Epipolar Geometry Improves Video Generation Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Video generation models have progressed tremendously through large latent diffusion transformers trained with rectified flow techniques. Yet, despite these advances, 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. This work explores how simple epipolar geometry constraints can improve modern video diffusion models trained on internet-scale datasets. Despite their massive training data, these models often fail to capture the fundamental geometric principles underlying all visual content. While traditional computer vision methods are often non-differentiable and computationally expensive, they provide reliable, mathematically grounded signals for 3D consistency evaluation. We demonstrate that aligning diffusion models through a preference-based optimization framework using pairwise epipolar geometry constraints yields videos with superior visual quality, enhanced 3D consistency, and significantly improved motion stability. Our approach offers an efficient alignment strategy that enforces established geometric principles without requiring end-to-end differentiability. Evaluation shows that our method outperforms baseline models and alternative alignment approaches across various metrics. By bridging the gap between data-driven deep learning and classical geometric computer vision, we present a practical method for generating more spatially consistent videos without compromising visual quality or requiring explicit 3D supervision.

1 Introduction

Video generation has witnessed remarkable progress in recent years, with newer models [1–6] producing increasingly realistic content from text and image conditions. This rapid advancement has spurred researchers to repurpose these powerful video models for broader applications, including animation [7], virtual worlds generation [8], and novel view synthesis [9].

Video diffusion models are trained on vast volumes of data, encoding rich priors about the visual world and its dynamics. Through extensive training, these models develop a strong understanding of object appearance, motion patterns, and scene composition. As a result, many recent works aim to utilize the priors from latent video diffusion models in various downstream tasks [10–12]. Despite this remarkable progress, these models still struggle to maintain perfect 3D consistency throughout generated sequences. Current video models often produce content with geometric inconsistencies, unstable motion, and perspective flaws, even though almost all training data is 3D consistent. Some approaches for enhancing 3D consistency rely on noise optimization [13], explicit guidance through point clouds [14, 15], or camera parameters [16]. Nevertheless, inaccurate control signals can constrain the model's generative capabilities, and the latent space optimization typical in diffusion training makes it difficult to compute direct geometric losses on the final outputs.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

With the rising popularity of reinforcement learning for model alignment [17–19], post-training alignment has recently gained more attention in diffusion model research as an alternative approach to improve model capabilities. Methods such as VideoReward [20] have finetuned vision-language models on a large-scale human preference data, enabling direct supervision through the reward model. However, it relies on human-annotated motion quality scores (1 to 5), which can introduce noisy signals into the training process and are expensive to collect. Human judgments about video quality are inherently subjective and may not consistently capture the geometric principles that ensure proper 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-differentiable geometric constraints as reward signals in a preference-based finetuning framework. Specifically, we leverage an epipolar geometry constraint: assessing 3D consistency between frames. By sampling multiple videos conditioned on the same prompt, we generate diverse camera trajectories that typically vary in geometric coherence. The quality of these trajectories is well-captured by epipolar geometry metrics, providing a reliable signal for identifying which generations better adhere to projective geometry principles. This insight enables us to rank videos based on their adherence to epipolar constraints, creating training pairs that guide the model toward improved geometric consistency.

Our method implements this through Direct Preference Optimization (DPO) [17], requiring only relative rankings rather than absolute reward values. This approach bypasses the difficulties of directly using non-differentiable computer vision algorithms in the training loop. DPO only needs to determine which output better adheres to the geometric constraints. By finetuning the model to prioritize generations that satisfy these classical geometric constraints, we guide it towards generating inherently more 3D-consistent videos, without restricting its creative capabilities or requiring explicit 3D supervision. As shown in Figure 1, this results in enhanced 3D consistency, smoother camera trajectories, and fewer artifacts compared to the baseline model.

While simple in nature, this paper shows that a basic geometric constraint, described in 1982 [21], 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 the Sampson distance to enhance 3D video consistency without needing differentiability. The models finetuned with the simple yet reliable signal from classical computer vision algorithms achieve superior consistency and quality, significantly reducing artifacts and unstable motion trajectories in generated content. Our approach demonstrates that aligning models with fundamental geometric principles leads to visually superior results while preserving the model's ability to generate diverse and creative content.

Comprehensive Evaluation Framework: We develop an extensive evaluation protocol that measures both perceptual quality and 3D consistency and adherence to projective geometry principles across diverse generation scenarios. We evaluate text and image-to-video finetuned models, exploring the impact of geometry-aware finetuning on a large set of metrics.

Large-Scale Preference Dataset: We create and release a large dataset of over 162,000 generated videos annotated with 3D scene consistency metrics, enabling further research in geometry-aware video generation. This dataset includes diverse prompts spanning natural landscapes, architectural scenes, and dynamic environments, each with multiple video generations.

2 Related Work

We structure the related work section into generative models and post-training methods to adapt them.

2.1 Video Generation Models

Recent advances in video generation have been dominated by closed-source models developed by well-resourced technology companies. These models, trained on large proprietary datasets with

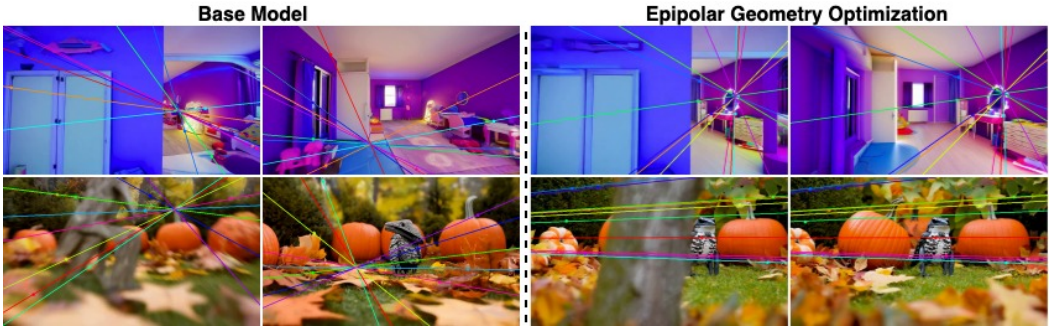

Figure 1: First and middle frame from videos generated by baseline and our epipolar-aligned model. The baseline model produces geometrically inconsistent outputs with artifacts and unnatural motion trajectories visible in distorted structures. Our model, finetuned with epipolar error, generates visibly improved results with smoother camera trajectories, reduced artifacts, and enhanced 3D consistency.

computational resources beyond academic reach, have demonstrated remarkable capabilities while revealing limited architectural details. Notable releases include OpenAI's Sora [1], which marked a significant leap in long-form video synthesis; Runway's Gen-2 and Gen-3 [22]; Luma AI's video models [23]; Pika Labs models [6]; and Google DeepMind's Veo series [2]. While these systems produce impressive results, their closed nature limits opportunities to finetune them or apply them to other vision tasks.

More recently, open-source large latent diffusion models have become available, increasing interest in improving video generators. Stable Video Diffusion [24] developed efficient training strategies for latent video diffusion. Hunyan-Video [5] presented a systematic approach to scaling models, LTX-Video [25] introduced optimizations for real-time generation, and Wan-2.1 [4] introduced an efficient 3D Variational Autoencoder [26] with expanded training pipelines. Wan-2.1 offers models for text-to-image and video-to-image in 1.3B and 14B parameter versions, enabling researchers to explore adaptation techniques for various downstream tasks.

These video diffusion models are trained on enormous data volumes covering more content variety than specific applications need, making domain-aware alignment valuable for specialized tasks. Geometry-aware finetuning allows general-purpose models to maintain creative flexibility while ensuring adherence to physical principles like 3D consistency. V3D [12] finetunes models to generate 360 orbit frames for 3D reconstruction, while VideoReward [20] introduced a framework for reinforcement learning-based video model alignment. However, prior methods rely on subjective human preferences or vision language models [27] trained to mimic them. In contrast, our approach optimizes against mathematical rules from epipolar geometry, providing a clean signal that aligns models with fundamental 3D consistency principles rather than subjective judgments.

2.2 Diffusion Models Alignment

Since image and video latent diffusion models are trained on internet-scale noisy data, efficient finetuning, and alignment strategies have emerged as an active research area. Latent image diffusion models [28, 29] finetune models on data highly ranked by the aesthetics classifier [30]. DRAFT [31] and AlignProp [32] further explore this paradigm by tuning the diffusion model to maximize the reward function directly. DPOK [33] and DDPO [34] expand the paradigm to introduce distributional constraints. Diffusion-DPO [35] introduces the Direct Preference Optimization algorithm into diffusion model alignment. In contrast to other approaches, DPO does not require direct access to the reward model and can be trained with only pairwise preference data. Additionally, this eliminates the need to decode the final denoised sample, which can be finetuned directly in latent space, significantly improving training efficiency. Recently, VideoReward [20] adapted Diffusion-DPO for video alignment, effectively aligning video generation with human preferences. Yet, all these approaches focus on optimizing for subjective and noisy human evaluation. Lately, DSO [36] employs DPO to align 3D generators with physical soundness, and PISA explicitly [37] improves the physical stability of video generators with a multi-component reward function. Our method leverages classical computer vision algorithms to provide objective, mathematically grounded preference signals based

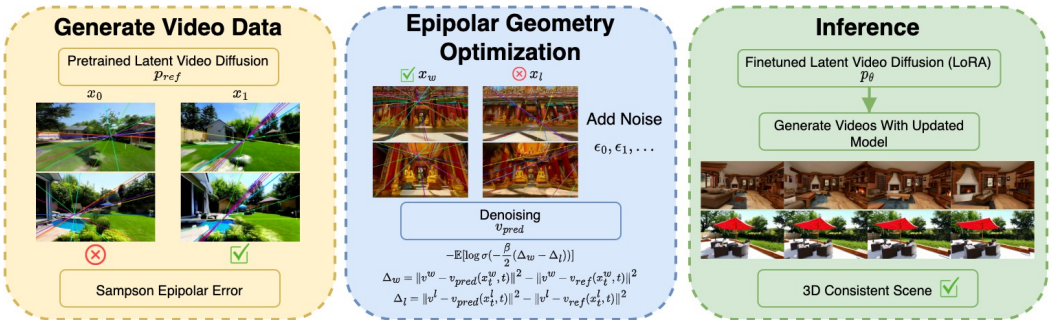

Figure 2: **Epipolar Geometry Optimization pipeline.** Our approach: (1) Generate diverse videos using pretrained generators [4] and leverage the Sampson epipolar error to identify 3D consistent vs. inconsistent samples; (2) Train policy p_θ using Flow-DPO [20] to prefer geometrically consistent outputs; (3) Apply the updated policy to enhance 3D consistency in the base video diffusion model.

on epipolar geometry, resulting in more reliable and consistent alignment with 3D physical principles. However, creating an explicit, robust, differentiable geometry reward model is challenging due to the complexity of accurately modeling and evaluating 3D consistency across diverse scenes. Our method addresses these challenges by leveraging classical computer vision algorithms to provide objective, physically grounded preferences based on epipolar geometry, resulting in consistent alignment.

3 Method

We aim to align pretrained video diffusion models to generate geometrically consistent 3D scenes from text or image prompts. To address this, we propose an alignment strategy that leverages classical epipolar geometry constraints within a preference-based optimization framework. Traditional reinforcement learning approaches [18, 38] require explicit reward functions and access to final samples which is impractical for video models due to the absence of robust differentiable reward models and the prohibitive computational cost of the denoising process. Our key observation is that while classical epipolar geometry constraints do not produce a smooth, globally comparable loss surface across different scene types (e.g., indoor vs. outdoor scenes may exhibit different absolute error magnitudes due to variations in matchable feature counts), the relative intra-prompt error measurements remain consistent. When generating multiple video sequences with fixed conditioning, the stochastic nature of diffusion sampling produces outputs with varying degrees of geometric consistency. Epipolar error metrics are an effective tool to quantify relative 3D consistency, with higher values reliably indicating lower geometric consistency. This finding aligns with the direct preference optimization (DPO) paradigm, which requires only a relative metric to determine preference between output pairs rather than absolute reward values. The pairwise comparison nature of DPO eliminates the need for a globally normalized reward function, instead leveraging the 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} that takes a text prompt and an optional first frame conditioning I and generates video samples $x_0 \sim p_{\text{ref}}(x_0|T, I^*)$, where $I^* \in \{I, \emptyset\}$ we want to learn the model p_θ which is optimized to generate 3D-consistent video sequences. The one choice would be to optimize it with the following objective:

$$\max_\theta \mathbb{E}_{(T, I^* \in \{I,\emptyset\}) \sim \mathcal{D}_c, x_0 \sim p_\theta(x_0|T,I^*)} \left[r(x_0) \right] - \beta \mathbb{D}_{\text{KL}} \left[p_\theta(x_0|T, I^*) \| p_{\text{ref}}(x_0|T, I^*) \right], \quad (1)$$

where $r(x_0)$ is a reward function that outputs 3D consistency scores, $\mathcal{D}_c$ are samples from the reference model. The reward is maximized while the optimized model p_θ is kept close to the reference model p_{ref} via a KL-divergence term weighted by the hyperparameter β. This formulation directly encourages the model to generate videos with improved geometric consistency. However,

this formulation presents a few critical practical challenges. First, the reward function $r(x_0)$ relies on classical computer vision algorithms that are non-differentiable, making direct gradient-based optimization infeasible. Second, evaluating this reward function requires complete video generation and subsequent geometric analysis, which is highly time-consuming for training large video diffusion models. These constraints make traditional reinforcement learning approaches impractical for our setting, and motivate our adoption of Direct Preference Optimization (DPO) [17, 35], which was originally designed for scenarios where direct reward optimization is similarly challenging.

Assuming a fixed dataset of $\mathcal{D}(\{x, x_0^w, x_0^l\})$ which consists of condition c (text, image), and a pair of samples from the p_{ref} such that x_0^w has higher reward value than x_0^l ($x_0^w \succ x_0^l$).

Diffusion-DPO [35] aligns diffusion models with human preferences by directly solving eq. (1) analytically. It interprets alignment as a classification problem and optimizes a policy to satisfy the preferences through supervised training. The Diffusion-DPO objective $\mathcal{L}_{\text{DD}}(\theta)$ is given by:

$$-\mathbb{E}\Bigg[\log\sigma\Bigg(-\frac{\beta}{2}\Big(\|\epsilon^w-\epsilon_\theta(\mathbf{x}_t^w,t)\|^2-\|\epsilon^w-\epsilon_{\text{ref}}(\mathbf{x}_t^w,t)\|^2 \\ -\big(\|\epsilon^l-\epsilon_\theta(\mathbf{x}_t^l,t)\|^2-\|\epsilon^l-\epsilon_{\text{ref}}(\mathbf{x}_t^l,t)\|^2\big)\Big)\Bigg)\Bigg], \tag{2}$$

where $x_t^* = (1-t)\, x_0^* + t\, \epsilon^*, \epsilon^* \sim \mathcal{N}(0, \mathbf{I})$. The superscript $^* \in \{w, l\}$ denotes either w for the sample with a higher score or l for a sample with a lower score, ϵ^* is a ground truth or predicted noise by a diffusion model. The expectation is taken over samples $\{\mathbf{x}_0^w, \mathbf{x}_0^l\} \sim \mathcal{D}$ and the noise schedule *t*.

For rectified flow models [39–41] the noise vector ϵ^* is related to the velocity field v^* following [20]:

$$\|\epsilon^*-\epsilon_{pred}(\mathbf{x}_t^*,t)\|^2 = (1-t)^2\|v^*-v_{pred}(\mathbf{x}_t^*,t)\|^2\,. \tag{3}$$

The final Flow-DPO loss [20] is formulated as:

$$-\mathbb{E}\Bigg[\log\sigma\Bigg(-\frac{\beta_t}{2}\Big(\|v^w-v_\theta(\mathbf{x}_t^w,t)\|^2-\|v^w-v_{\text{ref}}(\mathbf{x}_t^w,t)\|^2 \\ -\big(\|v^l-v_\theta(\mathbf{x}_t^l,t)\|^2-\|v^l-v_{\text{ref}}(\mathbf{x}_t^l,t)\|^2\big)\Big)\Bigg)\Bigg], \tag{4}$$

where $\beta_t = \beta(1-t^2)$.

Intuitively, minimizing this loss encourages the model to improve its denoising performance on preferred samples $\mathbf{x}_t^w$ relative to less preferred samples $\mathbf{x}_t^l$ [20, 35]. This guides the predicted velocity field v_θ to align more closely to videos exhibiting better 3D consistency while diverging from those with poorer geometric coherence.

3.2 3D Consistency Metric

We evaluate the 3D consistency of generated videos by validating how well they satisfy epipolar geometry constraints. Epipolar geometry represents the intrinsic projective relationship between two views of the same scene, depending only on the camera's internal parameters and relative positions. In perfectly consistent 3D scenes, corresponding points across different viewpoints must adhere to these geometric constraints.

For any two corresponding points $\mathbf{x}$ in one frame and $\mathbf{x}'$ in another, the epipolar constraint $\mathbf{x}'^T\mathbf{F}\mathbf{x} = 0$ must be satisfied, where $\mathbf{F}$ is the fundamental matrix. This constraint ensures that a point in one view must lie on its corresponding epipolar line in the other view. The fundamental matrix encapsulates the geometric relationship between the two camera poses. It can be formulated as $\mathbf{F} = [\mathbf{e}']_\times \mathbf{P}'\mathbf{P}^+$, where $\mathbf{P}$ and $\mathbf{P}'$ are the camera projection matrices, $\mathbf{P}^+$ is the pseudo-inverse of $\mathbf{P}$, and $\mathbf{e}'$ is the epipole in the second view.

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 [42] feature matching. While we validate the method with a simple, robust

handcrafted descriptor, the pipeline can also leverage more recent learned descriptors [43–46]. 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 [47] framework to handle outliers.

Once we have estimated the fundamental matrix, we can measure the geometric consistency using the Sampson epipolar error [21]:

$$S_E = \frac{(\mathbf{x}'^T \mathbf{F} \mathbf{x})^2}{(\mathbf{F}\mathbf{x})_1^2 + (\mathbf{F}\mathbf{x})_2^2 + (\mathbf{F}^T\mathbf{x}')_1^2 + (\mathbf{F}^T\mathbf{x}')_2^2} \tag{5}$$

The Sampson error 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.

3.3 Implementation Details

We conduct experiments with a state-of-the-art open-source video diffusion model called Wan2.1 [4], which possesses 1.3 billion parameters. Our approach is validated in text-to-video and image-to-video generation setups to demonstrate versatility across conditioning types.

Offline Dataset Generation Since our method focuses on 3D-consistent scene generation, we require videos of static scenes with dynamic camera movements. We extract text prompts from the DL3DV [48] and RealEstate10K [49] datasets, provided by [50], containing a wide variety of indoor and outdoor scenes. We generate three videos per caption to ensure sufficient variation in 3D consistency quality, as our preliminary experiments showed that pairs generated from just two samples often lacked meaningful geometric differences. This approach balances computational efficiency with training data quality. We filter put near-static videos to prevent the model from learning a degenerate solution of minimizing camera movement to satisfy epipolar constraint. In total, we generate 24,000 videos for text-to-video and 30,000 videos for image-to-video training, requiring approximately 1,980 GPU hours on NVIDIA A6000s.

Training Configuration Given the computational demands of fine-tuning large video diffusion models, we implement our approach using Low-Rank Adaptation (LoRA) [51] with rank $r = 64$ and $\alpha = 128$. This strategy offers the additional benefit of eliminating the need to store the reference model separately 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 [52] 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 assess the effectiveness of our epipolar-aligned video diffusion model compared to baseline approaches and evaluate its impact on scene consistency, visual quality, and prompt alignment. Our evaluation setup consists of 200 videos extracted from the test sets of DL3DV [48] and RealEstate10K [49] datasets, covering a diverse range of indoor and outdoor scenes. To thoroughly test geometric consistency under challenging conditions, we amplify the complexity of camera motion by augmenting prompts with motion-specific phrases (e.g., "orbiting around," "zooming in," "panning across"). We evaluate our model across three complementary benchmarks: (1) the VideoReward benchmark [20], which measures general video generation quality; (2) VBench [53], which provides standardized metrics for temporal consistency and visual fidelity; and (3) our custom suite of 3D consistency metrics based on epipolar geometry constraints. This multi-protocol evaluation approach allows us to comprehensively assess the generated videos' perceptual quality and geometric consistency.

Figure 3 shows some qualitative examples. Before our fine-tuning, the videos often contain morphing objects or inconsistent geometry.

4.1 VideoReward Benchmark Evaluation

The VideoReward [20] benchmark evaluates videos across Visual Quality, Motion Quality, and Text Alignment dimensions using a vision language model [27] finetuned on human preferences.

Figure 3: **Qualitative Evaluation:** Visual comparison between the videos generated by the base and finetuned model. First two rows: Wan-2.1-T2V [4], 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.

Table 1: **Win-rate vs. original model** on the VideoReward [20] benchmark compared to a learned metric [54].

Text-to-Video				
Method	Visual Quality	Motion Quality	Text Alignment	Overall
DPO-MET3R [54]	56.5%	64.5%	44.0%	55.0%
DPO-Epipolar	**72.0%**	**71.0%**	**55.0%**	**73.0%**
Image-to-Video				
Method	Visual Quality	Motion Quality	Text Alignment	Overall
DPO-MET3R [54]	47.02%	51.19%	**54.76%**	48.21%
DPO-Epipolar	**51.35%**	**56.08%**	49.32%	**52.02%**

Annotators select preferences between video pairs, and a VLM simulates these judgments. We use the resulting pairwise scores to compute win rates of our finetuned model versus the baseline. We also compare against a model aligned with MET3R [54] to assess how our epipolar geometry metric compares to modern 3D vision metrics [55]. Table 1 presents the results of this evaluation. Our text-to-video model significantly outperforms both the baseline and MET3R-based models across all metrics, with win rates of 72.0%, 71.0%, and 55.0% for Visual Quality, Motion Quality, and Text Alignment respectively. This demonstrates that alignment with epipolar constraints enhances not only motion quality but also visual fidelity by reducing artifacts. The image-to-video model, trained with more conservative hyperparameters to minimize baseline deviation, still shows meaningful improvements over both the baseline and MET3R-aligned models in most categories.

4.2 VBench Benchmark Evaluation

VBench [53] introduces a comprehensive benchmark suite for video generative models. It consists of a large set of metrics across multiple dimensions, facilitating fine-grained and objective evaluation. We provide the results on five metrics related to visual and motion quality. **Background Consistency** evaluates the temporal consistency of the background scenes by calculating CLIP [56] feature similarity across frames. **Aesthetic Quality** evaluates the artistic and beauty value humans perceive towards each video frame using the LAION aesthetic predictor [30], measuring such concepts as layout and photo-realism. **Temporal Flickering** extracts static frames and computes the mean absolute difference across frames. **Motion Smoothness** validates whether generated motion follows

Table 2: Results on the VBench [53] metrics comparing our epipolar-aligned model against the original model.

Text-to-Video					
Method	Background Consistency	Aesthetic Quality	Temporal Flickering	Motion Smoothness	Dynamic Degree
Baseline	0.930	0.541	0.958	0.981	**0.815**
Ours	**0.942**	**0.551**	**0.969**	**0.984**	0.627
Image-to-Video					
Method	Background Consistency	Aesthetic Quality	Temporal Flickering	Motion Smoothness	Dynamic Degree
Baseline	**0.955**	0.498	**0.981**	**0.992**	**0.378**
Ours	**0.955**	**0.499**	0.980	**0.992**	0.343

Table 3: 3D consistency metrics comparing our epipolar-aligned model against the baseline and MET3R [54] approach.

Text-to-Video				
Method	Motion (mean SSIM)	Perspective Fields	Sampson Distance	MET3R
Baseline	0.233	0.426	0.190	0.050
DPO-MET3R [54]	0.232	**0.438**	0.176	**0.049**
DPO-Epipolar	**0.223**	0.428	**0.127**	**0.049**
Image-to-Video				
Method	Motion (mean SSIM)	Perspective Fields	Sampson Distance	MET3R
Baseline	0.239	0.504	0.215	**0.048**
DPO-MET3R [54]	**0.220**	**0.517**	0.202	0.049
DPO-Epipolar	0.239	0.515	**0.197**	0.049

the physical law of the real world. It utilizes the motion priors in the video frame interpolation model [57] to evaluate the smoothness of generated motions. Finally, **Dynamic Degree** employs RAFT [58] to estimate the degree of dynamics in synthesized videos. The results are presented in Table 2. We compare the finetuned Wan-2.1 [4] models to the baseline. The text-to-video model improves the scores across all metrics; however, it sacrifices the dynamic degree by a bit. Nevertheless, the other benchmarks Table 3 and Table 1 demonstrate that the finetuned model generates comparable or superior dynamics. The image-to-video model, being finetuned, reduces the amount of edge cases that perform comparably to the baseline.

4.3 3D Geometry Evaluation

Last, we evaluate the direct impact of the finetuned models on 3D geometry metrics. Table 3 shows results across multiple geometric consistency measures. We assess Sampson error (our primary optimization target), MET3R score [54], the realism of Perspective Fields [59] and Motion Level. Perspective Fields classifier [60] evaluates the realism of image perspective fields. Since the metric is image-level, we compute the mean metric across all frames. Additionally, we validate whether the models tend to produce nearly static videos by computing the mean SSIM score between the first and all the other video frames, hence, high scores for static content. Models finetuned with epipolar geometry constraints show significant improvement in Sampson distance (33% reduction in text-to-video), while matching MET3R-optimized models on their own metric. This confirms that classical epipolar geometry provides a cleaner optimization signal than learned metrics, which show only modest self-improvement due to noise when evaluating generated content.

Table 4: Win-rate on the VideoReward [20] benchmark comparing different finetuning strategies.

Method	Visual Quality	Motion Quality	Text Alignment	Overall
sup. finetuning (SFT)	66.0%	63.0%	54.0%	64.5%
Flow-RWR [20]	63.5%	60.5%	**57.0%**	64.0%
DRO [36]	65.0%	54.0%	50.5%	64.5%
DPO [17]	**72.0%**	**71.0%**	55.0%	**73.0%**

4.4 Comparison with Other Fine-tuning Techniques

We compare four finetuning strategies as shown in Table 4: Supervised Finetuning (SFT), Flow-based Reward-Weighted Regression (Flow-RWR) [20], Direct Reward Optimization (DRO) [36], and our proposed DPO with Sampson Error. SFT directly optimizes for minimal epipolar error but struggles without negative samples to distinguish consistency levels. Flow-RWR weights samples by reward values but suffers from inconsistent absolute metrics, while DRO eliminates reference model queries but deviates substantially from the baseline capabilities. Our approach outperforms all alternatives, achieving the highest win rates in Visual Quality (72.0%), Motion Quality (71.0%), and Overall score (73.0%). This demonstrates that preference-based optimization with geometric constraints provides more effective guidance than approaches relying on absolute metrics or unconstrained optimization. Notably, our method achieves these improvements while maintaining the generative flexibility of the original model, allowing it to produce diverse outputs that satisfy both creative and geometric requirements simultaneously.

5 Limitations and Broader Impact

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. 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.

6 Conclusion

We have presented a novel approach for enhancing 3D consistency in video diffusion models by leveraging classical epipolar geometry constraints as preference signals. Our work demonstrates that aligning modern generative models with fundamental geometric principles can significantly improve the spatial coherence of generated content. The robust, mathematically grounded signal from simple Sampson error calculations provides clear guidance without requiring complex 3D supervision or differentiable rewards. The resulting models generate videos with notably fewer geometric inconsistencies and more stable camera trajectories while preserving creative flexibility. This work highlights how classical computer vision algorithms can effectively complement deep learning approaches, addressing limitations in purely data-driven systems and improving generated content quality through adherence to fundamental physical principles.

References

[1] OpenAI. Video generation models as world simulators, 2024. Accessed: 2024.

[2] Google DeepMind. Veo 2, 12 2024. Accessed: 2024.

[3] Adam Polyak et al. Movie gen: A cast of media foundation models, 2025.

[4] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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

[6] PikaLabs. Pika 1.5, 10 2024. Accessed: 2024.

[7] Haibo Yang, Yang Chen, Yingwei Pan, Ting Yao, Zhineng Chen, Chong-Wah Ngo, and Tao Mei. Hi3d: Pursuing high-resolution image-to-3d generation with video diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6870–6879, 2024.

[8] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*, 2025.

[9] Jensen Jinghao Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025.

[10] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction. *arXiv preprint arXiv:2504.07961*, 2025.

[11] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024.

[12] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024.

[13] Chao Liu and Arash Vahdat. Equivdm: Equivariant video diffusion models with temporally consistent noise. *arXiv preprint arXiv:2504.09789*, 2025.

[14] Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. *arXiv preprint arXiv:2412.01821*, 2024.

[15] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024.

[16] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

[17] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[18] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[20] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

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

[22] Runway. Gen-3, 06 2024. Accessed: 2024.

[23] LumaLabs. Dream machine, 06 2024. Accessed: 2024.

[24] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[25] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

[26] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[27] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[30] Christoph Schuhmann. Laion-aesthetics, 2022. Accessed: 2023-11-10.

[31] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

[32] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. 2023.

[33] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. 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.

[34] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

[35] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[36] Ruining Li, Chuanxia Zheng, Christian Rupprecht, and Andrea Vedaldi. Dso: Aligning 3d generators with simulation feedback for physical soundness. *arXiv preprint arXiv:2503.22677*, 2025.

[37] Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. *arXiv preprint arXiv:2503.09595*, 2025.

[38] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

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

[40] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[41] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

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

[43] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023.

[44] Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and Andre Araujo. Omniglue: Generalizable feature matching with foundation model guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19865–19875, 2024.

[45] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[46] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2691, 2024.

[47] MA Fischler. 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*, pages 381–395.

[48] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024.

[49] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[50] Guangcong Zheng, Teng Li, Xianpan Zhou, and Xi Li. Realcam-vid: High-resolution video dataset with dynamic scenes and metric-scale camera movements. *arXiv preprint arXiv:2504.08212*, 2025.

[51] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[52] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[53] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[54] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. *arXiv preprint arXiv:2501.06336*, 2025.

[55] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[57] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023.

[58] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[59] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17307–17316, 2023.

[60] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. 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*, pages 28140–28149, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: section 3 introduce and describe our approach in details while section 4 evaluates it's effectiveness.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the approach are discussed in section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The Implementation Details are discussed in section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We aim to release the code and preferences dataset upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The details are presented in section 3

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer:[Yes]

Justification: The paper evaluates the approach on common video evaluation metrics as well as a custom set of metrics that validate 3D consitency.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The analysis is provided in section 3

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research does not involve human subjects or participants. All experiments are conducted with publicly available models and datasets.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The analysis is provided in section 5

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The finetuned generators are based on publicly available video diffusion models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets used in our research are of MIT License.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, we plan to additionally release the code and preference data.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.