinely enhanced 3D structure rather than superficial improvements.

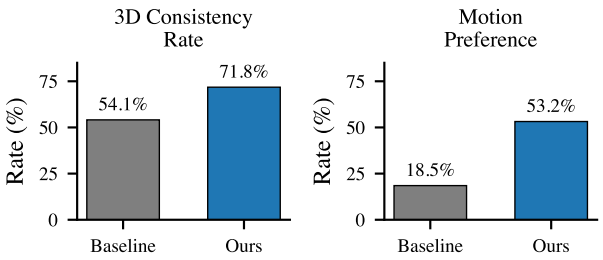
Geometric Consistency Metrics: We directly measure adherence to projective geometry principles using classical computer vision algorithms. The Sampson epipolar error shows a dramatic 31% reduction from 0.190 to 0.131, verifying that our alignment successfully optimizes the metric used for preference selection and confirming that classical epipolar geometry provides clean optimization signals. Additionally, perspective realism, measured by a model trained to evaluate whether image frames contain realistic perspective (Sarkar et al., 2024) improves from 0.426 to 0.428, demonstrating positive impact on adjacent geometric metrics despite this metric's inherent noise.

Human Evaluation: While numerical metrics capture geometric aspects, 3D inconsistencies often manifest as artifacts, jitter, or unnatural changes that humans excel at detecting (Figure 4a). Annotators evaluated videos for scene consistency, realism, and artifact-free content. Our method generates significantly more plausible scenes, with 71.8% of videos labeled as geometrically consistent compared to only 54.1% for baseline. This 17.7 percentage point improvement demonstrates that geometric alignment benefits are apparent to human observers.

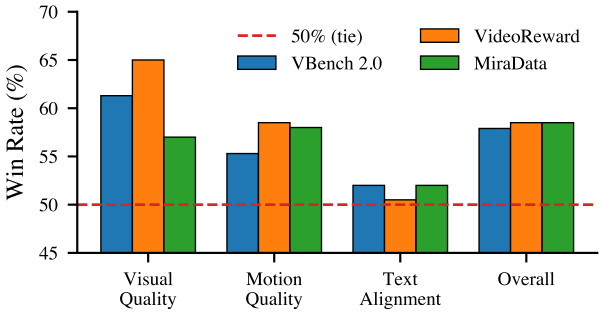
4.3 Motion Quality

Table 2: **Motion Quality Evaluation:** Epipolar aligned model improves motion stability and is preferred by human evaluators despite dynamics-consistency tradeoffs.

	VBench Motion Metrics			**VideoReward**	**Motion Level**	**Human Eval**
Method	Motion Smoothness ↑	Dynamic Degree ↓	Temporal Flickering ↑	Motion Quality ↑	Mean SSIM ↓	Motion Preference Rate
Baseline	0.981	**0.751**	0.958	50.0%	0.233	18.5%
Ours	**0.984**	0.710	**0.969**	**69.5%**	**0.223**	**53.2%**



(a) **Human Evaluation:** Our method improves 3D consistency (54%→72% of videos rated consistent) and motion preference (53% prefer ours vs. 19% baseline in pairwise comparison).



(b) **Generalization to Dynamic Scenes:** Win rates vs baseline across three benchmarks. Despite training only on static scenes, our method generalizes to dynamic content.

Table 3: Visual Quality and Aesthetic Fidelity Results (Text-to-Video)

	VBench Visual Metrics		**VideoReward**	**Human Eval**
Method	Background Consistency ↑	Aesthetic Quality ↑	Visual Quality ↑	Visual Preference Rate
Baseline	0.930	0.541	-	15.0%
Ours	**0.942**	**0.551**	**72.0%**	**52.8%**

While geometric alignment enforces smooth, consistent motion and reduced jitter, we also verify that our alignment preserves the model's ability to generate diverse motions. We evaluate motion quality using various metrics, focusing on different aspects, including temporal dynamics, perceptual assessment, and human preference.

Temporal Dynamics: VBench motion metrics show mixed results reflecting the dynamics-consistency tradeoff. Motion smoothness and temporal stability improve, indicating more stable frame-to-frame transitions. However, dynamic degree decreases from 0.751 to 0.710, suggesting reduced motion amplitude. Mean SSIM between first and remaining frames decreases as well, confirming the model's ability to generate dynamic scenes. While we acknowledge the tradeoff, single neural network metrics can exhibit bias, motivating multi-metric evaluation.

Perceptual Quality Assessment: VideoReward motion quality evaluation shows substantial improvement with our method, achieving a 69.5% win rate. This human-distilled assessment validates that geometric consistency training produces motion aligning better with human preferences for natural, stable dynamics, despite some reduction in motion amplitude.

Human Preference: Direct human evaluation further validates the motion quality of our method, with annotators preferring our approach at 53.2% rate across all video types. Since annotators focus on jitter and motion artifacts, our high preference rate shows that stability improvements outweigh motion amplitude reductions. This preference is particularly strong in initially inconsistent videos (60.4% vs 7.5% for baseline).

4.4 Visual Quality Evaluation

Generating more geometrically consistent scenes with fewer artifacts naturally leads to higher overall visual quality of the generated content. To validate this connection, we evaluate visual fidelity across multiple assessment frameworks.

Aesthetic Metrics: VBench Huang et al. (2024) visual quality assessment shows consistent improvements across multiple dimensions. For example, background consistency increases from 0.930 to 0.942 and aesthetic quality improves from 0.541 to 0.551. These metrics confirm that geometric training enhances visual stability and perceived quality.

Table 4: **Win-rate vs. Wan-2.1-14B Wang et al. (2025a)** on the VideoReward Liu et al. (2025) benchmark. The Baseline and Epipolar-Aligned Model contain only 1.3B parameters.

Text-to-Video				
Method	Visual Quality	Motion Quality	Text Alignment	Overall
Baseline	13.3%	14.4%	24.2%	8.6%
DPO-Epipolar	**18.1%**	**21.8%**	**25.0%**	**13.8%**

Table 5: **Comparison vs. Supervised Finetuning on Multi-View Data.** Finetuning on real multi-view static scenes Zhou et al. (2018); Ling et al. (2024) propagates dataset bias such as scene type or specific camera motion. As a result the model overfits to dataset specifics rather than learning general principles.

Benchmark	Method	Visual Quality	Motion Quality	Text Alignment	Overall
VideoReward	Supervised Finetuning	37.25%	46.75%	**53.0%**	44.75%
	DPO-Epipolar	**65.0%**	**58.5%**	50.5%	**58.5%**
VBench	Supervised Finetuning	35.8%	53.0%	39.4%	35.2%
	DPO-Epipolar	**61.3%**	**55.3%**	**52.0%**	**57.9%**
MiraData9K	Supervised Finetuning	41.5%	42.0%	51.0%	47.0%
	DPO-Epipolar	**57.0%**	**58.0%**	**52.0%**	**58.5%**

Perceptual Assessment: VideoReward Liu et al. (2025) visual quality evaluation demonstrates substantial improvement with a 72.0% win rate, indicating that human-distilled quality assessment strongly favors our geometrically-aligned approach. This suggests that geometric consistency contributes significantly to overall visual appeal.

Human Validation: Human preference evaluation shows a 52.8% preference rate for our method's visual quality across all video types, further validating that geometric improvements translate to perceptually superior results that human evaluators can identify and prefer.

4.5 Supervised Finetuning Evaluation

We compare the effectiveness of epipolar-aware alignment with standard video model finetuning on multi-view static datasets Ling et al. (2024); Zhou et al. (2018). Directly finetuning the model on perfectly 3D-consistent videos of static scenes could be seen as an alternative approach to improve generation consistency. Yet this naive approach only trains the model to reproduce observed frames, thereby optimizing average risk. It is not directly optimizing task-level objectives, such as multi-view geometric consistency, and instead might replicate biases present in the dataset and regress toward "mean" solutions for ambiguous problems. We observe that models finetuned with this objective overfit to dataset-specific factors, such as a narrow set of camera trajectories in RealEstate-10K and DL3DV videos and specific scene types, which hurts the generalization capabilities of the base model. Table 5 demonstrates that a model directly finetuned on multi-view datasets does not generalize well to dynamic scenes and, on average, tends to amplify camera motion, mimicking the average training set trajectories.

4.6 Generalization

Our approach demonstrates strong generalization beyond its training domain of static scenes (Figure 4b). Despite training only on static camera motion, our method achieves consistent improvements across VBench 2.0, VideoReward, and MiraData benchmarks, with overall win rates of 57.9%, 58.5%, and 58.5%, respectively.

This generalization occurs because aligning models with geometrically consistent camera trajectories improves video quality even when objects move independently. The primary error sources in dynamic scenes: unstable trajectories, artifacts, and flickering, are amplified by object movement. By learning to produce stable

Figure 5: **Qualitative Evaluation:** Comparison of baseline and epipolar-aligned models on dynamic scenes featuring both camera movement and object motion. Our approach maintains improved geometric consistency and smoother trajectories, demonstrating generalization beyond static scene training. Best seen in the supplementary video.

camera motion and reducing geometric inconsistencies, our approach addresses these fundamental issues. VBench metrics confirm that geometric consistency training effectively transfers across diverse scenarios, with improvements in background consistency, temporal stability, and motion smoothness validating our core insight that classical geometric constraints enhance overall 3D understanding.

4.7 Ablation Study

Table 6: Win-rate on the VideoReward benchmark comparing different finetuning strategies with geometric consistency metrics and VLM-based metrics (Visual Quality, Motion Quality and Text Alignment).

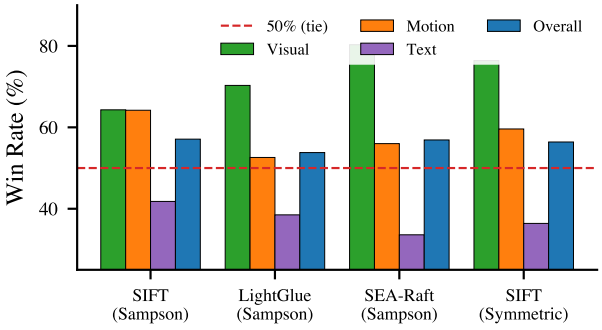
	VideoReward Metrics				**Consistency Metrics**		
Method	VQ	MQ	TA	Overall	Perspective ↑	Sampson ↓	Dynamics ↓
SFT	66.0%	63.0%	54.0%	64.5%	0.427	0.161	0.225
Flow-RWR Liu et al. (2025)	63.5%	60.5%	**57.0%**	64.0%	**0.434**	0.174	0.229
DRO Li et al. (2025b)	65.0%	54.0%	50.5%	64.5%	0.410	**0.068**	**0.195**
Epipolar-DPO (Ours)	**72.0%**	**71.0%**	55.0%	**73.0%**	0.428	0.127	0.223

We ablate descriptor choices, geometric metrics, alignment methods, and design components. For efficiency the ablations are done on a subset of data.

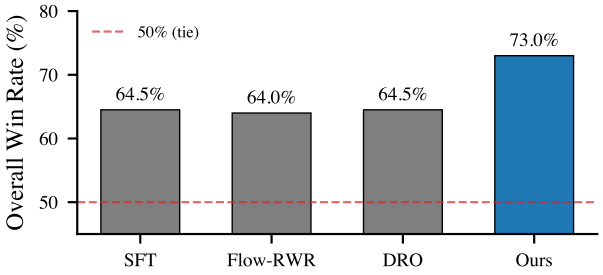
Descriptor and Metric Analysis: While SEA-Raft achieves highest visual quality (80.3%), it hacks the reward by preferring oversaturated scenes. LightGlue finds correspondences in clean areas when videos contain artifacts, resulting in misleadingly low epipolar error, whereas we want correspondences across the entire scene so artifacts anywhere produce high error. All setups are comparable, but classical geometric constraints provide cleaner optimization signals than sophisticated alternatives that can miss global inconsistencies.

Comparison with Learnable Metrics: Models trained with VideoReward Motion Quality achieve 61.3% VideoReward but 0.179 Sampson error, while ours achieve 64.3% VideoReward and 0.131 Sampson error, superior on both. Similarly, MET3R-trained models achieve 0.049 MET3R but 0.176 Sampson error, while Sampson-trained models match this MET3R score with far better Sampson performance (0.131). This confirms learnable metrics produce insufficient preference signals compared to classical geometry constraints.

Static Penalty Analysis: The temporal variation penalty achieves superior motion dynamics (dynamic degree 0.710 vs 0.627) while maintaining comparable geometric consistency and motion quality. This



(a) **Descriptor Ablation:** Win rate vs Baseline. Simple SIFT with Sampson error achieves balanced performance across all metrics, while sophisticated descriptors (SEA-Raft) overfit to visual quality.



(b) **Alignment Method Comparison:** Overall win rate on VideoReward benchmark. Our epipolar-based DPO significantly outperforms competitive alternative alignment strategies.

Table 7: **Static Penalty Ablation:** Adding static penalty significantly improves dynamic degree while only slightly sacrificing the consistency.

Method	Dynamic Degree	Motion Quality	Sampson Error
Ours	0.627	**71.5%**	**0.127**
Ours + Static Penalty	**0.710**	69.5%	0.131

component effectively prevents degenerate static solutions while preserving the geometric alignment. We observe that large λ causes the model to always increase camera motion, hurting both 3D consistency and generalization. Similarly, decreasing λ forces the model to optimize for a naive solution of increasing 3D consistency by reducing motion.

Alignment Method Comparison: Our approach outperforms all alternatives with highest win rates on VideoReward (Figure 6b), validating effectiveness of DPO for video model optimization. We compare the alignment against two simple methods: Supervised Finetuning on Win samples, and Flow-RWR (Liu et al., 2025) which weights the loss with the normalized reward. All strategies are trained with the same number of steps and target metric. We also find DRO (Li et al., 2025b) interesting as it does not require regularization with the reference model, but in our case, it quickly leads to degenerate solutions deviating from the base model and producing strong artifacts. The other two setups, while still effective, can not learn from the gap between consistent and inconsistent trajectories, resulting in less 3D consistent outputs. The evaluation also underscores the need to evaluate the model with multiple metrics. For example, learnable metrics are subject to overfitting: a model with strong artifacts still can get high scores from a VLM-based metric.

5 Conclusion

We present an approach for enhancing 3D consistency in video diffusion models by leveraging epipolar geometry constraints as preference signals. Our work shows that classical geometric constraints provide more stable optimization signals than learned metrics, which produce noisy targets that compromise alignment quality. Training on static scenes generalizes effectively to diverse dynamic content, demonstrating the broad applicability of geometric principles. The resulting models generate videos with fewer geometric inconsistencies and more stable camera trajectories while preserving creative flexibility. This work shows that classical computer vision algorithms effectively complement deep learning, addressing limitations of purely data-driven methods through mathematically grounded geometric principles.

References

Ackermann, J., Cai, S., Deng, B., Kuang, Z., Peng, S., and Wetzstein, G. Geoflow: Enforcing implicit geometric consistency in video generation. *arXiv preprint arXiv:2605.18365*, 2026.

Albergo, M. S. and Vanden-Eijnden, E. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

An, Z., Kupyn, O., Uscidda, T., Colaco, A., Ahuja, K., Belongie, S., Gonzalez-Franco, M., and Gazulla, M. T. Vggrpo: Towards world-consistent video generation with 4d latent reward. *arXiv preprint arXiv:2603.26599*, 2026.

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023a.

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023b.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Chen, Z., Wang, Y., Wang, F., Wang, Z., and Liu, H. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024.

Cheng, J., Lyu, R., Gu, X., Liu, X., Xu, J., Lu, Y., Teng, J., Yang, Z., Dong, Y., Tang, J., et al. Vpo: Aligning text-to-video generation models with prompt optimization. *arXiv preprint arXiv:2503.20491*, 2025.

Clark, K., Vicol, P., Swersky, K., and Fleet, D. J. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

Du, H., Ye, J., Cong, X., Li, R., Ni, J., Agarwal, A., Zhou, Z., Li, Z., Balestriero, R., and Wang, Y. Videogpa: Distilling geometry priors for 3d-consistent video generation. *arXiv preprint arXiv:2601.23286*, 2026.

Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., and Lee, K. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS) 2023*. Neural Information Processing Systems Foundation, 2023.

Fischler, M. Rc bolles random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography., 1981, 24. *DOI: https://doi. org/10.1145/358669.358692*, pp. 381–395.

Google DeepMind. Veo 2, 12 2024. URL `https://deepmind.google/technologies/veo/veo-2/`. Accessed: 2024.

HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

He, H., Yang, C., Lin, S., Xu, Y., Wei, M., Gui, L., Zhao, Q., Wetzstein, G., Jiang, L., and Li, H. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*, 2025.

Hou, C., Wei, G., Zeng, Y., and Chen, Z. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024.

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.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Jiang, Z., Zheng, C., Laina, I., Larlus, D., and Vedaldi, A. Geo4d: Leveraging video generators for geometric 4d scene reconstruction. *arXiv preprint arXiv:2504.07961*, 2025.

Ju, X., Gao, Y., Zhang, Z., Yuan, Z., Wang, X., Zeng, A., Xiong, Y., Xu, Q., and Shan, Y. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

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. *arXiv preprint arXiv:2509.13414*, 2025.

Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Li, C., Michel, O., Pan, X., Liu, S., Roberts, M., and Xie, S. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. *arXiv preprint arXiv:2503.09595*, 2025a.

Li, R., Zheng, C., Rupprecht, C., and Vedaldi, A. Dso: Aligning 3d generators with simulation feedback for physical soundness. *arXiv preprint arXiv:2503.22677*, 2025b.

Li, Z., Guo, H., Teng, R., Shen, C., and He, T. Geo-align: Video generation alignment via metric geometry reward, 2026. URL `https://arxiv.org/abs/2605.23903`.

Lindenberger, P., Sarlin, P.-E., and Pollefeys, M. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17627–17638, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, C. and Vahdat, A. Equivdm: Equivariant video diffusion models with temporally consistent noise. *arXiv preprint arXiv:2504.09789*, 2025.

Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Qin, W., Xia, M., et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022a.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lowe, D. G. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pp. 1150–1157. Ieee, 1999.

LumaLabs. Dream machine, 06 2024. URL `https://lumalabs.ai/dream-machine`. Accessed: 2024.

OpenAI. Video generation models as world simulators, 2024. URL `https://openai.com/index/video-generation-models-as-world-simulators/`. Accessed: 2024.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

PikaLabs. Pika 1.5, 10 2024. URL https://pika.art/. Accessed: 2024.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Polyak, A. et al. Movie gen: A cast of media foundation models, 2025. URL https://arxiv.org/abs/2410.13720.

Potje, G., Cadar, F., Araujo, A., Martins, R., and Nascimento, E. R. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2682–2691, 2024.

Prabhudesai, M., Goyal, A., Pathak, D., and Fragkiadaki, K. Aligning text-to-image diffusion models with reward backpropagation. 2023.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Runway. Gen-3, 06 2024. URL https://runwayml.com/. Accessed: 2024.

Sampson, P. D. Fitting conic sections to "very scattered" data: An iterative refinement of the bookstein algorithm. *Computer graphics and image processing*, 18(1):97–108, 1982.

Sarkar, A., Mai, H., Mahapatra, A., Lazebnik, S., Forsyth, D. A., and Bhattad, A. Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28140–28149, 2024.

Schuhmann, C. Laion-aesthetics, 2022. URL https://laion.ai/blog/laion-aesthetics/. Accessed: 2023-11-10.

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 arXiv:2402.03300*, 2024.

Sun, J., Shen, Z., Wang, Y., Bao, H., and Zhou, X. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8922–8931, 2021.

Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–12, 2023.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Voleti, V., Yao, C.-H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., and Jampani, V. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pp. 439–457. Springer, 2024.

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., and Naik, N. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025a.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., and Novotny, D. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025b.

Wiedemer, T., Li, Y., Vicol, P., Gu, S. S., Matarese, N., Swersky, K., Kim, B., Jaini, P., and Geirhos, R. Video models are zero-shot learners and reasoners, 2025. URL `https://arxiv.org/abs/2509.20328`.

Wu, T., Si, C., Jiang, Y., Huang, Z., and Liu, Z. Freeinit: Bridging initialization gap in video diffusion models. In *European Conference on Computer Vision*, pp. 378–394. Springer, 2024.

Yang, H., Chen, Y., Pan, Y., Yao, T., Chen, Z., Ngo, C.-W., and Mei, T. Hi3d: Pursuing high-resolution image-to-3d generation with video diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6870–6879, 2024.

Yin, T., Shi, J., Guo, H., and Wang, X. Vigor: Video geometry-oriented reward for temporal generative alignment. *arXiv preprint arXiv:2603.16271*, 2026.

Zhang, Q., Zhai, S., Bautista, M. A., Miao, K., Toshev, A., Susskind, J., and Gu, J. World-consistent video diffusion with explicit 3d modeling. *arXiv preprint arXiv:2412.01821*, 2024.

Zheng, G., Li, T., Jiang, R., Lu, Y., Wu, T., and Li, X. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

Zheng, G., Li, T., Zhou, X., and Li, X. Realcam-vid: High-resolution video dataset with dynamic scenes and metric-scale camera movements. *arXiv preprint arXiv:2504.08212*, 2025.

Zhou, J. J., Gao, H., Voleti, V., Vasishta, A., Yao, C.-H., Boss, M., Torr, P., Rupprecht, C., and Jampani, V. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025.

Zhou, T., Tucker, R., Flynn, J., Fyffe, G., and Snavely, N. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Table 8: **Image-to-Video Alignment:** Despite image conditioning constraints, epipolar alignment shows consistent improvements across multiple metrics.

	VideoReward		3D Reconstruction			3D Consistency		VBench Metrics				
Method	Visual Quality	Motion Quality	PSNR ↑	SSIM ↑	LPIPS ↓	Motion (SSIM ↓)	Sampson Error ↓	Background Consistency	Aesthetic Quality	Temporal Flickering	Motion Smoothness	Dynamic Degree
Baseline	-	-	**21.08**	0.686	0.408	0.239	0.215	0.955	0.498	**0.981**	0.992	**0.378**
Ours	**51.35%**	**56.08%**	20.99	**0.700**	**0.377**	**0.239**	**0.197**	**0.955**	**0.499**	0.980	**0.992**	0.343

A Image-to-Video Evaluation

Image-to-video alignment presents unique challenges due to the strong conditioning signal from the input image. The image conditioning is integrated into intermediate layers of the diffusion process, creating additional constraints that naturally reduce output variance and make alignment more challenging. Despite these limitations, our epipolar geometry optimization demonstrates consistent positive impact across multiple evaluation dimensions.

The 3D reconstruction results validate the geometric improvements: SSIM improves from 0.686 to 0.700, and LPIPS decreases from 0.408 to 0.377. These gains, while more modest than text-to-video results, confirm that enhanced geometric consistency translates to better downstream 3D understanding even under image conditioning constraints. The Sampson epipolar error improvement from 0.215 to 0.197 further validates the effectiveness of classical geometric alignment.

VideoReward metrics show meaningful improvements in motion quality (56.08% vs 43.92%) and visual quality (51.35% vs 48.65%). VBench metrics remain stable with slight improvements in aesthetic quality, demonstrating that geometric optimization preserves overall generation quality while enhancing 3D consistency.

While the input image provides strong structural guidance, it also constrains the model's ability to adapt toward geometrically optimal solutions. Nevertheless, consistent positive trends across reconstruction, consistency, and quality metrics validate that classical geometric constraints provide reliable optimization signals even in constrained generation scenarios.

B Prompt Optimization Evaluation

We further compare our method with VPO Cheng et al. (2025), a video prompt optimization technique, which is complementary to our approach since it optimizes prompts rather than model weights. We evaluate VPO alone and in combination with our method. Results are reported in Table 9.

Table 9: **Prompt Optimization:** VPO optimizes prompts while our method improves geometry alignment. The two approaches are complementary and can be combined to achieve both high visual quality and geometric consistency.

Method	**Visual Quality ↑**	**Motion Quality ↑**	**Overall ↑**	**Dynamic Degree ↑**	**Motion (mean SSIM) ↓**
Ours	63.1%	65.8%	59.1%	**0.80**	**0.211**
VPO	59.1%	70.6%	82.7%	0.65	0.235
VPO + Ours	**67.0%**	**71.9%**	**83.6%**	0.61	0.234

We observe that VPO tends to reduce camera motion and restructure prompts while optimizing for general video quality. However, such prompt optimization methods can be efficiently combined with geometry-aligned models like ours to simultaneously achieve high visual quality and geometric consistency.

C Scaling Analysis

To understand how our geometric alignment performs across different model scales, we compare both the baseline and epipolar-aligned 1.3B parameter models against the much larger Wan-2.1-14B model Wang et al.

(2025a). As shown in Table 4, while the performance gap remains substantial due to the 14B model's higher resolution (720p) and superior base capabilities, our epipolar alignment helps close this gap meaningfully. The aligned 1.3B model achieves win rates of 18.1%, 21.8%, and 25.0% for Visual Quality, Motion Quality, and Text Alignment respectively, compared to 13.3%, 14.4%, and 24.2% for the baseline 1.3B model. Notably, the 14B model requires approximately 10× longer inference time than the 1.3B variant, making our alignment approach particularly valuable for applications where computational efficiency is critical. This suggests that geometric consistency improvements can partially compensate for scale limitations, offering a practical path toward better video quality without the computational overhead of significantly larger models.

D Qualitative Evaluation

For comprehensive assessment of video quality and geometric consistency, we include an interactive webpage in the supplementary materials where readers can view the full video sequences and directly compare the baseline and epipolar-aligned model outputs.

E Limitations

Our approach primarily focuses on static scenes with dynamic camera movements, aligning well with applications in 3D reconstruction and novel view synthesis. Adapting this method to scenes with dynamic objects would require modifying the training pipeline to separately model and evaluate object motion and camera movement. Additionally, epipolar geometry constraints assume point correspondences coming from a static scene under camera motion, limiting effectiveness for scenes with independent object movement or non-rigid deformations where a single fundamental matrix cannot explain all correspondences. The data mining relies on feature matching and epipolar geometry scoring. However, it can produce false positives (assigning low error to geometrically inconsistent videos) and false negatives (high error to consistent ones) when scenes have repetitive textures, lack distinctive features, or contain significant motion blur. In addition, while aigned model significantly improves geometric consistency, it does not solve all failure modes. Complex dynamic scenes with extreme camera motion or highly ambiguous content remain challenging for both the baseline and our approach.

Impact Statement

Video generation models may be misused to produce realistic but deceptive content, contributing to the spread of misinformation, political manipulation, and erosion of public trust. Furthermore, the computational resources required to train such models raise environmental concerns and may exacerbate inequalities in access to advanced AI technologies. Geometry-aware video generation can facilitate various 3D vision tasks, including scene reconstruction, SLAM, and visual odometry. By improving geometric consistency in generated videos, our method produces more realistic and usable synthetic data for training computer vision systems. This advances applications in robotics and autonomous navigation, where accurate spatial understanding is crucial. The integration of classical geometry principles with modern generative models represents a promising direction for enhancing AI systems with stronger physical world understanding.