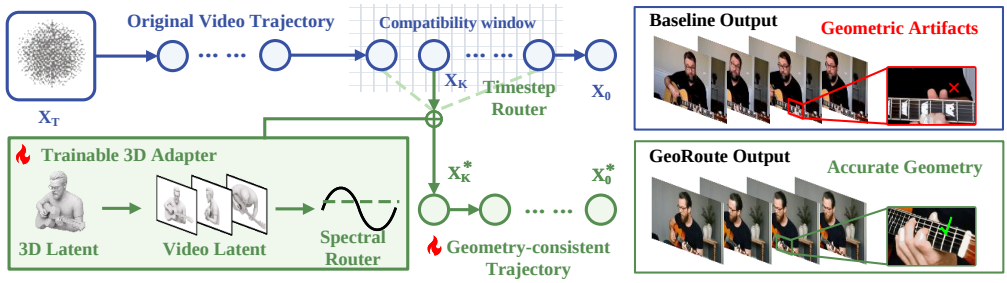
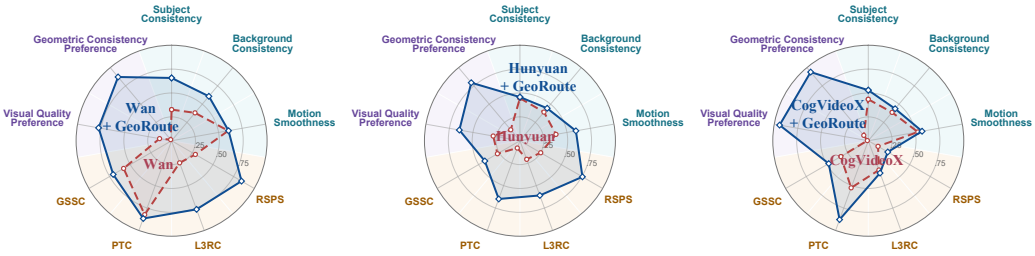
GEOROUTE: REROUTING VIDEO DIFFUSION TRAJECTORIES VIA SINGLE-STATE GEOMETRY INJECTION

Anonymous authors
Paper under double-blind review



(a) Trajectory rerouting and qualitative comparison.



(b) Cross-backbone quantitative comparison.

Figure 1: **Overview and effectiveness of GeoRoute.** (a) **Method:** With a compatibility 3D prior, GeoRoute adaptively selects a geometry-compatible route point and spectral guidance to redirect denoising toward a more geometry-consistent trajectory. (b) **Results:** Across three representative video-generation backbones, GeoRoute consistently improves geometric consistency (GeoVideoBench) while preserving overall video quality (VBench), with these gains also reflected in human preferences (user study).

ABSTRACT

Recent video diffusion models generate visually compelling videos, yet often fail to preserve coherent object geometry over time. Existing methods alleviate this problem by persistently incorporating 3D priors throughout denoising. However, as the video representation evolves, its compatibility with the 3D prior may vary substantially across denoising states. Persistent 3D injection may therefore introduce geometric information at poorly matched states, suggesting that 3D guidance could be applied more selectively. Motivated by this, we trace 3D-video compatibility along the denoising trajectory and find that higher compatibility is concentrated within a compact range of intermediate states, while the best-matched state varies across samples. We further find that, even at the best-matched state, different frequency components show different levels of compatibility. Together, these observations suggest that both when to inject the 3D prior and what 3D information to transfer should be determined adaptively. Based on these findings, we propose **GeoRoute**, a plug-in framework for adaptive 3D injection. Specifically, a Timestep Router selects a sample-specific compatible state, while a Spectral Router determines which frequency components to transfer. We further introduce a decoupled training strategy that separates 3D-video alignment from post-intervention denoising for stable optimization. In addition, we present GeoVideoBench, a reference-free benchmark for evaluating temporal geometric consistency. Extensive experiments across multiple video diffusion models show that GeoRoute consistently improves geometric consistency while preserving overall video quality. Code and website are available in the supplementary material.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

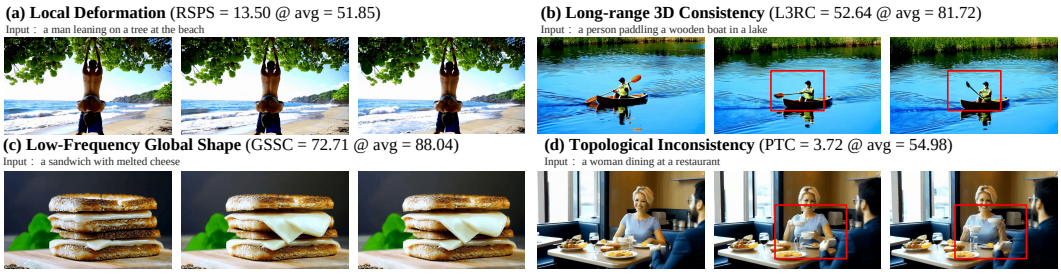


Figure 2: **Four typical geometric errors in video generation.** (a) implausible local deformation of the human body reduces rigid shape preservation; (b) the paddle partially disappears during motion, breaking long-range 3D consistency; (c) the low-frequency global shape of the cheese changes abruptly over time; and (d) an additional cup suddenly appears, causing topological inconsistency; these errors are quantified by the RSPS, L3RC, GSSC, and PTC metrics of our proposed GeoVideoBench, respectively, and all examples rank in the bottom 3% for the corresponding metric.

1 Introduction

Recent video diffusion models have achieved impressive visual fidelity, motion quality, and text alignment (Ho et al., 2022; Blattmann et al., 2023; Yang et al., 2024b). However, visually plausible motion does not necessarily imply geometrically coherent evolution. As an object moves, changes viewpoint, or undergoes deformation, its underlying structure should remain consistent over time, yet current models often fail to preserve this consistency. As illustrated in Figure 2, generated objects may exhibit local geometric deformation, structural inconsistency during motion, global shape drift, or even unintended topological changes. These failures highlight a remaining limitation of current video generation: maintaining coherent object geometry throughout dynamic evolution.

Generative 3D models offer a natural source of complementary geometric knowledge, as they explicitly model object structure across viewpoints (Chan et al., 2022; Wang et al., 2026). This has motivated recent approaches to incorporate 3D priors into video generation, often by persistently injecting geometric information throughout the denoising process (Ren et al., 2025; Wang et al., 2026). However, as the video representation evolves along the denoising trajectory, its compatibility with the 3D prior may vary substantially across denoising states. Persistent 3D injection may therefore introduce geometric information at poorly matched states, where the prior is less compatible with the current video representation. This observation suggests that 3D guidance could be applied more selectively according to the evolving compatibility between the two representations.

Motivated by this, we trace 3D–video compatibility along the denoising trajectory. We find that higher compatibility is concentrated within a compact range of intermediate denoising states, while the best-matched state within this range varies across samples. This sample-dependent variation motivates adaptive timestep selection rather than relying on a globally fixed intervention point. We further find that, even at the best-matched state, different frequency components of the 3D prior exhibit different levels of compatibility with the video representation. Together, these observations suggest that both *when* to inject the 3D prior and *what* 3D information to transfer should be determined adaptively.

Based on these findings, we propose **GeoRoute**, a trainable plug-in framework for adaptive 3D injection, as illustrated in Figure 1. The **Timestep Router** determines *when to intervene* by selecting a sample-specific compatible state for 3D injection. At the selected state, the **Spectral Router** determines *what to transfer* by adapting the frequency range and strength of tri-plane 3D features mapped into the video latent space. The selected geometric information is injected at the routed state, after which denoising continues from the modified latent representation. In this way, GeoRoute adapts both the intervention state and the transferred geometric information according to the compatibility between the 3D prior and the evolving video representation.

Training this intervention requires both aligning the heterogeneous 3D–video representations and enabling the video denoiser to evolve from the modified latent state. We therefore introduce a **decoupled training strategy** that separates 3D–video alignment from post-intervention denoising for stable optimization. Specifically, a latent-alignment objective learns a timestep-conditioned map-

ping from the 3D tri-plane features to the corresponding video latent space, providing compatible geometric features for intervention. A diffusion objective separately optimizes the post-intervention denoising trajectory to evolve from the modified latent state. Stop-gradient decouples these two optimization paths, preventing undesirable co-adaptation between the cross-modal mapping and the denoising trajectory and thereby stabilizing training.

Existing video benchmarks primarily assess perceptual quality, prompt alignment, and general temporal consistency (Huang et al., 2024; Han et al., 2025; Zheng et al., 2025), while providing limited measurement of temporal geometric consistency. We therefore introduce **GeoVideoBench**, a reference-free benchmark that evaluates local rigidity, long-range 3D consistency, low-frequency shape stability, and object topology from an RGB video and its caption. To verify its reliability, GeoVideoBench is validated through controlled geometric corruptions and blinded human judgments, which respectively assess its sensitivity to geometric degradation and agreement with human perception.

We evaluate GeoRoute on three representative video diffusion models: Wan (Wan et al., 2025), HunyuanVideo (Kong et al., 2024), and CogVideoX (Yang et al., 2024b). Across all three backbones, GeoRoute consistently improves temporal geometric consistency, achieving an average gain of 5.84 points on GeoVideoBench, while further improving VBench by 1.52 points. These results demonstrate the consistent effectiveness of GeoRoute in improving geometric consistency while preserving overall video quality.

Our main contributions are:

- We reveal that 3D–video compatibility is concentrated within a compact range of intermediate denoising states, while the best-matched state varies across samples and frequency components exhibit different levels of compatibility.
- We propose **GeoRoute**, a plug-in framework for adaptive 3D injection that determines *when* to intervene and *what* 3D information to transfer through a Timestep Router and a Spectral Router, together with a **decoupled training strategy** for stable 3D–video alignment and post-intervention denoising.
- We introduce **GeoVideoBench** for reference-free geometric evaluation, on which GeoRoute achieves an average gain of 5.84 points across three video diffusion models, with a further 1.52-point gain on VBench.

2 Related Work

Video Diffusion Models and Evaluation. Video diffusion models have evolved from spatio-temporal U-Nets (Ho et al., 2022) to transferable temporal modules (Guo et al., 2024) and large-scale latent diffusion (Blattmann et al., 2023). Recent diffusion-transformer systems, including CogVideoX (Yang et al., 2024b) and HunyuanVideo (Kong et al., 2024), further improve visual fidelity, motion, and prompt alignment. These models primarily learn correlations in projected video representations and remain susceptible to progressive geometric deformation. Existing benchmarks mainly assess perceptual quality, temporal smoothness, subject consistency, and text alignment. VBench (Huang et al., 2024), Video-Bench (Han et al., 2025), and VBench-2.0 (Zheng et al., 2025) provide increasingly comprehensive evaluations of video quality, condition consistency, and physical plausibility. PDI-Bench (Wu et al., 2026b) further evaluates projective consistency related to rigidity and 3D motion. GeoVideoBench focuses on object deformation, global structural inconsistency, and unintended merging or splitting as separate temporal geometric failure modes.

Bridging 3D to Video Generation. Prior works introduce geometry through mesh correspondences (Cai et al., 2024), 3D tracking videos (Gu et al., 2025), joint RGB–XYZ prediction (Zhang et al., 2025), depth-and-pose regularization (Bai et al., 2025), or persistent point-cloud renderings (Ren et al., 2025). Other approaches align video features with geometry models (Wu et al., 2026a), learn view-consistent representations (Danier et al., 2025), or use geometric preferences for post-training (Du et al., 2026). Most related to our work, Orbital (Wang et al., 2026) injects volumetric 3D features through multi-scale adapters for image-conditioned orbital video synthesis. GeoRoute differs by identifying a compatible denoising timestep, injecting low-frequency tri-plane features at that stage, and directly rerouting the trajectory of pretrained video diffusion models.

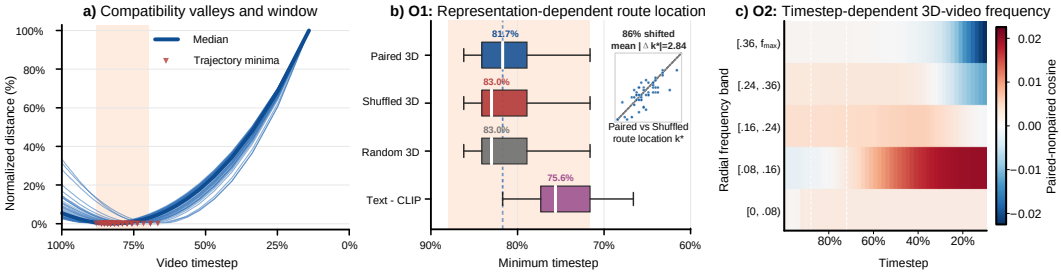


Figure 3: **Observations on 3D–video latent compatibility.** (a) **Overall validation:** paired 3D exhibits an interior low-discrepancy valley, defining a compact candidate window. (b) **O1: Minimum location by representation.** Route locations vary across representations and, more importantly, across individual paired trajectories. (c) **O2: 3D-video frequency specificity.** The matched-over-nonpaired 3D correspondence varies across both denoising timestep and frequency band.

3 GeoRoute

3.1 Preliminaries and Problem Formulation

We consider geometry-preserving generation with a pretrained text-conditioned video diffusion model and a pretrained generative 3D model. Under the flow-matching formulation, a clean video latent $\mathbf{x}_0$ is interpolated with Gaussian noise as

$$\mathbf{x}_t = (1-\sigma_t)\mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \mathbf{u}_t = \boldsymbol{\epsilon} - \mathbf{x}_0, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where σ_t is the noise level and $\mathbf{u}_t$ is the target velocity. The video model $\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c})$ predicts this velocity conditioned on text prompt $\mathbf{c}$.

At inference, integrating the predicted velocity field produces the reverse trajectory

$$\mathcal{T}_\theta = \{\mathbf{x}_T, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_0\}. \tag{2}$$

The complete text prompt $\mathbf{c}$ is shared by the video model and the generative 3D model. The 3D model produces $\mathbf{z}^{\mathrm{3D}} = \mathcal{G}_{\mathrm{3D}}(\mathbf{c})$, which we use as a 3D prior derived from the prompt, providing a view-consistent geometric reference. Our objective is to identify a compatible route point k^* and learn a transition $\mathcal{R}$ that maps the current video state to a geometry-aware state:

$$\mathbf{x}^*_{k^*} = \mathcal{R}\left(\mathbf{x}_{k^*}, \mathbf{z}^{\mathrm{3D}}\right), \qquad \mathcal{T}^*_\theta = \{\mathbf{x}_T, \ldots, \mathbf{x}_{k^*+1}, \mathbf{x}^*_{k^*}, \ldots, \mathbf{x}^*_0\}. \tag{3}$$

The goal is to improve geometric consistency along the post-route trajectory while preserving the pretrained model's motion quality and appearance fidelity. This requires determining a suitable route point and learning a stable geometry-aware transition, which we address in the following sections.

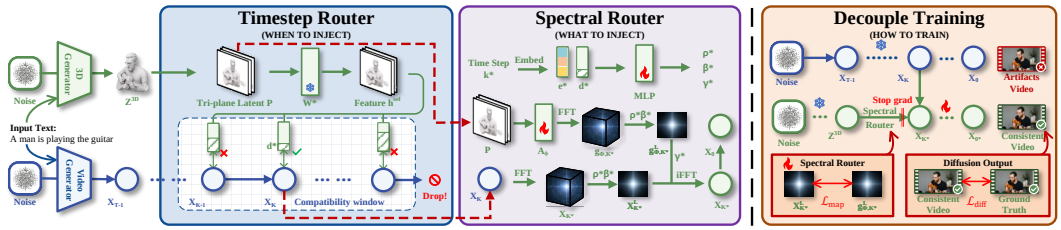
3.2 3D–Video Latent Compatibility Analysis

We first analyze 3D–video compatibility along the denoising trajectory to identify a compact transition region. We then make two observations: **O1** studies how the compatible route point varies across samples, while **O2** studies how 3D–video correspondence varies across frequency bands.

We fit a shared bias-free map from paired tri-plane latent $\mathbf{P}_i$ to the temporally averaged video latent $\bar{\mathbf{X}}_{i,t}$, and measure their out-of-fold discrepancy

$$\begin{aligned} d_i(t) &= \frac{\|\mathbf{P}_i\mathbf{W}^* - \bar{\mathbf{X}}_{i,t}\|_F^2}{HWC_v}, \\ k_i^* &= \arg\min_t d_i(t). \end{aligned} \tag{4}$$

Here, $\mathbf{P}_i \in \mathbb{R}^{HW\times C_p}$ contains the resized tri-plane tokens, $\bar{\mathbf{X}}_{i,t} \in \mathbb{R}^{HW\times C_v}$ is the video latent averaged over its temporal axis and flattened over spatial positions, and $\mathbf{W}^* \in \mathbb{R}^{C_p\times C_v}$ is the bias-free map fitted on the training folds. We use the temporally averaged latent only for compatibility



(a) Pipeline of GeoRoute (b) Training Strategy

Figure 4: Overview of GeoRoute. **(a) Pipeline of GeoRoute:** Timestep Router decides *when to inject* and selects a sample-compatible state inside Compatibility Window. Spectral Router decides *what to inject* and adapts the frequency range and strength of the mapped tri-plane latent P. **(b) Training Strategy:** To avoid training collapse caused by cross-trajectory training, GeoRoute adopts a decoupled optimization strategy that separates cross-modal latent mapping from post-route video adaptation.

scoring, while the full spatiotemporal latent is retained for subsequent geometry injection. Thus, HWC_v normalizes the squared Frobenius distance by the number of compared video features.

Overall validation: a compatible transition region exists. Figure 3a reveals a non-monotonic interior low-discrepancy valley, with trajectory-wise minima concentrated in a compact intermediate window. The spread of these minima inside the valley motivates a sample-adaptive router and leads to the two analyses below.

O1: The route point is representation- and sample-dependent. As shown in Figure 3b, different modalities induce different route-location distributions, and the paired 3D optimum varies across individual trajectories. The coarse valley localizes the routing interval, and the sample-wise spread motivates the Timestep Router to determine *When to inject* for each sample.

O2: Paired-3D specificity is timestep- and frequency-dependent. We further partition the video and mapped 3D features into radial 3D Fourier bands and measure the matched-minus-nonpaired cosine correspondence after feature normalization. Figure 3c shows that the relative 3D specificity of these bands changes throughout denoising: both the most compatible frequency band and its correspondence strength vary across candidate states. This motivates the Spectral Router to determine *what to inject* at the selected route point through a route-conditioned spectral configuration. More details are provided in Appendix A.

3.3 When to Inject: Timestep Router

Motivated by **O1**, GeoRoute determines *When* geometric guidance should be introduced by selecting a sample-compatible state within a compact, model-specific window. For each video backbone m, we estimate a model-specific compatibility window $\mathcal{W}_{.95}^{(m)}$, defined as the central 95% interval of the sample-wise optimal route points on the training set. This window is computed once and kept fixed during training and inference. The frozen shared mapping $\mathbf{W}^*$ projects the resized tri-plane tokens $\mathbf{P}$ into the video latent space and serves as the timestep selector during training and inference. Let $\{t_1 > \cdots > t_J\}$ denote the scheduler states in $\mathcal{W}_{.95}^{(m)}$, ordered along the reverse denoising process. At each state, we compute

$$d_{t_j} = \frac{1}{HWC_v} \left\| \bar{\mathbf{X}}_{t_j} - \mathbf{P}\mathbf{W}^* \right\|_F^2, \qquad j^* = \min\{j \geq 2 : d_{t_j} > d_{t_{j-1}}\}, \qquad k^* = t_{j^*-1}. \tag{5}$$

The Timestep Router uses these compatibility scores to select a sample-adaptive route point k^* within the estimated window. If no increase occurs, the final candidate t_J is selected.

3.4 What to Inject: Spectral Router

Motivated by **O2**, GeoRoute determines *what* geometric information to inject by adapting its spectral range and strength to the selected route state. We reuse the tri-plane tokens $\mathbf{P}$ and map them into the

video latent space with a timestep-conditioned 3D–video adapter $\mathcal{A}_\phi$, parameterized by ϕ, yielding $\mathbf{g}_{\phi,k^*} = \mathcal{A}_\phi(\mathbf{P}, \mathbf{e}_{k^*})$. Given the compatibility score d_{k^*} at the selected state, the Spectral Router $\mathcal{S}_\eta$, parameterized by η, predicts the cutoff ρ_{k^*}, boundary sharpness β_{k^*}, and injection strength γ_{k^*}:

$$\begin{aligned}(\rho_{k^*}, \beta_{k^*}, \gamma_{k^*}) &= \mathcal{S}_\eta\left(d_{k^*}, \mathbf{e}_{k^*}\right), \\ \mathbf{m}_{k^*}(\boldsymbol{\omega}) &= \mathrm{Sigmoid}\left[\beta_{k^*}\left(\rho_{k^*} - \|\boldsymbol{\omega}\|_2\right)\right].\end{aligned} \tag{6}$$

The resulting soft spectral mask $\mathbf{m}_{k^*}$ is applied in the Fourier domain to obtain the selected components $\mathbf{x}^L_{k^*}$ and $\mathbf{g}^L_{\phi,k^*}$, while $\mathbf{x}^H_{k^*}$ denotes the complementary video component. GeoRoute then performs

$$\mathbf{x}^*_{k^*} = \mathbf{x}^H_{k^*} + \mathrm{MM}\left(\mathbf{x}^L_{k^*} + \gamma_{k^*}\mathbf{g}^L_{\phi,k^*}; \mathbf{x}^L_{k^*}\right), \tag{7}$$

where MM matches the statistics of the modified component to those of the original selected video component.

3.5 How to Train: Decoupled Training

Since the 3D–video adapter and the post-route video branch serve different objectives, we optimize them in a decoupled manner. Specifically, GeoRoute is trained with a latent-alignment objective and a standard diffusion objective:

$$\begin{aligned}\mathcal{L}_{\mathrm{map}} &= \left\|\mathbf{g}_{\phi,k^*} - \mathrm{sg}\left(\mathbf{x}^L_{k^*}\right)\right\|_2^2, \\ \mathcal{L} &= \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{map}}\mathcal{L}_{\mathrm{map}},\end{aligned} \tag{8}$$

where $\mathcal{L}_{\mathrm{diff}}$ is the standard video diffusion objective evaluated on the routed latents, and sg denotes stop-gradient. The two objectives optimize different modules: $\mathcal{L}_{\mathrm{map}}$ trains the 3D–video adapter, while $\mathcal{L}_{\mathrm{diff}}$ trains the Spectral Router and video LoRA. More details are provided in Appendix B.

4 GeoVideoBench

To systematically evaluate geometry-related temporal failures, we introduce GeoVideoBench, which assesses two aspects of object consistency: *3D structure* and *object topology*. The former is characterized by RSPS, L3RC, and GSSC, while the latter is measured by PTC. GeoVideoBench operates from a generated video and its text prompt, with no ground-truth videos or 3D annotations.

Rigid Shape Preservation Score (RSPS). RSPS evaluates local 3D rigidity and detects nonuniform stretching, compression, and local collapse. We construct a nearest-neighbor graph over foreground pseudo-3D points and measure the scale-normalized logarithmic strain of each edge:

$$e^{\mathrm{rigid}}_{t,ij} = \left|\log \frac{d_{t,ij}/(m_t + \epsilon) + \epsilon}{d_{0,ij}/(m_0 + \epsilon) + \epsilon}\right|, \tag{9}$$

where $d_{t,ij}$ is the edge length, m_t is the median edge length at frame t, and the scalar $\epsilon > 0$ ensures numerical stability. Lower strain indicates better local shape preservation.

Long-range 3D Reprojection Consistency (L3RC). L3RC evaluates maintenance of a coherent global 3D structure over distant frames. We estimate a 3D similarity transformation from a subset of foreground tracks and evaluate its prediction on held-out points using the normalized reprojection error

$$r_t = \underset{i \in \mathcal{Q}_t}{\mathrm{median}} \frac{\|\widehat{\mathbf{u}}_{t,i} - \mathbf{u}_{t,i}\|_2}{\sqrt{H^2 + W^2}}. \tag{10}$$

Here, $\mathcal{Q}_t$ is the held-out track set, $\mathbf{u}_{t,i}$ is the observed image coordinate, and $\widehat{\mathbf{u}}_{t,i}$ is its prediction under the fitted 3D similarity transform. A lower error indicates that a single global 3D motion can better explain the object's long-term evolution.

Global Shape Spectrum Consistency (GSSC). GSSC captures coarse changes in global shape and depth organization. We construct a canonical foreground shape–depth field, extract its low-frequency Fourier descriptor $\mathbf{s}_t$, and compare each frame with the anchor:

$$E_{\mathrm{GSSC}} = \underset{t>0}{\mathrm{median}}\left(1 - \mathbf{s}_0^\top \mathbf{s}_t\right). \tag{11}$$

This low-frequency comparison emphasizes global 3D structure while suppressing texture and boundary variations.

Part Topology Consistency (PTC). PTC evaluates silhouette topology and detects hole changes, disappearance of thin or articulated parts, and abnormal endpoint or branch structure. We construct a signature $\mathbf{h}_t^{\text{topo}}$ from connected components, holes, skeleton endpoints, branch junctions, solidity, aspect ratio, and skeleton length, and measure its normalized deviation from the anchor:

$$e_t^{\text{topo}} = \frac{\sum_j \lambda_j^{\text{topo}} \min\left(|h_{t,j}^{\text{topo}} - h_{0,j}^{\text{topo}}|/\kappa_j,\, 3\right)}{\sum_j \lambda_j^{\text{topo}}}. \tag{12}$$

Here, κ_j and λ_j^{topo} are the fixed normalization scale and importance weight of signature dimension j, respectively.

Reliability-Calibrated Aggregate. Because the component metrics have different operational sensitivities, we define a reliability-calibrated aggregate using a separate synthetic controlled set. For metric m, we define its targeted geometric response G_m, worst nuisance response N_m, reliability R_m, and normalized weight w_m as

$$\begin{aligned} G_m &= \mathbb{E}[s_m^{\text{clean}} - s_m^{\text{target,max}}], \quad & N_m &= \max_{c \in \mathcal{C}_{\text{nuis}}} \mathbb{E}[|s_m^c - s_m^{\text{clean}}|], \\ R_m &= \frac{G_m}{G_m + N_m + \epsilon}, & w_m &= \frac{R_m}{\sum_\ell R_\ell}. \end{aligned} \tag{13}$$

This ratio is invariant to linear rescaling within a metric and downweights scores whose nuisance response is large relative to their targeted geometric response. Using $\widetilde{s}_m = s_m/100$, we compute

$$\text{Overall}_{\text{RC}} = 100 \exp\left(\sum_m w_m \log(\max(\widetilde{s}_m, \epsilon))\right). \tag{14}$$

For model-level comparisons, s_m is the mean of the valid video scores for metric m. The geometric mean prevents a severe failure in one dimension from being fully compensated by another. We therefore continue to report all four component scores and provide equal-weight aggregation results in Appendix C.4.

5 Experiments

5.1 Settings

Datasets. We use TRELLIS-500K (Xiang et al., 2025) for 3D-video latent alignment and timestep selection, and a subset of OpenVidHD (Nan et al., 2025) for training GeoRoute. **Baselines.** We evaluate GeoRoute on three representative open-source video diffusion models with diverse architectures: *Wan* (Wan et al., 2025), *CogVideoX* (Yang et al., 2024b), and *HunyuanVideo* (Kong et al., 2024), to assess its plug-in generalizability. **Training Details.** GeoRoute is fine-tuned for 10,000 steps with LoRA, a global batch size of 8, and AdamW with a learning rate of 1×10^{-5}. The latent mapping module and video denoiser are jointly optimized through decoupled training. All variants share identical training and sampling settings. **Evaluation Details.** Following prior geometry-aware video generation works (Bai et al., 2025; Du et al., 2026), we adopt three commonly used *VBench* (Huang et al., 2024) metrics and 4 metrics of *GeoVideoBench* to evaluate video quality and geometric consistency, respectively. We additionally conduct a pairwise *User Study* to assess both overall visual quality and geometric consistency. Further details are provided in Appendix D

5.2 Main Results

Quantitative Results. To evaluate the effectiveness of GeoRoute, we integrate it into three representative video diffusion backbones and compare each enhanced model with its original counterpart under identical evaluation settings. As shown in Table 1, GeoRoute consistently improves VBench Avg., with an average gain of 1.52 points. Subject and background consistency improve consistently, while motion smoothness remains comparable or improves, showing that geometric guidance preserves general visual quality and temporal generation capability. The improvements

Table 1: Quantitative comparison between baseline models and GeoRoute. Higher is better.

Model	VBench				GeoVideoBench				
	Subject Consistency	Background Consistency	Motion Smoothness	Avg.	RSPS	L3RC	GSSC	PTC	Overall (RC)
Wan	93.18	95.07	98.59	95.61	43.57	74.12	86.86	50.96	65.37
Wan + GeoRoute	**97.92** +4.74	**97.89** +2.82	**98.63** +0.04	**98.15** +2.54	**53.20** +9.63	**83.31** +9.19	**88.76** +1.90	**51.89** +0.93	**68.96** +3.59
Hunyuan	94.89	95.20	95.85	95.31	42.95	73.43	82.21	34.97	55.58
Hunyuan + GeoRoute	**95.06** +0.17	**95.85** +0.65	**98.50** +2.65	**96.47** +1.16	**51.65** +8.70	**80.56** +7.13	**84.51** +2.30	**47.20** +12.23	**64.85** +9.28
CogVideoX	94.72	95.15	97.65	95.84	40.65	75.54	83.11	44.52	61.12
CogVideoX + GeoRoute	**96.09** +1.37	**95.77** +0.62	**98.21** +0.56	**96.69** +0.85	**42.68** +2.03	**76.14** +0.60	**85.32** +2.21	**52.15** +7.63	**65.78** +4.65

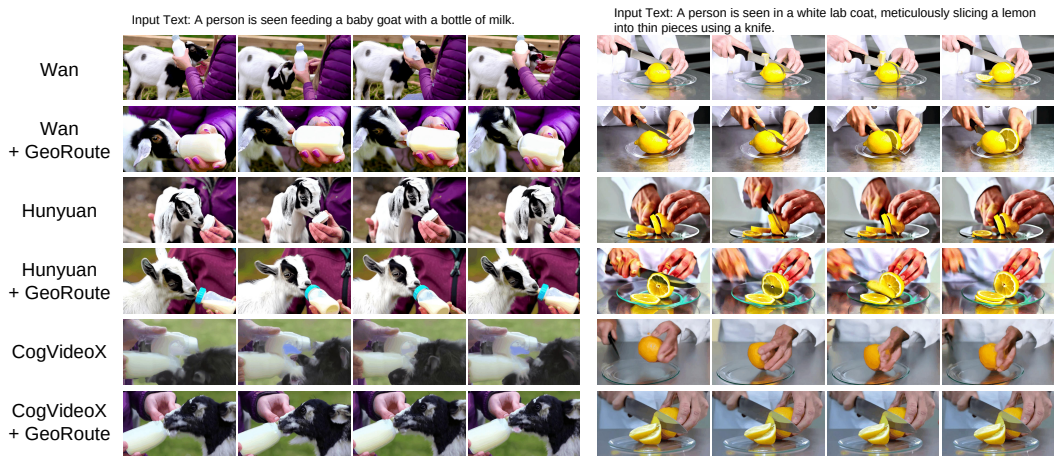


Figure 5: Qualitative comparison between baseline models and GeoRoute.

are more pronounced on GeoVideoBench, where the overall score increases by an average of 5.84 points. Improvements in RSPS and L3RC indicate better preservation of local rigidity and long-range 3D structure; gains in GSSC show more stable low-frequency global shape; and higher PTC reflects fewer unintended merging, splitting, or part-disappearance events. The consistent improvements across all four geometric dimensions and three architecturally different video diffusion models demonstrate that GeoRoute addresses complementary geometric failure modes and generalizes as a model-agnostic plug-in.

To evaluate whether GeoRoute improvements align with human preferences, we conduct a user study on visual quality and geometric consistency. As shown in Table 2, GeoRoute is consistently preferred over its corresponding baseline across all three backbones. Full details are provided in Appendix F.

Table 2: Pairwise human preference results (%).

Backbone	Visual Quality			Geometric Consistency		
	Baseline	Tie	GeoRoute	Baseline	Tie	GeoRoute
Wan	23.86	3.78	**72.36**	18.52	10.23	**71.25**
HunyuanVideo	35.56	1.96	**62.48**	26.82	6.75	**66.43**
CogVideoX	15.05	0.12	**84.83**	22.10	2.51	**75.39**

Qualitative Results. To qualitatively compare temporal geometric consistency, we visualize representative results from three video-generation backbones in Figure 5. The baseline models show clear 3D structural errors, such as distorted hands and fingers in CogVideoX, and topology changes, including the sudden appearance of additional lemon slices in Wan and the merging of multiple slices in HunyuanVideo. Incorporating GeoRoute substantially reduces these errors, yielding more stable object structures and topology throughout generation.

Comparison with Other SOTAs. To compare GeoRoute with existing SOTA methods, we evaluate both general video-generation quality and temporal geometric consistency on the common CogVideoX backbone. GeoVideo (Bai et al., 2025) introduces depth prediction and multi-view geometric regularization, while VideoGPA (Du et al., 2026) distills geometric preferences through DPO. As shown in Table 3, GeoRoute achieves the highest VBench Avg. and GeoVideoBench Overall scores, suggesting that injecting 3D priors into a selected intermediate noisy latent state improves geometric consistency while preserving general video quality.

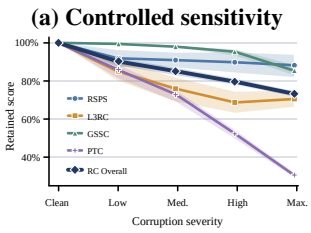
Table 3: Comparison with other SOTAs on VBench and GeoVideoBench using the CogVideoX.

Method	VBench				GeoVideoBench				
	Subject Consistency	Background Consistency	Motion Smoothness	Avg.	RSPS	L3RC	GSSC	PTC	Overall (RC)
CogVideoX	94.72	95.15	97.65	95.84	40.65	75.54	83.11	44.52	61.12
GeoVideo	95.80 +1.08	95.18 +0.03	97.32 -0.33	96.10 +0.26	41.57 +0.92	74.62 -0.92	84.15 +1.04	47.64 +3.12	62.91 +1.79
VideoGPA	95.98 +1.26	95.53 +0.38	97.74 +0.09	96.42 +0.58	42.05 +1.40	75.33 -0.21	84.92 +1.81	52.10 +7.58	65.43 +4.31
GeoRoute	**96.09** +1.37	**95.77** +0.62	**98.21** +0.56	**96.69** +0.85	**42.68** +2.03	**76.14** +0.60	**85.32** +2.21	**52.15** +7.63	**65.78** +4.65

Table 4: Ablation study of GeoRoute on Wan. $\times$ denotes training collapse.

Variant	Prior	Adapter	LoRA	Stop-grad.	Timestep Router	Spectral Router	VBench				GeoVideoBench				
							Subject Consistency	Background Consistency	Motion Smoothness	Avg.	RSPS	L3RC	GSSC	PTC	Overall (RC)
Base model	None	–	–	–	–	–	93.18	95.07	98.59	95.61	43.57	74.12	86.86	50.96	65.37
+ LoRA	None	–	✓	–	–	–	96.36	95.28	97.33	96.32	45.29	73.45	84.92	51.33	65.11
Direct 3D injection	Correct 3D	–	–	–	Fixed @ 81.7%	Fixed	91.44	93.68	96.55	93.89	39.97	72.14	82.21	50.88	63.32
+ Adapter (joint opt.)	Correct 3D	✓	–	–	Fixed @ 81.7%	Fixed		×					×		
+ LoRA (joint opt.)	Correct 3D	✓	✓	–	Fixed @ 81.7%	Fixed		×					×		
Fixed-routing baseline	Correct 3D	✓	✓	✓	Fixed @ 81.7%	Fixed	95.91	96.45	97.36	96.57	44.58	76.10	87.13	52.02	66.41
+ Adaptive timestep	Correct 3D	✓	✓	✓	Adaptive	Fixed	96.53	96.67	97.26	96.82	45.52	77.37	87.64	53.11	67.40
GeoRoute	Correct 3D	✓	✓	✓	Adaptive	Adaptive	**97.92**	**97.89**	**98.63**	**98.15**	**53.20**	**83.31**	**88.76**	51.89	**68.96**
GeoRoute (wrong prior)	Wrong 3D	✓	✓	✓	Adaptive	Adaptive	94.35	97.43	97.94	96.57	40.05	71.34	78.80	48.19	61.01
GeoRoute (blank prior)	Blank	✓	✓	✓	Adaptive	Adaptive	88.49	90.17	92.82	90.49	39.59	69.22	75.04	44.83	57.99
GeoRoute (text prior)	CLIP Text	✓	✓	✓	Adaptive	Adaptive	90.04	92.44	96.18	92.89	42.43	74.14	73.47	39.52	56.00
GeoRoute (all-step)	Correct 3D	✓	✓	✓	Fixed @ All	Adaptive	96.32	96.28	96.33	96.31	50.79	80.78	82.57	45.24	63.28
GeoRoute (window)	Correct 3D	✓	✓	✓	Fixed @ $\mathcal{W}_{95}^{(m)}$	Adaptive	96.23	96.32	96.02	96.19	52.66	81.27	80.46	53.46	67.00

Validation of GeoVideoBench. We validate the geometric sensitivity and human alignment of GeoVideoBench in Figure 6. Specifically, we construct geometric corruption videos with all metrics decreasing, indicating sensitivity to geometric errors. The overall score $\rho = 0.275$ is comparable to others and achieves 66.0% pairwise agreement with human judgments of geometric consistency, which supports GeoVideoBench as a meaningful and complementary measure of geometric correctness.



(b) Human alignment

Evaluator	Human ρ	Agree. %
RC Overall	0.275	66.0
RSPS	0.334	68.9
L3RC	0.226	68.6
GSSC	0.283	68.3
PTC	0.231	64.6
CLIP-Score	0.217	–
SSIM-sim	0.269	–
GPT-4o	0.287	–

Figure 6: GeoVideoBench validation.

5.3 Ablation Study

To verify the effectiveness of GeoRoute, we conduct extensive ablation studies in Table 4. Since GeoRoute is trained with additional data, we first introduce a LoRA baseline under the same training setting to establish a fair comparison. LoRA improves general video quality but brings little gain in geometric consistency, while direct 3D injection degrades both metrics. Jointly optimizing the 3D adapter and denoiser further leads to training collapse. Applying stop-gradient decouples cross-modal alignment from video adaptation and yields a stable fixed-routing baseline. Starting from this baseline, adaptive Timestep Router and Spectral Router progressively improve performance, with the full GeoRoute achieving the best results. Finally, replacing the matched 3D prior with an incorrect 3D prior, blank input, or CLIP text consistently reduces performance, confirming that the gain mainly comes from correctly matched geometric information. Additionally, to verify the effectiveness of transient injection, we compare GeoRoute with persistent geometry injection throughout denoising or at every step within the compatibility window. Persistent injection degrades both geometric consistency and video quality. Restricting injection to the window partially alleviates visual artifacts but still underperforms GeoRoute, supporting the benefit of a single transient intervention.

6 Conclusion and Limitations

In this work, we introduce GeoRoute for integrating generative 3D priors into video diffusion trajectories and GeoVideoBench for evaluating temporal geometric consistency. Nevertheless, GeoRoute requires an additional 3D generation stage, which increases computational cost. Future work will reduce this overhead and extend trajectory routing to more heterogeneous structures.

AI USE STATEMENT

In this work, generative AI tools were used solely to improve the language and readability of the manuscript. They were not used to develop the core research ideas, formulate mathematical claims, derive theoretical proofs, design the methodology or experiments, generate or clean datasets, interpret experimental results, or create figures, tables, or visualization code. All theoretical proofs were developed and verified by the authors. All AI-assisted language edits were manually reviewed by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

Our user study involved subjective comparisons of generated videos in terms of visual quality and geometric consistency. All participants provided informed consent, participated voluntarily, and were free to withdraw at any time. Responses were anonymized and reported only in aggregate.

REPRODUCIBILITY STATEMENT

We provide anonymized reproducibility code for both GeoRoute and GeoVideoBench as part of the supplementary material. The released code includes the training, inference, data-processing, and evaluation pipelines, together with the configurations and scripts used to reproduce the reported experiments. All datasets used in this work are publicly available, and their sources and processing procedures are described in the main paper and appendix. Detailed information about the computational environment, model configuration, optimization parameters, training objectives, and inference settings is provided in Appendix D. These materials enable independent reproduction of both the GeoRoute results and the GeoVideoBench evaluation.

REFERENCES

Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. GeoVideo: Introducing geometric regularization into video generation model. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/536d18fbb454f80221465f1a42c6f389-Abstract-Conference.html.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Paul Huang, Tuanfeng Yang Wang, and Gordon Wetzstein. Generative rendering: Controllable 4d-guided video generation with 2d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7611–7620, 2024.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16123–16133, 2022.

Duolikun Danier, Ge Gao, Steven McDonagh, Changjian Li, Hakan Bilen, and Oisin Mac Aodha. View-consistent diffusion representations for 3d-consistent video generation, 2025.

Hongyang Du, Junjie Ye, Xiaoyan Cong, Runhao Li, Jingcheng Ni, Aman Agarwal, Zeqi Zhou, Zekun Li, Randall Balestriero, and Yue Wang. VideoGPA: Distilling geometry priors for 3d-consistent video generation, 2026.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *ACM SIGGRAPH 2025 Conference Papers*, pp. 62:1–62:12, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li-jia Li, and Yongxin Ni. Video-Bench: Human-aligned video generation benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18858–18868, 2025.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024. doi: 10.18653/v1/2024.emnlp-main.127.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 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.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Xinran Ling, Chen Zhu, Meiqi Wu, Hangyu Li, Xiaokun Feng, Cundian Yang, Aiming Hao, Jiashu Zhu, Jiahong Wu, and Xiangxiang Chu. VMBench: A benchmark for perception-aligned video motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13087–13098, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 38–55, 2024. doi: 10.1007/978-3-031-72970-6_3.

Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, volume 2, pp. 674–679, 1981.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

Jianbo Shi and Carlo Tomasi. Good features to track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593–600, 1994. doi: 10.1109/CVPR.1994.323794.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, April 1991. doi: 10.1109/34.88573.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Rong Wang, Ruyi Zha, Ziang Cheng, Jiayu Yang, Pulak Purkait, and Hongdong Li. Towards realistic and consistent orbital video generation via 3d foundation priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18417, 2026.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Jiaxin Wu, Yihao Pi, Yinling Zhang, Yuheng Li, and Xueyan Zou. Quantitative video world model evaluation for geometric-consistency, 2026b.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21469–21480, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.

Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21685–21695, 2025.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Appendix Contents

A GeoRoute Mechanism and Efficiency Analysis **14**

A.1 O1: Timestep Compatibility and Sample-Specific Route Points 14

A.2 O2: Timestep-Dependent Spectral Compatibility 15

A.3 Trajectory Analysis 17

A.4 Inference Latency 17

B GeoRoute Design and Optimization Details **18**

B.1 Tri-Plane Geometry Representation 18

B.2 Adaptive Timestep Router 19

B.3 Adaptive Spectral Router 19

B.4 Inference 20

B.5 Decoupled Cross-Trajectory Training 21

C GeoVideoBench Metrics and Validation **23**

C.1 Benchmark Overview and Evaluation Dimensions 23

C.2 Shared Evaluation Pipeline 24

C.3 Metric Definitions 25

C.4 Controlled Validation and Human Alignment 28

D Experimental Implementation Details **32**

E Extended Ablation Studies **34**

F User Study **35**

G Additional Qualitative Results **37**

H Supplementary Materials **39**

The appendix reports the analyses, implementation details, and extended evaluations for GeoRoute and GeoVideoBench.

A GeoRoute Mechanism and Efficiency Analysis

This section provides additional evidence for the design of GeoRoute, including timestep-dependent compatibility, sample-specific route selection, frequency-dependent correspondence, trajectory rerouting, and inference efficiency.

A.1 O1: Timestep Compatibility and Sample-Specific Route Points

We first estimate a shared bias-free compatibility map on the training split:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \sum_{i\in\mathcal{I}_{\mathrm{tr}}} \sum_{t\in\mathcal{K}} \left\| \mathbf{P}_i\mathbf{W} - \bar{\mathbf{X}}_{i,t} \right\|_F^2, \tag{15}$$

where $\mathcal{I}_{\mathrm{tr}}$ denotes the training samples and $\mathcal{K}$ contains the evaluated scheduler states. We normalize the scheduler position as $q_t = 100t/T$, where T is the maximum scheduler timestep. The sample-wise optimal route point is obtained by minimizing the compatibility distance. We select the normalized interval $[q_{\mathrm{L}}, q_{\mathrm{R}}]$ where these optima are concentrated, and define $\mathcal{W}^{(m)}$ as the evaluated scheduler states whose normalized positions fall in this interval. We report window bounds in normalized coordinates throughout.

For a routed representation $\mathbf{z}_i$ and frozen readout $g(\cdot)$, the compatibility distance can be written as

$$d_i^{(\mathbf{z})}(t) = \frac{1}{HWC_v}\left[\left\|\bar{\mathbf{X}}_{i,t}\right\|_F^2 + \left\|g(\mathbf{z}_i)\right\|_F^2 - 2\left\langle \bar{\mathbf{X}}_{i,t}, g(\mathbf{z}_i)\right\rangle_F \right], \tag{16}$$

showing that the preferred route point depends on both the scheduler state and the sample-specific correspondence between the 3D representation and video latent.

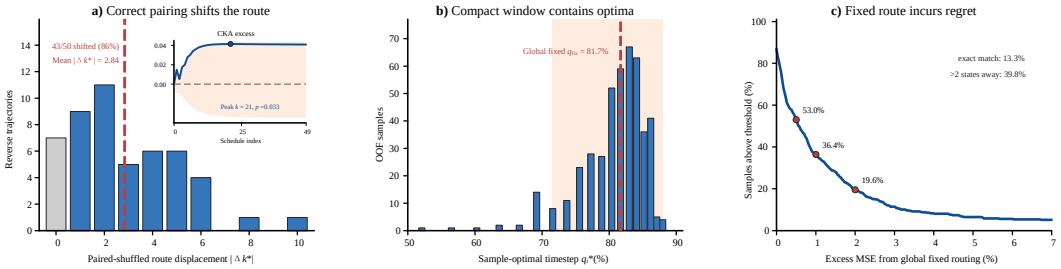


Figure 7: **Additional evidence for sample-adaptive route selection.** (a) Route-point displacement between paired and shuffled 3D representations over 50 reverse trajectories. (b) Distribution of sample-wise optimal route points. The blue region denotes the candidate route window and the dashed line denotes the globally optimal fixed point. (c) Additional compatibility error introduced by replacing the sample-wise route point with the global fixed point.

Route-point evidence from paired correspondence. We evaluate 445 HSSD objects with their TRELLIS captions and corresponding Wan trajectories. The compatibility map is learned from the training samples and kept fixed during evaluation. To examine whether the selected state reflects actual 3D–video correspondence rather than only a common preference along the diffusion trajectory, we further compare correctly paired 3D representations with shuffled ones on 50 completed reverse trajectories.

Correspondence-specific route-point selection. Replacing the paired 3D representation with a shuffled one changes the selected scheduler state for 43 of the 50 trajectories, with an average displacement of 2.84 states, as shown in Figure 7a. This indicates that the preferred route point is sensitive to sample-level 3D–video correspondence rather than being determined solely by the scheduler. Therefore, using a single fixed injection point for all samples can overlook differences in their latent compatibility.

Compact-window adaptive routing. Figure 7b shows that the sample-wise optima are concentrated in a relatively narrow region of the reverse trajectory. For Wan, the normalized bounds of $\mathcal{W}^{(\mathrm{Wan})}$ are $[71.6\%, 88.0\%]$; the window contains 12 of the 50 evaluated scheduler states and covers most sample optima. The globally optimal fixed point is located at $q_{\mathrm{fix}} = 81.7\%$, but only 13.26% of samples select exactly this state, while 39.78% differ from it by more than two scheduler states.

Figure 7c further shows that replacing the sample-specific optimum with the global fixed point introduces non-negligible compatibility error for a substantial portion of samples. These observations motivate the routing strategy adopted by GeoRoute: we first identify a compact, model-specific candidate region and then let the Timestep Router select the route point for each sample within this region. This preserves the efficiency of searching over a small number of states while accounting for sample-dependent 3D–video compatibility.

A.2 O2: Timestep-Dependent Spectral Compatibility

We analyze the spectral profile of geometry-compatible information across candidate route states. Using the same 50 completed Wan reverse trajectories, we measure both the generic spectral evolution of the video latent and its frequency-resolved correspondence with the matched 3D feature. We partition each latent into five disjoint radial 3D Fourier bands over its temporal and spatial axes:

$$\mathcal{B} = \big\{[0, .08), [.08, .16), [.16, .24), [.24, .36), [.36, f_{\max}]\big\}. \tag{17}$$

Here, $f_{\max}$ is the largest normalized radial frequency. At each timestep, we record the fraction of total latent energy contained in each band and separately measure matched-3D correspondence. For band b, let $\Psi_b(\cdot)$ denote the corresponding band-pass feature, represented by pooled signed responses and local RMS magnitudes. After centering each modality across samples, define the per-sample normalized video and 3D features as

$$\mathbf{f}^{v}_{i,t,b} = \mathrm{Norm}(\Psi_b(\mathbf{x}_{i,t})), \qquad \mathbf{f}^{3\mathrm{D}}_{i,b} = \mathrm{Norm}(\Psi_b(\mathbf{g}_i)). \tag{18}$$

The paired-3D specificity is

$$\Gamma_i(t,b) = \cos\big(\mathbf{f}^{v}_{i,t,b}, \mathbf{f}^{3\mathrm{D}}_{i,b}\big) - \frac{1}{N-1}\sum_{j\neq i}\cos\big(\mathbf{f}^{v}_{i,t,b}, \mathbf{f}^{3\mathrm{D}}_{j,b}\big)\,, \tag{19}$$

where $\mathbf{g}_i$ is the mapped 3D feature paired with sample i. A positive value means that the video feature in band b identifies its matched 3D geometry more strongly than the nonpaired 3D features; this comparison controls for timestep-dependent latent norms and generic spectral statistics.

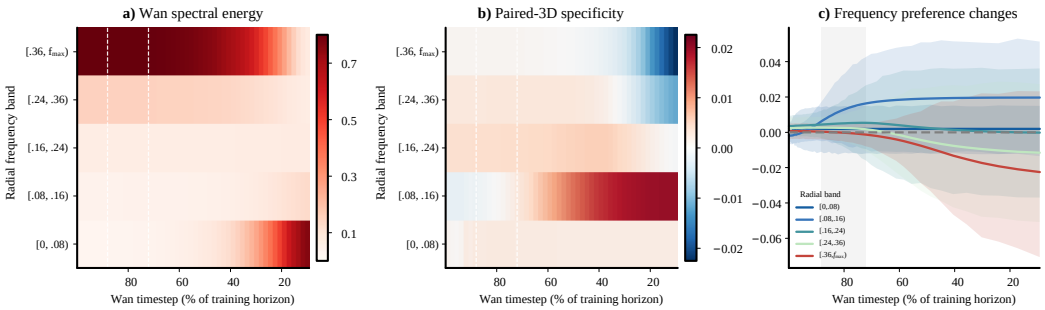


Figure 8: **Frequency-resolved analysis of Wan reverse trajectories.** (a) Total latent-energy fraction in each disjoint Fourier band. (b) Dataset-mean paired-3D specificity from Eq. (19). (c) The same specificity shown as band-wise curves.

Frequency-specific 3D correspondence. Figure 8a reports band-wise latent energy, whereas Figure 8b–c reports the correspondence-controlled specificity $\Gamma_i(t,b)$. The Wan latent is spectrally nonstationary during reverse denoising, and the matched-over-nonpaired result further shows that geometry correspondence changes with the spectrum: the strongest band changes from $[.16, .24)$ over the early reverse states to $[.08, .16)$ from approximately 89.8% of the training horizon onward. Within the routing interval, the relative strengths of the bands continue to evolve, while the higher-frequency bands become progressively less compatible with the matched 3D feature. Therefore,

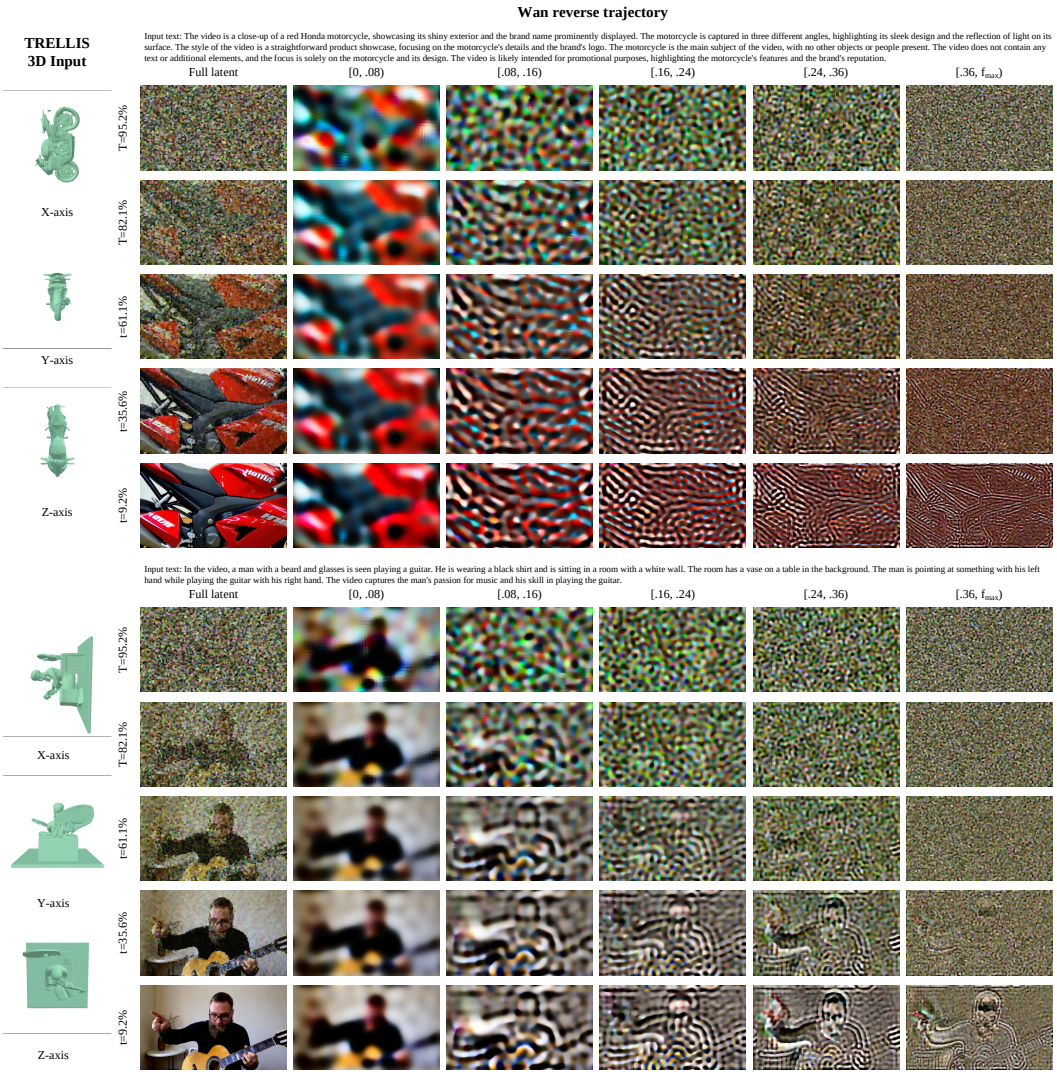


Figure 9: **Qualitative frequency evolution along Wan reverse denoising.** We visualize a single-object example (motorcycle) and a multi-object human–object interaction (man playing a guitar). For each example, the paired TRELLIS prior is shown together with the full-latent reconstruction and five frequency-isolated reconstructions at selected reverse timesteps. During the early reverse interval, from $q_t = 95.2\%$ to 61.1%, the coarse object layout and geometry are progressively established, primarily within the two lowest bands $[0, .08)$ and $[.08, .16)$. After $q_t = 61.1\%$, these low-frequency structures remain visually stable, while subsequent changes mainly refine higher-frequency appearance, color, texture, and boundaries. Each isolated band is RMS-normalized for visibility, and the displayed intensity follows this normalization.

both the useful spectral composition and the strength of geometric correspondence depend on the routed timestep.

Cross-probe spectral consistency. Figure 9 shows that, in both examples, the dominant silhouette, part arrangement, and spatial organization emerge mainly in the $[0, .16)$ range during the early reverse process and are largely established by $q_t = 61.1\%$. The low-frequency reconstruction then undergoes moderate changes, whereas later denoising increasingly refines appearance and fine-scale details. The Wan compatibility window $\mathcal{W}^{(\mathrm{Wan})}$, with normalized bounds $[71.6\%, 88.0\%]$, lies within this geometry-formation stage, where the structural latent is organized and remains responsive to 3D guidance. This quantitative interaction motivates conditioning the Spectral Router on the

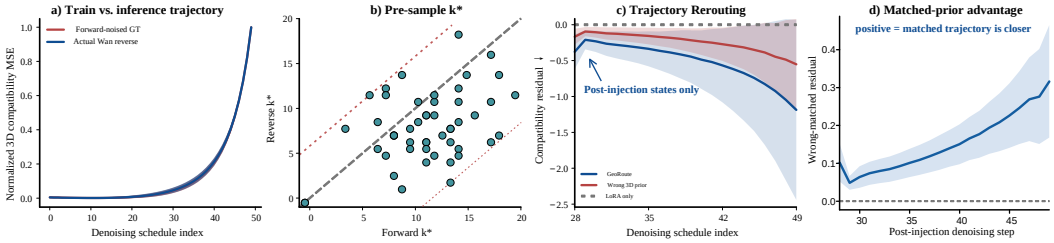


Figure 10: **Training–inference trajectory similarity and persistent rerouting ($N = 50$).** (a) Mean sample-normalized compatibility profiles for forward-noised training states and actual reverse-inference states. (b) Per-sample forward and reverse optima. (c) Post-injection compatibility residuals relative to the LoRA baseline for matched and wrong 3D priors. (d) The stepwise wrong-minus-matched residual, where positive values favor the matched prior.

selected route state, concentrating intervention in the geometry-bearing low-frequency range while adapting the precise cutoff and guidance strength. The controlled fixed-to-adaptive comparison in Table 4 provides empirical verification of adaptive Spectral Router.

A.3 Trajectory Analysis

We use 50 trajectories for two trajectory-level analyses: training–inference compatibility similarity and the persistence of rerouting after a geometry transition.

Training–inference trajectory similarity. Let $\mathbf{x}_{i,k}^{\chi}$ be the video latent of sample i at scheduler index k, where k counts evaluated states in reverse-denoising order (increasing toward the clean endpoint), and $\chi \in \{\mathrm{fwd}, \mathrm{rev}\}$ denotes the forward-noised training trajectory or actual reverse trajectory. With $\mathbf{g}_i^{\mathrm{LF}} = \mathrm{LPF}(\mathbf{g}_i)$ denoting the aligned low-frequency 3D feature and D the number of compared latent entries, we compute

$$\mathcal{C}_i^{\chi}(k) = \frac{1}{D}\left\|\mathrm{LPF}(\mathbf{x}_{i,k}^{\chi}) - \mathbf{g}_i^{\mathrm{LF}}\right\|_2^2, \qquad k_i^{*,\chi} = \arg\min_k \mathcal{C}_i^{\chi}(k). \tag{20}$$

The mean compatibility profiles follow similar trajectories (Figure 10a), while their sample-wise optima have a mean absolute displacement of 5.6 states in the 50-step schedule (Figure 10b). This forward–reverse comparison motivates the inference-time rule used by GeoRoute: selecting the final route point from the actual reverse states.

Persistent rerouting after geometry injection. For method a, let $\mathcal{C}_i^a(k)$ denote Eq. (20) evaluated on its post-injection trajectory; the matched 3D prior remains the target under both matched and shuffled-prior injections. We define

$$\Delta\mathcal{C}_i^a(k) = \mathcal{C}_i^a(k) - \mathcal{C}_i^{\mathrm{LoRA}}(k), \qquad k > k_{\mathrm{inj}}, \tag{21}$$

so a negative residual denotes lower matched-3D discrepancy than the LoRA baseline. Figure 10c directly visualizes the subsequent trajectories: after the injection state is excluded, matched and wrong priors produce distinct continuations relative to the LoRA baseline, with the matched prior remaining closer to the target geometry. Averaged over all post-injection states, the matched prior has a lower residual than the wrong prior for 76% of samples. The stepwise advantage $\Delta\mathcal{C}_i^{\mathrm{wrong}}(k) - \Delta\mathcal{C}_i^{\mathrm{matched}}(k)$ remains positive throughout the evaluated continuation (Figure 10d), with an overall mean of 0.153×10^{-3}. Evaluation begins at $k = 27$, after the injection state, so these curves measure the propagated denoising continuation. The persistent post-injection separation is the trajectory-level effect analyzed here; final video-quality and geometry improvements are evaluated in Sec. 5.

A.4 Inference Latency

We measure inference latency on a single NVIDIA A800-SXM4-80GB GPU using 81 frames at 832×480 resolution and 50 denoising steps. For each backbone, we sequentially generate three

videos under identical settings, treat the first run as warm-up, and report the mean latency of the remaining two runs. CUDA synchronization is applied at each timing boundary. Total latency includes text encoding and VAE decoding, both shared by the corresponding baseline and GeoRoute. Overhead is computed relative to each pretrained backbone.

Table 5: Mean inference latency per video on one NVIDIA A800-SXM4-80GB. All times are reported in seconds.

Method	TRELLIS generation	Router	Video denoising	Total latency	Overhead
Wan2.1	–	–	203.91	209.99	–
Wan2.1 + GeoRoute	7.08	< 0.01	236.24	249.21	$+18.7\%$
CogVideoX	–	–	211.32	232.10	–
CogVideoX + GeoRoute	9.04	< 0.01	254.62	278.44	$+20.0\%$
HunyuanVideo	–	–	1174.81	1213.15	–
HunyuanVideo + GeoRoute	17.19	< 0.01	1270.03	1310.65	$+8.0\%$

The routing operation itself is negligible, requiring less than 0.01 seconds per video. Including online TRELLIS generation and routed video denoising, GeoRoute introduces an average end-to-end overhead of 15.6% across the three backbones, with a maximum overhead of 20.0%. The runtime decomposition attributes the added computation primarily to 3D-prior generation and geometry-aware denoising.

B GeoRoute Design and Optimization Details

This section provides additional implementation and formulation details of GeoRoute, including tri-plane prior encoding, sample-adaptive Timestep Router, adaptive Spectral Router, and decoupled cross-trajectory optimization. We use θ for frozen backbone parameters and ψ for trainable video LoRA parameters; ϕ and η parameterize the 3D–video adapter and Spectral Router, respectively.

B.1 Tri-Plane Geometry Representation

Let the sparse representation produced by the pretrained generative 3D model be

$$\mathbf{z}^{\mathrm{3D}} = \{(\boldsymbol{\xi}_j, \mathbf{f}_j)\}_{j=1}^{N_s}, \tag{22}$$

where $\boldsymbol{\xi}_j$ denotes the spatial coordinate of the j-th active voxel and $\mathbf{f}_j$ denotes its feature vector.

We first scatter the sparse features into a dense voxel feature volume

$$\mathbf{V} \in \mathbb{R}^{C_s \times D \times H_3 \times W_3}. \tag{23}$$

The volume is projected onto three orthogonal planes:

$$\mathbf{p}_{xy} = \mathrm{Avg}_z(\mathbf{V}), \qquad \mathbf{p}_{xz} = \mathrm{Avg}_y(\mathbf{V}), \qquad \mathbf{p}_{yz} = \mathrm{Avg}_x(\mathbf{V}), \tag{24}$$

which are concatenated as

$$\mathbf{p} = \mathrm{Cat}[\mathbf{p}_{xy}, \mathbf{p}_{xz}, \mathbf{p}_{yz}]. \tag{25}$$

The tri-plane representation preserves structural information along all three principal spatial axes while providing a compact representation suitable for interaction with the video latent.

For compatibility evaluation, we resize the tri-plane and flatten its spatial positions into tokens:

$$\mathbf{P} = \mathrm{Tokens}(\Pi(\mathbf{p})) \in \mathbb{R}^{HW \times C_p}, \qquad C_p = 3C_s. \tag{26}$$

This deterministic representation is used to construct the dataset-level compatibility curve and its geometry-compatible window.

For geometry injection, GeoRoute additionally employs a lightweight trainable adapter $\mathcal{A}_\phi$ to transform the tri-plane representation into a feature compatible with the full video latent:

$$\mathbf{g}_{\phi,t} = \mathcal{A}_\phi\left(\mathcal{U}(\mathbf{p}), \mathbf{e}_t\right), \tag{27}$$

where $\mathcal{U}$ spatially resizes, temporally expands, and channel-aligns the tri-plane feature and $\mathbf{e}_t$ denotes the timestep embedding.

We separate timestep selection from geometry injection. The shared mapping $\mathbf{W}^*$ scores candidate states for the first-increase selection rule inside the model-specific compatibility window, whereas $\mathbf{g}_{\phi,t}$ is used after the route state has been selected.

B.2 Adaptive Timestep Router

The compatibility window defines a compact set of candidate states with high 3D–video correspondence. GeoRoute evaluates these states in denoising order and selects a local low-discrepancy state online from the observed decrease–increase pattern.

The frozen shared mapping first produces the selection feature

$$\mathbf{h}^{\mathrm{sel}} = \mathbf{P}\mathbf{W}^*. \tag{28}$$

Let $\{t_1 > \cdots > t_J\}$ denote the ordered scheduler states within $\mathcal{W}^{(m)}$. For every candidate, we compute

$$d_{t_j} = \frac{1}{HWC_v} \left\| \bar{\mathbf{X}}_{t_j} - \mathbf{h}^{\mathrm{sel}} \right\|_F^2, \qquad j = 1, \ldots, J. \tag{29}$$

The online stopping rule is

$$j^* = \min\{j \geq 2 : d_{t_j} > d_{t_{j-1}}\}, \qquad k^* = t_{j^*-1}. \tag{30}$$

Once the distance rises, the preceding state is selected as the route point. This first-increase rule detects a local minimum; it need not recover the global minimum of a non-unimodal profile. If the sequence remains decreasing until the end of the window, we set $k^* = t_J$. The empirical window restricts the scan to a compact compatible region, and $\mathbf{W}^*$ provides the same learned metric in training and inference.

Routing during training. During training, the availability of the clean latent $\mathbf{x}_0$ allows all candidate states in $\mathcal{W}^{(m)}$ to be constructed analytically:

$$\mathbf{x}_t = (1 - \sigma_t)\mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad t \in \mathcal{W}^{(m)}. \tag{31}$$

We use the same Gaussian noise realization $\boldsymbol{\epsilon}$ for all candidate timesteps of a sample. This makes variation of d_t primarily reflect the evolution of the latent representation along the noise schedule.

We evaluate the analytically constructed states in denoising order and apply Eq. (30) at the first increase of d_t. Since the comparison-based stopping decision is discrete, timestep selection is treated as non-differentiable and kept outside back-propagation.

Routing during inference. During inference, video latent states are naturally obtained as the reverse process traverses $\mathcal{W}^{(m)}$. For each candidate state, GeoRoute evaluates

$$d_t = \frac{1}{HWC_v} \left\| \bar{\mathbf{X}}_t - \mathbf{P}\mathbf{W}^* \right\|_F^2 \tag{32}$$

using the already evaluated backbone states for scoring. GeoRoute retains the preceding latent and score while traversing the window. When the current distance first exceeds the preceding one, it rolls back to the stored state $\mathbf{x}_{k^*}$, applies the spectral transition once, and resumes denoising from $\mathbf{x}^*_{k^*}$.

B.3 Adaptive Spectral Router

After determining *where* geometric guidance should enter the denoising trajectory, GeoRoute further determines *which* frequency components should be modified and *how strongly* the 3D representation should influence them.

At the selected route point k^*, the trainable adapter produces

$$\mathbf{g}_{\phi,k^*} = \mathcal{A}_\phi\left(\mathcal{U}(\mathbf{p}), \mathbf{e}_{k^*}\right). \tag{33}$$

The adaptive Spectral Router uses both the selected timestep and its compatibility distance as routing signals:

$$(\widehat{\rho}, \widehat{\beta}, \widehat{\gamma}) = \mathcal{S}_\eta \left(d_{k^*}, \mathbf{e}_{k^*} \right). \tag{34}$$

We transform the predicted variables into valid ranges:

$$\rho_{k^*} = \rho_{\min} + (\rho_{\max} - \rho_{\min}) \, \mathrm{Sigmoid}(\widehat{\rho}), \tag{35}$$

$$\beta_{k^*} = \beta_{\min} + \mathrm{Softplus}(\widehat{\beta}), \qquad \gamma_{k^*} = \gamma_{\max} \, \mathrm{Sigmoid}(\widehat{\gamma}). \tag{36}$$

Here, ρ_{k^*} specifies the effective spectral range of geometric intervention, β_{k^*} controls the sharpness of the frequency boundary, and γ_{k^*} determines the strength of geometric guidance.

For a normalized temporal–spatial frequency coordinate $\boldsymbol{\omega}$, the adaptive mask is defined as

$$\mathbf{m}_{k^*}(\boldsymbol{\omega}) = \mathrm{Sigmoid} \left[\beta_{k^*} \left(\rho_{k^*} - \|\boldsymbol{\omega}\|_2 \right) \right]. \tag{37}$$

The selected video component is

$$\mathbf{x}_{k^*}^L = \mathcal{F}^{-1} \left(\mathbf{m}_{k^*} \odot \mathcal{F}(\mathbf{x}_{k^*}) \right), \tag{38}$$

with complementary component

$$\mathbf{x}_{k^*}^H = \mathbf{x}_{k^*} - \mathbf{x}_{k^*}^L. \tag{39}$$

The corresponding low-frequency geometric feature is

$$\mathbf{g}_{\phi,k^*}^L = \mathcal{F}^{-1} \left(\mathbf{m}_{k^*} \odot \mathcal{F}(\mathbf{g}_{\phi,k^*}) \right). \tag{40}$$

GeoRoute performs the trajectory transition as

$$\mathbf{x}_{k^*}^* = \mathbf{x}_{k^*}^H + \mathrm{MM} \left(\mathbf{x}_{k^*}^L + \gamma_{k^*} \mathbf{g}_{\phi,k^*}^L ; \mathbf{x}_{k^*}^L \right), \tag{41}$$

where

$$\mathrm{MM}(\mathbf{a}; \mathbf{b}) = \frac{\mathbf{a} - \mu(\mathbf{a})}{\mathrm{std}(\mathbf{a}) + \epsilon} \, \mathrm{std}(\mathbf{b}) + \mu(\mathbf{b}) \tag{42}$$

matches the statistics of the modified structural feature to those of the original video latent.

B.4 Inference

GeoRoute decomposes timestep selection into two stages:

$$\underbrace{\{d_i(t)\}_{i,t} \to \mathcal{W}^{(m)}}_{\text{window estimation}} \qquad + \qquad \underbrace{\{d_t\}_{t \in \mathcal{W}^{(m)}} \to k^*}_{\text{online minimum selection}}. \tag{43}$$

The first stage estimates a compact compatibility window for each pretrained video backbone. The second stage reuses the frozen shared mapping $\mathbf{W}^*$ and applies the first-increase stopping rule in Eq. (30). This formulation selects the first detected local minimum in one pass and applies an identical timestep-selection rule during training and inference.

During inference, we follow the original denoising trajectory until it enters $\mathcal{W}^{(m)}$ and evaluate the compatibility score along the window. Once the route point k^* is selected, GeoRoute performs a single adaptive spectral transition and resumes normal denoising:

$$\mathbf{x}_{k^*} \xrightarrow{\mathcal{R}_{\phi,\eta}} \mathbf{x}_{k^*}^*, \qquad \mathbf{x}_{k^*}^* \xrightarrow{\mathcal{D}_{\theta,\psi}^{k^* \to 0}} \mathbf{x}_0^*. \tag{44}$$

Thus, GeoRoute introduces the generative 3D prior a single time, while adaptively determining *where*, *at which frequencies*, and *with what strength* it should alter the video denoising trajectory.



Figure 11: **Qualitative evidence of joint-optimization collapse.** Each row shows six uniformly sampled frames from a validation video generated at training step 2,500 with joint adapter optimization. Both examples exhibit severe persistent blockwise/mosaic corruption; in (b), global corruption persists through most frames. These outputs define the empirical collapse criterion for joint-optimization rows.

B.5 Decoupled Cross-Trajectory Training

GeoRoute contains two trainable processes with different objectives. The 3D adapter establishes correspondence between the generative 3D representation and the geometry-related video latent, whereas the video diffusion model learns to denoise from the routed latent state. Direct end-to-end optimization couples these objectives and lets the adapter and denoiser co-adapt, disrupting the geometry-compatible trajectory identified by the routing analysis.

Figure 11 operationalizes the collapse marker ($\times$) used in Table 4: under the matched training protocol, joint variants that produce severely corrupted validation videos are marked as collapse consistently.

At the selected route point k^*, the 3D adapter transforms the generative 3D representation into the video latent space:

$$\mathbf{g}_{\phi,k^*} = \mathcal{A}_\phi\left(\mathcal{U}(\mathbf{p}), \mathbf{e}_{k^*}\right). \tag{45}$$

The adapter establishes a cross-modal bridge between the 3D prior and the geometry-related component of the video latent. We optimize it using

$$\mathcal{L}_{\text{map}} = \left\|\mathbf{g}_{\phi,k^*} - \text{sg}\left(\mathbf{x}^L_{k^*}\right)\right\|_2^2, \tag{46}$$

where $\mathbf{x}^L_{k^*}$ denotes the selected structural component of the video latent and $\text{sg}(\cdot)$ denotes stop-gradient. The role of $\mathcal{L}_{\text{map}}$ is therefore explicit:

$$\mathbf{g}_{\phi,k^*} \to \mathbf{x}^L_{k^*}. \tag{47}$$

The mapped geometric feature is subsequently injected into the video latent through the routing module. Here, $\mathcal{R}_\eta$ denotes spectral fusion conditioned on an already mapped feature; together with $\mathcal{A}_\phi$, it constitutes the full transition $\mathcal{R}_{\phi,\eta}$ introduced in Sec. 3.1:

$$\mathbf{x}^*_{k^*} = \mathcal{R}_\eta\left(\mathbf{x}_{k^*}, \text{sg}(\mathbf{g}_{\phi,k^*})\right), \tag{48}$$

from which the remaining denoising process follows a modified trajectory. For the flow-matching backbones, the video model is optimized with the detached target velocity defined below:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_t\left[\left\|\mathbf{v}_{\theta,\psi}\left(\mathbf{x}^*_t, t, \mathbf{c}\right) - \text{sg}(\mathbf{u}^*)\right\|_2^2\right]. \tag{49}$$

Then, given the routed state $\mathbf{x}^*_{k^*}$, we construct the remaining training path while retaining the original clean video latent $\mathbf{x}_0$ as the supervised endpoint. For the linear flow-matching parameterization

$$\mathbf{x}_s = (1-s)\mathbf{x}_0 + s\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{50}$$

let $s^* = \sigma_{k^*} > 0$ denote the noise level at the selected route point. We define an effective noise endpoint

$$\boldsymbol{\epsilon}^* = \frac{\mathbf{x}^*_{k^*} - (1-s^*)\mathbf{x}_0}{s^*}. \tag{51}$$

The post-routing training states are then constructed as

$$\mathbf{x}_s^* = (1-s)\mathbf{x}_0 + s\boldsymbol{\epsilon}^*, \qquad 0 \le s \le s^*. \tag{52}$$

This path passes through the routed state at $s = s^*$ and reaches the original clean latent at $s = 0$. Its target velocity is

$$\mathbf{u}^* = \boldsymbol{\epsilon}^* - \mathbf{x}_0 = \frac{\mathbf{x}_{k^*}^* - \mathbf{x}_0}{s^*}. \tag{53}$$

We sample a scheduler timestep t uniformly from the discrete post-routing timestep set $\mathcal{T}_{\text{post}}(k^*)$ and evaluate the corresponding state at $s = \sigma_t$. The diffusion objective is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\left\|\mathbf{v}_{\theta,\psi}\left(\mathbf{x}_t^*, t, \mathbf{c}\right) - \text{sg}\left(\mathbf{u}^*\right)\right\|_2^2\right], \tag{54}$$

where the expectation is over training samples, sampled noise, and post-routing timesteps. Here $\mathbf{x}_t^* = \mathbf{x}_s^*|_{s=\sigma_t}$ relates scheduler-indexed states to noise-level-indexed states. The target velocity is detached, while gradients through $\mathbf{x}_t^*$ update the Spectral Router. The adapter output remains detached before fusion, as described above.

The effective endpoint $\boldsymbol{\epsilon}^*$ is determined by the routed state rather than sampled independently from a Gaussian distribution. This construction trains the video model on states following the routing intervention while preserving the original clean-video supervision. These interpolation and velocity-target equations describe the flow-matching backbones; CogVideoX retains its native diffusion parameterization and training objective.

Gradient separation. The mapping loss and diffusion loss optimize different roles in the routed system. The mapping loss aligns $\mathbf{g}_{\phi,k^*}$ with the structural video component $\mathbf{x}_{k^*}^L$, while the diffusion loss trains the Spectral Router and video LoRA to denoise from the routed state $\mathbf{x}_{k^*}^*$. With joint optimization, the adapter receives gradients from both objectives:

$$\nabla_\phi \mathcal{L} = \nabla_\phi \mathcal{L}_{\text{diff}} + \lambda_{\text{map}} \nabla_\phi \mathcal{L}_{\text{map}}. \tag{55}$$

The two gradient terms point to different targets. The mapping gradient encourages

$$\mathbf{g}_{\phi,k^*} \to \mathbf{x}_{k^*}^L, \tag{56}$$

whereas the diffusion gradient encourages a feature that minimizes the downstream generation objective:

$$\mathbf{g}_{\phi,k^*} \to \arg\min_{\mathbf{g}} \mathcal{L}_{\text{diff}}\left(\mathcal{R}_\eta(\mathbf{x}_{k^*}, \mathbf{g})\right). \tag{57}$$

Decoupled training assigns these updates to separate parameter groups: ϕ is updated by $\mathcal{L}_{\text{map}}$, whereas η and ψ are updated by $\mathcal{L}_{\text{diff}}$. This keeps geometry alignment anchored to the selected video-latent component while the video branch learns the post-route denoising continuation.

Let $\mathbf{g}^{\text{geo}}$ denote the desired geometry-aligned feature. Under joint optimization, the adapter produces

$$\mathbf{g}_{\phi,k^*} = \mathbf{g}^{\text{geo}} + \Delta\mathbf{g}, \tag{58}$$

where $\Delta\mathbf{g}$ denotes a deviation favored by the downstream diffusion objective. The resulting routed latent becomes

$$\mathbf{x}_{k^*}^* = \mathcal{R}_\eta\left(\mathbf{x}_{k^*}, \mathbf{g}^{\text{geo}} + \Delta\mathbf{g}\right). \tag{59}$$

The trainable denoiser then adapts to the shifted latent distribution, so the adapter and denoiser co-adapt: the former changes the routed representation, while the latter learns the corresponding continuation. GeoRoute separates these roles so that the adapter preserves cross-modal correspondence and the denoiser adapts after the routed state is formed.

Local Jacobian view. The effect of adapter deviation on the routed trajectory follows from a local first-order analysis. Let

$$\mathbf{x}_{k^*}^{\text{geo}} = \mathcal{R}_\eta\left(\mathbf{x}_{k^*}, \mathbf{g}^{\text{geo}}\right) \tag{60}$$

denote the intended geometry-routed state. For a small perturbation $\Delta\mathbf{g}$, the routing operation can be approximated as

$$\mathbf{x}_{k^*}^* \approx \mathbf{x}_{k^*}^{\text{geo}} + \mathbf{J}_{\mathcal{R}}\Delta\mathbf{g}, \tag{61}$$

where

$$\mathbf{J}_{\mathcal{R}} = \left.\frac{\partial \mathcal{R}_\eta(\mathbf{x}_{k^*}, \mathbf{g})}{\partial \mathbf{g}}\right|_{\mathbf{g}=\mathbf{g}^{\text{geo}}} \tag{62}$$

is the Jacobian of the routing operation with respect to the geometric feature. Hence,

$$\Delta \mathbf{x}_{k^*} = \mathbf{x}^*_{k^*} - \mathbf{x}^{\text{geo}}_{k^*} \approx \mathbf{J}_{\mathcal{R}} \Delta \mathbf{g}. \tag{63}$$

The shifted state is subsequently processed by the trainable video denoiser. Locally, the change in the predicted velocity is approximated by

$$\Delta \mathbf{v} \approx \mathbf{J}_x \Delta \mathbf{x}_{k^*} + \mathbf{J}_\psi \Delta \psi, \tag{64}$$

where

$$\mathbf{J}_x = \frac{\partial \mathbf{v}}{\partial \mathbf{x}}, \qquad \mathbf{J}_\psi = \frac{\partial \mathbf{v}}{\partial \psi}. \tag{65}$$

Substituting the routing perturbation gives

$$\Delta \mathbf{v} \approx \mathbf{J}_x \mathbf{J}_{\mathcal{R}} \Delta \mathbf{g} + \mathbf{J}_\psi \Delta \psi. \tag{66}$$

If the adapter-induced change lies approximately in the range of $\mathbf{J}_\psi$, a denoiser update can satisfy

$$\mathbf{J}_\psi \Delta \psi \approx -\mathbf{J}_x \mathbf{J}_{\mathcal{R}} \Delta \mathbf{g}. \tag{67}$$

Such an update offsets the adapter-induced velocity change, yielding

$$\Delta \mathbf{v} \approx \mathbf{0} \quad \text{with} \quad \Delta \mathbf{g} \neq \mathbf{0}. \tag{68}$$

This illustrates a possible local mechanism for joint co-adaptation, rather than a guarantee of collapse: the denoising prediction can remain nearly unchanged while the adapter drifts from the geometry-aligned feature. The stop-gradient paths described above prevent the diffusion loss from driving this adapter drift.

Trajectory interpretation. The compatibility probe measures discrepancy between a mapped 3D feature and the temporally averaged video latent along sampled trajectories. It does not establish a distance to the full video-latent manifold or guarantee that the routed state remains on that manifold. The analysis instead motivates a localized intervention at a compatible state. Decoupled optimization preserves the intended roles: the adapter is anchored to $\mathbf{x}^L_{k^*}$ by the mapping objective, and the video model learns the continuation from the routed state. The collapse examples and ablations provide empirical evidence for this optimization choice.

C GeoVideoBench Metrics and Validation

C.1 Benchmark Overview and Evaluation Dimensions

Conventional video-generation benchmarks primarily measure appearance quality, semantic alignment, and generic temporal consistency. GeoVideoBench targets a common failure mode in which the foreground object remains recognizable while its geometry gradually stretches, collapses, drifts, merges, or changes topology. We introduce **GeoVideoBench**, a lightweight, reference-free benchmark for evaluating the temporal geometric integrity of the principal foreground object.

GeoVideoBench takes a generated RGB video and its text prompt as input; it requires no ground-truth video, calibrated camera, reference mesh, or multi-view supervision. It evaluates four complementary dimensions, summarized in Table 6.

The four dimensions are reported separately because they capture different failure modes and rely on estimators with different sensitivities. For compact model-level comparisons, GeoVideoBench additionally reports a reliability-calibrated geometric aggregate in Eq. 14. Its weights are estimated once from the controlled geometric and nuisance interventions and reused unchanged across all evaluated methods. Equal-weighted results are retained as an aggregation sensitivity analysis in Tables 10 and 11. The resulting scores measure 3D-aware temporal consistency in the monocular reference-free setting used throughout our experiments. Since the pipeline relies on monocular estimates and principal-object masks, legitimate articulation, occlusion, and viewpoint changes can affect its scores.

Table 6: The four evaluation dimensions of GeoVideoBench. Each metric lies in $[0, 100]$, and a higher value indicates better temporal geometric consistency.

Evaluation dimension	Metric	Geometric evidence	Primary failure modes
Local metric rigidity	RSPS	Scale-normalized local pseudo-3D distances	Nonuniform stretching, compression, and local collapse
Long-range 3D explainability	L3RC	Held-out reprojection under one global $\mathrm{Sim}(3)$ motion	Structural drift across temporally distant frames
Global low-frequency shape–depth stability	GSSC	Spectrum of a canonical silhouette–depth field	Coarse silhouette change and depth-mass redistribution
Part topology consistency	PTC	Holes and skeleton structure of the principal component	Part disappearance, hole changes, and abnormal branches

C.2 Shared Evaluation Pipeline

Video sampling and anchor frame. Let a video contain N decoded RGB frames $\{\mathbf{I}^{(n)}\}_{n=0}^{N-1}$, each of spatial resolution $H \times W$. We uniformly sample

$$N_{\mathrm{eval}} = \min(N, 16) \tag{69}$$

frames over the complete temporal extent. The sampled frames are denoted by $\{\mathbf{I}_t\}_{t=0}^{N_{\mathrm{eval}}-1}$. Frame $t = 0$ is used as the anchor frame.

Foreground isolation. We extract the text prompt from the video filename and use it as the grounding query for Grounding DINO (Liu et al., 2024). The highest-confidence, valid-area detection in the anchor frame is selected as the principal-object box. The box is enlarged by 6% along each axis to reduce truncation at object boundaries. If no valid prompt-conditioned detection is found, we issue a generic main-subject query over common object categories. The resulting anchor box initializes SAM 2 (Ravi et al., 2025), which propagates a binary foreground mask $\mathbf{M}_t \in \{0, 1\}^{H \times W}$ through the sampled frames. We apply a 5×5 morphological closing operation and retain the largest connected foreground component in each frame. Videos with median foreground-area ratio below 0.2% are marked invalid.

Relative depth estimation. For every sampled frame, Depth Anything V2 (Yang et al., 2024a) predicts a dense inverse-depth-like proximity map $\widehat{\mathbf{Z}}_t$. To suppress extreme predictions, we normalize each map using its 2nd and 98th percentiles:

$$\mathbf{Z}_t(\mathbf{u}) = \mathrm{clip}\left(\frac{\widehat{\mathbf{Z}}_t(\mathbf{u}) - q_{2,t}}{q_{98,t} - q_{2,t} + \epsilon},\, 0,\, 1\right), \tag{70}$$

where $\mathbf{u} = (u, v)^\top$ is an image coordinate, $q_{p,t}$ is the p-th percentile of $\widehat{\mathbf{Z}}_t$, and ϵ is a small constant. We convert proximity to a positive relative-depth field through

$$\mathbf{D}_t(\mathbf{u}) = \frac{1}{0.1 + 0.9\mathbf{Z}_t(\mathbf{u})}. \tag{71}$$

This parameterization bounds the effective relative depth and improves numerical stability when lifting tracked points.

Foreground correspondences. We first detect up to $K_{\max} = 256$ Shi–Tomasi features (Shi & Tomasi, 1994) inside $\mathbf{M}_0$. If fewer than half of the requested points are detected, the remaining locations are filled by approximately uniform foreground-grid sampling. We track the points through the sampled frames using pyramidal Lucas–Kanade optical flow (Lucas & Kanade, 1981). A correspondence is accepted when (i) both forward and backward tracking succeed, (ii) the forward–backward error is below two pixels, (iii) the point remains inside the image, and (iv) the point lies inside the propagated foreground mask. Once a track becomes invalid, it remains invalid in subsequent frames. Let

$$\mathbf{u}_{t,i} = (u_{t,i}, v_{t,i})^\top, \qquad a_{t,i} \in \{0, 1\} \tag{72}$$

denote the image position and validity of track i in frame t.

Pseudo-3D representation. We use a centered pinhole approximation with

$$f = 0.8\max(H, W), \qquad c_x = \frac{W-1}{2}, \qquad c_y = \frac{H-1}{2}. \tag{73}$$

The relative depth of a subpixel track location is obtained by bilinear sampling, $z_{t,i} = \mathbf{D}_t(\mathbf{u}_{t,i})$. Each valid track is lifted to a pseudo-3D point

$$\mathbf{y}_{t,i} = \begin{bmatrix} (u_{t,i} - c_x) z_{t,i}/f \\ (v_{t,i} - c_y) z_{t,i}/f \\ z_{t,i} \end{bmatrix} \in \mathbb{R}^3. \tag{74}$$

The resulting pseudo-3D points are scale-normalized before scoring, and the metrics below remove global scale where possible and use robust statistics.

Common error-to-score mapping. For error-based metrics, we use the monotonic mapping

$$\Phi_\tau(E) = 100 \exp\left(-\frac{\max(E, 0)}{\tau}\right), \tag{75}$$

where τ is fixed for each metric across all compared methods. The mapping assigns 100 to zero error, preserves the raw-error ranking, and smoothly compresses large outliers. The raw errors define the scores through the formulas below.

C.3 Metric Definitions

All four metrics use the shared representations from Sec. C.2, but each isolates a different geometric dimension from Table 6.

C.3.1 RSPS: Local Metric Rigidity

What it measures. RSPS measures the temporal stability of local pairwise distances within the object. Translation, rotation, and global scale changes preserve these normalized distances, whereas nonuniform stretching, compression, or local collapse changes them.

Anchor graph construction. Let

$$\mathcal{V}_0 = \{i \mid a_{0,i} = 1\} \tag{76}$$

be the set of valid anchor-frame tracks. We construct a symmetric k-nearest neighbor graph in anchor pseudo-3D space with $k = 6$:

$$\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0). \tag{77}$$

For every edge $(i, j) \in \mathcal{E}_0$, its frame-t pseudo-3D length is

$$d_{t,ij} = \|\mathbf{y}_{t,i} - \mathbf{y}_{t,j}\|_2. \tag{78}$$

At time t, edges with two valid endpoints are evaluated:

$$\mathcal{E}_t = \{(i, j) \in \mathcal{E}_0 \mid a_{t,i} a_{t,j} = 1\}. \tag{79}$$

Scale-normalized local deformation. We normalize edge lengths by a robust frame-level characteristic scale:

$$m_0 = \underset{(p,q)\in\mathcal{E}_0}{\text{median}}\, d_{0,pq}, \tag{80}$$

$$m_t = \underset{(p,q)\in\mathcal{E}_t}{\text{median}}\, d_{t,pq}, \tag{81}$$

$$\widetilde{d}_{0,ij} = \frac{d_{0,ij}}{m_0 + \epsilon}, \qquad \widetilde{d}_{t,ij} = \frac{d_{t,ij}}{m_t + \epsilon}. \tag{82}$$

The logarithmic strain of an edge is

$$e^{\text{rigid}}_{t,ij} = \left| \log \frac{\widetilde{d}_{t,ij} + \epsilon}{\widetilde{d}_{0,ij} + \epsilon} \right|. \tag{83}$$

The log ratio treats an expansion by a factor α and a contraction by $1/\alpha$ symmetrically. Normalizing by m_t removes global pseudo-depth scale fluctuations as well as legitimate changes in object size caused by camera motion.

We retain frames with at least 16 valid anchor edges and pool their edge errors. The video-level rigidity error and score are

$$E_{\mathrm{RSPS}} = \underset{\substack{t=1,\dots,N_{\mathrm{eval}}-1 \\ (i,j)\in\mathcal{E}_t}}{\mathrm{median}} e_{t,ij}^{\mathrm{rigid}}, \tag{84}$$

$$\mathrm{RSPS} = \Phi_{0.20}(E_{\mathrm{RSPS}})\,. \tag{85}$$

C.3.2 L3RC: Long-Range 3D Explainability

What it measures. L3RC evaluates the extent to which one global 3D similarity transform explains the object's evolution over temporally distant frames. It fits the transform on a subset of tracks and evaluates it on held-out object points, exposing structural drift beyond local edge changes.

Long-range frame set. We evaluate target frames in the final two-thirds of the video:

$$\mathcal{I}_{\mathrm{long}} = \{t \,|\, t = \max(1, \lfloor N_{\mathrm{eval}}/3 \rfloor), \dots, N_{\mathrm{eval}} - 1\}\,. \tag{86}$$

For each t, the common-track set is

$$\mathcal{C}_t = \{i \mid a_{0,i}a_{t,i} = 1\}. \tag{87}$$

Frames with fewer than 16 common tracks are excluded.

Held-out similarity alignment. Using a deterministic random permutation, we split $\mathcal{C}_t$ into a fitting set $\mathcal{F}_t$ containing approximately 25% of the tracks and a disjoint testing set $\mathcal{Q}_t = \mathcal{C}_t \setminus \mathcal{F}_t$. At least six fitting and six testing points are required. From $\mathcal{F}_t$, we estimate the least-squares similarity transformation

$$(s_t, \mathbf{R}_t, \mathbf{b}_t) = \arg\min_{s>0,\, \mathbf{R}\in\mathrm{SO}(3),\, \mathbf{b}\in\mathbb{R}^3} \sum_{i\in\mathcal{F}_t} \|\mathbf{y}_{t,i} - (s\mathbf{R}\mathbf{y}_{0,i} + \mathbf{b})\|_2^2\,, \tag{88}$$

using the closed-form Umeyama solution (Umeyama, 1991). The predicted target-frame pseudo-3D location of a held-out point is

$$\widehat{\mathbf{y}}_{t,i} = s_t\mathbf{R}_t\mathbf{y}_{0,i} + \mathbf{b}_t, \qquad i \in \mathcal{Q}_t. \tag{89}$$

Held-out reprojection error. Let $\widehat{\mathbf{y}}_{t,i} = (\widehat{y}_{t,i}^{(1)}, \widehat{y}_{t,i}^{(2)}, \widehat{y}_{t,i}^{(3)})^\top$. We project the prediction using the same pinhole approximation:

$$\widehat{\mathbf{u}}_{t,i} = \begin{bmatrix} f\widehat{y}_{t,i}^{(1)} / \max(\widehat{y}_{t,i}^{(3)}, \epsilon) + c_x \\ f\widehat{y}_{t,i}^{(2)} / \max(\widehat{y}_{t,i}^{(3)}, \epsilon) + c_y \end{bmatrix}. \tag{90}$$

The frame-level held-out reprojection error is normalized by the image diagonal:

$$r_t = \underset{i\in\mathcal{Q}_t}{\mathrm{median}} \frac{\|\widehat{\mathbf{u}}_{t,i} - \mathbf{u}_{t,i}\|_2}{\sqrt{H^2 + W^2}}. \tag{91}$$

Because the points in $\mathcal{Q}_t$ are held out from transformation estimation, a low error requires the fitted global 3D motion to generalize across the object.

Relative depth-order consistency. Monocular depth has frame-dependent scale and offset, whereas ordinal depth is typically more stable. We therefore compute Spearman's rank correlation

$$\rho_t^{\mathrm{rank}} = \mathrm{Spearman}\left(\{\widehat{y}_{t,i}^{(3)}\}_{i\in\mathcal{Q}_t}, \{z_{t,i}\}_{i\in\mathcal{Q}_t}\right) \tag{92}$$

and define the depth-order penalty

$$o_t = \frac{1 - \rho_t^{\mathrm{rank}}}{2}. \tag{93}$$

When the correlation is undefined, we set $o_t = 0.5$.

The complete long-range error and score are

$$E_{\mathrm{L3RC}} = \underset{t\in\mathcal{I}_{\mathrm{valid}}}{\mathrm{median}}\, r_t + 0.02\, \underset{t\in\mathcal{I}_{\mathrm{valid}}}{\mathrm{median}}\, o_t, \tag{94}$$

$$\mathrm{L3RC} = \Phi_{0.05}(E_{\mathrm{L3RC}})\,, \tag{95}$$

where $\mathcal{I}_{\mathrm{valid}} \subseteq \mathcal{I}_{\mathrm{long}}$ contains the frames satisfying the correspondence requirements.

C.3.3 GSSC: Global Low-Frequency Shape–Depth Stability

What it measures. GSSC measures the stability of the object's coarse silhouette and internal depth organization. It compares the low-frequency spectrum of a dense, canonical object-centric shape–depth field, making the metric less sensitive to texture flicker and fine boundary noise.

Canonical object-centric field. For each mask $\mathbf{M}_t$, we compute the covariance matrix of its foreground pixel coordinates. The dominant eigenvector defines a principal orientation θ_t. We rotate both $\mathbf{M}_t$ and $\mathbf{Z}_t$ around the foreground centroid by θ_t, crop the tight foreground bounding box with 8% padding along each dimension, and resize the result to $S \times S$, where $S = 96$. This canonicalization reduces sensitivity to translation, image-space scale, and dominant in-plane orientation.

Within the aligned foreground, the proximity values are normalized using their 5th and 95th percentiles, producing $\overline{\mathbf{Z}}_t \in [0,1]^{S\times S}$. We then define the dense shape–depth field

$$\mathbf{F}_t(x,y) = \mathbf{M}_t^{\mathrm{a}}(x,y)\left(0.5 + 0.5\overline{\mathbf{Z}}_t(x,y)\right), \tag{96}$$

where $\mathbf{M}_t^{\mathrm{a}}$ is the aligned and resized mask. The constant term preserves silhouette information, while the proximity-dependent term encodes coarse internal depth organization.

Low-frequency spectral descriptor. We compute the centered log-amplitude spectrum

$$\mathbf{A}_t = \log\left(1 + |\mathrm{fftshift}\left(\mathrm{FFT2}(\mathbf{F}_t)\right)|\right). \tag{97}$$

For a spectral coordinate (x,y), define the normalized radial frequency

$$\nu(x,y) = \sqrt{\left(\frac{y - S/2}{S}\right)^2 + \left(\frac{x - S/2}{S}\right)^2}. \tag{98}$$

We retain the annular low-frequency region

$$\Omega_{\mathrm{LF}} = \left\{(x,y) \;\middle|\; \frac{1}{S} < \nu(x,y) \le 0.18\right\}, \tag{99}$$

excluding the DC coefficient. After vectorizing the retained coefficients, we subtract their mean and apply ℓ_2 normalization:

$$\mathbf{s}_t = \frac{\mathrm{vec}(\mathbf{A}_t|_{\Omega_{\mathrm{LF}}}) - \mu_t \mathbf{1}}{\|\mathrm{vec}(\mathbf{A}_t|_{\Omega_{\mathrm{LF}}}) - \mu_t \mathbf{1}\|_2 + \epsilon}. \tag{100}$$

We compare every valid frame with the anchor using cosine similarity

$$c_t = \mathrm{clip}\left(\mathbf{s}_0^\top \mathbf{s}_t, -1, 1\right), \tag{101}$$

and define

$$E_{\mathrm{GSSC}} = \underset{t=1,\ldots,N_{\mathrm{eval}}-1}{\mathrm{median}} (1 - c_t), \tag{102}$$

$$\mathrm{GSSC} = 100\left(1 - \frac{E_{\mathrm{GSSC}}}{2}\right). \tag{103}$$

The score is clipped to $[0, 100]$.

C.3.4 PTC: Part Topology Consistency

What it measures. PTC measures discrete structural changes that distance- or spectrum-based metrics can miss, including disappearance of thin structures, opening or closing of holes, and creation or removal of branches. It uses a silhouette-derived topology signature of the principal foreground component and requires no semantic part annotations.

Topology signature. Each binary mask is first resized to 128×128 using nearest-neighbor interpolation. We compute the seven-dimensional signature

$$\mathbf{h}_t^{\mathrm{topo}} = [C_t,\, H_t,\, E_t,\, B_t,\, Q_t,\, A_t,\, L_t]^\top. \tag{104}$$

Its components are defined as follows.

- C_t is the number of 8-connected foreground components with an area of at least 16 pixels. Since the common mask-cleaning stage retains the principal component, this term is normally one and primarily acts as an integrity check; the discriminative topology signal is supplied by the remaining signature dimensions.
- Let χ_t denote the Euler number under 8-connectivity. The number of holes is

$$H_t = \max(0, C_t - \chi_t). \tag{105}$$

- We skeletonize the foreground mask. A skeleton pixel with exactly one occupied neighbor in its 3×3 neighborhood is an endpoint; E_t is the total number of such endpoints.
- A skeleton pixel with at least three occupied neighbors is marked as a branch pixel. Spatially adjacent branch pixels are grouped into one junction, and B_t is the number of connected junction groups. Grouping avoids over-counting a single thick junction as multiple branch points.
- The silhouette solidity is

$$Q_t = \frac{\mathrm{Area}(\partial\Omega_t)}{\mathrm{Area}(\mathrm{Hull}(\partial\Omega_t)) + \epsilon}, \tag{106}$$

  where $\partial\Omega_t$ is the largest external contour and $\mathrm{Hull}(\cdot)$ is its convex hull.
- For the width w_t and height h_t of the contour bounding box, the log aspect ratio is

$$A_t = \log \frac{\max(w_t, 1)}{\max(h_t, 1)}. \tag{107}$$

- The area-normalized skeleton length is

$$L_t = \frac{\#\{\text{skeleton pixels at } t\}}{\sqrt{\max(\#\{\text{foreground pixels at } t\}, 1)}}. \tag{108}$$

Normalized topology drift. The seven signature dimensions have different numerical ranges. We use fixed normalization scales

$$\boldsymbol{\kappa} = [1,\ 1,\ 4,\ 3,\ 0.25,\ 0.7,\ 2]^\top \tag{109}$$

and importance weights

$$\boldsymbol{\lambda}^{\mathrm{topo}} = [1.2,\ 1.2,\ 1.0,\ 1.0,\ 0.7,\ 0.7,\ 0.5]^\top. \tag{110}$$

For dimension j, the anchor-relative normalized deviation is

$$\delta_{t,j} = \min\left(\frac{|h_{t,j}^{\mathrm{topo}} - h_{0,j}^{\mathrm{topo}}|}{\kappa_j}, 3\right). \tag{111}$$

Clipping at three prevents a single segmentation or skeletonization outlier from dominating the complete metric. The frame-level drift, video-level error, and final score are

$$e_t^{\mathrm{topo}} = \frac{\sum_{j=1}^{7} \lambda_j^{\mathrm{topo}} \delta_{t,j}}{\sum_{j=1}^{7} \lambda_j^{\mathrm{topo}}}, \tag{112}$$

$$E_{\mathrm{PTC}} = \underset{t=1,\ldots,N_{\mathrm{eval}}-1}{\mathrm{median}} e_t^{\mathrm{topo}}, \tag{113}$$

$$\mathrm{PTC} = \Phi_{0.35}(E_{\mathrm{PTC}}). \tag{114}$$

C.4 Controlled Validation and Human Alignment

Controlled dataset. We construct a synthetic controlled set containing six articulated assets, two camera paths (static and orbital), and 25 configurations per asset–camera pair, yielding 300 videos in total. Each video contains 24 frames at 320×320 resolution and 12 fps. For every targeted failure, the five motion-corruption levels correspond to clean, low, medium, high, and maximal corruption. The clean video is shared within each matched asset–camera group, while the four corrupted videos progressively increase the prescribed geometric intervention.

Table 7: Controlled perturbations used to validate GeoVideoBench.

Target	Perturbation	Controlled modification
RSPS	Local stretch	The length of one articulated limb changes progressively while the remaining object structure and base animation are fixed.
L3RC	Part drift	A connected subset of parts undergoes an independent 3D displacement that violates a single global similarity transform.
GSSC	Global warp	The complete object undergoes a temporally varying anisotropic deformation that changes its low-frequency global shape.
PTC	Topology change	Distal and proximal parts are progressively removed, modifying silhouette endpoints, branches, and skeleton structure.
Control	Lighting	Illumination intensity changes while geometry, camera path, object appearance, and base animation remain fixed.
Control	Appearance	Object colors vary while geometry, projected support, camera path, and base animation remain fixed.
Control	Camera path	The same clean asset and animation are rendered using static and orbital cameras.

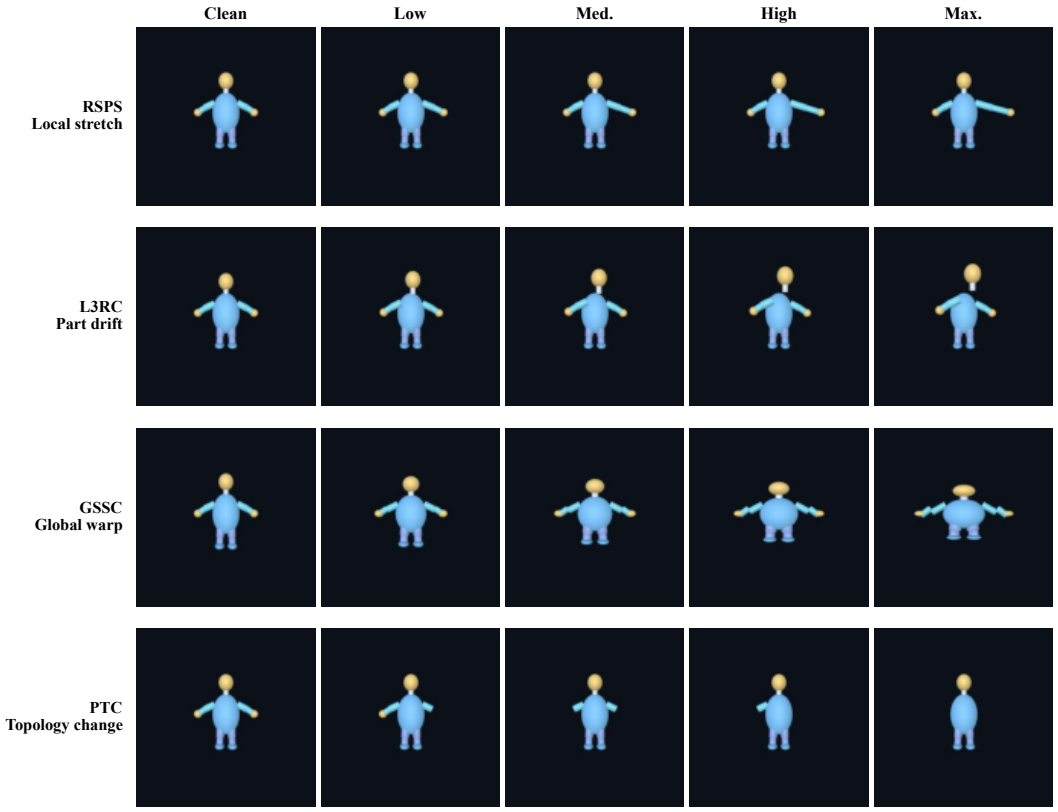


Figure 12: Representative controlled sequences. Columns show the five motion-corruption levels, from clean to maximal corruption, and rows show the perturbation associated with each GeoVideoBench metric. Asset identity, camera path, rendering configuration, and base animation are matched within each row.

Intervention design. Table 7 associates each metric with a controlled failure family. Local stretch tests relative-scale preservation, part drift violates locally rigid part relationships, global warp changes low-frequency shape, and progressive part removal changes topology. Lighting, appearance, and camera-path interventions vary video formation while preserving the targeted geometry.

Figure 12 visualizes the prescribed severity ordering. The interventions are deliberately simple and localized, providing known geometric changes and clean-to-corrupted orderings independent of sub-

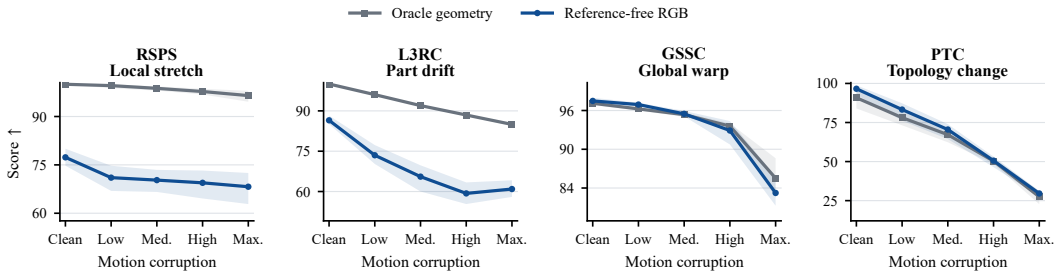


Figure 13: Targeted responses from oracle geometry and the complete reference-free RGB pipeline. Curves show the mean score at each of the five motion-corruption levels. Per-panel ordinate ranges are metric-specific.

Table 8: Quantitative controlled-validity results. The Spearman coefficient ρ measures the association between corruption severity and score, for which a more negative value indicates stronger degradation sensitivity. Δ is the paired clean-to-maximal score decrease from the reference-free RGB pipeline. For RC Overall, the four targeted failure families are balanced for the pooled correlation and averaged within each asset–camera pair for Δ. Absolute decreases are reported in each metric's own scale.

Metric	Oracle ρ	RGB ρ	Δ score
RSPS	-0.816	-0.223	9.12
L3RC	-0.980	-0.587	25.57
GSSC	-0.670	-0.714	14.25
PTC	-0.891	-0.960	67.02
RC Overall	-0.672	-0.708	24.82

jective quality labels. The four metrics use different score calibrations, so absolute score differences are interpreted within each metric. GSSC is evaluated by its ordered response to corruption severity.

Oracle and reference-free evaluation. We evaluate each sequence in two modes. The oracle mode computes the four scores directly from the rendered 3D geometry and therefore measures the response of the metric definitions to their intended failures. The reference-free mode applies the complete RGB-video pipeline and therefore additionally includes errors from foreground extraction, landmark recovery, part correspondence, and geometric estimation.

As shown in Figure 13, oracle scores decrease under their corresponding interventions, confirming that the four definitions encode the intended geometric failures. The reference-free curves retain the same clean-to-maximal degradation tendency, showing that the controlled signal remains observable after the complete RGB-to-geometry estimation pipeline. L3RC has a strong net decline from clean to maximal corruption, with one local high-to-maximal fluctuation.

Quantitative analysis. For each severity level, we first average the two camera paths within each asset and then compute the cross-asset mean. Table 8 reports the Spearman association between prescribed severity and per-video score, the paired clean-to-maximal score decrease over the 12 asset–camera pairs.

All four component metrics and RC Overall decrease from clean to maximal corruption. RSPS exhibits a positive paired decrease under progressive local stretching, while L3RC, GSSC, and PTC respond consistently to their corresponding controlled interventions. RC Overall retains the prescribed severity ordering after balancing the four failure families, with $\rho = -0.708$ in the complete RGB pipeline.

Reliability-calibrated aggregation. The controlled set provides a model-independent basis for a reliability-calibrated aggregate. Table 9 reports the targeted geometric response G_m, the mean absolute responses to the three nuisance controls, and the resulting reliability and normalized weight from Eq. 13. We set N_m to the largest lighting, appearance, or camera-path response. This weighting emphasizes metrics whose targeted geometric response dominates nuisance variation.

Table 9: Derivation of the fixed weights used in the reliability-calibrated GeoVideoBench aggregate. G_m is the clean-to-maximal response to the targeted geometric intervention. Lighting, appearance, and camera columns report mean absolute score changes under the corresponding nuisance controls, and N_m is their maximum. $R_m = G_m/(G_m + N_m + \epsilon)$ measures the reliability of the targeted response relative to the largest nuisance response, and $w_m = R_m / \sum_\ell R_\ell$ is its normalized aggregation weight.

Metric	G_m	Lighting	Appearance	Camera	N_m	R_m	w_m
RSPS	9.12	39.76	17.89	38.61	39.76	0.187	0.078
L3RC	25.57	23.79	9.43	26.85	26.85	0.488	0.203
GSSC	14.25	0.41	0.53	2.84	2.84	0.834	0.347
PTC	67.02	2.20	7.92	5.10	7.92	0.894	0.372

Table 10: Sensitivity of the GeoRoute improvement to the Overall definition. Each entry is the GeoRoute gain over its matched backbone. Eq. denotes the unweighted arithmetic mean, and RC denotes Eq. 14.

Backbone	Δ Overall (Eq.)	Δ Overall (RC)
Wan	+5.41	+3.59
HunyuanVideo	+7.59	+9.28
CogVideoX	+3.12	+4.65
Average	+5.37	+5.84

The resulting weights reflect the measured balance between each metric's targeted geometric response and its responses to lighting, appearance, and camera-path controls. They are estimated once on the synthetic controlled set and then applied unchanged to every baseline, competing method, and ablation variant. All four component metrics remain visible in the result tables, and their correspondence with human judgments is evaluated below.

Table 10 shows that GeoRoute improves all three backbones under both equal weighting and reliability calibration, indicating that the observed improvement is consistent across the two reported aggregation rules.

The SOTA comparison in Table 11 yields the same ranking for the original arithmetic mean and the reliability-calibrated aggregate, with GeoRoute ranking first under both definitions.

Cross-asset repeatability. We assess cross-asset repeatability by separately plotting the trajectory of each asset after averaging its two camera paths.

Figure 14 shows consistent clean-to-maximal trends across renderer assets. Across assets, all four metrics retain their mean clean-to-maximal decrease under the corresponding controlled geometric intervention.

Photometric and camera-path sensitivity. We finally measure score changes when geometry is preserved while illumination, appearance, or camera path changes. These experiments quantify the operational sensitivity introduced by the reference-free RGB front end.

As shown in Figure 15, the reference-free metrics exhibit different responses to illumination, appearance, and camera-path changes. RSPS and L3RC use fine-scale masks, landmarks, correspondences, and relative-depth estimates; maximal lighting variation changes the reference-free RSPS score by 39.8 points. This measurement characterizes the complete RGB-to-geometry front end alongside the oracle-geometry validation above.

Human alignment. We measure metric–human agreement with the same eight participants, blinded interface, and response protocol as the main user study in Appendix F. The alignment block uses 180 videos generated by the baseline and GeoRoute-enhanced variants of Wan, HunyuanVideo, and CogVideoX. Its comparison graph contains 360 pairs, each video appears in exactly four comparisons, and 52 comparisons are prompt-matched. All eight participants independently judge preservation of the main object's geometry; method identities and automatic metric scores are hidden. Ties

Table 11: Aggregation sensitivity for the CogVideoX-based SOTA comparison. Higher is better. GeoRoute ranks first under both Overall definitions.

Method	Overall (Eq.)	Overall (RC)
CogVideoX	60.96	61.12
GeoVideo	62.00	62.91
VideoGPA	63.60	65.43
GeoRoute	**64.07**	**65.78**

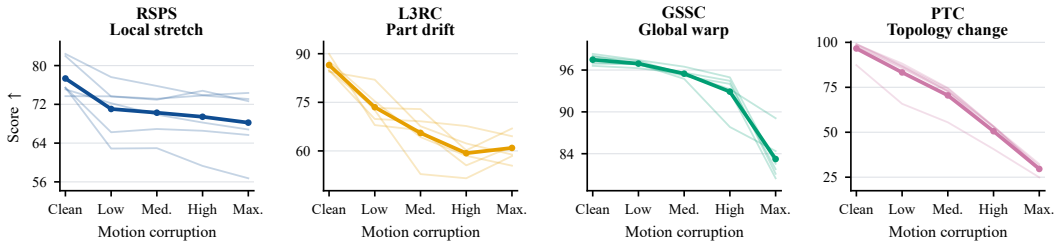


Figure 14: Asset-level repeatability of the reference-free pipeline. Thin curves denote individual assets after averaging the two camera paths, and thick curves denote the cross-asset mean. The five positions correspond to clean, low, medium, high, and maximal motion corruption.

are retained as valid judgments, and *Cannot judge* or missing responses are excluded. The metric–human alignment block and the main method-preference block are two reporting blocks from the same eight-participant study.

We fit Bradley–Terry scores from all judgeable participant-level comparisons, treating a human tie as half a win for each video, and compute Spearman correlation between the resulting aggregate human ranking and each automatic metric. Pairwise agreement is computed over judgeable participant-level comparisons with a strict human preference and measures the fraction for which the automatic metric selects the same video; exact metric ties are also omitted. Table 12 summarizes representative human correlations reported by open-domain video evaluators, including CLIP-, SSIM-, optical-flow-, warping-, and MLLM-based signals. The VideoScore/EvalCrafter CLIP-Score, SSIM-sim, and GPT-4o entries provide the reference values for contextualizing human alignment used in the main text.

GeoVideoBench obtains consistent positive correlations (0.226–0.334) and above-chance pairwise agreement (64.6%–68.9%) across all four geometric dimensions. RC Overall ($\rho = 0.275$) falls within the VideoScore/EvalCrafter range spanned by CLIP-Score (0.217), SSIM-sim (0.269), and GPT-4o (0.287), while RSPS attains the highest pairwise agreement among our component metrics (68.9%). These results place GeoVideoBench on a standard human-alignment scale and connect its geometry-focused scores with human judgments.

Controlled validation summary. These controlled interventions test known geometric failures with prescribed severity orderings. Together with real-model comparisons and human evaluation, they show that each metric reacts predictably when its targeted geometric property is progressively degraded.

D Experimental Implementation Details

CogVideoX and HunyuanVideo follow the same training configuration as Wan, while retaining their backbone-specific objectives. The flow-matching derivation applies to Wan and HunyuanVideo; CogVideoX retains its native diffusion objective.

Compute environment. We train the Wan-based GeoRoute model on a single node with eight NVIDIA A800-SXM4 GPUs, each with 80 GB of memory. The node contains a 56-core Intel Xeon Gold 6348 CPU and 1 TB of system memory. Training uses Python 3.10.20, PyTorch 2.4.1 with CUDA 12.4, PyTorch Lightning 2.6.5, Transformers 4.43.3, and Diffusers 0.38.0. We use eight-

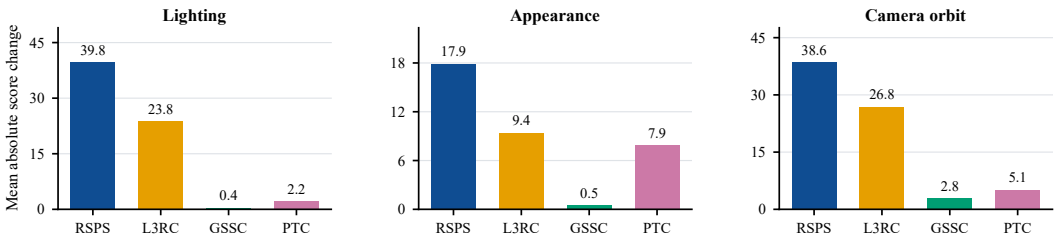


Figure 15: Photometric and camera-path responses of the reference-free pipeline. Bars report the mean absolute score change under maximal lighting and appearance perturbations and between clean static and orbital camera paths. Lower values indicate stronger invariance to the corresponding non-geometric change.

Table 12: Human alignment of GeoVideoBench and representative open-domain video-generation evaluators. Higher is better. Values for prior work are transcribed from the corresponding papers, and ranges summarize results across evaluation dimensions. Pairwise agreement is reported for GeoVideoBench.

Work	Metric or evaluator	Evaluation set / human criterion	Spearman ρ	Pairwise agreement (%)
GeoVideoBench (ours)	RC Overall	Our study / geometric consistency	0.275	66.0
	RSPS	Our study / geometric consistency	0.334	68.9
	L3RC	Our study / geometric consistency	0.226	68.6
	GSSC	Our study / geometric consistency	0.283	68.3
	PTC	Our study / geometric consistency	0.231	64.6
VideoScore (He et al., 2024)	CLIP-Score	EvalCrafter / multi-aspect quality	0.217	–
	SSIM-sim	EvalCrafter / multi-aspect quality	0.269	–
	GPT-4o	EvalCrafter / multi-aspect quality	0.287	–
Video-Bench (Han et al., 2025)	Conventional components	Video-Bench / dimension-specific ratings	0.260–0.545	–
	ComBench evaluator	Video-Bench / condition alignment	0.611–0.696	–
	Video-Bench evaluator	Video-Bench / quality and condition alignment	0.402–0.750	–
T2V-CompBench (Sun et al., 2024)	CLIP	T2V-CompBench / seven compositional categories	−0.014–0.486	–
	B-CLIP	T2V-CompBench / seven compositional categories	−0.089–0.356	–
	B-BLEU	T2V-CompBench / seven compositional categories	−0.198–0.358	–
	ViCLIP	T2V-CompBench / seven compositional categories	0.007–0.612	–
	D-LLaVA	Dynamic attribute binding	0.506	–
	DOT	Motion binding	0.537	–
	G-Dino	Numeracy / spatial relationship	0.538 / 0.706	–
	Grid-LLaVA	Interaction / action / consistent attribute	0.593 / 0.735 / 0.789	–
VMBench (Ling et al., 2025)	SSIM	VMBench / motion quality	0.016	–
	RAFT	VMBench / motion quality	−0.017	–
	DINO	VMBench / motion quality	0.094	–
	CLIP	VMBench / motion quality	0.150	–
	DOVER Technical	VMBench / motion quality	0.206	–
	AMT	VMBench / motion quality	0.036	–
	Warping Error	VMBench / motion quality	−0.060	–
	Prompted MLLMs	VMBench / motion quality	0.153–0.269	–
	PMM (model–rule synergy)	VMBench / motion quality	0.622	–

process distributed data parallelism with one process per GPU, bfloat16 mixed precision, gradient checkpointing, and `find_unused_parameters=True`.

Training data and preprocessing. We train on the OpenVidHD subset used in our main experiments. The training split contains approximately 26.3K text–video pairs. For each sample, we randomly select a contiguous clip of 81 frames; videos shorter than 81 frames are padded by repeating the last frame. Frames are resized to 832×480 and normalized to $[-1, 1]$. We use a per-GPU batch size of one with no gradient accumulation, giving an effective global batch size of eight. Each process uses four data-loading workers.

Base models and trainable parameters. The video generator is initialized from Wan2.1-T2V-1.3B, while the text-to-3D branch is initialized from TRELLIS-text-xlarge. The TRELLIS model, UMT5 text encoder, Wan VAE, and original Wan parameters remain frozen throughout training. We add LoRA adapters to the query, key, value, and output projections of both self-attention and text cross-attention, as well as the two linear layers of the feed-forward network in all 30 Wan transformer blocks. The LoRA rank is 32, the scaling parameter is 64, and dropout is disabled. This produces 300 adapted linear layers and approximately 43.7M trainable LoRA parameters. The 3D mapping network uses two point-wise $1 \times 1 \times 1$ Conv3D layers with channel dimensions $16 \to 32 \to 16$. A

64-dimensional sinusoidal embedding of $1000t/T$ is passed through a $64 \to 64 \to 64$ conditioning MLP with SiLU between its linear layers. Its output is split into 32-channel scale and shift vectors, broadcast over temporal and spatial positions. For the hidden feature $\mathbf{h}$ from the first convolution, the adapter computes $\mathrm{SiLU}((1 + \mathbf{a}_t) \odot \mathbf{h} + \mathbf{b}_t)$ before the output convolution. The conditioning MLP's final layer and the output convolution have zero-initialized weights and biases, so the initial geometric injection is zero.

Spectral Router configuration. The Spectral Router concatenates the same 64-dimensional timestep embedding with $\log(1 + \max(d_{k^*}, 0))$ and processes the resulting 65-dimensional vector with a $65 \to 64 \to 3$ MLP, using SiLU after the first linear layer. The sinusoidal embedding uses a maximum period of 10,000. No additional normalization layer is applied to the router input. We set $\rho_{\min} = 0.08$, $\rho_{\max} = 0.24$, $\beta_{\min} = 4$, and $\gamma_{\max} = 0.10$ in the output transformations of Appendix B.3. The final linear weights are initialized to zero, and its biases are chosen using the inverse sigmoid and inverse softplus transforms so that the initial outputs are $(\rho, \beta, \gamma) = (0.18, 16, 0.05)$. The remaining layers use the default PyTorch initialization.

3D latent construction and injection. For every sample, we feed the complete caption directly to TRELLIS and generate one sparse 3D latent online. This latent serves as a prompt-level dominant geometry prior. We voxelize it on a 64^3 grid and average along the three spatial axes to obtain a tri-plane representation, which is normalized and resized for timestep selection and geometry injection. Within $\mathcal{W}^{(m)}$, the frozen shared mapping $\mathbf{W}^*$ scores the candidate video states and selects k^* using Eq. (30). The timestep-conditioned adapter then produces $\mathbf{g}_{\phi,k^*}$, and the Spectral Router predicts the cutoff ρ_{k^*}, boundary sharpness β_{k^*}, and injection strength γ_{k^*}. After fusion, we match the mean and standard deviation of the modified low-frequency component to those of the original video low-frequency component before recombining it with the unchanged complementary component. The prior controls in Table 4 further show that performance depends on prompt–geometry correspondence.

Training objective and gradient separation. Let $\mathbf{x}_0$ denote the encoded video latent and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. After selecting the route point k^*, we construct $\mathbf{x}_{k^*} = (1 - \sigma_{k^*})\mathbf{x}_0 + \sigma_{k^*}\boldsymbol{\epsilon}$ and apply the spectral transition to obtain $\mathbf{x}^*_{k^*}$. For each training sample, we uniformly select one of the discrete scheduler timesteps at or below k^* and use the interpolation and target defined in Appendix B.5. The diffusion branch detaches the adapter output before spectral fusion, updating the Spectral Router and Wan LoRA. The adapter is trained separately by minimizing the MSE between its output and the detached low-frequency component of $\mathbf{x}_{k^*}$. We set $\lambda_{\text{map}} = 1$, so the two losses are summed with equal weights. This separation allows the LoRA and Spectral Router branches to learn the post-injection denoising trajectory while the adapter learns to align the 3D and video latent spaces.

Optimization and inference. We optimize the LoRA adapters, Spectral Router, and 3D mapping MLP through the decoupled gradient paths for 10,000 optimizer steps using AdamW with a constant learning rate of 1×10^{-5}, $\beta_1 = \beta_2 = 0.9$, weight decay 0.03, and gradient-norm clipping at 1.0. The random seed is 42. We generate validation videos every 500 steps and save checkpoints every 1,000 steps. Unless otherwise stated, inference uses the 10,000-step checkpoint, 50 Flow-UniPC sampling steps, a scheduler shift of 5.0, classifier-free guidance scale 6.0, and the standard Wan negative prompt. As denoising traverses $\mathcal{W}^{(m)}$, GeoRoute evaluates the compatibility distance at each candidate state, selects k^*, applies one adaptive spectral transition, and resumes denoising from $\mathbf{x}^*_{k^*}$. The generated videos contain 81 frames at 832×480 and are saved at 16 fps.

E Extended Ablation Studies

Transient versus persistent geometry injection. The quantitative results confirm that persistently injecting geometry into noisy video latents is less effective than GeoRoute's single transient intervention. All-step injection lowers the GeoVideoBench aggregate to 63.28 because it continues to modify late denoising states after high-level appearance and structure have formed. Restricting persistent injection to the compatibility window partially recovers geometric consistency to 67.00 and reduces the colored flare and temporal flicker visible in Figure 16, but its overall visual quality remains below the single-injection GeoRoute.

Ablation Study of Injection Strategies. Table 14 compares GeoRoute with cross-attention-based injection. Cross-attention injection applied throughout denoising or restricted to the geometry-

Table 13: Detailed configuration for training the Wan-based GeoRoute model.

Configuration	Value
Video backbone	Wan2.1-T2V-1.3B
3D backbone	TRELLIS-text-xlarge
Training resolution / frames	832×480 / 81
Wan latent shape	$16 \times 21 \times 60 \times 104$
GPUs / distributed strategy	$8\times$ A800-SXM4-80GB / DDP
Numerical precision	bfloat16 mixed precision
Per-GPU / global batch size	1 / 8
Gradient accumulation	1
Optimizer	AdamW
Learning rate	1×10^{-5}, constant
AdamW β_1, β_2 / weight decay	$(0.9, 0.9)$ / 0.03
Gradient clipping	1.0
Training length	10,000 optimizer steps
LoRA rank / scaling / dropout	32 / 64 / 0
LoRA target layers	Self-attention, cross-attention, and FFN linear layers
3D mapping MLP	Conv3D $16 \rightarrow 32 \rightarrow 16$, kernel size 1
Timestep Router	Single adaptive route point k^* selected within $\mathcal{W}^{(m)}$
Spectral Router	Sample-specific ρ_{k^*}, β_{k^*}, and γ_{k^*}
Loss weights	Diffusion MSE: 1; 3D mapping MSE: 1
Validation / checkpoint interval	500 / 1,000 steps
Inference sampler	Flow-UniPC, 50 steps, shift 5.0
Classifier-free guidance	6.0
Random seed	42

compatible window $\mathcal{W}^{(m)}$ remains close to the pretrained Wan model. The result highlights the importance of modifying spatial latent components at the selected route point.

In contrast, GeoRoute directly modifies the spatial latent at the selected route point and redirects the subsequent denoising trajectory using the aligned 3D prior. This produces substantially higher subject and background consistency while preserving motion smoothness, resulting in the best three-metric VBench average.

Table 14: Ablation of different injection strategies on VBench.

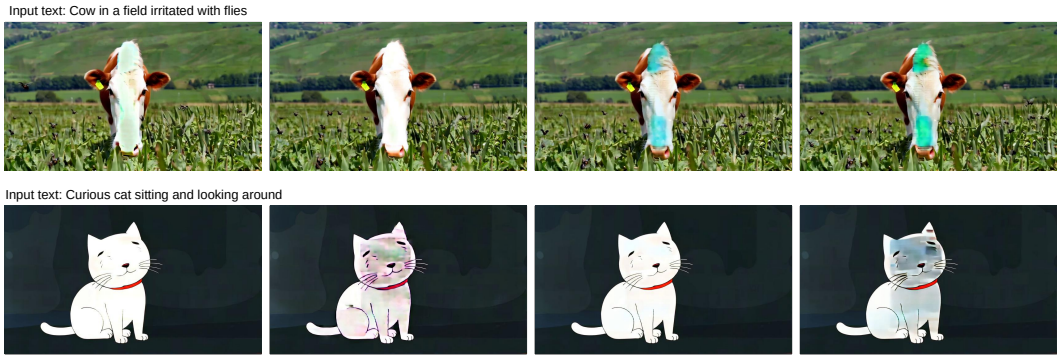
Method	Subject Consistency	Background Consistency	Motion Smoothness	Avg.
Wan	93.18	95.07	98.59	95.61
Cross-Attention Injection (All Timesteps)	93.21	95.38	96.91	95.17
Cross-Attention Injection (At $\mathcal{W}^{(m)}$)	93.27	94.94	98.26	95.49
Noise Injection (GeoRoute)	**97.92**	**97.89**	**98.63**	**98.15**

F User Study

We conducted a pairwise human preference study with eight volunteers. All participants were Master's or Ph.D. students with at least three years of experience in multimodal understanding and generation or art-related fields. Each participant provided informed consent before the study and received compensation of USD 200 or an equivalent amount in another currency. The same participants, interface, blinding procedure, and geometry-judgment criterion are used for the metric–human alignment analysis in Appendix C.4.

For each of the three backbone pairs, i.e., Wan versus Wan + GeoRoute, HunyuanVideo versus HunyuanVideo + GeoRoute, and CogVideoX versus CogVideoX + GeoRoute, each participant evaluated 360 video pairs randomly sampled from the outputs generated using the VBench prompt set. Each pair contained one video generated by the original baseline and one generated by its GeoRoute-enhanced counterpart using the same text prompt. The method identities were hidden from the participants. The presentation order of the video pairs was randomized, and the left–right placement of the two videos was independently randomized for each comparison.

Participants independently assessed each pair along two criteria: *overall visual quality* and *geometric consistency*. For visual quality, they selected the video with better overall perceptual quality. For



a) All-step Injection



b) Window Injection

Figure 16: **Visualization of persistent geometry injection.** Each row contains four uniformly sampled frames. For each prompt, persistent injection at every denoising step is followed by window-restricted persistent injection at the 12 steps within $\mathcal{W}^{(m)}$.

geometric consistency, they selected the video that better preserved object shape, structure, and topology over time. Each criterion provided four response options—left video, tie, right video, or *Cannot judge*. Ties were treated as valid responses, and *Cannot judge* was reserved for ambiguous pairs.

The main method-preference block comprised $8 \times 3 \times 360 = 8{,}640$ video-pair trials, with two criterion-specific judgments requested in each trial. After removing 373 trials with *Cannot judge* or missing criterion responses, 8,267 complete and judgeable trials were retained, corresponding to a valid-trial rate of 95.68% and 16,534 valid criterion-specific judgments. Tie responses were retained as valid judgments. The geometry metric-alignment block is analyzed separately in Appendix C.4.

For each backbone and criterion, valid judgments were categorized as preferring the baseline, indicating a tie, or preferring GeoRoute. We first compute these three proportions separately for each participant using that participant's valid trials, then report their arithmetic means across the eight participants. Thus, participants receive equal weight, and the reported percentages are not pooled vote proportions. As reported in Table 2, GeoRoute receives visual-quality preference rates of 72.36%, 62.48%, and 84.83%, with tie rates of 3.78%, 1.96%, and 0.12%, for Wan, HunyuanVideo, and CogVideoX, respectively. For geometric consistency, GeoRoute receives preference rates of 71.25%, 66.43%, and 75.39%, with corresponding tie rates of 10.23%, 6.75%, and 2.51%.

Additionally, Figure 18 shows the informed-consent form used for the study.

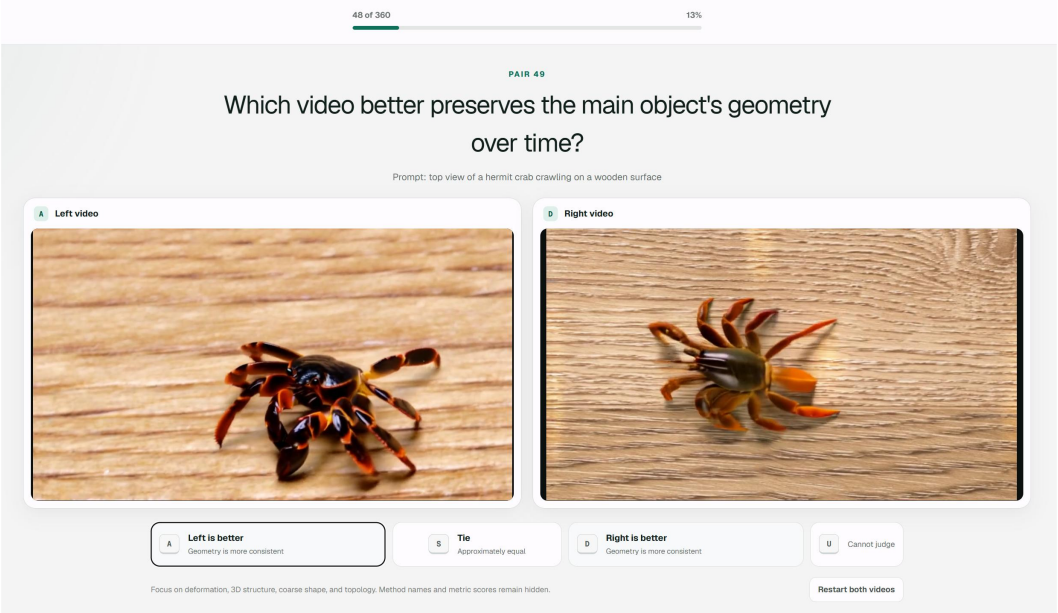


Figure 17: A sample interface of the user-study website. Method identities are hidden, and the left–right placement of the baseline and GeoRoute outputs is randomized.

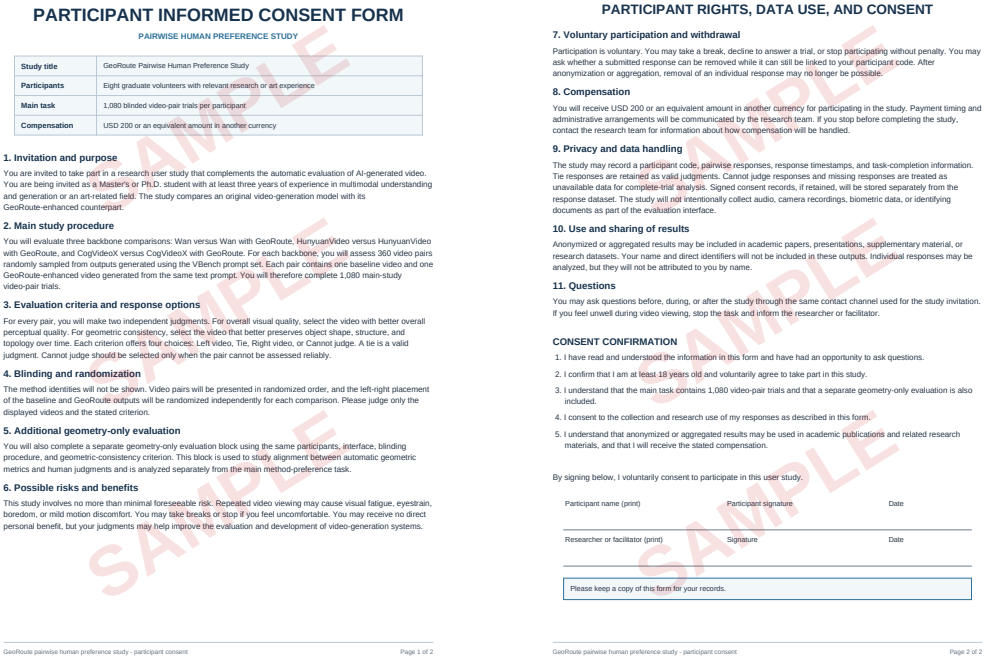


Figure 18: Sample informed-consent form for the pairwise user study.

G Additional Qualitative Results

More Visual Results for GeoVideoBench. To qualitatively examine GeoVideoBench failure retrieval, we visualize videos ranked in the bottom 3% by GeoVideoBench Overall (RC) in Figure 19. These examples cover inter-object fusion, spontaneous object and part appearance, implausible rota-

(a) A girl pushes a chair, and her younger sister sits on it.



(b) One person took a cucumber and cut it.



(c) A worker is sawing wood



(d) A cook is plating



(e) Slicing loofah



Figure 19: **Videos ranked in the bottom** 3% **by GeoVideoBench Overall (RC).** From top to bottom: (a) the girl merges with the exercise equipment, and the background girl temporarily merges with the foreground girl; (b) a cucumber appears spontaneously and rotates implausibly in the person's hand; (c) numerous branches suddenly appear while the worker saws the log and subsequently disappear; (d) the plate splits into two objects; and (e) a single knife cut produces many implausible vegetable fragments.

tion, object splitting, and severe fragmentation, demonstrating that GeoVideoBench surfaces diverse temporal geometric and topological failures from generated videos.

More Qualitative Results for GeoRoute. To complement the main qualitative comparison, we provide four additional CogVideoX results in Figure 20. The examples span object manipulation, articulated object structure, human–vehicle interaction, mixing, and liquid pouring. Across these diverse motions and object relations, GeoRoute produces more temporally stable geometry than the original CogVideoX backbone. The prompts used to generate the examples in Figure 20 are listed below:

(a) In the video, a woman and a child are seen preparing food together on a kitchen counter. The woman is cutting green peppers with a knife, while the child watches and learns. The counter is cluttered with various kitchen items, including a red bowl, a bottle of olive oil, and a cutting board. The scene is set in a home kitchen, with a white countertop and a wooden table. The woman and child are focused on their task, creating a warm and inviting atmosphere. The video captures a moment of bonding and learning between the two, as they prepare a meal together.

(b) The video captures a man in a red racing suit walking towards a red racing car on a track. The car is parked on the side of the track, and the man is approaching it with purpose. The track itself is wide and well-maintained, with a clear sky overhead. The man's racing suit is a vibrant red, matching the color of the car, and he is carrying a helmet under his arm. The overall style of the video is dynamic and action-packed, capturing the anticipation and excitement of a race about to begin.

(c) In the video, a person is seen pouring a liquid from a black kettle into a white coffee cup. The kettle has a long spout and a black handle. The coffee cup is placed on a wooden table. The person is wearing a white shirt and a black tie. The background of the video is blurred, but it appears to be a kitchen setting. The liquid being poured is likely coffee, given the context of the scene. The style of the video is simple and straightforward, focusing on the action of pouring the coffee.

(d) In the video, a chef is seen preparing a dish in a modern kitchen. The chef, dressed in a white chef's coat, is using a silver mixer to blend ingredients in a clear bowl. The kitchen is well-equipped with a white countertop and a stainless steel oven. A potted plant adds a touch of greenery to the space. The chef's focused expression and the precise movements suggest that he is in the process of creating a culinary masterpiece. The overall style of the video is clean and professional, highlighting the chef's skill and the kitchen's modern design.

Additional qualitative results and complete video sequences are available on the supplementary website.

H SUPPLEMENTARY MATERIALS

We provide source code and a companion website to facilitate reproducibility and further illustrate the motivation, design, and empirical behavior of GeoRoute.

Code. The code includes implementations of all GeoRoute modules and a complete workflow for integrating GeoRoute into Wan, covering environment setup, router preparation, training, inference, and evaluation. These resources detail how the proposed components are implemented and incorporated into an existing video diffusion model.

Website. The website provides additional qualitative results and side-by-side video comparisons that reveal temporal behavior beyond the static figures in the paper. It includes video examples of model collapse observed in the ablation experiments, as well as representative geometric failure modes characterized by GeoVideoBench. These examples illustrate the challenges motivating GeoRoute and provide complementary visual evidence for its design choices and its effects on geometric consistency.

(a) In the video, a woman and a child are seen preparing food together on a kitchen counter.



(b) The video captures a man in a red racing suit walking towards a red racing car on a track.



(c) In the video, a person is seen pouring a liquid from a black kettle into a white coffee cup.



(d) In the video, a chef is seen preparing a dish in a modern kitchen.



Figure 20: **Additional qualitative comparisons on CogVideoX.** For each prompt, the top row shows CogVideoX and the bottom row shows CogVideoX + GeoRoute; columns progress from earlier to later video frames. In (a), CogVideoX generates spurious extra fingers during vegetable cutting, whereas GeoRoute preserves a normal hand structure. In (b), multiple left-front wheels appear on the red sports car, while GeoRoute maintains a coherent car body with the correct wheel topology. In (c), the CogVideoX kettle loses its spout and liquid emerges directly from its body; GeoRoute retains a complete spout and a physically plausible pouring stream. In (d), the mixer generated by CogVideoX becomes severely distorted during stirring, whereas GeoRoute preserves its rigid structure throughout the interaction.