GeCo: Evaluating Geometric Consistency for Video Generation via Motion and Structure

Leslie Gu[1] Junhwa Hur[2] Charles Herrmann[2] Fangneng Zhan[3]
Todd Zickler[1] Deqing Sun[2] Hanspeter Pfister[1]

[1]Harvard University [2]Google DeepMind [3]MIT

https://GeCo-GeoConsistency.github.io

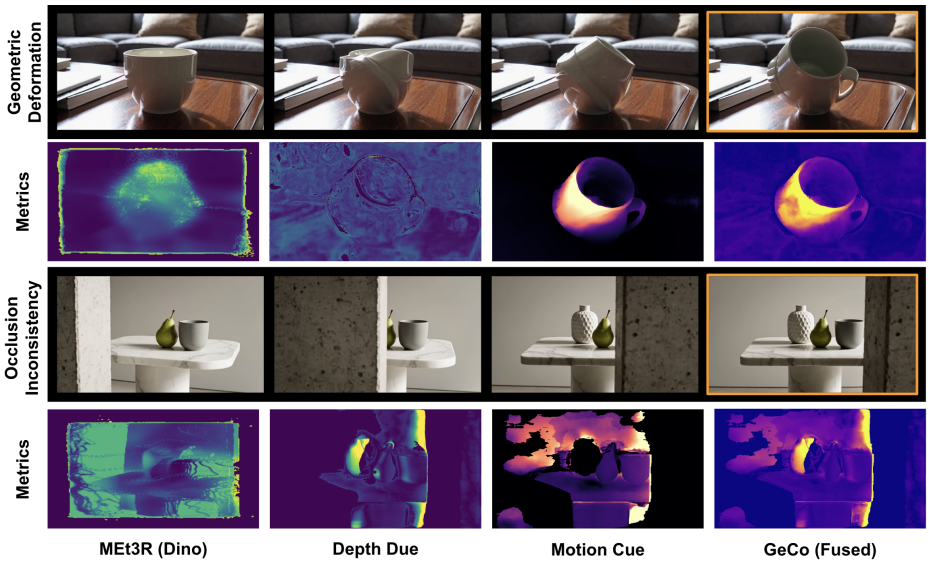


Fig. 1: Artifact videos: a cup undergoes geometric deformation; a vase is hallucinated behind the pear after occlusion; orange boxes mark the target frames. **Comparison of evaluation methods**: MEt3R [2] fails to pinpoint geometric deformation and gives a vague signal on occlusion inconsistency. Depth cue captures occlusion inconsistency but does not localize surface deformation. Motion cue highlights non-rigid deformation artifacts, but provides no signal in occluded regions and thus misses occlusion inconsistency artifacts. GeCo (Fused) integrates motion and depth cues to accurately localize both types of artifacts in a single interpretable map.

Abstract. Visual generation models can produce photorealistic videos yet violate multiview geometry, exhibiting non-rigid deformations and occlusion inconsistencies (e.g., hallucinated content in disoccluded regions). These failures hinder downstream applications such as video world model development and 3D asset creation, and are poorly captured by existing metrics. We introduce GeCo, a geometry-grounded metric for jointly detecting geometric deformation and occlusion-inconsistency artifacts in static scenes. By fusing residual motion and depth priors, GeCo produces interpretable, dense consistency maps that localize these artifacts. Using GeCo, we systematically benchmark recent video generation

models, revealing common geometric failure modes. We further show that GeCo provides an actionable signal by applying it as a training-free guidance loss that substantially reduces geometric artifacts during generation.

Keywords: geometric consistency · video generation · static scenes

1 Introduction

Recent generative video models have achieved remarkable photorealism [6,12,26, 51] and are increasingly used in applications that require geometric consistency across viewpoints, such as video world models [17,20,24,32,40,57] and 3D asset creation [1,14,42,44]. However, even when genearting nominally static scenes, these models frequently produce perceptually salient artifacts that violate multiview geometry.

We identify two primary, complementary geometric failure modes in static scenes: (i) Geometric deformation: co-visible rigid surfaces stretch, bend, or "melt" under camera motion. (ii) Occlusion inconsistency: previously occluded objects reappear altered, or disoccluded regions suffer from hallucinatory content shifts, violating object permanence. Because these artifacts arise in different visibility regimes—co-visible surfaces versus newly revealed regions—reliable geometric evaluation must detect both jointly.

Yet existing metrics capture only part of this picture. General video benchmarks [22,23] lack explicit 3D geometric evaluation. Epipolar constraints [1,53] and reprojection-based world-model suites [10] rely on sparse point matching, which inherently fails in occluded regions [7]—precisely where occlusion inconsistency manifests. Conversely, dense multi-view metrics based on 3D warped semantic features [2] expose disocclusion errors but lack the geometric precision to penalize subtle surface deformations. As summarized in Table 1, no existing method offers a dense, interpretable diagnostic for both artifact types.

To bridge this gap, we propose GeCo, a dense, geometry-grounded metric that jointly localizes geometric deformation and occlusion inconsistency. GeCo leverages feed-forward estimators for optical flow, depth, and camera pose to compute two complementary signals: a *motion cue* measuring flow-based rigidity (for deformation), and a *structure cue* tracking depth reprojection consistency (for occlusion errors). GeCo fuses the cues into a unified, scale-invariant, per-pixel error map. Crucially, because all components are differentiable, GeCo serves as both an interpretable evaluation metric and an inference-time guidance signal.

We validate GeCo using two controlled datasets: WarpBench (thin-plate-spline warps isolating geometric deformation) and OccluBench (composited edits simulating occlusion inconsistencies). Experiments confirm our motion and structure cues are highly complementary, robust to off-the-shelf estimator noise, and their fusion substantially outperforms semantic baselines [2]. We demonstrate GeCo's versatility in two scenarios: first, as a metric benchmarking state-of-the-art video models across diverse static scenes (*e.g.*, object-centric, indoor, outdoor), revealing consistent geometric failure patterns; second, as a training-free

Table 1: Comparison of GeCo and existing metrics. We factor metrics into a two-stage pipeline: (i) correspondence matching and (ii) cue comparison. We report whether each method yields a dense, interpretable consistency map and whether it is diagnostic for subtle geometric deformation and occlusion inconsistency artifacts.

Method	Correspondence	Cue	Dense	Deform.	Occlusion
TSED [53]	SIFT [30]	Symmetric epipolar distance [16]	✗	✓	✗
Cosmos [1]	SuperPoint [9] & LightGlue [27]	Sampson error [39]	✗	✓	✗
VBench [22, 23]	None (frame-wise)	CLIP feature [35]	✗	✗	✗
MEt3R [2]	3D warping (DUSt3R [46])	DINO feature [7]	✓	✗	✓
WorldScore [10]	Learnt flow (DROID-SLAM [41])	Reprojection error	✓	✓	✗
GeCo - Motion	Optical flow (UFM [58])	Residual motion	✓	✓	✗
GeCo - Structure	3D warping (VGGT [45])	Reprojection error	✓	✗	✓
GeCo - Fused	Flow & 3D warping	Fused	✓	✓	✓

inference guidance objective that suppresses spurious motion and significantly improves both geometric consistency and downstream 3D reconstruction.

In summary, our contributions are:

- GeCo: A differentiable, dense metric that jointly detects geometric deformation and occlusion inconsistency, producing interpretable error maps.
- GeCo-Eval: A geometric consistency evaluation suite for static scenes to benchmark video generation models, along with a systematic comparison of state-of-the-art models.
- Inference-Time Guidance: A training-free inference-time optimization that uses GeCo as a guidance objective during sampling to reduce geometric artifacts and improve consistency.
- Failure-Mode Analysis: An analysis of spurious motion artifacts in recent text-to-video models, showing that GeCo-based guidance reduces erroneous motion in otherwise static regions.

2 Related Works

2.1 Video Generation Benchmarks

Comprehensive benchmarks such as VBench [22, 23] and EvalCrafter [29] evaluate video generation along perceptual axes (e.g., temporal coherence, text–video alignment) but do not explicitly target multi-view geometric consistency. Another line of work measures physical plausibility [4, 33, 49], focusing on realistic dynamics rather than static-scene geometry under camera motion. WorldScore [10] includes a depth-reprojection-based 3D-consistency score, but remains limited under occlusion and therefore less sensitive to occlusion inconsistency. To fill this gap, we target geometric consistency in static scenes, evaluating both deformation and occlusion inconsistency.

2.2 Geometric Consistency Metrics

Most geometric consistency metrics follow a two-stage pipeline: (i) establish inter-frame correspondences and (ii) compare cues on the matched content. Ta-

ble 1 summarizes representative designs in terms of dense interpretability, deformation sensitivity, and occlusion-inconsistency awareness.

Correspondence matching. Correspondences are typically obtained via (a) sparse keypoints [1,53], (b) dense flow [10], or (c) 3D reconstruction-based warping [2] (*i.e.*, reprojecting co-visible 3D points into the target view). Keypoint- and flow-based pipelines are inherently undefined in occlusion regions where no reliable matches exist; these regions are therefore commonly masked, directly ignoring occlusion inconsistency artifacts. Reconstruction-based warping is visibility-aware: it can identify co-visible pixels across frames, enabling comparison in disoccluded regions where hallucinated content may appear.

Cue comparison. Given correspondences, methods compare either semantic features or geometric errors. Semantic features (*e.g.*, DINO [7], CLIP [35]) may under-penalize subtle geometric distortions when semantics are preserved. Epipolar-constraint cues (*e.g.*, symmetric epipolar distance [16], Sampson error [39]) are less sensitive when epipolar lines have similar orientations, as the epipolar error is insensitive to correspondence errors parallel to the epipolar line [53]. Depth cues depend on the accuracy of the underlying depth model and, when paired with 3D warping, may under-penalize deformations that remain close in 3D (*i.e.*, small depth residuals despite incorrect correspondences).

Representative methods. TSED [53] uses SIFT [30] with thresholded symmetric epipolar distance [16]; Cosmos (Geometric Consistency) [1] uses SuperPoint [9] and LightGlue [27] with Sampson error [39]; and VBench (Background Consistency) [22,23] compares global CLIP features [35] without explicit correspondence. MEt3R [2] uses DUSt3R [46] warping with DINO features [7], while WorldScore (3D Consistency) [10] uses DROID-SLAM [41] to compute dense reprojection error from learned-flow correspondences. Our method fuses dense flow [58] and 3D warping [45] to produce dense, interpretable maps that capture both subtle deformation and occlusion inconsistency within a unified metric.

2.3 Improving Geometric Consistency in Video Generation

Training-based approaches improve 3D consistency through joint video-geometry estimation [21,55] or explicit 3D caches [38], but require large-scale 3D annotations (depth, pointmaps, camera poses) obtained from off-the-shelf estimators [5,41,46,50], which are often costly to prepare. Orthogonally, training-free approaches steer diffusion sampling at inference time without updating model parameters, using guided sampling [3,15], structural control [54,56], or flow-based objectives [11,34] to impose motion or structural consistency. In this work, we apply GeCo as training-free guidance at inference time, showcasing that it provides an effective signal for improving geometric consistency during sampling.

3 GeCo: A Differentiable Geometric Consistency Metric

We introduce GeCo, a differentiable metric for geometric consistency in generated videos and apply it as guidance to improve 3D consistency for generation.

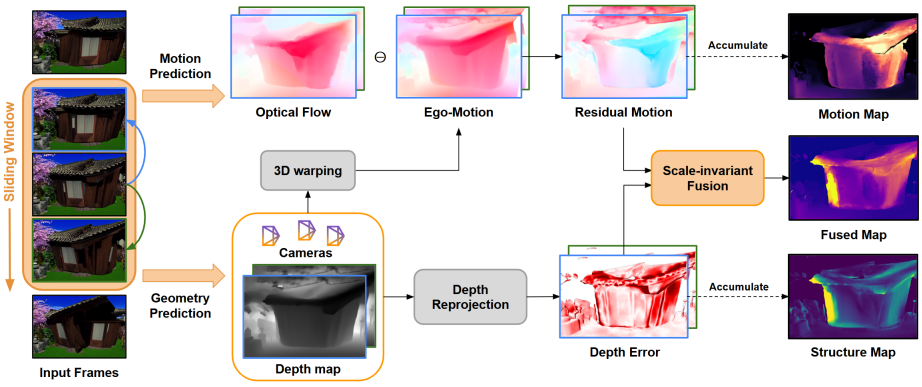


Fig. 2: GeCo pipeline. Within a sliding window, we jointly estimate dense optical flow and 3D geometry (depth and camera pose) for frame pairs. We compute residual motion and depth errors and fuse them into scale-invariant inconsistency maps. Aggregation over the window localizes artifacts in the target frame, while motion and structure maps provide complementary diagnostics.

3.1 Measuring Geometric Consistency with GeCo

To quantify geometric consistency in static scenes under camera motion, GeCo uses dense per-pixel correspondences and two complementary cues. Motion consistency measures multi-view rigidity via the residual between optical flow and camera-induced rigid flow in co-visible regions (*i.e.*, geometric deformation). Structure consistency measures object permanence when visibility changes via depth reprojection error, capturing occlusion inconsistency where flow is unreliable. Both cues are then normalized via perspective geometry and fused into a unified, scale-invariant per-pixel inconsistency map.

Fig. 2 shows the pipeline for metric computation. GeCo processes an input video in a sliding window fashion. Given N frames in a window, we compute the metric for the center frame $\mathbf{I}_c$. For each pair $(\mathbf{I}_c, \mathbf{I}_i)$ with $i \in \{1, \dots, N\} \setminus \{c\}$, it computes two consistency metrics, motion consistency and structure consistency, and fuses them into a per-pixel error map for the center frame. This per-pixel map can localize where the deformation occurs in each image.

Motion consistency. Motion consistency measures the residual between optical flow and rigid motion induced by camera in the pixel space. Given two frames ($\mathbf{I}_c$ and $\mathbf{I}_i$), dense optical flow $\mathbf{F}_{\text{flow}}^{c\to i}$ is obtained by an off-the-shelf model [58]. For camera-induced motion, we use a geometry foundation model [45] to predict depth map $\mathbf{D}_c$, camera intrinsics $\mathbf{K}_c$ (with focal lengths f_x, f_y), and relative pose $\mathbf{P}_{c\to i}$=($\mathbf{R}, \mathbf{T}$). Then we get residual motion by subtracting the rigid motion from the optical flow:

$$\mathbf{F}_{\text{residual}}(\mathbf{p}) = \mathbf{F}_{\text{flow}}^{c\to i}(\mathbf{p}) - \mathbf{F}_{\text{rigid}}(\mathbf{p}) \tag{1}$$

$$\mathbf{F}_{\text{rigid}}(\mathbf{p}) = \pi_{\mathbf{K}_i}\big(\mathbf{R}\,[\mathbf{D}_c(\mathbf{p})\,\mathbf{K}_c^{-1}\tilde{\mathbf{p}}] + \mathbf{T}\big) - \mathbf{p}, \tag{2}$$

with $\tilde{\mathbf{p}}$ being the homogeneous coordinate of pixel $\mathbf{p}$ and perspective projection $\pi_{\mathbf{K}}(\cdot)$ with intrinsics $\mathbf{K}$. The residual motion is computed only for co-visible

pixels between the center frame and the rest frames, discarding occluded regions where optical flow is not reliable.

Structure consistency. Structure consistency evaluates geometric coherence to handle regions where optical flow is unreliable (*e.g.* occlusions). Given two frames $\mathbf{I}_c$ and $\mathbf{I}_i$, we reproject the depth map $\mathbf{D}_i$ into the viewpoint of $\mathbf{I}_c$ using the estimated relative pose $\mathbf{P}_{i\to c}$ and intrinsics $\mathbf{K}_i$. It yields a warped depth map $\mathbf{D}_{i\to c}$, utilizing z-buffering to resolve self-occlusions. The structure consistency at pixel $\mathbf{p}$ is defined as the residual between the predicted and reprojected depth:

$$\Delta z(\mathbf{p}) = \mathbf{D}_c(\mathbf{p}) - \mathbf{D}_{i\to c}(\mathbf{p}). \tag{3}$$

Scale-invariant fusion. Direct fusion of the two signals is non-trivial due to unit mismatch: motion consistency is in pixels, while structure consistency is in depth unit up to scale. To combine them, we normalize both signals into a dimensionless, scale-invariant space via the pinhole camera model. Let $(\Delta u, \Delta v) = \mathbf{F}_{\text{residual}}(\mathbf{p})$ be the flow residual and $\Delta z(\mathbf{p})$ be the depth residual. We define the normalized motion consistency $\mathbf{m}(\mathbf{p})$ and structure consistency $\mathbf{s}(\mathbf{p})$ as:

$$\mathbf{m}(\mathbf{p}) = \sqrt{\left(\frac{\Delta u}{f_x}\right)^2 + \left(\frac{\Delta v}{f_y}\right)^2}, \quad \mathbf{s}(\mathbf{p}) = \left|\frac{\Delta z(\mathbf{p})}{\mathbf{D}_c(\mathbf{p})}\right|. \tag{4}$$

We fuse these cues into a unified error map $\mathbf{M}_{\text{geo}}$ for the center frame $\mathbf{I}_c$:

$$\mathbf{M}_{\text{geo}}(\mathbf{p}) = \sqrt{\big(\mathbb{1}_{\text{cov}}(\mathbf{p}) \cdot \mathbf{m}(\mathbf{p})\big)^2 + \mathbf{s}(\mathbf{p})^2}, \tag{5}$$

where $\mathbb{1}_{\text{cov}}$ is the co-visibility mask. This allows structure consistency to dominate in occluded regions (where $\mathbb{1}_{\text{cov}}$=0), while utilizing both metrics in co-visible areas to robustly localize deformation.

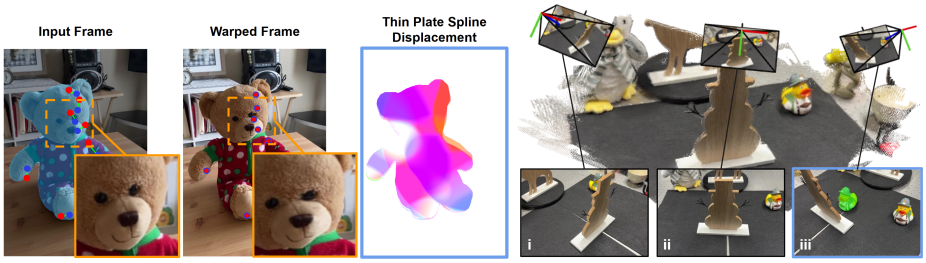
Temporal aggregation. For each center frame $\mathbf{I}_c$ in a video, we average the pairwise error maps within its sliding window to obtain aggregated frame-level error maps $\{\mathbf{m}_c, \mathbf{s}_c, \mathbf{M}_{\text{geo},c}\}$. We then compute the spatial mean of each map to define scalar frame-level scores. Finally, video-level metrics are obtained by averaging these frame-level scores over all frames in the sequence, yielding three per-video scalars, which we denote by $\mathcal{M}_{\text{motion}}$, $\mathcal{M}_{\text{struct}}$, and $\mathcal{M}_{\text{geo}}$ for motion, structure, and fused geometric consistency, respectively.

3.2 Training-free Guidance with GeCo

As all components are differentiable, GeCo can serve as a guidance term to improve geometric consistency during video generation, without requiring model fine-tuning. Building on CogVideoX-5B [51] and Frame Guidance [25], we update the latent z_t at timestep t by minimizing our GeCo metric during sampling time:

$$\mathbf{z}_t \;\leftarrow\; \mathbf{z}_t - \eta_t \nabla_{\mathbf{z}_t} \mathcal{L}_{\text{geo}}\big(\hat{\mathbf{x}}_{0|t}^{\mathcal{I}}\big), \tag{6}$$

where η_t is the guidance scale and $\hat{\mathbf{x}}_{0|t}^{\mathcal{I}}$ represents the approximated clean frames. The loss $\mathcal{L}_{\text{geo}}$ aggregates the error map $\mathbf{M}_{\text{geo}}$ over a subset of frames $\mathcal{I}$ and



(a) WarpBench deformation process **(b)** OccluBench sequence

Fig. 3: Validation datasets. (a) WarpBench: (Left) Input frame with segmentation mask (cyan), sampled thin-plate spline (TPS) control points (red), and destination points (blue). (Middle) Warped frame after TPS deformation. (Right) Ground-truth displacement field from the deformation. **(b) OccluBench**: An example sequence where a region in the image center is (i) visible and empty, (ii) occluded, and (iii) re-revealed with a new object, forming a controlled occlusion inconsistency artifact.

temporal offsets $\mathcal{K}$:

$$\mathcal{L}_{\text{geo}}(\hat{\mathbf{x}}^{\mathcal{I}}) = \frac{1}{|\mathcal{I}||\mathcal{K}|} \sum_{c\in\mathcal{I}} \sum_{k\in\mathcal{K}} \sum_{\mathbf{p}\in\Omega} \mathbf{M}_{\text{geo}}^{(c,k)}(\mathbf{p}), \tag{7}$$

where $\mathcal{I}$ denotes the subset of guidance frames, $\mathcal{K}$ defines the temporal offsets relative to each center frame $c \in \mathcal{I}$, and Ω represents the pixel domain.

As our metric requires clean input images, we estimate the clean latent $\hat{\mathbf{z}}_{0|t}$ from the current noisy state $\mathbf{z}_t$ and predicted velocity $\mathbf{v}_\theta$, then decode it via the decoder $D(\cdot)$:

$$\hat{\mathbf{z}}_{0|t} = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\,\mathbf{v}_\theta(\mathbf{z}_t, t) \tag{8}$$

$$\hat{\mathbf{x}}_{0|t}^{\mathcal{I}} = D(\hat{\mathbf{z}}_{0|t})^{\mathcal{I}}, \tag{9}$$

Gradients are backpropagated through the frozen decoder and the GeCo pipeline to update $\mathbf{z}_t$, improving the geometric consistency (*i.e.*, lowering $\mathcal{L}_{\text{geo}}$) throughout the sampling process. All pretrained backbones used to compute depth [45] and flow [58] are frozen.

To ensure efficiency and stability during sampling, we employ three strategies:

- Latent slicing [25] to decode only a small temporal window $\mathcal{I}$ for guidance $\mathcal{L}_{\text{geo}}$ computation, instead of using the full sequence.
- Recursive denoising [3,31,48] to repeat updates R times per step to improve convergence
- Time-travel [18] to re-noise the latent after updates, to mitigate accumulated sampling error.

4 GeCo Validation

We validate GeCo in two steps. First, using controlled datasets to isolate specific artifacts, we demonstrate that its motion and structure cues are complementary

and their fusion outperforms baselines (§4.1–4.2). Second, we confirm the robustness of GeCo against errors from its underlying off-the-shelf estimators (§4.3). Detailed setups and qualitative visualizations are in the supplementary material.

4.1 WarpBench for Geometric Deformation

WarpBench development. We construct WarpBench using clips from CO3D [37] (*CO3D-Warp*, object-centric) and ScanNet++ [8, 52] (*ScanNet-Warp*, indoor scenes). To simulate non-rigid deformation artifacts, we inject temporally smooth thin-plate-spline (TPS) warps into foreground regions (Fig. 3a), producing ground-truth dense displacement fields. WarpBench contains 200 clips (4,000 frames).
Task1: Frame-level anomaly detection. We randomly replace a single frame in a clip with its warped version, and the goal is to identify the anomalous frame. For each clip, we treat each metric as a frame-wise anomaly score and predict the anomalous frame as the one with the largest score in the clip (*i.e.*, $\hat{t} = \arg\max_t \mathcal{M}(t)$). Tab. 2 shows that motion cue is most reliable on CO3D-Warp and fused cue on ScanNet-Warp. The structure cue is informative for scenes but noisier for object-centric data. MEt3R performs poorly, underscoring the difficulty of detecting subtle deformations with semantic features. WorldScore and TSED improve over MEt3R but trail our cues by a wide margin. WorldScore and TSED improve over MEt3R but trail our cues by a wide margin. WorldScore measures reprojection error, performing well on scene-centric ScanNet-Warp (81.54%) but insensitive to subtle object-level deformation in CO3D-Warp (32.95%). TSED relies on sparse keypoint matching, which lacks dense surface coverage, and performs poorly on both subsets.

Table 2: WarpBench: anomaly detection accuracy in %. Higher is better. Both of our motion and fused metric shows their competitiveness on both benchmark datasets while MEt3R performs poorly.

Method	CO3D-Warp	ScanNet-Warp
MEt3R [2]	6.82	15.38
WorldScore [10]	32.95	81.54
TSED [53]	36.78	38.46
Structure $\mathcal{M}_{\text{struct}}$	42.05	84.62
Motion $\mathcal{M}_{\text{motion}}$	**71.59**	89.23
Fused $\mathcal{M}_{\text{geo}}$	55.68	**92.31**

Task 2: Deformation localization. Given a real and warped frame pair, each metric outputs an inconsistency map, which we compare against the ground-truth deformation mask. We report AP (%), IoU (%), and Spearman's rank correlation coefficient (SRCC) (See supplementary for details) to measure detection quality, mask overlap, and rank correlation with the true deformation magnitude. As shown in Tab. 3 (WarpBench columns), motion is the strongest cue, achieving the highest AP and IoU. MEt3R yields negative SRCC, indicating anti-correlation with deformation magnitude. WorldScore attains positive SRCC and reasonable AP but markedly lower IoU than motion cue (29.56% vs. 44.48% on CO3D-Warp). Fusing depth due does not surpass motion cue alone for pure deformations but remains competitive.

Table 3: Combined results on WarpBench and OccluBench. We validate our metrics on both benchmarks, WarpBench for deformation and OccluBench for occlusion inconsistency. TSED relies on sparse correspondence without dense prediction, so we report it on the detection task only. Higher is better for all metrics.

Method	WarpBench						OccluBench		
	CO3D-Warp			ScanNet-Warp			Inconsistency Segmentation		
	AP (%) ↑	IoU (%) ↑	SRCC ↑	AP (%) ↑	IoU (%) ↑	SRCC ↑	AP (%) ↑	IoU (%) ↑	F1 (%) ↑
MEt3R [2]	16.26	15.95	-0.176	30.34	33.13	-0.351	46.51	33.97	48.57
WorldScore [10]	48.70	29.56	0.253	77.51	32.45	0.332	5.36	5.77	10.61
Structure $\mathcal{M}_{\text{struct}}$	25.64	3.20	0.079	51.71	14.11	0.183	62.36	51.46	66.96
Motion $\mathcal{M}_{\text{motion}}$	**64.90**	**44.48**	**0.581**	**87.12**	**52.36**	**0.706**	4.43	5.85	8.62
Fused $\mathcal{M}_{\text{geo}}$	60.63	41.69	0.554	82.70	48.87	0.547	**83.48**	**69.83**	**81.74**

Table 4: Robustness evaluation. GeCo consistency scores on artifact-free videos. Fused score divergence between predicted (Pred) and ground-truth (GT) geometry is marginal ($\Delta \approx 0.008$). Crucially, the noise floor on both synthetic and real data is an order of magnitude below generated-video scores (Tab. 5).

Data	**Motion↓**		**Structure↓**		**Fused↓**		**Mean Motion (%/s)**
	Pred	**GT**	**Pred**	**GT**	**Pred**	**GT**	
TartanAir V2 Easy	0.0060	0.0009	0.0075	0.0038	0.0066	0.0023	32.09±21.76
TartanAir V2 Hard	0.0110	0.0010	0.0088	0.0037	0.0109	0.0024	64.48±38.09
DL3DV	0.0090	–	0.0172	–	0.0223	–	121.63±80.56

4.2 OccluBench for Occlusion Inconsistency

OccluBench development. We construct OccluBench to isolate occlusion inconsistency artifacts: each clip contains a region that is (i) initially visible and empty, (ii) temporarily occluded, and (iii) re-revealed with a new, inconsistent object (Fig. 3b). The dataset comprises 30 clips (5,165 frames), with ground-truth artifact masks for 1,654 re-revealed frames annotated using SAM 2 [36].
Task 3: Inconsistent object localization. We evaluate artifact localization in re-revealed frames, reporting AP (%), IoU (%), and F1 (%), where AP is threshold-free and implemented as ranking average precision, IoU and F1 are computed at the per-image best threshold, then macro-averaged across N=60 anchors. Results in Tab. 3 (OccluBench columns) confirm the design: motion cue and WorldScore fail almost entirely, as flow cannot track content across occlusion boundaries. In contrast, MEt3R performs far better here than on deformation, while the structure cue effectively localizes artifacts and the fused cue achieves the best overall performance, validating the necessity of fusing both cues.

4.3 GeCo Robustness Evaluation

As GeCo builds on off-the-shelf estimators for optical flow, depth, and camera pose, we evaluate the system's robustness by (i) comparing scores computed from predicted versus ground-truth geometry, (ii) measuring the noise floor on real videos, and (iii) stress-testing sensitivity to controlled noise injection.

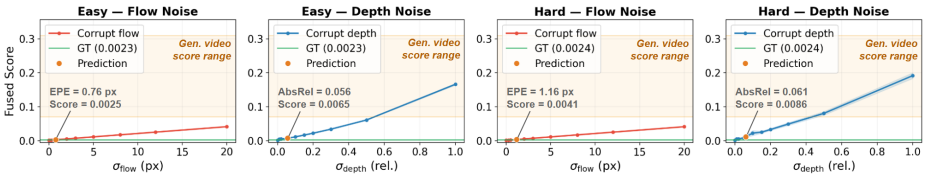


Fig. 4: Sensitivity to input noise. Gaussian noise is added to ground-truth flow (σ_{flow}, additive, in pixels) and depth (σ_{depth}, multiplicative, in log-space) independently on TartanAir. Orange dots mark the prediction operating point. The score remains near zero well beyond the estimator error level, rising only at substantially larger noise.

Component evaluation. We compare GeCo scores computed from predicted versus ground-truth geometry on TartanAir v2 [47] Easy and Hard splits (Tab. 4). Evaluating the components directly, the underlying estimators incur 0.757/1.155 flow EPE (px), 0.056/0.061 depth AbsRel, and 0.513/0.748° pose angular error on Easy/Hard, respectively. Despite these degraded predictions on the Hard split, the fused score divergence between predicted and ground-truth remains marginal ($\Delta \approx 0.004/0.008$). The predicted fused score ($\sim$0.007/0.011) establishes a practical synthetic noise floor.

Noise floor on real videos. We further run GeCo on DL3DV [28] (Tab. 4), a dataset of real videos with predominantly rigid camera motion, yielding a predicted fused score of 0.0223. Because real-world data inherently contains more physical noise (e.g., motion blur, complex lighting) than synthetic environments, it is reasonable that this score is slightly higher than the synthetic baseline. Crucially, this real-world noise floor remains an order of magnitude below typical generated-video levels (0.1–0.3 in Tab. 5). This confirms that elevated scores reflect genuine model-induced geometric failures rather than underlying estimator noise.

Sensitivity to input noise. We further stress-test GeCo by injecting controlled Gaussian noise into ground-truth flow and depth independently on TartanAir v2 (Fig. 4). Flow is perturbed additively in pixel space (σ_{flow}), and depth multiplicatively in log-space (σ_{depth}) so that errors scale with distance. The metric is particularly robust to flow noise: even at $\sigma_{\text{flow}} = 20$ px—over 26×/17× the prediction EPE on Easy/Hard—the fused score remains below the generated-video range. For depth, the score stays below the generated-video range up to $\sigma_{\text{depth}} \approx 0.4/0.5$ on Easy/Hard (7–8× the prediction AbsRel). This wide tolerance margin confirms that GeCo is not brittle to backbone imperfections.

5 Experiments

We use GeCo in two settings: as an evaluation metric to benchmark state-of-the-art video generation models, and as a training-free guidance to reduce geometric artifacts during inference time. Detailed experimental setups and implementation details are provided in the supplementary material.

Table 5: GeCo-Eval across scenarios. Per-model motion inconsistency (Mot.), structure inconsistency (Str.), and fused GeCo scores (lower is better) together with Total Motion (TM, %) and Mean Motion (MM, %/s; mean±std over clips). Commercial models (top block) and open-source models (bottom block) are compared across four scenarios. Darker green cells indicate the best scores among commercial models, lighter green cells indicate the best scores among open-source models, and red cells highlight models with very low TM and MM, where a high consistency score may be partially achieved by under-shooting motion (near-static generations).

(a) Object-centric

Model	Mot. ↓	Str. ↓	Fused ↓	TM (%)	MM (%/s)
Sora 2	0.016	**0.063**	**0.069**	24.97 ± 42.68	6.29 ± 10.76
Veo 3.1	**0.012**	0.087	0.090	52.48 ± 34.84	6.59 ± 4.38
WAN 2.2	0.027	0.131	0.140	87.11 ± 91.10	17.42 ± 18.22
HunyuanVideo	0.014	0.333	0.336	34.29 ± 35.06	6.43 ± 6.57
CogVideoX1.5	0.008	0.054	0.056	9.16 ± 11.55	1.83 ± 2.31
CogVideoX-5B	0.027	0.073	0.080	59.81 ± 77.73	19.94 ± 25.91
CogVideoX-2B	0.018	0.053	0.059	30.57 ± 34.77	10.19 ± 11.59
LTX-Video	**0.005**	**0.024**	**0.026**	13.01 ± 20.07	3.25 ± 5.02

(b) Indoor

Model	Mot. ↓	Str. ↓	Fused ↓	TM (%)	MM (%/s)
Sora 2	**0.026**	**0.162**	**0.169**	71.12 ± 123.63	17.93 ± 31.17
Veo 3.1	0.030	0.223	0.233	136.06 ± 113.56	17.10 ± 14.27
WAN 2.2	0.039	0.269	0.279	106.79 ± 111.28	21.36 ± 22.26
HunyuanVideo	0.027	0.274	0.281	122.06 ± 139.96	22.89 ± 26.24
CogVideoX1.5	0.010	0.115	0.117	14.22 ± 14.12	2.84 ± 2.82
CogVideoX-5B	0.025	0.128	0.136	57.14 ± 68.53	19.05 ± 22.84
CogVideoX-2B	0.022	0.117	0.124	47.14 ± 43.65	15.71 ± 14.55
LTX-Video	**0.009**	**0.070**	**0.073**	21.27 ± 33.21	5.32 ± 8.30

(c) Outdoor

Model	Mot. ↓	Str. ↓	Fused ↓	TM (%)	MM (%/s)
Sora 2	0.032	0.258	0.265	63.62 ± 97.34	16.04 ± 24.54
Veo 3.1	**0.026**	**0.209**	**0.217**	161.61 ± 137.62	20.31 ± 17.29
WAN 2.2	0.054	0.423	0.437	159.98 ± 179.82	32.00 ± 35.96
HunyuanVideo	0.028	0.232	0.240	118.64 ± 150.64	22.25 ± 28.24
CogVideoX1.5	**0.009**	**0.043**	**0.047**	24.44 ± 35.76	4.89 ± 7.15
CogVideoX-5B	0.032	0.172	0.180	81.52 ± 86.97	27.17 ± 28.99
CogVideoX-2B	0.015	0.063	0.068	44.15 ± 39.60	14.72 ± 13.20
LTX-Video	0.011	0.068	0.071	27.32 ± 37.89	6.83 ± 9.47

(d) Appearance-complex

Model	Mot. ↓	Str. ↓	Fused ↓	TM (%)	MM (%/s)
Sora 2	0.035	**0.195**	**0.206**	56.58 ± 90.84	14.26 ± 22.90
Veo 3.1	**0.030**	0.200	0.210	90.61 ± 80.47	11.39 ± 10.11
WAN 2.2	0.047	0.193	0.212	104.36 ± 123.70	20.87 ± 24.74
HunyuanVideo	0.031	0.180	0.193	53.65 ± 58.44	10.06 ± 10.96
CogVideoX1.5	0.012	0.176	0.181	11.28 ± 10.04	2.26 ± 2.01
CogVideoX-5B	0.041	0.110	0.126	53.35 ± 54.24	17.78 ± 18.08
CogVideoX-2B	0.024	0.080	0.091	39.47 ± 38.52	13.16 ± 12.84
LTX-Video	**0.007**	**0.039**	**0.041**	13.51 ± 29.38	3.38 ± 7.34

5.1 GeCo-Eval: Benchmark Suite for T2V Models

Benchmark design. We first evaluate the geometric consistency of text-to-video models by developing GeCo-Eval benchmark. To mirror downstream 3D applications, we design four scenarios: *(i)* object-centric scenes, *(ii)* indoor scenes, *(iii)* outdoor scenes, and *(iv)* appearance-complex scenes (*e.g.*, characterized by high-frequency patterns, reflections, and thin structures). The suite consists of total 80 prompts describing static scenes, split into slow and fast camera trajectories to test both subtle breathing artifacts and multi-view coherence under large viewpoint changes.

Models. We evaluate recent, state-of-the-art models: 2 commercial models (Sora 2 [6], Veo 3.1 [12]) and 6 open-source models (WAN 2.2 [43], HunyuanVideo [26], LTX-Video [13], and CogVideoX variants [19,51]). We generate 4 videos per each prompt across 8 models, yielding 2,560 videos in total.

Evaluation protocol. To ensure fair comparison across different models with their own native generation settings, we compute GeCo scores on overlapping $\approx$ 3-second windows resampled to a maximum of f_{eval}=8 FPS. Within each window, we compute motion, structure, and fused scores averaged over pixels with valid depth and covisibility. Final clip-level scores are derived from the frame-weighted average of these windows, and we report the mean across all 320 clips (80 prompts $\times$ 4 seeds) per model.

Motion magnitude. High consistency scores can stem from valid geometry or simply a lack of motion in generated videos. Therefore, geometric consistency metrics should be jointly analyzed with motion to distinguish truly consistent generation from degenerate, static outputs. We first quantify motion using the

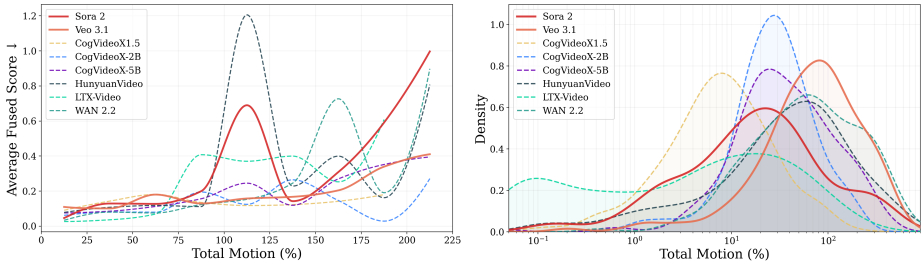


Fig. 5: Motion-dependent consistency and dynamism distribution. Left: Average GeCo Fused Score versus Total Motion. At lower motion levels (<100%), most models maintain stable scores, but in higher motion levels (>100%), several models struggle with obvious artifacts, leading to sharply degraded consistency scores. **Right:** Motion distribution of generated videos. Most models concentrate motion between 10%–100%, while Veo 3.1 notably tends to generate videos with higher dynamism.

per-frame optical flow magnitude normalized by the image diagonal (m_t), and introduce two metrics: Total Motion is the cumulative flow ($\sum m_t$) over the clip, representing the overall dynamism and amount of visual change independent of FPS. Mean Motion divides total motion by duration (Total Motion/S), representing the average rate of visual change (*i.e.*, apparent camera speed).

Motion–consistency analysis. Tab. 5 details the quantitative benchmark scores across scenarios. As some open-source models achieve better consistency scores simply by generating near-static videos, we contextualize these results using a motion–consistency Pareto front (Fig. 6). Among large-scale models, Veo 3.1 and Sora 2 achieve similarly strong consistency, though Veo sustains notably higher dynamism. Conversely, models like LTX-Video anchor the lower-error end of the frontier primarily by operating in a highly conservative, low-motion regime. Notably, CogVideoX-5B acts as a strong middle ground for open-weight models, achieving surprisingly good consistency while maintaining an intermediate dynamic range. Furthermore, finer-level analysis (Fig. 5) reveals that while Veo leans heavily toward high dynamism and LTX-Video toward conservatism, most models suffer from severely degraded consistency scores when video dynamics increase (TM>100%). Against this trend, Veo 3.1 and CogVideoX1.5 prove to be the most stable under highly dynamic conditions.

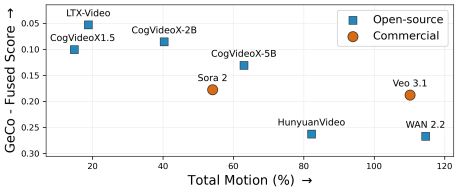


Fig. 6: Motion–consistency Pareto analysis. Fused GeCo score versus Total Motion, averaged across all four evaluation scenarios from Tab. 5. The inverted y-axis places generations with high dynamism and geometric consistency in the top-right quadrant.

Real vs. generated videos. We compare different metric score averages on 30 real videos from DL3DV and 30 generated videos (5 per model from GeCo-Eval), with comparable motion magnitudes (Total Motion, %: real 47.88 ± 35.05 vs. generated 52.44 ± 46.68). We use Cohen's d to measure the standardized gap between real and generated scores, where positive values indicate real videos

Table 6: Real vs. generated score gap. Average scores on 30 real (DL3DV) and 30 generated videos; Cohen's d measures the standardized gap, positive when real videos score as more consistent. ↓/↑: lower/higher score is more consistent.

Metric	Real (n=30)	Generated (n=30)	Cohen's d
MEt3R ↓	0.9329 ± 0.0308	0.9119 ± 0.0763	-0.361
WorldScore ↓	0.4592 ± 0.2572	0.6357 ± 0.6565	$+0.354$
TSED ↑	1.0000 ± 0.0000	0.9870 ± 0.0573	$+0.322$
GeCo ↓	0.0999 ± 0.1887	0.1809 ± 0.2682	$+0.349$

Table 7: GeCo-guided sampling improves geometric consistency. We compare CogVideoX-5B with and without GeCo guidance on a set of 45 prompts. Lower is better for consistency metrics. Guided videos exhibit slightly higher motion (TM, MM), confirming that GeCo improves consistency without hurting dynamism.

Method	MEt3R [2] ↓	Structure ↓	Motion ↓	Fused ↓	TM (%)	MM (%/s)
Without guidance	0.100	0.136	0.032	0.147	74.35±68.09	24.78±22.70
With guidance	**0.098**	**0.126**	**0.029**	**0.135**	**83.16±98.36**	**27.72±32.79**

score as more consistent. As shown in Tab. 6, GeCo, WorldScore, and TSED all yield positive d, whereas MEt3R is negative ($d = -0.361$). We do not expect a metric to separate real from generated videos, since generators can produce geometrically consistent results; rather, a valid metric should not systematically rate generated videos as more consistent than real ones. MEt3R's negative d likely stems from its focus on semantic feature consistency, which is less sensitive to the geometric artifacts.

5.2 GeCo-Guided Inference

We apply GeCo as training-free guidance to CogVideoX-5B (Sec. 3.2). Using 45 benchmark prompts, we generate paired videos with and without guidance under identical settings (details in the supplementary) and evaluate them using our metrics alongside MEt3R [2] to quantify the change in multiple metrics.

As reported in Tab. 7, GeCo guidance consistently improves all scores by reducing both structure and motion errors. Notably, guided videos exhibit slightly higher motion, indicating that GeCo enforces geometric consistency without sacrificing dynamism. Finally, we perform 3D reconstruction on the generated videos using VGGT. Qualitative results (Fig. 7) demonstrate that guided videos yield significantly cleaner reconstructions, featuring reduced drifting and fewer artifacts around thin structures.

5.3 Findings on Video Generation Models

The globe that can't be stopped. We identify a persistent failure mode where video models struggle to generate static objects under camera motion, likely due to training data bias (*e.g.*, spinning globes). As shown in Fig. 8, models consistently animate the globe despite static prompts, while GeCo correctly identifies this spurious object motion as a geometric inconsistency. We believe a

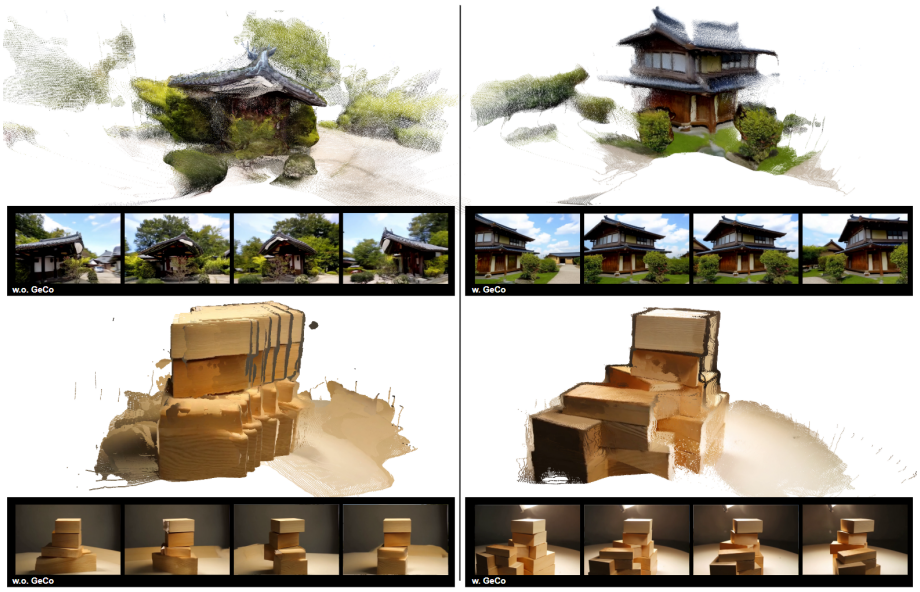


Fig. 7: GeCo guidance improves geometric consistency for 3D reconstruction. Top: 3D reconstructions from videos generated by CogVideoX-5B without (left) and with GeCo guidance (right). Bottom: corresponding video frames. Both guided videos yield cleaner geometry with fewer deformation and drifting artifacts across views, which enables higher reconstruction quality.

Fig. 8: "The globe that can't be stopped." A common failure mode that models consistently make the globe rotate despite static prompts. GeCo localizes this spurious object motion on the globe surface, clearly separating it from the intended egocentric camera motion.

possible reason for the failure mode is the bias in training data, where most examples of globes appear in motion, leading models to conflate object persistence with deformation or drift.

Freezing the spurious motion. To empirically evaluate our assumption, we test prompts that should be sparse in video training data such as "a camera orbiting a dog that has no motion." Models typically fail to keep the subject rigid. However, by using GeCo as a guidance term, we can strongly suppress this non-ego motion. As illustrated in Fig. 9, our guidance effectively freezes the subject, creating a bullet-time effect where the camera orbits while the object remains rigid.

Fig. 9: Stopping the globe and freezing the dog with GeCo-guidance. We compare video generations for prompts specifying a camera orbiting a nominally static globe (left) and a static dog (right). Without guidance, the model introduces spurious object motion, causing the globe to spin and the dog to move. Applying GeCo guidance effectively suppresses this non-ego motion, enforcing geometric consistency where the subject remains rigid while the camera orbits, producing a "bullet-time" effect.

6 Discussion

Limitations. GeCo's assumption of a mostly rigid scene means dynamic contents can trigger false positives by penalizing true object motion. Addressing this by applying motion masks to static backgrounds is a promising direction for future work. Beyond scene constraints, deploying GeCo for training-free guidance increases generation time (e.g., from 2 to 19 minutes on an H200 GPU). Because this overhead stems from iterative sampling and backpropagation rather than the metric itself, integrating efficient samplers or lightweight estimators could substantially reduce it. Finally, our metric inherits the limitations of its off-the-shelf estimators. Although we observed mild prediction noise, it remains well below typical artifact severity, meaning GeCo will seamlessly yield more accurate localization as geometry foundation models advance.

Conclusion. We present GeCo, an interpretable metric that fuses optical flow and depth priors to localize geometric artifacts in generated videos. After validating robustness on controlled datasets, we use it to benchmark state-of-the-art models at scale and identify persistent failure modes. We further show GeCo can act as a training-free guidance loss, improving geometric consistency, reducing spurious motion, and benefiting downstream 3D reconstruction. As GeCo is fully differentiable, it also supports finetuning and distillation. Overall, GeCo provides actionable diagnostics for geometric consistency and supports applications such as 3D asset creation and world simulation.

Acknowledgements

This project is supported in part by NSF grant CRCNS-2309041. The views and conclusions contained herein are those of the authors and do not represent the official policies or endorsements of these institutions.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)

2. Asim, M., Wewer, C., Wimmer, T., Schiele, B., Lenssen, J.E.: MEt3R: Measuring multi-view consistency in generated images. In: CVPR. pp. 6034–6044 (2025)
3. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: CVPRW. pp. 843–852 (2023)
4. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation (2025)
5. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: ZoeDepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
6. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), OpenAI technical report
7. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9630–9640 (2021)
8. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: CVPR (2017)
9. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-supervised interest point detection and description. CVPRW pp. 337–33712 (2017)
10. Duan, H., Yu, H.X., Chen, S., Fei-Fei, L., Wu, J.: WorldScore: A unified evaluation benchmark for world generation. In: ICCV (2025)
11. Geng, D., Owens, A.: Motion guidance: Diffusion-based image editing with differentiable motion estimators. In: ICLR (2024)
12. Google DeepMind: Veo: a text-to-video generation system. Tech. rep. (2025), Veo 3 technical report
13. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: LTX-Video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
14. Han, J., Kokkinos, F., Torr, P.: VFusion3D: Learning scalable 3D generative models from video diffusion models. In: ECCV (2024)
15. Han, X., Zickler, T., Nishino, K.: Multistable shape from shading emerges from patch diffusion. NeurIPS **37**, 34686–34711 (2024)
16. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2003)
17. He, X., Peng, C., Liu, Z., Wang, B., Zhang, Y., Cui, Q., Kang, F., Jiang, B., An, M., Ren, Y., Xu, B., Guo, H.X., Gong, K., Wu, C., Li, W., Song, X., Liu, Y., Li, E., Zhou, Y.: Matrix-Game 2.0: An open-source, real-time, and streaming interactive world model. arXiv preprint arXiv:2508.13009 (2025)
18. He, Y., Murata, N., Lai, C.H., Takida, Y., Uesaka, T., Kim, D., Liao, W.H., Mitsufuji, Y., Kolter, J.Z., Salakhutdinov, R., Ermon, S.: Manifold preserving guided diffusion. In: ICLR (2024)
19. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: CogVideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
20. Hong, Y., Mei, Y., Ge, C., Xu, Y., Zhou, Y., Bi, S., Hold-Geoffroy, Y., Roberts, M., Fisher, M., Shechtman, E., Sunkavalli, K., Liu, F., Li, Z., Tan, H.: RELIC: Interactive video world model with long-horizon memory (2025)
21. Huang, C.H.P., Mitra, N., Jeong, H., Yoon, J.S., Ceylan, D.: JOG3R: Towards 3D-consistent video generators. In: BMVC (2025)

22. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench: Comprehensive benchmark suite for video generative models. CVPR pp. 21807–21818 (2023)
23. Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., Wang, Y., Chen, X., Chen, Y., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench++: Comprehensive and versatile benchmark suite for video generative models. ArXiv (2024)
24. HunyuanWorld, T.: HY-World 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. arXiv preprint (2025)
25. Jang, S.S., Ki, T., Jo, J., Yoon, J., Kim, S.Y., Lin, Z.L., Hwang, S.J.: Frame guidance: Training-free guidance for frame-level control in video diffusion models. ArXiv **abs/2506.07177** (2025)
26. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: HunyuanVideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
27. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: LightGlue: Local feature matching at light speed. In: ICCV. pp. 17581–17592 (2023)
28. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., Li, X., Sun, X., Ashok, R., Mukherjee, A., Kang, H., Kong, X., Hua, G., Zhang, T., Benes, B., Bera, A.: DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In: CVPR. pp. 22160–22169 (2023)
29. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: Evalcrafter: Benchmarking and evaluating large video generation models. In: CVPR. pp. 22139–22149 (2024)
30. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV. vol. 2, pp. 1150–1157 vol.2 (1999)
31. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Gool, L.V.: RePaint: Inpainting using denoising diffusion probabilistic models. CVPR pp. 11451–11461 (2022)
32. Mao, X., Li, Z., Li, C., Xu, X., Ying, K., He, T., Pang, J., Qiao, Y., Zhang, K.: Yume-1.5: A text-controlled interactive world generation model. arXiv preprint arXiv:2512.22096 (2025)
33. Meng, F., Liao, J., Tan, X., Shao, W., Lu, Q., Zhang, K., Cheng, Y., Li, D., Qiao, Y., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. arXiv preprint arXiv:2410.05363 (2024)
34. Nam, H., Kim, J., Lee, D., Ye, J.C.: Optical-flow guided prompt optimization for coherent video generation. CVPR pp. 7837–7846 (2024)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
36. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
37. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3D: Large-scale learning and evaluation of real-life 3d category reconstruction. In: ICCV (2021)
38. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3C: 3D-informed world-consistent video generation with precise camera control. In: CVPR. pp. 6121–6132 (2025)

39. Sampson, P.D.: Fitting conic sections to "very scattered" data: An iterative refinement of the bookstein algorithm. Computer graphics and image processing (1982)
40. Sun, W., Zhang, H., Wang, H., Wu, J., Wang, Z., Wang, Z., Wang, Y., Zhang, J., Wang, T., Guo, C.: Worldplay: Towards long-term geometric consistency for real-time interactive world model. arXiv preprint (2025)
41. Teed, Z., Deng, J.: DROID-SLAM: Deep visual slam for monocular, stereo, and RGB-D cameras. NeurIPS **34**, 16558–16569 (2021)
42. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. In: ECCV. pp. 439–457. Springer (2024)
43. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
44. Wang, H., Liu, F., Chi, J., Duan, Y.: VideoScene: Distilling video diffusion model to generate 3d scenes in one step. In: CVPR. pp. 16475–16485 (2025)
45. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual geometry grounded transformer. In: CVPR (2025)
46. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3D vision made easy. In: CVPR. pp. 20697–20709 (2024)
47. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.A.: TartanAir: A dataset to push the limits of visual slam. In: IROS. pp. 4909–4916 (2020)
48. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. In: ICLR (2023)
49. Xue, Q., Yin, X., Yang, B., Gao, W.: Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In: CVPR. pp. 18826–18836 (2025)
50. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything V2. NeurIPS **37**, 21875–21911 (2024)
51. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: CogVideoX: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
52. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: ScanNet++: A high-fidelity dataset of 3D indoor scenes. In: ICCV (2023)
53. Yu, J.J., Forghani, F., Derpanis, K.G., Brubaker, M.A.: Long-term photometric consistent novel view synthesis with diffusion models. In: ICCV (2023)
54. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
55. Zhang, Q., Zhai, S., Martin, M.A.B., Miao, K., Toshev, A., Susskind, J., Gu, J.: World-consistent video diffusion with explicit 3D modeling. In: CVPR. pp. 21685–21695 (2025)
56. Zhang, Y., Wei, Y., Jiang, D., ZHANG, X., Zuo, W., Tian, Q.: ControlVideo: Training-free controllable text-to-video generation. In: ICLR (2024)
57. Zhang, Y., Peng, C., Wang, B., Wang, P., Zhu, Q., Kang, F., Jiang, B., Gao, Z., Li, E., Liu, Y., Zhou, Y.: Matrix-Game: Interactive world foundation model. arXiv preprint arXiv:2506.18701 (2025)
58. Zhang, Y., Keetha, N., Lyu, C., Jhamb, B., Chen, Y., Qiu, Y., Karhade, J., Jha, S., Hu, Y., Ramanan, D., Scherer, S., Wang, W.: UFM: A simple path towards unified dense correspondence with flow. In: arXiV (2025)