GEOFIX: GEOMETRY FIXING FOR 4D CONSISTENT GENERATION

Anonymous authors
Paper under double-blind review

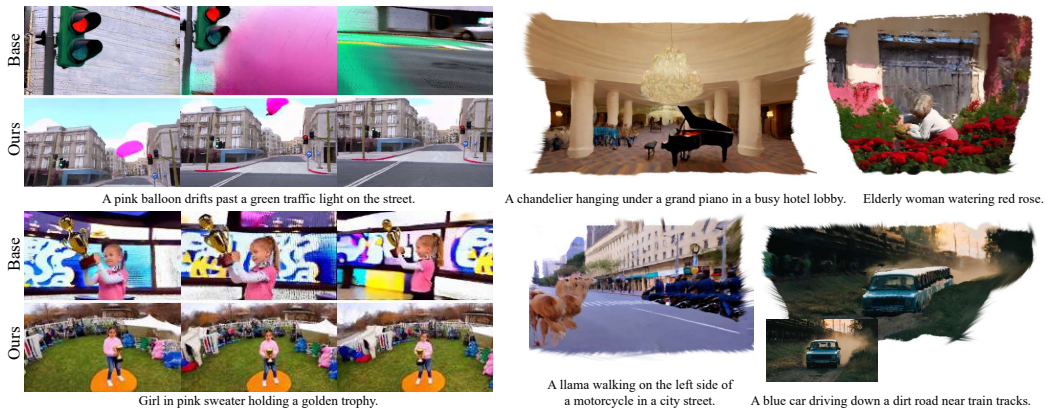

Figure 1: **Generated 4DGS and rendered videos from GeoFix**. Our geometric consistency reward yields more coherent 4D representations through both training and inference-time alignment.

ABSTRACT

Video diffusion models increasingly serve as priors for 4D generation, but the generated videos often lack geometric consistency. Recent approaches address this through post-training by designing geometric rewards, yet this can degrade texture and fine detail, requiring auxiliary appearance rewards and careful reward balancing. To address this challenge, we propose GeoFix, a unified framework that improves geometric consistency and visual quality with a single geometric reward, applicable to both training and inference-time reward alignment. Our reward measures cross-view depth consistency in static scene regions, targeting structural inconsistencies without directly constraining appearance or penalizing valid object motion. Crucially, depth agreement increases with camera field-of-view overlap even in real videos, allowing a smaller camera movement to be mistaken for better geometry. We account for this effect by calibrating the reward against the overlap-dependent trend. GeoFix achieves a 39% improvement in WorldScore 3D consistency and a 7% improvement in general video quality measured by VBench.

1 INTRODUCTION

Video diffusion models (VDMs) (Wan et al., 2025; Yang et al., 2024) have achieved impressive visual fidelity by learning from large-scale video data. However, their outputs can contain geometric inconsistencies, such as background structures deforming or drifting as the camera moves (Duan et al., 2025). These artifacts limit their use as priors for 4D generation (Wu et al., 2024; Wang et al., 2025; Yang et al., 2026), which requires scene structure and dynamics to remain coherent across time and viewpoints. Improving this geometric coherence is essential for turning visually compelling videos into reliable dynamic scene representations.

Recent work has therefore explored reward-based post-training to improve the geometric consistency of pretrained video models. One line of work provides indirect geometric feedback by assess-

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.

ing the visual quality or plausibility of rendered views using learned preference models or VLMs (Go et al., 2026; Wang et al., 2026). Such feedback is informative, but it conflates geometric inconsistency with appearance and rendering fidelity, which limits its ability to target geometric errors.

Other rewards evaluate consistency more directly through RGB reconstruction errors (Go et al., 2026; Du et al., 2026), epipolar constraints (Kupyn et al., 2026), depth agreement (An et al., 2026), or agreement between geometric reprojections and tracked correspondences (Yin et al., 2026). However, optimizing these rewards can favor outputs that are easier to match or reconstruct, improving consistency at the expense of visual detail (Go et al., 2026; Kupyn et al., 2026). To mitigate these effects, existing methods introduce additional rewards that penalize nearly static videos (Kupyn et al., 2026; Yin et al., 2026) or reward visual quality (Go et al., 2026), or use explicit preference-data filtering (Du et al., 2026). This reliance on auxiliary objectives or preference-data filtering complicates alignment design, requiring additional choices about reward weights or sample-selection thresholds.

Therefore, current rewards that evaluate 3D consistency fall short of the broader goal of geometric alignment: correcting structural inconsistencies without harming the visual richness of pretrained models. Since geometric inconsistencies are themselves visual defects, we argue that correcting them should also improve video quality. This motivates our central question: *can a single geometric reward improve both geometric consistency and video quality without auxiliary rewards?*

In this paper, we introduce **GeoFix**, a framework that improves both geometric consistency and video quality using a single geometric reward. GeoFix combines three components: our **FO**V-**CAL**ibrated geometric reward (**FOCAL**), a stitched 4D reconstructor, and alignment methods.

At the core of GeoFix, FOCAL measures cross-view depth consistency to target scene structure without directly requiring colors or textures to match. We restrict the comparison to static scene regions to avoid penalizing valid object motion as geometric inconsistency. A key observation behind FOCAL is that depth agreement is not determined by geometry alone. Even clean real videos score higher when the camera field-of-view (FOV) overlap is larger. FOCAL therefore calibrates depth agreement against the score expected for a given overlap, using the relationship estimated from clean real videos and kept fixed during optimization. This reduces the extent to which greater overlap alone is rewarded as better geometry.

To compute FOCAL directly from generated video latents, we adapt model stitching (Go et al., 2026) to a pretrained 4D reconstructor (Yang et al., 2026). This bypasses RGB decoding and the reconstructor's early layers, and is more robust to the latent noise introduced during generation (Go et al., 2026). The resulting model supplies depth and camera estimates for reward computation while producing an explicit 4D Gaussian representation from the latent.

FOCAL can be integrated into existing methods for both post-training and inference-time alignment. We demonstrate this within GeoFix using DiffusionNFT (Zheng et al., 2025) for post-training, and best-of-N and DPS-style gradient-based guidance (Chung et al., 2023) at inference time. Experiments on WorldScore (Duan et al., 2025) and VBench (Huang et al., 2023) show that GeoFix outperforms baselines in geometric consistency and video quality, while also improving camera accuracy. We further demonstrate its effectiveness in 4D Gaussian scene generation, producing more stable background geometry and coherent object motion.

2 Related Work

Diffusion Model Alignment. Reward alignment adapts pretrained models using rewards or preferences through post-training or inference-time steering. Post-training methods include differentiable reward optimization (Clark et al., 2023), reward-weighted regression (Lee et al., 2023), policy-gradient methods (Black et al., 2024; Fan et al., 2023; Liu et al., 2025), and direct preference optimization (Wallace et al., 2023). Inference-time approaches include best-of-N selection (Ma et al., 2025), particle resampling (Singhal et al., 2025), gradient-free guidance (Kim et al., 2025), and sampling-noise optimization (Tang et al., 2024). Our work focuses on geometric reward design, whose effectiveness we evaluate through DiffusionNFT post-training (Zheng et al., 2025), best-of-N selection and DPS-style gradient guidance (Chung et al., 2023).

Geometric Rewards for Video and 4D Generation. Building on diffusion alignment methods, recent works develop rewards to improve geometric consistency in video and 4D generation. These

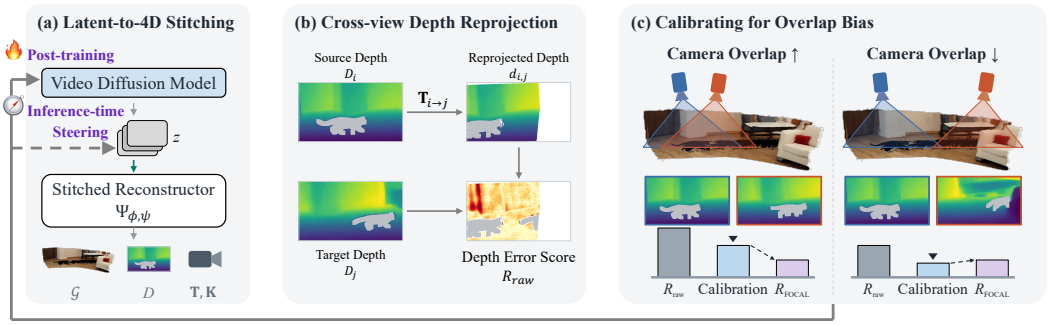

Figure 2: **Overview of GeoFix.** (a) FOCAL is computed directly from video latents using a stitched 4D reconstructor. (b) We measure cross-view depth reprojection error in static regions and (c) calibrate it to remove FOV bias.

approaches draw on two broad types of feedback: indirect quality assessments and direct geometric consistency measurements. Indirect feedback assesses the perceived quality or plausibility of generated and rendered views using preference models or VLMs (Go et al., 2026; Wang et al., 2026). Such feedback provides useful perceptual supervision, but its dependence on appearance and rendering fidelity limits its ability to isolate structural errors.

Direct feedback measures reconstruction errors (Go et al., 2026; Du et al., 2026; Park et al., 2025; Wang et al., 2026; Ge et al., 2026), epipolar constraints (Kupyn et al., 2026), geometric correspondence consistency (Yin et al., 2026; Gu et al., 2025), or depth agreement (An et al., 2026). However, directly optimizing consistency does not ensure better visual quality. For example, VIST3A (Go et al., 2026) reports blurred outputs when reconstruction consistency is optimized without a quality reward. Epipolar-DPO (Kupyn et al., 2026) reports oversaturation with a particular correspondence estimator. Existing methods use auxiliary objectives (Go et al., 2026; Wang et al., 2026; An et al., 2026; Kupyn et al., 2026; Yin et al., 2026; Li et al., 2026) or preference-data filtering (Du et al., 2026), complicating alignment design, requiring additional choices about reward weights or sample-selection thresholds.

We address this challenge by designing a single geometric reward that improves both geometric consistency and video quality without auxiliary rewards, eliminating the need to balance multiple reward objectives.

3 Method

Our goal is to improve video quality by correcting geometric inconsistencies without unnecessarily constraining the rich textures and fine visual details of pretrained video models. To this end, we present **GeoFix**, a unified framework that aligns video diffusion models with a geometric reward computed directly on video latents (Figure 2). The proposed reward, **FOCAL** (**FO**V-**CAL**ibrated geometric reward), measures cross-view depth agreement in static scene regions. We compute FOCAL using geometry obtained from video latents through a stitched 4D reconstructor (Section 3.1). We identify an overlap bias in raw depth-consistency scores (Section 3.2). FOCAL removes this bias by calibrating the score against the depth agreement expected at a given overlap (Section 3.3). Finally, GeoFix applies to *any* reward alignment method without modification, which we demonstrate across training and inference-time alignment (Section 3.4).

3.1 Latent-to-4D Model Stitching

Our geometric reward requires reliable depth and camera estimates, which must be computed repeatedly during video alignment. We obtain these estimates directly from video latents by adapting model stitching (Go et al., 2026; An et al., 2026) to a pretrained 4D reconstructor (Yang et al., 2026). This bypasses the VAE decoder and the initial reconstruction layers, avoiding their computational cost. The choice is further motivated by VIST3A's finding that stitched 3D reconstruction is more robust to latent perturbations than sequential RGB decoding and reconstruction (Go et al.,

2026). Beyond providing geometric feedback, the stitched model produces an explicit 4D Gaussian representation from the same video latent.

Let $\mathcal{F}$ denote the pretrained 4D reconstructor with L layers. A lightweight stitching layer $\mathcal{S}_\phi$ maps a video latent z into the feature space at the output of layer k. We initialize a trainable reconstruction module $\mathcal{R}_\psi$ from the remaining layers $\mathcal{F}_{k+1:L}$. The stitched reconstructor is then defined as

$$\Psi_{\phi,\psi}(z) = \mathcal{R}_\psi\big(\mathcal{S}_\phi(z)\big) = \Big(\mathcal{G}, \{D_f, \mathbf{T}_f, \mathbf{K}_f\}_{f=1}^{F}\Big), \tag{1}$$

where $\mathcal{G}$ denotes the reconstructed 4D Gaussian representation, F is the number of frames, and D_f, $\mathbf{T}_f$, and $\mathbf{K}_f$ denote the depth map, camera-to-world pose, and camera intrinsics of frame f, respectively.

We train ϕ and ψ jointly by output distillation. For a video x, the frozen teacher $\mathcal{F}$ provides the target $\mathcal{F}(x)$. The stitched reconstructor predicts from the latent $\mathcal{E}(x)$, where $\mathcal{E}$ is the frozen VAE encoder. We minimize

$$\mathcal{L}_{\text{stitch}}(\phi,\psi) = \mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{q\in\mathcal{Q}} \lambda_q \left\|\Psi^q_{\phi,\psi}(\mathcal{E}(x)) - \mathcal{F}^q(x)\right\|_1\right], \tag{2}$$

where $\mathcal{D}$ contains multi-view image sets and monocular videos of dynamic scenes. Ψ^q and $\mathcal{F}^q$ denote the q-th output tensor (depth, poses, intrinsics, or Gaussian attributes) and λ_q weights its distillation loss. The resulting stitched reconstructor achieves higher reconstruction fidelity than the original RGB-input reconstructor (see Table 6).

3.2 Cross-View Depth Consistency Reward

We base our geometric reward on cross-view depth consistency to constrain scene structure without directly restricting the model's freedom to generate rich textures and fine visual details. View-dependent effects such as specular highlights make RGB an unreliable signal of geometric consistency (Asim et al., 2025), and optimizing it can favor blurred outputs that are easier to match across views (Go et al., 2026). Instead, depth reprojection evaluates cross-view agreement of the 3D structure, without requiring colors or textures to match or relying on point-tracking correspondences.

For each frame i, we unproject the depth $D_i(\mathbf{p})$ with $\mathbf{K}_i$ to obtain the center of the pixel-aligned Gaussians $\mathbf{X}_i(\mathbf{p})$ in camera i's coordinates. For a frame pair (i,j), we transform these points into the target camera using the relative pose $\mathbf{T}_{i\to j} = \mathbf{T}_j^{-1}\mathbf{T}_i$ and project them as

$$\big(\mathbf{y}_{i,j}(\mathbf{p}), d_{i,j}(\mathbf{p})\big) = \Pi\big(\mathbf{T}_{i\to j}\cdot\mathbf{X}_i(\mathbf{p}), \mathbf{K}_j\big), \tag{3}$$

where Π is the forward projection function, $\mathbf{y}_{i,j}(\mathbf{p})$ is the projected pixel location in frame j and $d_{i,j}(\mathbf{p})$ is the target-camera depth. Depth agreement is measured by comparing $d_{i,j}(\mathbf{p})$ with the target depth map D_j, bilinearly sampled at $\mathbf{y}_{i,j}(\mathbf{p})$.

Reprojection using camera transformations alone does not account for independent object motion. To avoid treating valid motion as geometric inconsistency, we introduce a dynamic-region mask and restrict the comparison to static content in both views. Specifically, we apply a pretrained motion segmentation network (Huang et al., 2025) to videos rendered from $\mathcal{G}$ to obtain per-frame dynamic masks. We aggregate the remaining depth discrepancies as

$$\text{MSE} = \frac{1}{|\mathcal{C}|}\sum_{(i,j)\in\mathcal{C}}\frac{1}{|\mathcal{V}_{i,j}|}\sum_{\mathbf{p}\in\mathcal{V}_{i,j}}\big(D_j\big(\mathbf{y}_{i,j}(\mathbf{p})\big) - d_{i,j}(\mathbf{p})\big)^2, \tag{4}$$

where $\mathcal{C}$ contains frame pairs with stride s. The set $\mathcal{V}_{i,j}$ contains source pixels with valid depths whose projections lie within the target image, have positive target-camera depth, and fall in static regions in both views. To obtain a reward that increases with depth agreement, we convert this error into a PSNR score:

$$R_{\text{raw}} = 10\log_{10}\left(\frac{D_{\max}^2}{\text{MSE}+\epsilon}\right), \tag{5}$$

where $D_{\max}$ is the maximum depth in the reconstructed clip and $\epsilon > 0$ ensures numerical stability.

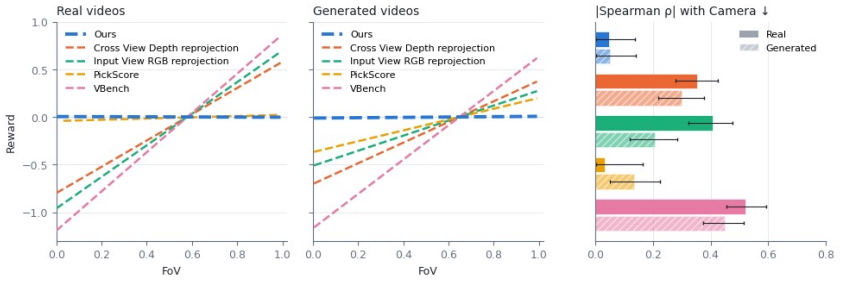

Figure 3: **Camera Correlation Analysis.** Existing direct and indirect geometric rewards vary with camera movement, creating a bias toward smaller viewpoint changes. Our FOV-calibrated reward remains invariant, providing a more reliable measure of geometric consistency.

Observation: Overlap Bias. Although R_{raw} evaluates structural agreement, a higher score does not necessarily indicate better geometry. We find that even clean real videos exhibit lower depth reprojection errors when camera overlap is larger (Figure 3). The same dependence holds for existing geometric rewards (Section 4.5). We refer to this effect as overlap bias. These scores cannot disentangle geometric quality from camera overlap, so maximizing them favors larger camera overlap without improvement in scene geometry.

To quantify camera overlap, we use the area-weighted IoU of the two cameras' fields of view (Lee et al., 2026). We sample a reference sphere centered at the first camera on a uniform azimuth–elevation grid, and let $\mathcal{A}_f$ denote the sampled points within camera f's field of view. The overlap between the first and last cameras is

$$\Omega = \frac{||\mathcal{A}_1 \cap \mathcal{A}_F||}{||\mathcal{A}_1 \cup \mathcal{A}_F||}, \tag{6}$$

where $||\cdot||$ denotes a count weighted by the cosine of elevation to account for spherical area.

3.3 FOCAL: FOV-Calibrated Geometric Reward

Motivated by this observation, we use clean real videos to estimate the raw depth consistency score expected at a given FOV overlap. We then evaluate each generated video relative to this overlap-dependent reference. We estimate the reference by fitting the relationship between the raw depth score and FOV overlap:

$$R_{\text{raw}} = \alpha + \beta \log_{10} \max(\Omega, \epsilon_{\text{fov}}) + \xi, \tag{7}$$

where α is the intercept, β is the slope, and ξ is the residual. The small constant $\epsilon_{\text{fov}} > 0$ ensures numerical stability at zero overlap. To compare a generated video's score with this reference, we subtract the fitted overlap-dependent term. The fitted intercept is constant and does not affect reward maximization, so we omit it from the correction:

$$R_{\text{FOCAL}} = R_{\text{raw}} - \hat{\beta} \log_{10} \max(\Omega, \epsilon_{\text{fov}}), \tag{8}$$

where $\hat{\beta}$ is the slope estimated from real videos. A high reward therefore indicates strong depth agreement relative to the reference for the video's camera overlap. Equivalently, defining the calibration coefficient $\gamma = \hat{\beta}/10$, we obtain FOCAL (**FOV-CAL**ibrated geometric reward):

$$R_{\text{FOCAL}} = 10 \log_{10} \left(\frac{D_{\max}^2}{(\text{MSE} + \epsilon) \max(\Omega, \epsilon_{\text{fov}})^{\gamma}} \right). \tag{9}$$

We estimate the calibration coefficient γ only from clean real videos and keep it fixed throughout alignment. On held-out real videos and generated videos, the corrected score–overlap trend has a near-zero regression slope, showing that the correction also reduces overlap bias beyond the calibration set. FOCAL thus accounts for viewing conditions within the geometric reward itself, without introducing auxiliary rewards.

3.4 GeoFix: Reward-Based Video Alignment

We refer to the resulting framework as GeoFix, a unified reward alignment scheme that uses FOCAL to fine-tune the VDM and to select or guide samples at inference time. For a video latent z,

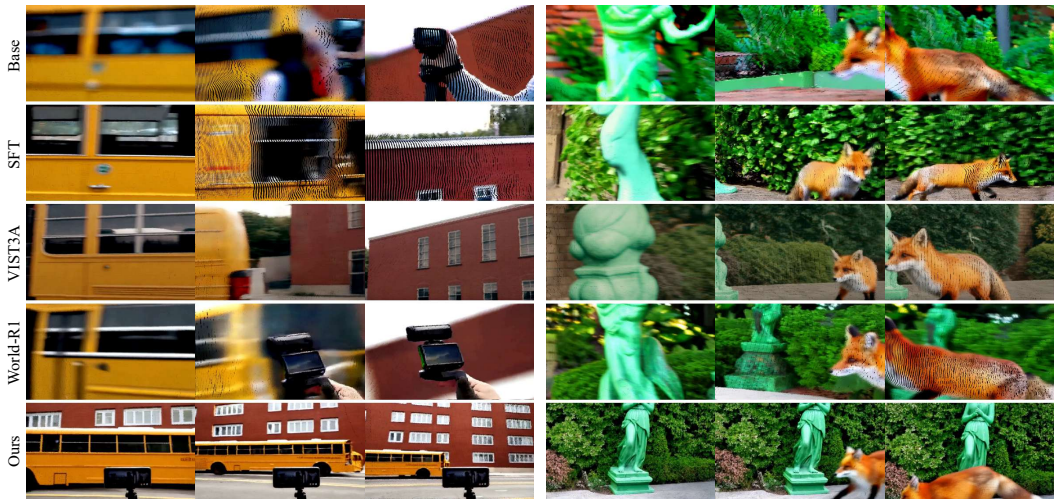

Figure 4: Qualitative Comparison of 4DGS-Rendered Text-to-Video Generation.

we write $R_{\mathrm{FOCAL}}(z)$ for the reward of Eq. 9 computed from the geometry $\Psi_{\phi,\psi}(z)$ predicted by the stitched reconstructor. $R_{\mathrm{FOCAL}}(z)$ operates in the latent space, requires no VAE decoding and is differentiable in z. GeoFix is therefore agnostic to the choice of alignment method, including methods that require reward gradients. To demonstrate this, we instantiate GeoFix with representative methods: reward fine-tuning (DiffusionNFT (Zheng et al., 2025)), zeroth-order sample selection (Best-of-N (Beirami et al., 2025)), and gradient guidance (DPS (Chung et al., 2023)). Ψ and γ stay fixed and no auxiliary reward is used in any setting.

Training-time Alignment with DiffusionNFT. We fine-tune the VDM using DiffusionNFT. We generate latents, score them with $R_{\mathrm{FOCAL}}(z)$, and use these scores as feedback for its forward-process flow-matching objective.

Inference-time Alignment. We keep the video model fixed and use the reward in two ways.

1) Best-of-N. We generate N complete video samples for each prompt, score their latents with $R_{\mathrm{FOCAL}}(z)$, and return the sample with the highest reward.

2) DPS-style Guidance. We use the reward gradient to guide sampling following the gradient-based approach of DPS. At each guided step, we evaluate the reward on the clean latent prediction $\hat{z}_{0|t}$ from z_t and adjust the next latent in a direction that locally increases the reward:

$$z_{t-1} = \tilde{z}_{t-1} + \eta_t \nabla_{z_t} R_{\mathrm{FOCAL}}\big(\hat{z}_{0|t}\big), \tag{10}$$

where $\tilde{z}_{t-1}$ is the next latent from the unguided sampling and $\eta_t \geq 0$ controls the guidance strength.

4 Experiments

4.1 Experimental Setup

For training-time alignment, we use Wan2.2-TI2V-5B (Wan et al., 2025) and jointly optimize a supervised loss on MultiCamVideo (Bai et al., 2025) and UltraVideo (Xue et al., 2025), and a DiffusionNFT (Zheng et al., 2025) loss using 64 rollouts per step generated from VideoUFO (Wang & Yang, 2025) prompts and scored by R_{FOCAL}. For inference-time alignment, we use Wan2.2-Fun-5B-Control-Camera. Since both models share the same VAE, we use a fixed stitched reconstructor fine-tuned from NeoVerse (Yang et al., 2026). Implementation details are provided in Appendix A.

Table 1: Quantitative comparison of training-time reward alignment methods.

	R_{FOCAL}	VBench	WorldScore							
			Scene Consistency		Dynamics			Visual Quality		Text Alignment
Method	Avg	Avg	3D Cons.	Photo. Cons.	Motion Acc.	Motion Mag.	Motion Smooth.	Style Cons.	Subj. Qual.	Content Align.
Base	29.99	0.6390	44.39	33.55	50.30	45.17	18.08	84.70	6.80	89.36
SFT	31.02	0.6422	50.11	38.47	52.76	**48.93**	19.90	89.00	9.80	87.17
VIST3A	31.06	0.6550	58.32	47.00	52.84	41.94	25.62	**96.63**	15.16	**95.47**
World-R1	30.49	0.6438	42.26	31.64	51.35	45.90	19.72	88.07	9.57	91.03
Ours	**33.01**	**0.6861**	**61.60**	**51.06**	**55.94**	35.44	**26.08**	93.30	**15.94**	91.40

Table 2: Quantitative comparison of training-time reward alignment against VGGRPO. Parentheses show relative improvements over the base model.

	VBench				
Method	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness
Base	0.9542	0.9528	0.5966	0.6733	0.9841
VGGRPO	0.9644 (+1.07%)	0.9583 (+0.58%)	0.5991 (+0.42%)	0.6861 (+1.90%)	0.9895 (+0.55%)
Base	0.8741	0.9389	0.4679	0.5523	0.9589
Ours	**0.9317 (+6.59%)**	**0.9601 (+2.25%)**	**0.5272 (+12.67%)**	**0.5979 (+8.25%)**	**0.9802 (+2.22%)**

4.2 VDM Training-time Alignment

Baselines, Evaluation Setting. We compare our method against all baselines on Wan2.2-TI2V-5B: base model, supervised fine-tuning (SFT), VIST3A (Go et al., 2026) and World-R1 (Wang et al., 2026). All fine-tuned models use the same SFT dataset, training budget and leverage the same stitched reconstructor. We generate a 4DGS scene for 200 T2V-CompBench (Sun et al., 2025) prompts and evaluate on the rendered videos. We report on VBench (Huang et al., 2023) for general video quality and WorldScore (Duan et al., 2025) for world generation benchmark. We also compare against VGGRPO (An et al., 2026), which leverages a direct geometric reward. Since VGGRPO has released neither code nor checkpoints, we compare with the results reported in its paper. Following their protocol, we score the VDM decoded outputs for the standard VBench prompts. We report the relative **improvement** each method achieves over its own base model's results.

Results. GeoFix achieves the strongest geometric consistency among all baselines while also improving general video quality, without an auxiliary reward. As shown in Table 1, it obtains the highest 3D consistency and photometric consistency on WorldScore. SFT alone yields only moderate gains, indicating that most of the improvement comes from the reward rather than from the supervised data. GeoFix also attains the best motion accuracy and motion smoothness, and remains superior or competitive on visual quality and text alignment. It further scores highest on VBench, supporting our claim that correcting geometric inconsistencies, which are themselves visual defects, also improves video quality. Furthermore, our method improves every VBench dimension by a larger relative margin than VGGRPO (Table 2), with the largest gaps in visual quality. Note that none of these video quality metrics is used as a reward during training.

Qualitatively, our method produces the cleanest 4D scene renderings, with fewer artifacts and substantially reduced background drift and flickering compared with the baselines. Figure 4 compares the videos rendered from the generated 4DGS scenes. Evaluating these renderings exposes geometric inconsistencies that can manifest as reconstruction and rendering artifacts. For the baselines, the reconstructor receives geometrically inconsistent videos, resulting in scenes with rendering artifacts.

4.3 Inference-time Alignment

Baselines, Evaluation Setting. We plug our reward into two inference-time steering methods. We use 50 NFEs, N=20 for Best-of-N and guiding step 3 to 17 for DPS. We pair each of the 355 VBench-I2V (Huang et al., 2024) images with one of ten target camera trajectories. We measure camera accuracy between the camera trajectory recovered by the reconstructor and the target trajectory. We evaluate WorldScore for world generation quality on the rendered videos.

Results. Our proposed reward improves both camera accuracy and 4D scene quality for both first and zeroth order inference-time steering methods. As shown in Table 3, the proposed method improves the reconstructed scene's camera predictions and 4D scene quality in both first and zeroth order guidance methods. This indicates that our reward even provides a useful gradient signal through the stitched model. As in training-time alignment results, GeoFix not only improves consistency metrics but also visual quality and text alignment metrics. These improvements are accompanied

Table 3: Quantitative comparison of inference-time reward alignment.

	Camera Accuracy			WorldScore							
				Scene Consistency		Dynamics			Visual Quality		Text Alignment
Method	Rot. Err. ↓	Trans. Err. ↓	ATE Norm. ↓	3D Cons.	Photo. Cons.	Motion Acc.	Motion Mag.	Motion Smooth.	Style Cons.	Sub. Qual.	Content Align.
None	2.04	34.97	0.2563	75.43	76.06	**44.81**	**31.93**	23.65	95.47	20.87	94.10
Best-of-N	**1.751**	**34.57**	**0.2507**	**78.00**	**78.33**	44.59	31.15	**24.61**	**96.08**	**21.86**	**94.66**
None	2.04	34.97	0.2563	75.43	76.06	**44.81**	**31.93**	23.65	95.47	20.87	94.10
DPS	**1.97**	**34.12**	**0.2500**	**75.82**	**76.96**	44.42	31.54	**24.02**	**95.80**	**20.95**	**94.45**



Figure 5: Qualitative comparison on Image-to-4D generation. We overlay 3D representations from multiple timesteps in a single view to visualize dynamic motion over time.

by only a marginal reduction in motion accuracy and motion magnitude, whereas every camera accuracy and every other WorldScore aspect improve at the same time.

4.4 4D Scene Generation

Baselines, Evaluation Setting. We compare image-to-4D generation with MoGe4D (Zhang et al., 2026) and 4DNeX (Chen et al., 2025) on 100 images from VBench-I2V (Huang et al., 2023). We render every generated 4D scene along an identical novel-view camera trajectory from ReCamMaster (Bai et al., 2025). Following the baseline protocol, we fill disoccluded regions using MoGe4D's 4D-ViSM identically across all methods. We then score the resulting novel-view videos with six complementary metrics. From VBench, we use subject consistency and temporal flickering for consistency and temporal stability of the rendered scene, dynamic degree for the amount of motion, aesthetic quality for visual quality. We use CLIP score (Radford et al., 2021) for text alignment, and physical commonsense from VideoPhy-2 (Bansal et al., 2025) for physical plausibility.

Results. GeoFix performs best overall and achieves the highest consistency scores (Figure 6), indicating stable geometry under novel camera trajectories. 4DNeX attains the highest dynamic degree, but this largely arises from frame-to-frame structural inconsistency rather than valid object motion (Figure 5), consistent with its poor consistency scores. In contrast, MoGe4D exhibits the lowest dynamic degree, indicating insufficient motion for 4D synthesis. Ours better preserves scene structure while maintaining meaningful dynamics.

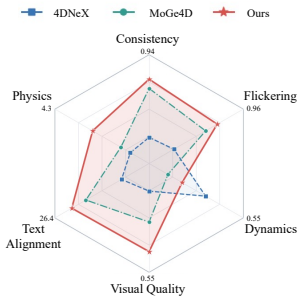

Figure 6: Quantitative comparison on 4D Scene Generation. Higher is better; axes use different scales.

Figure 5 shows that GeoFix produces coherent 4D scenes over time. We overlay the 3D representation of multiple timesteps, so a geometrically consistent scene appears as a sharp static background with the moving object tracing its trajectory through it. MoGe4D generates almost no motion. In 4DNeX, background drift produces duplicated and smeared structure, while discontinuous object motion contributes to its high dynamic degree. Our overlays keep the background aligned across all timesteps while the dynamic content moves through it.

4.5 Ablation Studies

Reward corruption test. Our proposed reward decreases monotonically to geometric corruptions, so it can serve as a reliable signal for geometric consistency. Following SGC (Dou et al., 2026), we reconstruct real videos, perturb the reconstruction, and evaluate our reward on it. We use four perturbations of the kind a valid camera motion cannot explain.

Warp shifts the depth by a sinusoidal displacement field, deforming the scene non-rigidly. Squeeze stretches the depth map vertically while leaving the intrinsics unchanged. Depth-layer warp shifts the depth horizontally, which inverts parallax. Pose drift rotates the camera poses while leaving the depth untouched. The static-region mask is transformed together with the depth, so that changes in mask quality do not affect the measurement. As shown in Figure 7, the reward decreases as all four corruptions increase, confirming that our proposed reward measures geometric inconsistency rather than image fidelity. Lines show mean rewards, shaded regions indicate 95% bootstrap confidence intervals of the mean.

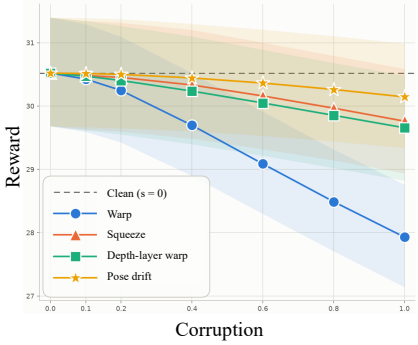

Figure 7: Analysis of reward sensitivity to geometric corruption.

Geometric reward vs. video-quality reward. VBench rewards capture general video quality but are not designed to assess geometric consistency, and directly optimizing them can lead to degenerate solutions. Following DDRL (Ye et al., 2025), we trained the VDM with direct reward fine-tuning on five VBench (Huang et al., 2023) rewards. The VBench-trained model produces saturated videos (Figure 8), and training longer adds blur and grid artifacts. The failure follows from what these rewards measure. VBench metrics reward high DINO or CLIP feature similarity between frames, collapsing inter-frame variation at the cost of blur and grid artifacts. Our geometric reward, in contrast, improves consistency and even video quality, which supports our claim that correcting geometric defects is an effective route to better video quality. Additional details are in the Appendix D.3.



Figure 8: VBench reward comparison.

Effect of FOV calibration. FOV calibration makes our reward invariant to camera motion, whereas other direct and indirect rewards correlate with it. Existing measures of geometric consistency are easier to satisfy when camera moves less, which can be increased without improving geometry at all. As in Figure 3, direct rewards such as cross-view depth reprojection and input-view RGB reprojection (Go et al., 2026; Du et al., 2026), as well as indirect rewards such as PickScore, all vary systematically with camera motion.

FOV calibration prevents reward optimization from suppressing camera movement during training. Removing the calibration (γ=0) loses dynamic motion, since the model reduces camera motion to collect the easier reward (Table 4). Our setting even matches the dynamic degree obtained with a larger (γ=1). Our calibrated setting achieves the highest overall video quality, indicating the best balance between geometric consistency and dynamic motion.

Table 4: Effect of FOV calibration.

γ	Overall	Dynamic
0	0.6802	0.395
0.7 (Ours)	**0.6861**	**0.425**
1	0.6823	**0.425**

5 Conclusion

In this paper, we presented FOCAL, a geometric consistency reward for video diffusion models, and GeoFix, which applies FOCAL to reward alignment methods to improve geometric consistency and visual quality. Our reward evaluates cross-view depth consistency in static scene regions, targeting structural inconsistencies while preserving valid object motion. The core of our method is calibrating depth agreement against the expected value at a given camera overlap, reducing the extent to which greater overlap alone is rewarded as an improvement in geometric consistency. FOCAL is computed directly from video latents through a stitched 4D reconstructor, without VAE decoding. Experiments across training and inference-time reward alignment demonstrate improvements in geometric consistency, video quality and camera accuracy. We further show that GeoFix produces coherent 4D scenes compared to baselines. Overall, GeoFix advances video diffusion models as priors for visually rich and geometrically coherent dynamic scene generation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature search, code implementation, drafting and editing parts of the manuscript. All AI-assisted content was reviewed and verified by the authors, including independently checking relevant literature against the original sources. We take responsibility for the final content of this work, including code, text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on improving the geometric consistency of video generative models and does not involve human-subject studies or the collection of personal or sensitive data. The proposed method inherits potential societal risks associated with generative video models, including the creation of synthetic or misleading visual content. Our method is designed specifically to improve geometric consistency and does not introduce capabilities explicitly intended for harmful or deceptive applications. We encourage the responsible use of the proposed method in accordance with applicable policies and ethical guidelines for generative media.

REPRODUCIBILITY STATEMENT

We provide details necessary for reproducing the proposed method and experiments throughout the main paper and appendix. The methodology and reward formulation are described in the method section, while implementation details, training configurations, hyperparameters, datasets, and evaluation protocols are provided in the experimental sections and appendix.

REFERENCES

Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge Belongie, Mar Gonzalez-Franco, and Marta Tintore Gazulla. Vggrpo: Towards world-consistent video generation with 4d latent reward, 2026. URL https://arxiv.org/abs/2603.26599.

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

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. Recammaster: Camera-controlled generative rendering from a single video, 2025. URL https://arxiv.org/abs/2503.11647.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation, 2025. URL https://arxiv.org/abs/2503.06800.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy, 2025. URL https://arxiv.org/abs/2401.01879.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy, 2025. URL https://arxiv.org/abs/2508.13154.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

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

Weijia Dou, Wenzhao Zheng, Weiliang Chen, Yu Zheng, Jie Zhou, and Jiwen Lu. Measuring 3d spatial geometric consistency in dynamic video generation, 2026. URL https://arxiv.org/abs/2603.19048.

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. URL https://arxiv.org/abs/2601.23286.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation, 2025. URL https://arxiv.org/abs/2504.00983.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Wenhang Ge, Guibao Shen, Jiawei Feng, Luozhou Wang, Hao Lu, Xingye Tian, Xin Tao, and Ying-Cong Chen. Campilot: Improving camera control in video diffusion model with efficient camera reward feedback. *arXiv preprint arXiv:2601.16214*, 2026.

Hyojun Go, Dominik Narnhofer, Goutam Bhat, Prune Truong, Federico Tombari, and Konrad Schindler. Text-to-3d by stitching a multi-view reconstruction network to a video generator, 2026. URL https://arxiv.org/abs/2510.13454.

Leslie Gu, Junhwa Hur, Charles Herrmann, Fangneng Zhan, Todd Zickler, Deqing Sun, and Hanspeter Pfister. Geco: Evaluating geometric consistency for video generation via motion and structure. *arXiv preprint arXiv:2512.22274*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Nan Huang, Wenzhao Zheng, Chenfeng Xu, Kurt Keutzer, Shanghang Zhang, Angjoo Kanazawa, and Qianqian Wang. Segment any motion in videos, 2025. URL https://arxiv.org/abs/2503.22268.

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, 2023. URL https://arxiv.org/abs/2311.17982.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models, 2024. URL https://arxiv.org/abs/2411.13503.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer, 2021. URL https://arxiv.org/abs/2108.05997.

Jaemin Kim, Bryan Sangwoo Kim, and Jong Chul Ye. Free^{2}guide: Training-free text-to-video alignment using image lvlm, 2025. URL https://arxiv.org/abs/2411.17041.

Orest Kupyn, Théo Uscidda, Marta Tintore Gazulla, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models, 2026. URL https://arxiv.org/abs/2510.21615.

Dohun Lee, Chun-Hao Paul Huang, Xuelin Chen, Jong Chul Ye, Duygu Ceylan, and Hyeonho Jeong. Memory-v2v: Memory-augmented video-to-video diffusion for consistent multi-turn editing, 2026. URL https://arxiv.org/abs/2601.16296.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback, 2023. URL https://arxiv.org/abs/2302.12192.

Zizun Li, Haoyu Guo, Runzhe Teng, Chunhua Shen, and Tong He. Geo-align: Video generation alignment via metric geometry reward. *arXiv preprint arXiv:2605.23903*, 2026.

Chenguo Lin, Yuchen Lin, Panwang Pan, Yifan Yu, Tao Hu, Honglei Yan, Katerina Fragkiadaki, and Yadong Mu. Movies: Motion-aware 4d dynamic view synthesis in one second, 2026. URL https://arxiv.org/abs/2507.10065.

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

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl, 2025. URL https://arxiv.org/abs/2505.05470.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2523–2534, June 2025.

Byeongjun Park, Hyojun Go, Hyelin Nam, Byung-Hoon Kim, Hyungjin Chung, and Changick Kim. Steerx: Creating any camera-free 3d and 4d scenes with geometric steering. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27326–27337. IEEE, 2025.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models, 2025. URL https://arxiv.org/abs/2501.06848.

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, 2025. URL https://arxiv.org/abs/2407.14505.

Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Inference-time alignment of diffusion models with direct noise optimization, 2024. URL https://arxiv.org/abs/2405.18881.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023. URL https://arxiv.org/abs/2311.12908.

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, 2025. URL https://arxiv.org/abs/2503.20314.

Chaoyang Wang, Ashkan Mirzaei, Vidit Goel, Willi Menapace, Aliaksandr Siarohin, Avalon Vinella, Michael Vasilkovsky, Ivan Skorokhodov, Vladislav Shakhrai, Sergey Korolev, Sergey Tulyakov, and Peter Wonka. 4real-video-v2: Fused view-time attention and feedforward reconstruction for 4d scene generation, 2025. URL https://arxiv.org/abs/2506.18839.

Weijie Wang, Xiaoxuan He, Youping Gu, Yifan Yang, Zeyu Zhang, Yefei He, Yanbo Ding, Xirui Hu, Donny Y. Chen, Zhiyuan He, Yuqing Yang, and Bohan Zhuang. World-r1: Reinforcing 3d constraints for text-to-video generation, 2026. URL https://arxiv.org/abs/2604.24764.

Wenhao Wang and Yi Yang. Videoufo: A million-scale user-focused dataset for text-to-video generation, 2025. URL https://arxiv.org/abs/2503.01739.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613*, 2024.

Zhen Xu, Zhengqin Li, Zhao Dong, Xiaowei Zhou, Richard Newcombe, and Zhaoyang Lv. 4dgt: Learning a 4d gaussian transformer using real-world monocular videos, 2025. URL https://arxiv.org/abs/2506.08015.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos, 2026. URL https://arxiv.org/abs/2601.00393.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images, 2024. URL https://arxiv.org/abs/2410.24207.

Haotian Ye, Kaiwen Zheng, Jiashu Xu, Puheng Li, Huayu Chen, Jiaqi Han, Sheng Liu, Qinsheng Zhang, Hanzi Mao, Zekun Hao, Prithvijit Chattopadhyay, Dinghao Yang, Liang Feng, Maosheng Liao, Junjie Bai, Ming-Yu Liu, James Zou, and Stefano Ermon. Data-regularized reinforcement learning for diffusion models at scale, 2025. URL https://arxiv.org/abs/2512.04332.

Tengjiao Yin, Jinglei Shi, Heng Guo, and Xi Wang. Vigor: Video geometry-oriented reward for temporal generative alignment, 2026. URL https://arxiv.org/abs/2603.16271.

Yanran Zhang, Ziyi Wang, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Geometry-aware single-image 4d synthesis via dense trajectory generation, 2026. URL https://arxiv.org/abs/2512.05044.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.

A Implementation Details

A.1 Experimental Setup

A.1.1 Tasks and Datasets

All three tasks produce a 4D Gaussian Splatting (4DGS) scene from a video diffusion model (VDM). In each case, a NeoVerse (Yang et al., 2026) 4D reconstructor, stitched to the VDM latent space, decodes the VDM latents into a 4DGS scene.

Text-to-4DGS (T2V-CompBench). We use 200 T2V-CompBench (Sun et al., 2025) prompts and generate one video per prompt. At generation time, each prompt is wrapped in a camera-motion sentence chosen at random from a fixed template set. This gives the 4D reconstructor parallax to recover geometry from. The augmentation is applied identically to every method.

Image-to-4D generation (VBench-I2V). We use the 100 image–caption pairs of the VBench-I2V benchmark (Huang et al., 2024). The input image is used as the first frame, and use camera prompt augmentation. For evaluation, we render every generated 4D scene along the same ReCamMaster novel-view camera trajectory.

Camera-controlled image-to-4D generation. We use the 355 image–caption pairs of the VBench-I2V benchmark (Huang et al., 2024) and assign each image one of ten target camera trajectories (Table 5). The pan trajectories translate the camera at a constant rate in the image plane. The zoom trajectories translate it along the optical axis at twice the pan rate. Every method uses the same image–trajectory assignment.

Table 5: Ten target camera trajectories used for camera-controlled image-to-4DGS evaluation on 355 VBench-I2V images.

Trajectory	Trajectory
Pan Left	Pan Left-Up
Pan Right	Pan Left-Down
Pan Up	Pan Right-Up
Pan Down	Pan Right-Down
Zoom In	Zoom Out

A.1.2 Baselines

Training-time alignment. We compare against the base Wan2.2-TI2V-5B (Wan et al., 2025) without fine-tuning and three fine-tuned baselines: supervised fine-tuning (SFT) only, VIST3A (Go et al., 2026) with SFT, and World-R1 (Wang et al., 2026) with SFT. For a controlled comparison, all fine-tuned models, including ours, use the same SFT dataset, and the same training GPU hours, and are decoded by the same frozen stitched 4D reconstructor. We use the same frozen stitched 4D reconstructor for reward computation in VIST3A and World-R1.

Image-to-4D generation. We use the publicly released checkpoints of MoGe4D (Zhang et al., 2026) and 4DNeX (Chen et al., 2025). Both baselines generate their respective 4D point-cloud representations at 512×512 resolution. Since their output representations differ from our 4DGS representation, we render all methods under a common target camera trajectory for evaluation. We derive the target camera trajectory from ReCamMaster. The predicted camera poses are first converted to the OpenCV camera convention and expressed as relative poses with respect to the first frame. The same relative camera motion is then applied consistently to all frames and methods. To account for the scale ambiguity of monocular 4D reconstruction, we normalize the translational magnitude of the camera trajectory such that it corresponds to 0.25 times the median scene depth. When projecting the 4D point clouds to the target views, multiple points may map to the same image pixel; in such cases, we use a standard depth test and retain the point closest to the camera. The rendering procedure is adapted to the representation produced by each baseline. 4DNeX predicts a separate point map for each frame together with a corresponding input-view camera. For each frame, we transform its point map from the original input-view camera to the target ReCamMaster camera and rasterize the warped points in the target view. In contrast, MoGe4D represents the dynamic scene as

a set of points initialized from the pixels of the first frame and tracked throughout the entire 41 frame sequence, rather than predicting an independent point map for every frame. As MoGe4D directly generates the dynamic point-cloud representation without an explicit per-frame input camera, we render the tracked 3D point positions at each frame directly from the corresponding target camera pose.

Test-time steering. Test-time steering uses no trained checkpoint. We apply every steering method to the base camera-control model Wan2.2-Fun-5B-Control-Camera and compare unguided sampling (None), Best-of-N (BoN) and DPS (Chung et al., 2023).

A.1.3 TRAINING

We fine-tune Wan2.2-TI2V-5B with LoRA (Hu et al., 2021) (rank 4, α=8) on the query, key, value, and output projections of both the self- and cross-attention layers. The stitched 4D reconstructor is kept frozen. Each optimizer step combines a supervised loss with an online reinforcement-learning loss based on DiffusionNFT (Zheng et al., 2025).

Supervised term. We use the MultiCamVideo (Bai et al., 2025) (120,460 multi-camera videos) and the short-clip split of UltraVideo (Xue et al., 2025) (42,184 clips). Each video is 13 frames at 480×832. With probability 0.5, the first frame is used as an image condition, so the model is trained for both text-to-video and image-to-video generation.

Reinforcement-learning term. Each step draws 8 prompts from the VideoUFO (Wang & Yang, 2025) training prompts and samples a group of 8 rollouts per prompt (64 rollouts per step). Each rollout has 13 frames and is sampled with 10 denoising steps and no classifier-free guidance. Each rollout is scored by our reward with γ=0.7. The KL coefficient is β_{KL}=5×10^{-3}.

Optimization. We use AdamW (Loshchilov & Hutter, 2019) with a learning rate of 1×10^{-4}, (β_1, β_2)=$(0.9, 0.999)$, a weight decay of 1×10^{-4}, gradient clipping at 1.0, and bf16 mixed precision, and we train for 1,000 steps.

A.1.4 INFERENCE

Unless stated otherwise, all videos are generated at a resolution of 832×480 with 41 frames, 50 sampling steps, a classifier-free guidance scale of 5.0, and a fixed random seed shared across methods. For image-to-4DGS generation, the input image is fixed to the first latent frame at every denoising step. The resulting VDM latents are decoded into a 4DGS representation by the stitched reconstructor at a resolution of 518×294.

A.1.5 TEST-TIME STEERING

All steering methods share one stochastic flow-DDIM sampler (η=1, 50 steps, 41 frames). The initial latent of each sample is derived deterministically from the run seed and the sample index. As a result, every method starts from the same noise and draws the same transition noise. All methods maximize the proposed reward FOCAL, with γ=0.7. We use the following settings:

- **BoN**: sample N=20 full trajectories and keep the highest-reward one.
- **DPS** (Chung et al., 2023): at each guided step, back-propagate the reward through the transformer's clean-latent prediction, the 4D reconstructor, and the depth-reprojection term, and update the latent with step size 3.0. We guide denoising steps 3–17 of 50 (15 guided steps). The field-of-view term is non-differentiable and enters the reward only as a constant offset.

A.1.6 STITCHED MODEL TRAINING

The stitched model feeds the latents of a frozen Wan 2.2 (Wan et al., 2025) video VAE encoder into NeoVerse (Yang et al., 2026). We initialize the stitching layer following VIST3A (Go et al., 2026), which gives the lowest regression error among the candidate encoder blocks. We train the stitched model by self-distillation from the frozen NeoVerse. The student matches all of the teacher's outputs with ℓ_1 losses: point maps, camera poses and intrinsics, depth, Gaussian parameters, and,

for dynamic scenes, velocities. We train the stitching layer and the prediction heads fully, and adapt the NeoVerse backbone with LoRA (rank 64).

We first train on static scenes from DL3DV (Ling et al., 2024). We then fine-tune on dynamic videos from UltraVideo (Xue et al., 2025). The number of input frames per iteration is randomly sampled from lengths of the form $4n+1$. We use AdamW with a learning rate of 10^{-4}.

A.2 Evaluation Details

A.2.1 Evaluated Video

Except for the VGGRPO comparison (Section A.2.2), we score rendered videos rather than raw VDM decoded outputs. For text-to-4DGS and test-time steering, the generated 4DGS scene is rendered using the input-view cameras predicted by the 4D reconstructor. For image-to-4D generation, we render every generated 4D scene along the same novel-view camera trajectory. Following the baseline protocol, we fill disoccluded regions with MoGe4D's 4D-ViSM, applied identically to every method. Scoring the rendered videos exposes geometry that is inconsistent across frames.

A.2.2 Comparison Protocol with VGGRPO

VGGRPO (An et al., 2026) has released neither code nor checkpoints, so we compare with the numbers reported in its paper, matching its prompt and generation protocol, on the VAE decoded videos. We apply the same setting identically to the base model and to ours. Because the VGGRPO evaluated on a different random seed, we compare the relative improvement of each method over the base model, not the absolute scores.

A.2.3 Metrics per Task

- **Text-to-4DGS**: FOCAL, VBench and WorldScore.
- **Image-to-4D generation**: VBench subject consistency, temporal flickering, dynamic degree, and aesthetic quality; CLIP score for text alignment; and VideoPhy-2 physical commonsense (PC) for physical plausibility.
- **Test-time steering**: camera accuracy and WorldScore.

A.2.4 Metric Definitions

We compute metrics with the official implementations[1][2][3]:

1. **VBench** (Huang et al., 2023): We use VBench in custom-input mode and report eight dimensions: subject consistency (DINO (Caron et al., 2021) feature similarity across frames), background consistency (CLIP feature similarity across frames), aesthetic quality (LAION aesthetic predictor), imaging quality (MUSIQ), overall consistency (ViCLIP text–video similarity), temporal flickering, motion smoothness (AMT-based frame interpolation prior), and dynamic degree (RAFT optical flow). We report each dimension and their unweighted mean.
2. **WorldScore** (Duan et al., 2025): We report eight aspects: content alignment (CLIP score), 3D consistency (DROID-SLAM reprojection error), photometric consistency (SEA-RAFT optical-flow AEPE), style consistency (Gram-matrix distance of VGG-19 features), subjective quality (CLIP-IQA+ and CLIP aesthetic score), motion accuracy, motion magnitude (optical-flow magnitude), and motion smoothness (VFIMamba frame-interpolation error measured with LPIPS and SSIM). Motion accuracy measures whether motion is confined to dynamic regions. Upstream WorldScore obtains the dynamic region with a text-grounded detector. Our prompts do not name a single dynamic object, so we instead use SegAnyMo (Huang et al., 2025) to segment moving content in each rendered video and

[1] https://github.com/Vchitect/VBench
[2] https://github.com/haoyi-duan/WorldScore
[3] https://github.com/Hritikbansal/videophy

score flow inside and outside the resulting mask. We omit camera control and object control, which require per-instance camera and object annotations that our benchmarks do not provide.

3. **VideoPhy-2** (Bansal et al., 2025): We use VideoPhy-2-AutoEval (`videophy_2_auto`), an mPLUG-Owl-Video model with a LLaMA-7B backbone, further fine-tuned from VideoCon-Physics on VideoPhy-2 human annotations. We evaluate 32 sampled frames with the official physical-commonsense template (Figure 9) using greedy decoding. We report the mean physical commonsense (PC) score on a 1–5 scale.

4. **CLIP score** (Radford et al., 2021): The cosine similarity between CLIP ViT-L/14 image and text embeddings, scaled by CLIP's logit scale, computed on every frame and averaged over frames.

5. **Camera accuracy**: We compare the per-frame cameras predicted by the 4D reconstructor with the target trajectory. Predicted poses are defined only up to a similarity transform, so we express both trajectories relative to their first frame and report three errors: rotation error (the mean geodesic distance between the relative rotations, in degrees), translation-direction error (the median angle between the relative translation vectors, in degrees), and normalized ATE (the RMSE of the camera centers after a least-squares scale fit, divided by the maximum extent of the target trajectory). We use the median for translation-direction error because the first frames have almost no translation, which makes their direction ill-conditioned.

Prompt for VideoPhy-2 Evaluation

Physical commonsense (PC).
The following is a conversation between a curious human and an AI assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
Human: <|video|>
Human: Does this video adhere to the physical laws? Rate the video on a scale from 1 to 5, where 5 means full compliance and 1 means significant violations.
AI:

Figure 9: Official VideoPhy-2 physical-commonsense (PC) query template used with `videophy_2_auto`.

B Stitched Model Reconstruction Results

Setup. We evaluate on a subset of MultiCamVideo (Bai et al., 2025) dataset, which provides 10 synchronized cameras per scene. *Novel-time synthesis*: given the 33 even-indexed frames of one camera, the model renders the 32 held-out odd-indexed frames. *Novel-view synthesis*: given 21 frames from a context camera, the model renders the synchronized frames of a different, randomly chosen target camera. We compare against 4DGT (Xu et al., 2025), MoVieS (Lin et al., 2026), and NeoVerse (Yang et al., 2026). NeoVerse and our model do not take ground-truth cameras as input, and they predict poses in their own coordinate frame. For both models we therefore follow the evaluation-time pose alignment of NoPoSplat (Ye et al., 2024).

Results. As shown in Table 6, our stitched model outperforms baselines in most of the metrics in both novel-time and novel-view synthesis evaluation. On novel-time synthesis, compared with NeoVerse, our model improves PSNR by $+1.68$ dB, SSIM by $+0.066$, and LPIPS by -0.031. Since both models share the same backbone, we attribute this gain to the spatio-temporally compressed video VAE latents, which yield a more temporally coherent reconstruction. Novel-view synthesis results show that operating on video latents does not degrade cross-view geometry.

Table 6: Quantitative comparison on novel-time and novel-view reconstruction.

Method	Novel Time			Novel View		
	PSNR	SSIM	LPIPS ↓	PSNR	SSIM	LPIPS ↓
4DGT	17.72	0.5510	0.5032	16.35	0.5015	0.5167
MoVieS	19.12	0.5819	0.3785	16.83	0.4904	0.4271
NeoVerse	20.30	0.5793	0.2917	17.23	0.5117	**0.3716**
Ours	**21.98**	**0.6451**	**0.2605**	**17.65**	**0.5187**	0.3748

C Additional Results

C.1 Additional 4DGS Qualitative Results

Figure 10 shows additional Text-to-4DGS results.



Plush teddy bear tumbling next to a leather couch.



A cat slinking to the right side of a living room.



A teacher explaining in front of a chalkboard.



A cyclist pedals from the right side to the left side on a street.



Wooden car speeding past a porcelain tree.



A tiger walking on a dirt path in the woods.

Figure 10: Additional qualitative results on Text-to-4DGS generation. We overlay 3D representations from multiple timesteps in a single view to visualize dynamic motion over time.

C.2 Additional Text-to-Video Qualitative Results

Figure 11 provides additional 4DGS-rendered Text-to-Video comparisons against baseline methods.

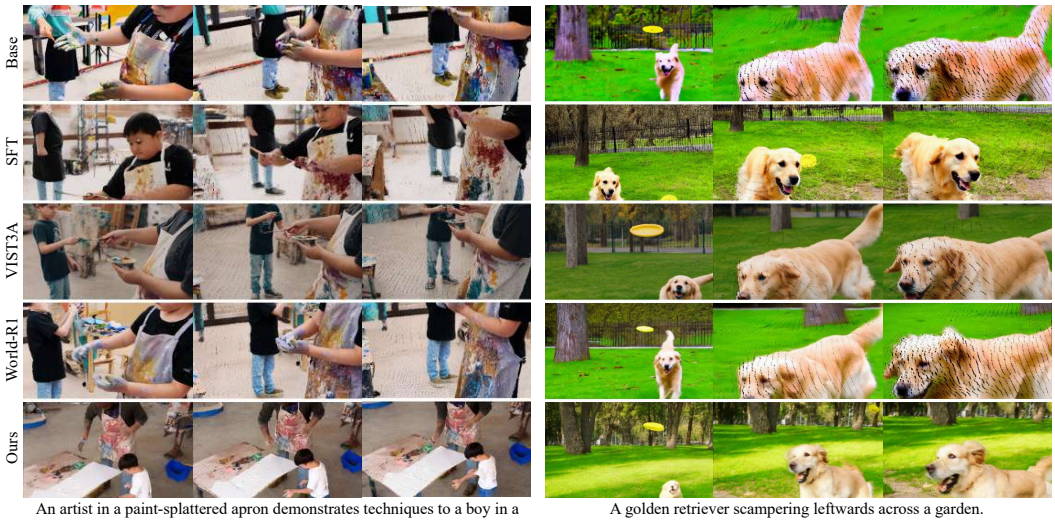

Figure 11: Additional qualitative comparison on Text-to-Video generation.

D Additional Analysis

D.1 FOV Bias in Depth Errors

Figure 12 visualizes the FOV bias discussed in Section 3.2 on the MultiCamVideo (Bai et al., 2025) dataset. For each clip, we compute the raw cross-view depth reprojection error (Eq. 4) with the stitched reconstructor and the FOV overlap Ω (Eq. 6). Clips with large camera motion, and thus a large change in viewing frustum between the first and last frames (low Ω), show substantially higher depth errors. Clips with small camera motion (high Ω) show low errors. This gap arises from the viewing configuration rather than geometric inconsistency: with small overlap, reprojected points land near image boundaries, at grazing angles, or on newly revealed and occluded surfaces, where depth prediction is inherently harder. A reward based on the raw score would therefore favor videos with little camera motion over videos that are more geometrically consistent, which motivates the calibration in Section 3.3.



Figure 12: **FOV bias in cross-view depth errors on real videos.** Real videos with a large FOV change (low overlap Ω) exhibit larger depth reprojection errors than those with a small FOV change (high Ω), even when their reconstruction quality is comparable. Thus, the raw depth-agreement score is influenced not only by geometric quality but also by camera motion.

D.2 FOV Overlap Computation

We describe how the FOV overlap Ω in Eq. 6 is computed from the stitched reconstructor's camera predictions. Figure 13 illustrates the procedure.

Camera FOV. For frame f with predicted intrinsics $\mathbf{K}_f$ and image resolution $W \times H$, the horizontal and vertical FOVs are

$$\theta_f^h = 2\arctan\left(\frac{W}{2f_x}\right), \qquad \theta_f^v = 2\arctan\left(\frac{H}{2f_y}\right), \tag{11}$$

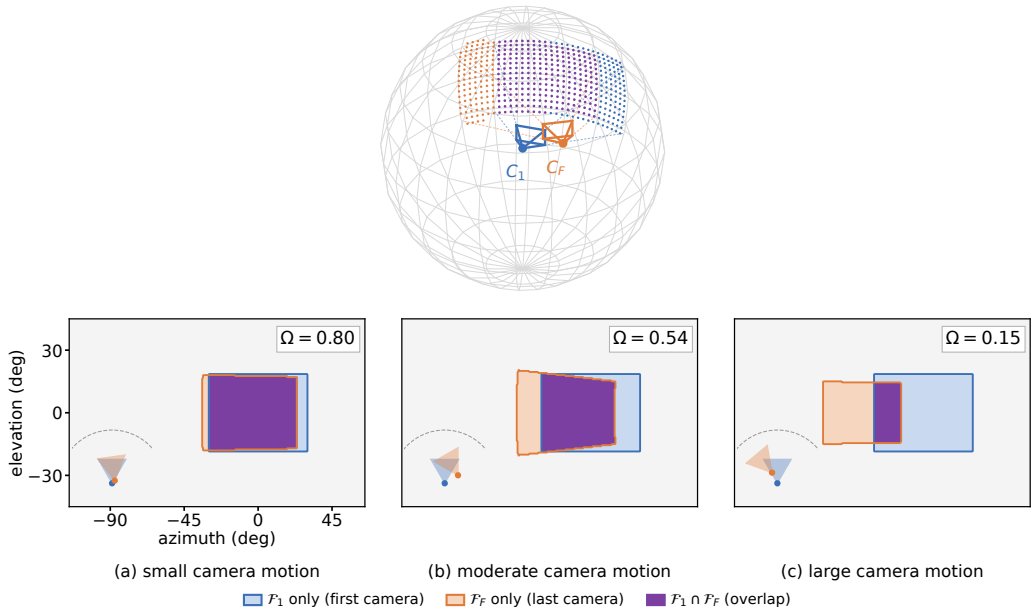

Figure 13: **FOV overlap computation. Top:** the first camera C_1 and the last camera C_F are expressed in the first camera's frame, and a reference sphere of radius r is centered at C_1. Sphere samples inside the FOV of C_1 only, of C_F only, and of both cameras are shown in blue, orange, and purple, respectively. **Bottom:** the same samples on the azimuth–elevation grid for (a) small, (b) moderate, and (c) large camera motion between the first and last frames, with the resulting overlap Ω. The top row corresponds to (b). Insets show the camera poses from above.

where f_x and f_y are the focal lengths in $\mathbf{K}_f$. We use only the first and last frames for efficiency, $f \in \{1, F\}$, which span the full camera trajectory of the clip.

Reference sphere. We express both cameras in the coordinate frame of the first camera, $\tilde{\mathbf{T}}_f = \mathbf{T}_1^{-1}\mathbf{T}_f$, so that C_1 lies at the origin. We then place a reference sphere of radius r centered at C_1 (Figure 13, top). We set r to the 75th percentile of the predicted depths of the clip, so that the sphere lies at a typical scene distance and camera translation is measured relative to the scene scale. We use a percentile instead of the maximum or the mean because depth distributions are often heavy-tailed: regions such as the sky or distant backgrounds receive very large, and often unreliable, depth values. A radius driven by these regions would place the sphere far beyond the actual scene content, so that even a large camera translation becomes negligible relative to r and Ω reflects almost only rotation. We sample the sphere on a uniform azimuth–elevation grid with 360×180 points,

$$\mathbf{s}_{a,e} = r\left(\cos e \cos a,\ \cos e \sin a,\ \sin e\right), \qquad a \in [-\pi, \pi],\ e \in [-\tfrac{\pi}{2}, \tfrac{\pi}{2}]. \tag{12}$$

Area-weighted IoU. A uniform grid in azimuth and elevation oversamples the sphere near the poles, since the area of a grid cell is proportional to $\cos e$. We therefore weight each point by $w(\mathbf{s}_{a,e}) = \cos e$ and compute

$$\Omega = \frac{\sum_{\mathbf{s} \in \mathcal{A}_1 \cap \mathcal{A}_F} w(\mathbf{s})}{\sum_{\mathbf{s} \in \mathcal{A}_1 \cup \mathcal{A}_F} w(\mathbf{s})}, \tag{13}$$

so that Ω approximates the ratio of the spherical areas of the intersection and union of the two FOVs. $\Omega = 1$ when the first and last cameras see the same region, and $\Omega = 0$ when their FOVs do not overlap on the sphere. As shown in Figure 13(a)–(c), Ω decreases as the last camera rotates and translates further away from the first camera.

D.3 Analysis of VBench Rewards

We provide additional analysis of the video-quality reward ablation in Section 4.5. Figure 14(a) shows additional qualitative results from training with the five VBench (Huang et al., 2023) re-

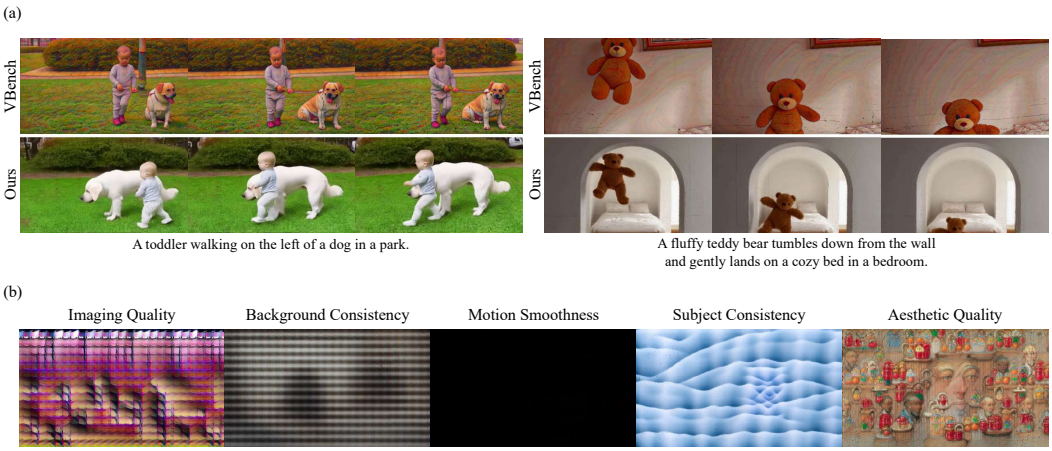

Figure 14: **Analysis of VBench rewards.** **(a)** Additional qualitative results for the video-quality reward ablation in Section 4.5, comparing training with VBench rewards against FOCAL. **(b)** Representative artifacts from optimizing each of the five VBench rewards individually.

wards used by DDRL (Ye et al., 2025). The VBench-trained model exhibits excessive saturation and unnatural color casts, whereas our geometric reward produces more natural colors and cleaner scene structure. To examine the contribution of each reward, Figure 14(b) shows results obtained by optimizing each of the five metrics individually in the direct reward fine-tuning setting. The resulting artifacts reveal distinct failure modes, which we interpret below in terms of what each metric measures.

Imaging quality. Optimizing imaging quality produces strong color banding and repetitive, high-contrast textures that obscure the scene. This reward averages the frame-wise predictions of MUSIQ (Ke et al., 2021), an image-quality predictor to assess distortions such as blur, noise, and exposure errors. The observed patterns suggest that direct optimization can exploit the predictor's learned responses to contrast and texture, amplifying artificial detail without improving perceptual quality. A high predicted image-quality score does not impose explicit constraints on scene geometry or temporal coherence, allowing such structured artifacts to emerge.

Background consistency. Optimizing background consistency produces blurred content overlaid with a regular grid. Despite its name, this metric compares global CLIP (Radford et al., 2021) image features extracted from whole frames, without explicitly isolating the background. It averages the CLIP feature similarity of each frame to the first frame and to its immediately preceding frame. Consequently, reducing frame-to-frame feature variation can be favored even at the expense of image fidelity: blur can suppress changing details, while a persistent grid can provide a common appearance across frames. This offers a shortcut to feature consistency without preserving a sharp, geometrically consistent background under camera motion.

Motion smoothness. Optimizing motion smoothness leads to an almost black, nearly featureless output. The metric interpolates intermediate frames from neighboring frames and rewards a small pixel-wise difference between the interpolated and actual frames. Temporally uniform, low-contrast content makes this prediction task easier: suppressing visible structures and their motion can reduce interpolation error without generating realistic dynamics. The observed collapse is consistent with this incentive, since the reward itself does not require nontrivial motion or recognizable scene content.

Subject consistency. Optimizing subject consistency replaces recognizable objects with repetitive, smooth blue patterns. This metric measures DINO (Caron et al., 2021) feature similarity to the first and preceding frames; it does not explicitly verify the presence or geometry of the intended subject. Repeating similar structures across frames can therefore stabilize the descriptor even when subject identity is lost. The resulting pattern illustrates how consistency in feature space can be

achieved by suppressing meaningful appearance variation, rather than maintaining a coherent object as its pose or viewpoint changes.

Aesthetic quality. Optimizing aesthetic quality yields a collage-like image with repeated faces and decorative fragments instead of a coherent scene. The reward averages a LAION (Schuhmann et al., 2022) aesthetic predictor applied to CLIP image embeddings of individual frames. Such optimization can amplify visual motifs associated with high predicted aesthetic scores while leaving their spatial arrangement and temporal evolution unconstrained.

These examples illustrate the gap between evaluating generated videos with perceptual proxies and directly optimizing those proxies. Temporal feature similarity can admit degenerate content, while frame-wise quality predictors can favor artifacts outside their intended image distribution. Our reward instead evaluates cross-view geometric agreement in static regions, providing an explicit constraint on scene structure. The comparison in Figure 14(a) supports the finding in Section 4.5 that improving geometric consistency can also improve visual quality.

E LIMITATIONS

Our method inherits the bias of the pre-trained VDM and the 4D reconstruction model. The depth and camera predictions are distilled from a pretrained 4D reconstructor, so failure cases of the teacher can produce inaccurate reward signals. Errors in the dynamic segmentation stage can penalize valid motion or leave inconsistent regions unconstrained. We expect that 4D reconstructors and dynamic tracking will further improve. Finally, computing the reward requires running the stitched reconstructor and the motion segmentation network, and gradient-based guidance additionally requires backpropagation through the reconstructor, which increases training and inference cost.