CAVE: A GENERATED WORLD SHOULD STILL EXPLAIN ITS PAST

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generation models can produce visually realistic sequences, yet object structures and spatial relationships may still drift over time. Existing geometric metrics compare frame pairs or jointly model the entire sequence under a shared interpretation, making conflicts difficult to localize or allowing anomalous observations to be absorbed into the overall solution. We introduce an observer-centric principle: video geometric consistency should be evaluated in observation order. The historical prefix provides an internal geometric reference, and a new observation should preserve geometric relationships already supported by history. Building on this principle, we develop CAVE, which maintains an evolving world state and adopts an "assimilate-then-look-back" verification strategy. At each timestep, CAVE queries identical historical views before and after the update and measures the resulting increase in historical explanation error. These responses form a geometric damage tensor that supports video scoring, temporal conflict localization, and current-view spatial attribution. On controlled geometric conflicts, CAVE improves temporal and spatial localization over prior metrics, with particularly strong gains under persistent conflicts. We further use the CAVE video score as preference supervision for DPO post-training, improving geometric consistency across multiple independent metrics while maintaining video quality and dynamics. CAVE therefore turns directional historical verification into a practical signal for both geometric diagnosis and post-training.

1 INTRODUCTION

Video generation models (Wan et al., 2025; Kong et al., 2024) have made substantial progress in visual quality, motion coherence, and controllability (Chen et al., 2024; Yang et al., 2025; Wang et al., 2024). Yet visually convincing frames do not guarantee a geometrically stable video. Object structures and spatial relationships may still drift, deform, or reorganize over time (Gu et al., 2025; Du et al., 2026). Reliable geometric evaluation is therefore needed to diagnose these failures and provide feedback for improving generation.

Existing geometric evaluations derive evidence from inter-frame correspondences, reprojection errors, or multi-view reconstruction (Yu et al., 2023; Asim et al., 2025; Gu et al., 2025). Pairwise methods can reveal that two observations are geometrically incompatible, but their symmetric relations do not identify which observation introduced the conflict. Full-sequence methods instead fit a shared geometric interpretation to all observations, allowing conflicting evidence to be absorbed into the overall solution. Both paradigms can quantify geometric incompatibility, yet neither explicitly attributes the conflict to its source.

A real observer understands a scene sequentially, building an evolving interpretation of the world from accumulated observations. For a static scene observed by a moving camera, this observation order provides a natural direction for geometric evaluation. Open-ended generation provides no unique target 3D world, but the preceding observations support a partial geometric interpretation (Wu et al., 2026b). A new observation may legitimately expand this interpretation by revealing previously unseen geometry; however, it should not invalidate structures or spatial relationships already supported by history (Xiao et al., 2026). We therefore characterize geometric inconsistency not as geometric change itself, but as the damage that a new observation causes to the world established by its past.

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.

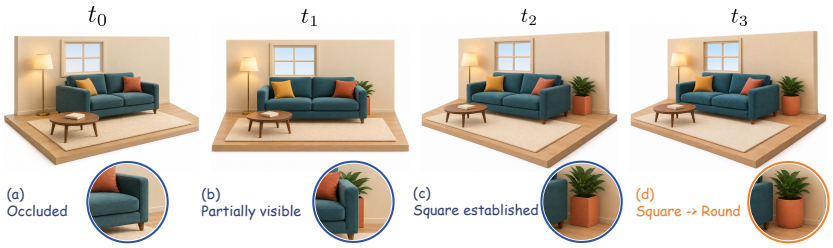

Figure 1: **Geometric updates under sequential observation.** As the camera moves, hidden parts of the scene become visible. The planter appearing from behind the sofa is a natural change in visibility, whereas its pot changing from square to round cannot be explained by viewpoint alone.

Based on this principle, we introduce **CAVE** (**C**onsistency **A**ssessment via **V**erification of **E**volving Worlds), which follows the same sequential process by first assimilating each new observation into the evolving world and then asking whether the updated world can still explain its past. Rather than judging consistency by the post-update error alone, CAVE measures how much the world's ability to explain geometry already supported by history deteriorates after the new observation is assimilated. This directional damage is organized into a geometric damage tensor, supporting video-level scoring, temporal and spatial localization, and preference-based post-training.

Our main contributions are as follows:

1. We formulate video geometric consistency from an observer-centric, sequential perspective. We view its history as an evolving interpretation of the world and characterize inconsistency by whether a new observation damages geometry already supported by history.
2. We introduce CAVE, an assimilate-then-look-back framework that compares the explanation of historical evidence before and after each update and organizes the newly introduced error into a geometric damage tensor. This representation supports both geometric measurement and spatial attribution.
3. We demonstrate the utility of CAVE for both diagnosis and optimization. On controlled geometric conflicts, CAVE improves temporal and spatial localization over existing metrics, while CAVE-guided preference post-training improves geometric consistency across multiple independent evaluators.

2 Related Work

2.1 Geometric Consistency Evaluation for Generated Videos

Generated-video geometry is commonly evaluated using cross-frame correspondences, projective constraints, or reconstructed 3D structure. TSED (Yu et al., 2023) measures pairwise projective consistency using symmetric epipolar distance over feature matches. Li et al. (2024) combine matching with projective constraints and 3D reconstruction to study the geometric stability of Sora videos across temporal intervals, while GeoCon-Bench (Xue et al., 2026) fits homography or fundamental-matrix models and reports inlier- and error-based diagnostics.

Depth, point maps, and reconstruction provide denser geometric evidence. MEt3R (Asim et al., 2025) aligns pairwise point maps and compares visual features after cross-view reprojection; WorldScore (Duan et al., 2025) evaluates SLAM-based dense reprojection; and GeCo (Gu et al., 2025) fuses motion residuals with depth reprojection to localize deformation and occlusion inconsistencies. SGC (Dou et al., 2026) separates dynamic foreground from static scene regions and measures 3D spatial inconsistency through divergence among locally estimated camera poses. GeoT2V-Bench (Fan & Favaro, 2026) and PDI-Bench (Wu et al., 2026a) further diagnose rigidity and projective violations using static-versus-deformable reconstruction and object-centric monocular 3D analysis, respectively. Existing methods use diverse geometric cues, but mainly assess consistency through relations among observations or a shared scene interpretation. CAVE instead evaluates how each new observation affects geometry already supported by history.

2.2 Feedback-Driven Post-Training for Video Generation

Post-training of diffusion and flow models incorporates evaluative feedback through differentiable reward optimization, offline preference learning, or online group-relative policy optimization (Clark et al., 2024; Wallace et al., 2024; Liu et al., 2026a).

For geometric feedback, Epipolar-DPO (Kupyn et al., 2025) and VideoGPA (Du et al., 2026) construct offline preference pairs from pairwise Sampson errors and scene-level reprojection signals, respectively. World-R1 (Wang et al., 2026) instead combines feedback from 3D foundation models and vision-language models within Flow-GRPO, while VGGRPO (An et al., 2026) performs GRPO using latent-space camera-motion and geometric reprojection rewards. These methods explore different ways of incorporating geometric feedback into video post-training. CAVE instead contributes a geometric feedback signal that can be used within an existing preference-learning framework.

3 A Sequential Formulation of Geometric Consistency

3.1 Historical Geometric Constraints in Open-Ended Generation

Given a text condition, object shapes, scene layouts, and camera trajectories may all admit multiple plausible realizations. A generated video therefore has no unique target 3D world. Let the video be a temporally ordered sequence of observations,

$$\mathbf{X}_{1:T} = (\mathbf{X}_1, \ldots, \mathbf{X}_T), \qquad \mathbf{X}_t \in \mathbb{R}^{H\times W\times 3}.$$

Each $\mathbf{X}_t$ is only a partial observation of the world depicted. The historical prefix $X_{1:t-1}$ does not specify a complete world, but it constrains the set of geometric interpretations consistent with the observed surfaces, object structures, and spatial relationships. The video's own history therefore serves as an internal reference for judging the geometric consistency of subsequent observations.

3.2 Sequential Attribution and Geometric Damage

At time t, $\mathbf{X}_t$ is the only newly introduced evidence relative to the historical prefix $\mathbf{X}_{1:t-1}$. If incorporating X_t deteriorates geometric relationships already supported by history, this change provides a directional measure of the conflict introduced at the current update. We define *geometric damage* as the deterioration in the explanation of historical geometry caused by incorporating a new observation, measured relative to the pre-update interpretation.

This judgment concerns not whether the geometric interpretation changes, but whether geometric relationships already supported by the history become invalid after an update. As illustrated in Figure 1, revealing previously occluded geometry expands the interpretation, whereas contradicting an already established structure constitutes damage.

For $t = 2, \ldots, T$, we represent the damage caused by observation $\mathbf{X}_t$ to the history as

$$\mathbf{D}_t = \mathcal{F}(\mathbf{X}_{1:t-1}, \mathbf{X}_t) \in \mathbb{R}_{\geq 0}^{(t-1)\times H\times W}.$$

Here, $\mathbf{D}_t[h, y, x]$ denotes the damage caused by observation $\mathbf{X}_t$ to the geometry supported at location (y, x) in historical observation $\mathbf{X}_h$. Organizing the results from all timesteps along the update order yields the complete geometric damage tensor

$$\mathbf{D} \in \mathbb{R}_{\geq 0}^{T\times T\times H\times W}, \qquad \mathbf{D}[t, h, :, :] = \begin{cases} \mathbf{D}_t[h, :, :], & h < t, \\ \mathbf{0}, & h \geq t. \end{cases}$$

Accordingly, $\mathbf{D}[t, h, y, x]$ records which observation introduced the conflict, which historical view was affected, and where within that view the geometry was affected. Each $\mathbf{D}_t$ is determined solely by the prefix $\mathbf{X}_{1:t}$ and does not depend on subsequent observations.

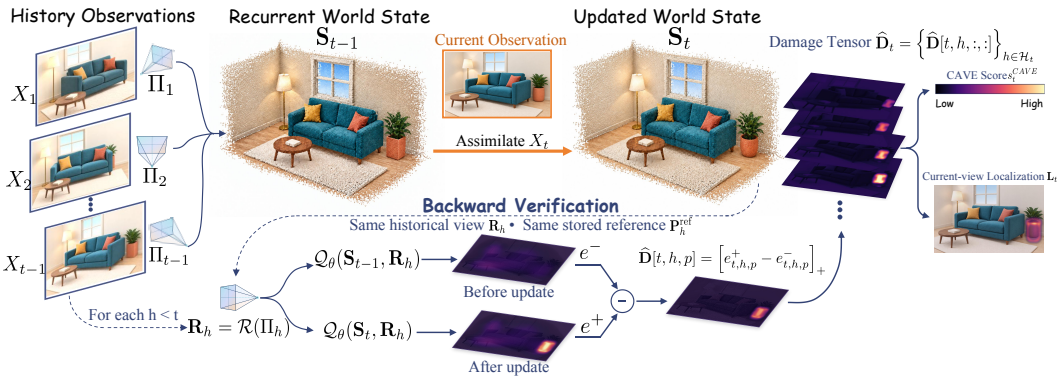

Figure 2: **Overview of CAVE.** CAVE queries the same historical view from both states, constructs a geometric damage tensor from the newly introduced error relative to a fixed historical reference, and further derives a video-level score and spatial attribution in the current view.

4 METHOD

4.1 RECURRENT WORLD STATE AND HISTORICAL REFERENCES

The world expands with observation. Our default implementation uses CUT3R (Wang et al., 2025b), with its parameters frozen, as a recurrent geometry observer. Given a video sequence $\mathbf{X}_{1:T}$, its state update and geometric query are written as

$$\mathbf{S}_t = \mathcal{U}_\theta(\mathbf{S}_{t-1}, \mathbf{X}_t), \quad (\mathbf{P}, \mathbf{\Gamma}) = \mathcal{Q}_\theta(\mathbf{S}_t, \mathbf{R}). \tag{1}$$

Here, $\mathcal{U}_\theta$ writes observation $\mathbf{X}_t$ into the world state, while $\mathcal{Q}_\theta$ reads a point map $\mathbf{P}$ and its normalized confidence $\mathbf{\Gamma}$ from view encoding $\mathbf{R}$ without updating the state.

When observation $\mathbf{X}_h$ is assimilated, we store its point map $\mathbf{P}_h^{\text{ref}}$, confidence $\mathbf{\Gamma}_h^{\text{ref}}$, and camera parameters $\mathbf{\Pi}_h$, comprising predicted camera pose and intrinsics from $\mathbf{P}_h^{\text{ref}}$, forming the archive

$$\mathcal{A}_h = \left(\mathbf{P}_h^{\text{ref}}, \mathbf{\Gamma}_h^{\text{ref}}, \mathbf{\Pi}_h\right). \tag{2}$$

The camera parameters further determine the corresponding query view $\mathbf{R}_h = \mathcal{R}(\mathbf{\Pi}_h)$. The world state continues to assimilate new observations, whereas the historical archive preserves the geometric reference established when that observation was first assimilated.

4.2 SEQUENTIAL ASSIMILATION AND HISTORICAL VERIFICATION

History constrains only what has already been observed. Testing a new observation directly against the historical geometric interpretation would leave newly revealed geometry and rewritten structure equally unexplained. CAVE thus reverses the direction of examination: admit the new observation into the world, then let the revised world re-explain its own past.

The only new observational evidence introduced at update t is $\mathbf{X}_t$. For each historical timestep $h < t$, CAVE retrieves the fixed archive $\mathcal{A}_h$ and queries the same historical view $\mathbf{R}_h$ from the states immediately before and after assimilating $\mathbf{X}_t$:

$$(\mathbf{P}_{t,h}^-, \mathbf{\Gamma}_{t,h}^-) = \mathcal{Q}_\theta(\mathbf{S}_{t-1}, \mathbf{R}_h), \tag{3}$$

$$(\mathbf{P}_{t,h}^+, \mathbf{\Gamma}_{t,h}^+) = \mathcal{Q}_\theta(\mathbf{S}_t, \mathbf{R}_h). \tag{4}$$

To compare the two predictions under a common geometric reference, CAVE independently aligns each point map to the fixed historical reference $\mathbf{P}_h^{\text{ref}}$ using weighted rigid $\mathrm{SE}(3)$ alignment:

$$\widetilde{\mathbf{P}}^{\eta}_{t,h} = \mathrm{Align}_{\mathrm{SE}(3)}\left(\mathbf{P}^{\eta}_{t,h}, \mathbf{P}^{\mathrm{ref}}_{h}; w_{t,h}\right), \qquad \eta \in \{-,+\}. \tag{5}$$

Both aligned point maps are expressed in the self-view camera coordinates of $\mathbf{X}_h$, whose optical center is the origin; alignment uses locations with finite positive depth and positive confidence in all three maps, weighted by the geometric mean of their predicted confidences:

$$w_{t,h,p} = \left(\Gamma^{-}_{t,h,p}\Gamma^{+}_{t,h,p}\Gamma^{\mathrm{ref}}_{h,p}\right)^{1/3}. \tag{6}$$

The aligned interpretations are then compared pointwise:

$$e^{\eta}_{t,h,p} = \frac{\left\|\widetilde{\mathbf{P}}^{\eta}_{t,h,p} - \mathbf{P}^{\mathrm{ref}}_{h,p}\right\|_2}{\max\left\{\frac{1}{2}\left(\left\|\widetilde{\mathbf{P}}^{\eta}_{t,h,p}\right\|_2 + \left\|\mathbf{P}^{\mathrm{ref}}_{h,p}\right\|_2\right), \epsilon\right\}}, \tag{7}$$

$$\widehat{\mathbf{D}}[t,h,p] = \left[e^{+}_{t,h,p} - e^{-}_{t,h,p}\right]_{+}, \qquad h < t. \tag{8}$$

Here, $\epsilon > 0$ is a numerical stability constant. The pre-update error $e^{-}_{t,h,p}$ provides the baseline of the interpretation. Only increases after assimilation are retained as damage; unchanged or improved historical explanations contribute zero. CAVE thus measures not how much the world changes during an update, but how much that change reduces the world's ability to explain its own history.

Repeating this test over all historical views and update times yields the concrete damage estimate $\hat{\mathbf{D}}$ corresponding to the formulation in Section 3. Each slice $\widehat{\mathbf{D}}[t,h,:,:]$ is determined solely by the prefix $\mathbf{X}_{1:t}$. The entire detection process uses a frozen geometry observer, draws its confidence weights directly from CUT3R, and introduces no additional learnable parameters.

4.3 Geometric Scoring and Current-View Attribution

Let $\mathrm{RMS}_{\mathbf{w}}(\mathbf{A}) = (\sum_p w_p A_p^2 / \sum_p w_p)^{1/2}$ denote the weighted spatial root mean square, and let $\mathcal{H}_t$ be the set of valid historical views for which alignment succeeds. The frame-level and video-level CAVE scores are defined as

$$\begin{aligned} s^{\mathrm{CAVE}}_t &= \frac{1}{|\mathcal{H}_t|}\sum_{h\in\mathcal{H}_t} \mathrm{RMS}_{\mathbf{w}_{t,h}}\left(\widehat{\mathbf{D}}[t,h,:,:]\right), \\ s^{\mathrm{CAVE}}(\mathbf{X}_{1:T}) &= \operatorname*{mean}_{t:\mathcal{H}_t\neq\varnothing} s^{\mathrm{CAVE}}_t. \end{aligned} \tag{9}$$

Damage responses are defined in historical views and therefore cannot directly provide localization in the current observation. Rather than transporting historical errors across views, CAVE reads the same state update in two complementary ways: backward verification in historical views determines the damage magnitude, while the state revision in the current view provides its spatial carrier.

CAVE queries $\mathcal{Q}_\theta(\mathbf{S}_{t-1}, \mathbf{R}_t)$ and $\mathcal{Q}_\theta(\mathbf{S}_t, \mathbf{R}_t)$ on the same current-view ray grid, and independently aligns the resulting point maps $\mathbf{P}^{-}_{t,t}$ and $\mathbf{P}^{+}_{t,t}$ to the fixed reference $\mathbf{P}^{\mathrm{ref}}_t$ stored when $\mathbf{X}_t$ was assimilated. Here, $\mathbf{P}^{\mathrm{ref}}_t$ serves only to remove branch-dependent rigid gauges. Their normalized difference defines the current-view state revision

$$\mathbf{r}_{t,p} = \frac{\widetilde{\mathbf{P}}^{+}_{t,t,p} - \widetilde{\mathbf{P}}^{-}_{t,t,p}}{\max\left\{\frac{1}{2}\left(\|\widetilde{\mathbf{P}}^{+}_{t,t,p}\|_2 + \|\widetilde{\mathbf{P}}^{-}_{t,t,p}\|_2\right), \epsilon\right\}}. \tag{10}$$

Because the two queries share the same current-view ray grid, we use the confidence-weighted ray-aligned revision as the spatial carrier:

$$a_{t,p} = \bar{w}_{t,p}\left|\mathbf{d}^{\top}_{t,p}\mathbf{r}_{t,p}\right|, \tag{11}$$

where $\mathbf{d}_{t,p}$ is the unit viewing ray and $\bar{w}_{t,p}$ is the geometric mean of the confidences from the two queries and the current reference.

For a nonzero carrier RMS, we normalize the spatial carrier to match the frame-level damage score:

$$\mathbf{L}_t[p] = s_t^{\mathrm{CAVE}} \frac{a_{t,p}}{\mathrm{RMS}_{\bar{\mathbf{w}}_t}(\mathbf{a}_t)}, \qquad \mathrm{RMS}_{\bar{\mathbf{w}}_t}(\mathbf{L}_t) = s_t^{\mathrm{CAVE}}. \tag{12}$$

Current-view attribution therefore redistributes the damage established by historical verification without altering its frame-level magnitude or CAVE score.

4.4 CAVE-DPO: From Geometric Measurement to Generation Optimization

CAVE-DPO constructs preference pairs using the video-level damage score. We adopt VideoGPA's DPO framework (Rafailov et al., 2023; Du et al., 2026; Wallace et al., 2024) for post-training.

Let ϕ and ϕ_0 denote the current model and the fixed reference model, respectively, and let ℓ^{w} and ℓ^{l} denote the velocity-prediction mean-squared errors for the preferred and dispreferred samples. The CAVE-DPO objective is

$$\mathcal{L}_{\text{CAVE-DPO}} = -\mathbb{E} \log \sigma\big(\beta \left[\big(\ell^{\mathrm{w}}_{\phi_0} - \ell^{\mathrm{w}}_{\phi}\big) - \big(\ell^{\mathrm{l}}_{\phi_0} - \ell^{\mathrm{l}}_{\phi}\big)\right]\big). \tag{13}$$

Here, β is the standard temperature coefficient of diffusion DPO. CAVE thereby closes the loop from measuring geometric inconsistency to providing an optimization signal for the generative model.

5 Experiments

We evaluate CAVE at two levels: whether it can identify when and where a geometric conflict enters a video, and whether the resulting signal can improve generation through post-training.

5.1 Experimental Setup

Experiment 1: Controlled geometric conflicts. We construct 100 test samples from CO3D (Reizenstein et al., 2021) videos by applying thin-plate-spline (TPS) deformations following GeCo's WarpBench protocol (Gu et al., 2025). Experiment 1 evaluates geometric conflict detection in both time and space. Experiment 1A replaces one randomly selected frame with its deformed version and evaluates localization of that frame. Experiment 1B continuously applies geometric perturbations to subsequent frames from the anomalous time step in Experiment 1A onward, and evaluates localization under persistent anomalies. Experiment 1C evaluates the spatial responses from Experiment 1A in the anomalous frames, using the injected deformations as ground truth.

Temporal localization is evaluated using exact-frame accuracy (Exact) and mean absolute frame error (MAE). Spatial localization is evaluated using average precision (AP), intersection over union under the optimal per-sample threshold (Best IoU), and the Spearman correlation between response intensity and deformation magnitude (Displacement SRCC).

We compare CAVE with MEt3R, VGGT-Reprojection, Epipolar, GeCo-Motion, and GeCo-Fused. These baselines cover pairwise-frame, centered-window, and joint full-sequence modeling schemes. Each method retains its original geometric computation.

Experiment 2: Video post-training. We use CogVideoX-5B (Yang et al., 2025) as the base model and follow the diffusion DPO framework of VideoGPA (Du et al., 2026). We compare VideoGPA, Epipolar-DPO (Kupyn et al., 2025), and CAVE-DPO. The three DPO variants share the same base model, data sources, optimization configuration, and training budget of 2,000 steps. Each geometric metric independently ranks the candidate videos, thereby producing its own preference pairs.

We evaluate on 200 DL3DV (Ling et al., 2024) scene descriptions focusing on scene structure and camera exploration, using the same sampling configuration for all methods. We use CAVE, 3D Consistency Score (3DCS) (Du et al., 2026), GeCo, Epipolar and SGC (Dou et al., 2026) to evaluate geometric consistency, and VideoReward (Liu et al., 2026b), VBench-Q, and Dynamic Degree (Huang et al., 2024) to evaluate generation quality and dynamics.

Table 1: Temporal and spatial localization of controlled geometric conflicts. **Best** Second best .

	1A: Single-frame		1B: Persistent		1C: Spatial		
Method	Exact ↑	MAE ↓	Exact ↑	MAE ↓	AP ↑	Best IoU ↑	SRCC ↑
MEt3R	36%	1.59	18%	2.13	0.248	0.343	-0.309
VGGT-Reprojection	63%	0.72	14%	1.78	0.425	0.355	0.119
Epipolar	68%	**0.50**	12%	1.50	–	–	–
GeCo-Motion	56%	0.98	27%	1.73	0.485	**0.411**	0.195
GeCo-Fused	45%	1.20	25%	1.87	0.441	0.399	0.143
CAVE	**71%**	0.57	**59%**	**1.02**	**0.515**	0.402	**0.234**

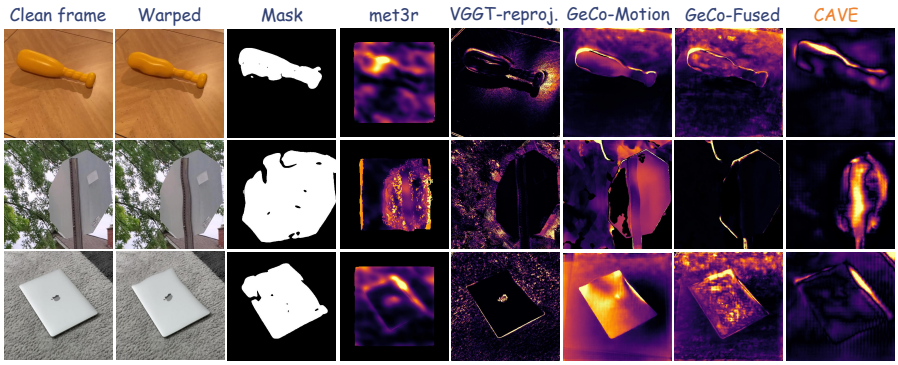

Figure 3: Qualitative comparison on controlled geometric conflicts.

5.2 Temporal and Spatial Localization of Geometric Conflicts

5.2.1 Which Observation Introduced the Conflict?

As shown in Table 1, under the single-frame perturbation setting, CAVE achieves the highest Exact accuracy of 71%. Its MAE is 0.57 frames, second only to the 0.50 of Epipolar. CAVE more frequently identifies the conflicting frame while maintaining comparable temporal localization error.

Experiment 1B evaluates temporal localization under persistent anomalies. Pairwise methods can detect a conflict between observations but struggle to determine which observation introduced it. Centered windows mix observations before and after the anomalous time step, while full-sequence modeling incorporates anomalous frames into reconstruction. Consequently, the Exact accuracy of Epipolar drops from 68% to 12%, while that of VGGT-Reprojection drops from 63% to 14%.

In contrast, CAVE evaluates each observation against the prefix world state before the update, attributing newly introduced damage to the update that first violates the existing world state. In Experiment 1B, CAVE achieves an Exact accuracy of 59%. The Exact advantage of CAVE grows from 3 percentage points in Experiment 1A to 32 percentage points in Experiment 1B.

5.2.2 Where Did the Damage Occur?

Experiment 1C extends the evaluation from temporal localization of the anomalous frame to spatial localization within that frame. CAVE achieves the highest AP of 0.515 and the highest Displacement SRCC of 0.234. The higher AP indicates that genuinely deformed regions are ranked more prominently across the spatial response, while the higher Displacement SRCC shows that response intensity more faithfully reflects the magnitude of local deformation.

CAVE obtains a Best IoU of 0.402, close to the best result of 0.411, while achieving the strongest threshold-independent ranking and deformation-magnitude correlation. Its response therefore provides accurate spatial localization together with a stronger correspondence between response intensity and deformation magnitude.

Table 2: Controlled post-training with different geometric preference signals. **Best** Second best .

	Geometry consistency ↓					VideoReward (%) ↑			Quality & dynamics	
Method	CAVE	3DCS	GeCo	Epipolar	SGC	VQ	MQ	TA	VBench-Q ↑	Dynamic (%)
Baseline	1.044	0.508	3.273	6.389	0.046	–	–	–	79.242	38.5
VideoGPA	1.033	0.503	3.612	5.951	0.040	66.25	61.75	**54.25**	79.366	37.5
Epipolar-DPO	1.076	0.502	3.296	**5.576**	0.041	58.75	60.50	49.75	79.254	36.0
CAVE-DPO	**0.991**	**0.501**	**2.969**	5.830	**0.038**	**73.25**	**70.00**	50.00	**79.456**	38.5

Figure 3 illustrates the spatial response patterns. CAVE concentrates its response on the deformed object while suppressing unrelated background activations. In comparison, MEt3R produces diffuse responses, and the reprojection- and GeCo-based methods are more easily dominated by object boundaries or background regions. These examples support the quantitative results that CAVE provides more accurate spatial localization of geometric conflicts. Appendix A examines CAVE across different geometry observers and analyzes its response using clean–conflict paired controls.

5.3 From Geometric Measurement to Video Generation Optimization

Experiment 1 establishes that CAVE can identify geometric damage in a video. We next examine whether this signal can directly guide a generative model toward more stable world evolution.

5.3.1 Quantitative Results

Table 2 compares different geometric preference signals under a unified framework. CAVE-DPO achieves the best results on CAVE, 3DCS, SGC, and GeCo and the second-best result on Epipolar. More importantly, it improves all four independent geometric metrics over the base model, whereas VideoGPA increases the GeCo error and Epipolar-DPO increases both CAVE and GeCo errors.

CAVE-DPO also achieves the highest VideoReward VQ and MQ preference rates of 73.25% and 70.00%, respectively, with a TA preference rate of 50.00%. It further obtains the highest VBench-Q score of 79.456 while maintaining the same Dynamic Degree as the base model at 38.5%. Cross-benchmark evaluation on the public GeCo-Eval prompt set is provided in Appendix B.

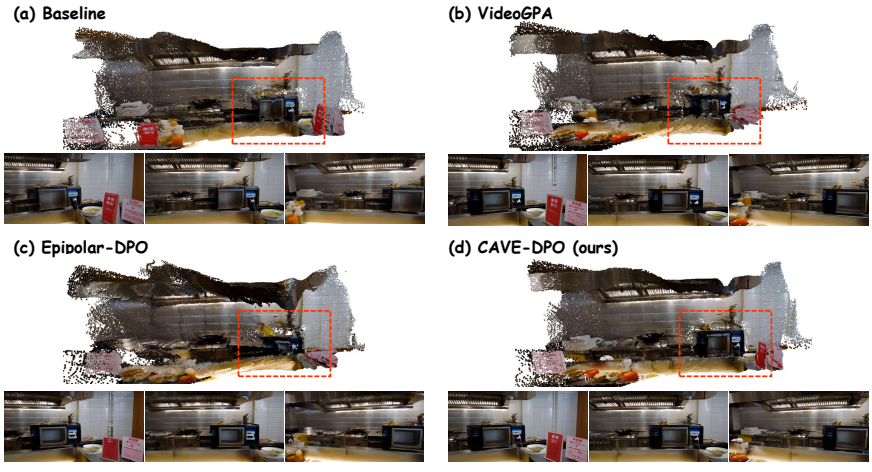

Figure 4: Qualitative comparison under the same scene description. Each panel shows sampled video frames (bottom) and the corresponding VGGT reconstruction (top).

5.3.2 Qualitative Results

Figure 4 presents videos generated by four methods from the same scene description, together with point clouds reconstructed from the corresponding videos using VGGT (Wang et al., 2025a). The 3D reconstructions reveal whether object structures remain consistent across different observations.

In the point clouds of Base, VideoGPA, and Epipolar-DPO, the microwave structure exhibits varying degrees of distortion, and its local surfaces fail to align consistently across multiple viewpoints. CAVE-DPO preserves a more complete rectangular outline of the microwave and its spatial relationship with the countertop, producing a more coherent 3D structure as the camera moves. This example visually demonstrates that preference supervision from CAVE reduces cross-frame structural drift and helps objects maintain more stable geometry under viewpoint changes.

5.4 Ablation Study

We compare four formulations of the local geometric signal. The post-update error e^+ measures inconsistency after assimilation without accounting for the pre-existing error. The direct state revision $d(\mathbf{P}^-, \mathbf{P}^+)$ measures the magnitude of the state update without reference to its effect on historical geometry. The unsigned change $|e^+ - e^-|$ compares the errors before and after assimilation but treats improvement and degradation identically. The complete CAVE definition, $[e^+ - e^-]_+$, retains only increases in historical error.

Table 3: Ablation of the geometric damage definition for temporal localization. **Best** <u>Second best</u> .

	Exp1A: Single-frame		Exp1B: Persistent	
Damage	Exact (%) ↑	MAE ↓	Exact (%) ↑	MAE ↓
e^+	14	2.22	20	2.25
$d(\mathbf{P}^-, \mathbf{P}^+)$	<u>59</u>	<u>0.80</u>	43	<u>1.12</u>
$\lvert e^+ - e^- \rvert$	58	0.93	<u>47</u>	1.21
$[e^+ - e^-]_+$	**71**	**0.57**	**59**	**1.02**

Table 3 isolates the effect of the damage definition while keeping the remaining pipeline fixed. Using only e^+ yields 14% and 20% Exact accuracy under single-frame and persistent perturbations, respectively, showing that absolute post-update error is a poor indicator of when a conflict is introduced. Both direct state revision and unsigned error change substantially improve localization, confirming the importance of measuring update-induced change. However, treating increases and decreases symmetrically remains suboptimal.

Retaining only error increases provides further gains over $|e^+ - e^-|$: Exact accuracy rises from 58% to 71% for single-frame perturbations and from 47% to 59% for persistent perturbations, while MAE decreases from 0.93 to 0.57 and from 1.21 to 1.02, respectively. The complete definition achieves the best results on all four temporal localization metrics, supporting the use of positive reference-relative error increments to identify when geometric conflicts enter a video.

6 Conclusion

The central idea of this work is that geometric consistency in generated video should be evaluated in observation order. A new frame may reveal previously unseen content, but it should not invalidate geometry already supported by earlier frames. We introduced CAVE to make this principle measurable. CAVE maintains an evolving world state and measures how each new observation affects its ability to explain historical geometry, producing directional damage for both scoring and localization. The same signal also provides effective geometric feedback for video post-training. Our results show that observation order offers a practical foundation for evaluating and improving the geometric consistency of generated videos.

AI USE STATEMENT

We used a large language model solely for language editing, including correcting grammar and improving the clarity and readability of the manuscript. All scientific contributions, including the research ideas, methodology, experimental design, and conclusions, are the authors' own work. The authors take full responsibility for the final content of this manuscript.

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. *arXiv preprint arXiv:2603.26599*, 2026.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6034–6044. IEEE, 2025.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320. IEEE, 2024.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. In *International Conference on Learning Representations*, volume 2026, pp. 50694–50718, 2026.

Kevin Clark, Paul Vicol, Kevin Swersky, and David Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, volume 2024, pp. 4793–4822, 2024.

Weijia Dou, Wenzhao Zheng, Weiliang Chen, Yu Zheng, Jie Zhou, and Jiwen Lu. Measuring 3D spatial geometric consistency in dynamic video generation. In *Computer Vision – ECCV 2026*, pp. 130–148, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-36846-1. doi: 10.1007/978-3-032-36846-1_8.

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. *arXiv preprint arXiv:2601.23286*, 2026.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 27713–27724, 2025.

Chenrui Fan and Paolo Favaro. Geot2v-bench: Benchmarking 3d consistency in text-to-video models via 3d reconstruction. *arXiv preprint arXiv:2606.24829*, 2026.

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. *arXiv preprint arXiv:2106.09685*, 2021.

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

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

Orest Kupyn, Théo Uscidda, Marta Tintore Gazulla, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models. *arXiv preprint arXiv:2510.21615*, 2025.

Xuanyi Li, Daquan Zhou, Chenxu Zhang, Shaodong Wei, Qibin Hou, and Ming-Ming Cheng. Sora generates videos with stunning geometrical consistency. *arXiv preprint arXiv:2402.17403*, 2024.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 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. *Advances in neural information processing systems*, 38:40783–40818, 2026a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2026b.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10901–10911, October 2021.

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

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

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

Weijie Wang, Xiaoxuan He, Youping Gu, Yifan Yang, Zeyu Zhang, Yefei He, Yanbo Ding, Xirui Hu, Donny Y Chen, Zhiyuan He, et al. World-r1: Reinforcing 3d constraints for text-to-video generation. *arXiv preprint arXiv:2604.24764*, 2026.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Jiaxin Wu, Yihao Pi, Yinling Zhang, Yuheng Li, and Xueyan Zou. Quantitative video world model evaluation for geometric-consistency. *arXiv preprint arXiv:2605.15185*, 2026a.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems*, 38:49371–49393, 2026b.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

Yifei Xue, Yuanchen Fei, Hao Zhang, Chenzhi Nie, Tie Ji, and Yizhen Lao. Illusion or integrity? geometrical consistency metric for AIGC video quality evaluation. *arXiv preprint arXiv:2608.09594*, 2026.

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. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Jason J Yu, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7094–7104, 2023.

A Discussion

A.1 CAVE with Different Geometry Observers

The central idea of CAVE is to observe an evolving world sequentially, as a real observer would. Accordingly, CAVE does not measure changes between observations themselves, but whether a new observation violates geometric relationships already supported by history.

We use CUT3R as the default geometry observer. CUT3R processes observations recurrently and maintains a persistent world state over time. It allows the historical view to be queried from the states before and after each update and compared against a fixed historical reference. This mechanism naturally matches the "sequential update–historical verification" process required by CAVE.

Beyond CUT3R, existing geometry models adopt different mechanisms for world modeling and state updating. For example, VGGT performs feed-forward multi-view geometric inference without maintaining a recurrent persistent state, whereas TTT3R (Chen et al., 2026) preserves the geometry representation and pretrained model of the CUT3R family while modifying the state-update mechanism at test time. Motivated by these differences, we further construct VGGT-CAVE and TTT3R-CAVE to examine CAVE under different world representations and state-update mechanisms.

A.1.1 Instantiation with Different Geometry Observers

VGGT-CAVE. Unlike CUT3R, VGGT does not maintain a persistent world state that is recurrently updated over time, but instead performs joint feed-forward geometric inference over the input views. If VGGT were rerun over all historical observations at every timestep, the arrival of a new observation would trigger a joint re-estimation of historical geometry. The resulting before–after difference would mix the effect of the current observation with changes caused by reinterpreting the history, and would no longer correspond to the single-observation update defined by CAVE.

We therefore construct an explicit append-only 3D world state $\mathcal{W}_t^{\mathrm{VGGT}}$ for VGGT. When a new observation $\mathbf{X}_t$ arrives, VGGT infers geometry over a short causal window ending at $\mathbf{X}_t$. The reconstruction is aligned to the existing world using the overlapping historical views, after which only the surface $\mathbf{P}_t^{\mathrm{VGGT}}$ contributed by the current observation is appended:

$$\mathcal{W}_t^{\mathrm{VGGT}} = \mathcal{W}_{t-1}^{\mathrm{VGGT}} \cup \mathcal{A}_t\big(\mathbf{P}_t^{\mathrm{VGGT}}\big)\,,$$

where $\mathcal{A}_t$ denotes the geometric alignment between the current surface and the existing world. Previously established historical surfaces remain frozen during subsequent updates. Consequently, $\mathcal{W}_{t-1}^{\mathrm{VGGT}}$ and $\mathcal{W}_t^{\mathrm{VGGT}}$ represent the world interpretations before and after assimilating the current observation, respectively, and can be directly used for CAVE's historical verification. All other components, including the historical references, geometric alignment, confidence weighting, damage computation, and score aggregation, remain unchanged.

TTT3R-CAVE. TTT3R provides a complementary variant. It preserves the geometry backbone, state representation, and pretrained parameters of CUT3R, but modifies the state-update process at test time. The recurrent update of CUT3R can be written as

$$\mathbf{S}_t = \mathcal{U}_{\mathrm{CUT3R}}(\mathbf{S}_{t-1}, \mathbf{X}_t),$$

where $\mathcal{U}_{\mathrm{CUT3R}}$ produces and commits a candidate global state from the incoming observation. TTT3R instead modulates the update magnitude of individual global state tokens using cross-attention responses between the existing state and the current observation. For exposition, its effect can be conceptually expressed as

$$\mathbf{S}_t = (1 - \boldsymbol{\beta}_t) \odot \mathbf{S}_{t-1} + \boldsymbol{\beta}_t \odot \widetilde{\mathbf{S}}_t,$$

where $\widetilde{\mathbf{S}}_t$ denotes the candidate global state proposed by the current observation, and the token-wise gate $\boldsymbol{\beta}_t$ controls the trade-off between preserving historical information and absorbing new information. CAVE applies historical verification to the states before and after the TTT3R update, while keeping the fixed references, geometric alignment, confidence weighting, damage definition, and score aggregation unchanged.

Table 4: Geometric damage localization with different geometry observers.

	1A: Single-frame		1B: Persistent		1C: Spatial		
Observer	Exact ↑	MAE ↓	Exact ↑	MAE ↓	AP ↑	Best IoU ↑	SRCC ↑
CUT3R-CAVE	71%	0.57	59%	1.02	0.515	0.402	0.234
VGGT-CAVE	68%	0.69	56%	1.10	0.509	0.378	0.212
TTT3R-CAVE	70%	0.55	59%	1.02	0.520	0.410	0.278

A.1.2 Geometric Damage Localization Across Observers

To isolate the effect of the geometry observer, we fully reuse the controlled geometric-conflict protocol from Experiment 1. Experiment 1A evaluates temporal localization under a single-frame geometric conflict, Experiment 1B evaluates localization of the first conflicting observation when the anomaly persists, and Experiment 1C evaluates spatial localization within the anomalous frame. Except for the observer and its world-state implementation, CAVE uses the same historical references, geometric alignment, confidence weighting, damage definition, score aggregation, and evaluation metrics across all variants. The experimental results are shown in Table 4.

Replacing CUT3R with VGGT, which uses a different geometry backbone and world representation, leads to only limited changes in localization performance. The Exact accuracy of VGGT-CAVE changes from 71% to 68% under the single-frame conflict setting and from 59% to 56% under persistent conflicts, while the spatial AP changes from 0.515 to 0.509. Although VGGT does not maintain CUT3R's recurrent latent state and instead uses an explicit append-only 3D surface representation, VGGT-CAVE retains comparable temporal and spatial localization performance. This shows that the sequential verification process can be instantiated with a different world representation.

TTT3R-CAVE uses the same geometry backbone and pretrained representation as CUT3R-CAVE but changes the memory-update mechanism at test time. Its Exact accuracy under the single-frame setting is 70%, close to the 71% of CUT3R-CAVE; under persistent conflicts, both achieve 59% Exact accuracy and an MAE of 1.02. Spatial localization also remains stable, with TTT3R-CAVE achieving an AP of 0.520 and a Best IoU of 0.410, while its Displacement SRCC increases from 0.234 to 0.278. Since TTT3R mainly modifies the state-update mechanism, these results indicate that CAVE's damage localization does not depend on the original state-transition rule of CUT3R.

VGGT-CAVE and TTT3R-CAVE provide complementary evidence: the former changes the geometry model and world representation, whereas the latter primarily changes the state-update mechanism. Their temporal and spatial localization performance indicates that CAVE's effectiveness mainly arises from the sequential verification principle of re-examining history after updating the world, rather than from a particular world-state implementation.

A.2 Clean–Conflict Paired Analysis

CAVE measures the deterioration of historical explainability after each state update. Because this response is computed through a learned geometry observer, even a geometrically valid observation may produce nonzero damage by refining an incomplete world state or exposing estimation errors. We therefore isolate the additional damage induced by a geometric conflict through a paired control.

We use all 100 samples from Experiment 1A. For each sample, we evaluate the clean sequence and its conflict counterpart, which differ only at the annotated onset frame; all other frames are identical. Both sequences contain 20 frames and are processed independently using the same fixed CAVE implementation. The clean counterpart is used only for this controlled analysis and is not required by CAVE at inference time.

Figure 5 shows both the absolute CAVE responses and their paired difference. At the onset frame, the mean score increases from 0.00282 for the clean sequences to 0.00407 for the conflict sequences, yielding a mean paired excess of 0.00125 with a bootstrap 95% CI of $[0.00093, 0.00162]$. The paired excess is exactly zero before the onset, peaks when the perturbation is introduced, and gradually decays over the following two to three frames, consistent with subsequent clean observations reducing the transient effect of the perturbed frame.

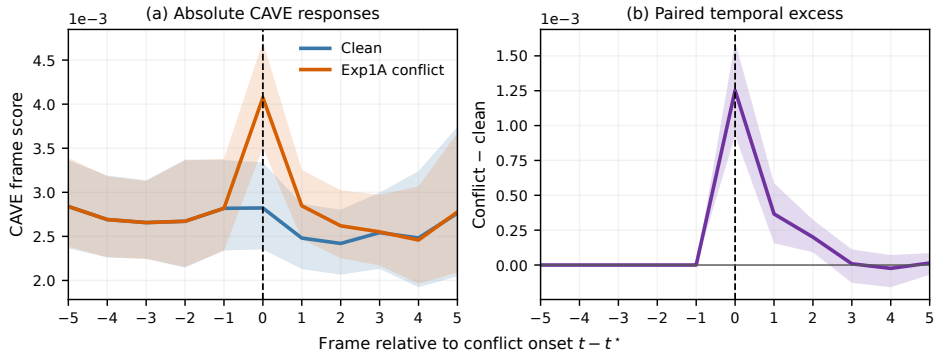

Figure 5: Paired analysis of CAVE responses. (a) Frame-level CAVE scores for clean and conflict sequences aligned to the perturbation onset. (b) Paired excess $\Delta s_t = s_t^{\text{conflict}} - s_t^{\text{clean}}$.

The analysis separates the normal response of the geometry observer from the effect of an injected conflict. Absolute CAVE scores exhibit a nonzero baseline due to ordinary state revision and observer error; nevertheless, under matched scene, camera, and observation history, the injected conflict produces a reliable and temporally localized excess. CAVE should therefore be interpreted as measuring directional deterioration in historical explainability rather than as a zero-baseline binary indicator of geometric conflict.

A.3 Historical Contribution to Current-View Attribution

We further examine whether historical damage evidence materially contributes to the final spatial attribution. To this end, we compare the normalized current-view carrier with the full CAVE attribution calibrated by the frame-level historical damage score.

We introduce clean–conflict paired AP. For all samples in Exp1C, we evaluate each conflict sequence together with its strictly matched clean control. The two sequences share the same scene, camera trajectory, and historical prefix, and differ only in whether the current frame contains a geometric perturbation. Over their common valid support, pixels in the conflict map are labeled according to the deformation mask, while all pixels in the clean map are treated as no-conflict controls. AP is then computed jointly over both maps. Unlike the within-frame spatial AP used in the main text, this metric evaluates both conflict-region localization and the suppression of normal state revisions.

Table 5: **Contribution of historical damage calibration.** Clean–conflict paired AP jointly evaluates conflict localization and response suppression on matched clean controls.

Spatial response	**Clean–Conflict Paired AP ↑**
Current-view carrier	25.3
Full CAVE attribution	**33.3**

Historical damage calibration increases paired AP from 25.3% to 33.3%, an average gain of 7.97 percentage points with a bootstrap 95% confidence interval of $[6.04, 9.89]$. As a consistency check, the maximum difference between the two responses in within-frame spatial AP is only 6.6×10^{-9}. The improvement therefore arises from distinguishing clean state revisions from conflict-induced revisions rather than altering the spatial ordering within a frame. These results show that the final CAVE attribution is not merely a visualization of current-view state revision: historical verification determines whether and how strongly a revision should be treated as damage, thereby suppressing benign geometric updates and producing clearer conflict attribution.

B Cross-Benchmark Evaluation on GeCo-Eval

B.1 Evaluation Protocol

We further evaluate the four controlled post-training variants on the prompt set released with GeCo. This benchmark provides an independent collection of scene descriptions and camera trajectories, allowing us to examine whether the improvements observed on the DL3DV evaluation set extend to a different prompt distribution. We generate all videos from scratch using the same CogVideoX-5B backbone, sampling configuration, and random seed for Base, VideoGPA, Epipolar-DPO, and CAVE-DPO. No additional training or model selection is performed on this benchmark.

The evaluation follows the same protocol as Experiment 2. We measure geometric consistency using CAVE, 3DCS, GeCo-Fused, Epipolar, and SGC. VideoReward reports the paired preference rate along visual quality (VQ), motion quality (MQ), and text alignment (TA). We additionally report VBench-Q and Dynamic Degree for video quality and motion preservation.

Table 6: Cross-benchmark evaluation on the GeCo-Eval prompt set. **Best** Second best .

	Geometry consistency ↓					VideoReward (%) ↑			Quality & dynamics	
Method	CAVE	3DCS	GeCo	Epipolar	SGC	VQ	MQ	TA	VBench-Q ↑	Dynamic (%)
Baseline	1.309	0.679	7.825	12.557	0.139	–	–	–	76.398	74.4
VideoGPA	1.281	0.665	7.541	12.525	0.134	63.95	56.67	46.15	76.497	74.4
Epipolar-DPO	1.268	0.676	7.789	**12.519**	0.132	56.41	51.28	49.36	75.998	69.2
CAVE-DPO	**1.259**	**0.663**	**7.322**	12.522	**0.122**	**69.23**	**64.10**	**49.59**	**76.571**	75.6

B.2 Quantitative Results

Table 6 reports the cross-benchmark results. CAVE-DPO ranks first on four of the five geometric criteria, achieving the lowest CAVE, 3DCS, GeCo-Fused, and SGC scores, while ranking second on Epipolar. The improvement therefore extends beyond the CAVE measure used to construct its preference data and is consistently recovered by multiple independent geometric evaluators.

CAVE-DPO also obtains the highest VideoReward preference rates among the three post-training methods across VQ, MQ, and TA. In particular, its VQ and MQ preference rates over Base reach 69.23% and 64.10%, respectively. It further achieves the highest VBench-Q and Dynamic Degree. Together, these results show that CAVE-based preference supervision improves geometric consistency while preserving the visual quality and motion required by video generation.

C Dynamic Generation after CAVE-DPO

CAVE-DPO is trained with geometric preferences derived primarily from scenes with persistent geometry. We further evaluate its behavior in generation with dynamic objects. To this end, we use PDI-Bench (Wu et al., 2026a), a public benchmark that evaluates dynamic objects in object-centric 3D space through scale–depth alignment, 3D motion consistency, and structural rigidity.

We use all 28 official generation prompts and generate videos with the CogVideoX-5B and CAVE-DPO under identical sampling configurations. Both sets are evaluated using the official PDI pipeline. We additionally report Dynamic Degree (Huang et al., 2024) on the same generated videos to measure whether changes in geometric consistency are accompanied by reduced video dynamics.

As shown in Table 7, CAVE-DPO reduces PDI from 0.651 to 0.591, demonstrating improved object-centric 3D consistency in videos with dynamic objects. Dynamic Degree remains unchanged at 82.1. These results indicate that the geometric benefits of CAVE-DPO also extend to dynamic objects.

The qualitative examples in Figure 6 provide a complementary scene-level view. Even when the generated videos contain clear object motion, the geometric improvements observed in our main experiments remain visible: compared with the base model, CAVE-DPO preserves more stable scene structures and reduces local geometric distortions while retaining the intended dynamics. Thus, the

Table 7: **Dynamic-scene evaluation on PDI-Bench.** CAVE-DPO improves object-level geometric consistency while preserving the overall degree of motion.

Method	PDI ↓	Dynamic Degree ↑
Base	0.651	82.1
CAVE-DPO	**0.591**	82.1

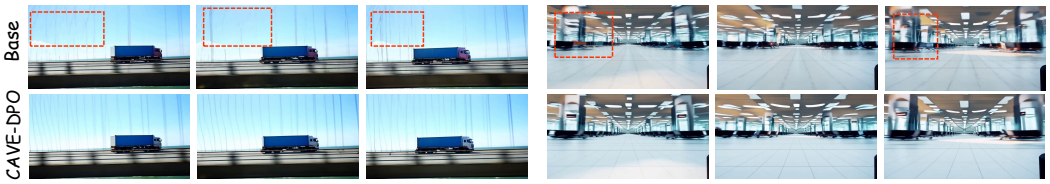

Figure 6: **Qualitative comparison on dynamic PDI-Bench prompts.** CAVE-DPO maintains more stable scene geometry than the base model in the presence of dynamic objects.

benefit of CAVE-DPO is not confined to static-scene generation; its scene-level geometric improvements remain observable in the presence of dynamic objects.

D Supplementary Details of CAVE

D.1 Valid Support and Rigid Alignment

Historical verification requires the pre-update and post-update queries to be compared against the same historical reference. This section specifies the valid support and alignment procedure used in this comparison.

For a historical view $h < t$, let $\mathbf{P}^{-}_{t,h}$, $\mathbf{P}^{+}_{t,h}$, and $\mathbf{P}^{\mathrm{ref}}_{h}$ denote the pre-update query, post-update query, and historical reference point maps, respectively. The raw confidence c predicted by CUT3R is converted to $[0, 1]$ as

$$\Gamma(c) = \mathrm{clip}\left(1 - \frac{1}{\max(c, 1)}, 0, 1\right).$$

The joint weight at location p is the geometric mean of the confidences from the three point maps:

$$w_{t,h,p} = \left(\Gamma^{-}_{t,h,p}\Gamma^{+}_{t,h,p}\Gamma^{\mathrm{ref}}_{h,p}\right)^{1/3}.$$

The valid support $\Omega_{t,h}$ consists of locations where all three point maps have finite coordinates and positive depth, and the joint weight is positive. No additional confidence threshold is applied. The pre-update and post-update query point maps are independently aligned to the historical reference through weighted rigid transformations:

$$(\mathbf{R}^{\eta}_{t,h}, \mathbf{b}^{\eta}_{t,h}) = \underset{\mathbf{R}\in\mathrm{SO}(3),\, \mathbf{b}\in\mathbb{R}^3}{\arg\min} \sum_{p\in\Omega_{t,h}} w_{t,h,p} \left\| \mathbf{R}\mathbf{P}^{\eta}_{t,h}[p] + \mathbf{b} - \mathbf{P}^{\mathrm{ref}}_{h}[p] \right\|_2^2, \qquad \eta \in \{-, +\}.$$

No scale transformation is estimated. Independent alignment removes coordinate-gauge differences between the two state queries while preserving their respective geometric errors relative to the fixed historical reference. A historical view is excluded from the valid set $\mathcal{H}_t$ when fewer than three jointly valid correspondences are available or the rigid alignment cannot be solved.

D.2 Contraction Consistency of Current-View Attribution

Using the current-view revision $\mathbf{r}_{t,p}$ defined in equation 10, let $\mathbf{C}_{t,p} = \mathbf{r}_{t,p}\mathbf{r}_{t,p}^\top$ be its energy tensor and $\mathbf{V}_{t,p} = \mathbf{d}_{t,p}\mathbf{d}_{t,p}^\top$ the projection tensor of the unit direction $\mathbf{d}_{t,p}$ of $\mathbf{R}_t[p]$. Their contraction gives the ray-aligned revision

$$\begin{aligned}\rho_{t,p} &= \sqrt{\mathbf{V}_{t,p} : \mathbf{C}_{t,p}} \\ &= \sqrt{\mathrm{tr}\big(\mathbf{d}_{t,p}\mathbf{d}_{t,p}^\top \mathbf{r}_{t,p}\mathbf{r}_{t,p}^\top\big)} = \left|\mathbf{d}_{t,p}^\top \mathbf{r}_{t,p}\right|.\end{aligned} \tag{14}$$

Let $\bar{w}_{t,p}$ be the geometric mean of the confidences from the pre-update query, post-update query, and current reference. The spatial carrier is $a_{t,p} = \bar{w}_{t,p}\rho_{t,p}$. Defining

$$\xi_t = \mathrm{RMS}_{\bar{\mathbf{w}}_t}(\mathbf{a}_t), \tag{15}$$

the attribution map for $\xi_t > 0$ satisfies

$$\begin{aligned}\mathbf{L}_t[p] &= s_t^{\mathrm{CAVE}} \frac{a_{t,p}}{\xi_t}, \\ \mathrm{RMS}_{\bar{\mathbf{w}}_t}(\mathbf{L}_t) &= \frac{s_t^{\mathrm{CAVE}}}{\xi_t}\,\mathrm{RMS}_{\bar{\mathbf{w}}_t}(\mathbf{a}_t) = s_t^{\mathrm{CAVE}}.\end{aligned} \tag{16}$$

Thus, current-view attribution redistributes the damage established by historical verification without changing its frame-level magnitude. Locations without common geometric support remain undefined. When $s_t^{\mathrm{CAVE}} = 0$, the response is zero over the valid support; when $s_t^{\mathrm{CAVE}} > 0$ but $\xi_t \le \epsilon_{64}$, where $\epsilon_{64} = 2^{-52} \approx 2.22 \times 10^{-16}$, no valid spatial localization is produced.

D.3 Computational Cost

For a T-frame video, exact CAVE performs $T(T-1)/2$ update–history verifications, giving $\mathcal{O}(T^2)$ pairwise geometric comparisons. Our implementation batches the world-state queries, evaluates independent alignment and residual computations with eight CPU workers, and contracts each dense response immediately rather than materializing the full $\mathcal{O}(T^2HW)$ damage tensor. On a 49-frame video, evaluating all update–history pairs with a single RTX 4090 takes 109 seconds and uses 5.02 GiB of peak GPU memory.

E Experimental Details

E.1 Controlled Benchmark and Evaluation Protocol

The controlled experiments follow the geometric perturbation protocol of GeCo and generate geometric changes with known spatial support on static, object-centric videos from CO3D. The resulting benchmark contains 100 sequences, each consisting of 20 consecutive frames. Foreground regions are determined using the object masks provided by CO3D. The crop is computed from the union of the foreground masks over the sequence and enlarged by 30% in each direction to retain scene context. A valid crop preserves at least 98% of the foreground, with the foreground area constrained to occupy between 0.08 and 0.85 of the image. The crop is further expanded to match the $1{:}1$ target aspect ratio before resizing to 512×512.

Geometric perturbations are generated using thin-plate spline transformations. Each sample uses 24 control points whose displacements evolve according to a first-order autoregressive process:

$$\boldsymbol{\delta}_t = 0.95\boldsymbol{\delta}_{t-1} + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, 0.6^2\mathbf{I}).$$

The TPS solver uses a regularization coefficient of 10^{-6}. The resulting displacement fields are exponentially smoothed with a coefficient of 0.8 and normalized according to their mean foreground displacement. The target displacement magnitude for each sample is drawn uniformly between 3 and 8 pixels.

The deformation is restricted by the foreground mask. The mask is eroded by 10 pixels to define the region receiving the full deformation, followed by a 20-pixel cosine transition band in which the displacement decays continuously toward the background. The final image is generated through backward coordinate mapping and bilinear interpolation.

Data filtering uses only detector-independent conditions based on foreground coverage and realized pixel changes, excluding samples in which the object is severely truncated or the applied perturbation is not visually present. All random processes are determined by a fixed global seed together with the sample identifier. Outputs from CAVE and the compared baselines are not used during data construction or filtering.

All methods are evaluated over the same temporal candidate interval. Exp1A selects the frame with the highest response, while Exp1B applies the same change-point rule to every method. The adaptation changes only how each method's native output is aggregated into a frame-level score or a dense spatial response; its underlying geometric computation remains unchanged.

E.2 Video Post-Training Protocol

Video post-training follows the diffusion DPO protocol of VideoGPA, using CogVideoX-5B as the base model. The candidate pool is generated from 3,147 prompts centered on scene structure and camera exploration. Three candidate videos are generated for each prompt using fixed random seeds. Each video contains 49 frames at a resolution of 720×480. All geometric preference signals operate on the same candidate videos and share the same motion-based filtering procedure.

VideoGPA, Epipolar, and CAVE independently perform validity checking and geometric scoring on the candidate videos. For each group containing at least two valid candidates, the candidate with the lowest geometric error is selected as the preferred sample and the candidate with the highest error as the dispreferred sample. No additional score-margin constraint is imposed. Because the three geometric measures have different validity conditions and rankings, their preference pairs are constructed independently rather than shared across methods.

All three DPO variants use the same base model, optimization budget, and training configuration; the geometric signal used to construct the preference pairs is the only experimental variable. Training uses LoRA (Hu et al., 2021) with rank 64 and scaling factor 128, a learning rate of 5×10^{-6}, and a DPO coefficient of 1.0. Each model is trained for 2,000 optimization steps. Training is performed on 8 GPUs with a per-device batch size of 1 and two gradient-accumulation steps, resulting in a global batch size of 16. All remaining optimizer and diffusion-training settings follow VideoGPA.

Table 8: Adaptation of each method to temporal detection and spatial localization in Exp1. All methods retain their original geometric computations.

Method	Frame-level response	Spatial response in Exp1C
MEt3R	Pairs the target frame with every other frame and averages the pairwise scores.	Averages the native dense response maps over valid frame pairs.
VGGT-Reproj.	Jointly reconstructs the complete sequence and measures the reprojection error at each target view.	Uses the per-pixel reprojection error in the target view.
Epipolar	Aggregates the normalized Sampson errors between the target frame and its preceding and following frames.	Not evaluated because its response is defined only at sparse correspondences.
GeCo-Motion	Applies the original sequence window and spatially averages the aggregated motion-residual map.	Uses the native dense motion-residual map.
GeCo-Fused	Applies the same sequence window and spatially averages the fused motion and depth response.	Uses the native dense fused response map.
CAVE	Contracts historical damage over space and valid preceding views.	Uses the score-calibrated current-view attribution map.

F Additional Qualitative Results

The quantitative results establish the aggregate behavior of the different preference signals. We complement them with additional frame-level comparisons and VGGT-based 3D reconstructions. For every case, all methods use the same prompt and sampling configuration. Frames are sampled at identical temporal positions, while the reconstructed point clouds are rendered using the same VGGT protocol, filtering parameters, and camera viewpoint across methods.

F.1 Qualitative Results on DL3DV

Figure 7 presents additional qualitative comparisons on the DL3DV evaluation set.

F.2 Qualitative Results on GeCo-Eval

Figure 8 presents additional qualitative comparisons on the GeCo-Eval prompt set.



Figure 7: **Additional qualitative results on DL3DV.** Each case compares sampled video frames and the corresponding VGGT reconstruction across the four methods.

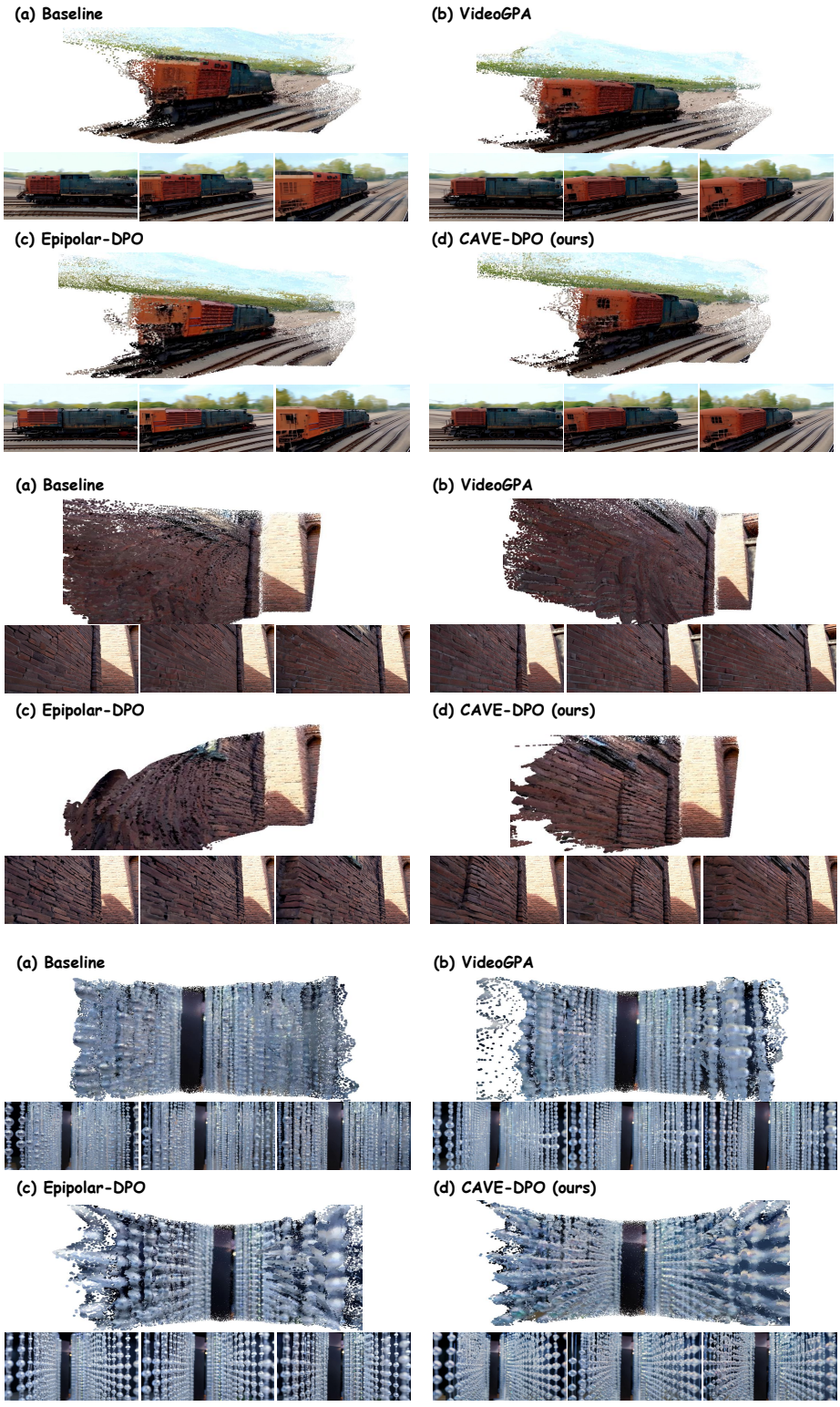

Figure 8: **Additional qualitative results on GeCo-Eval.** Each case compares sampled video frames and the corresponding VGGT reconstruction across the four methods.