Measuring 3D Spatial Geometric Consistency in Dynamic Video Generation

Weijia Dou[1,2,*], Wenzhao Zheng[1,3,*,†], Weiliang Chen[1], Yu Zheng[1], Jie Zhou[1], and Jiwen Lu[1,✉]

[1]Tsinghua University [2]Tongji University [3]University of California, Berkeley
Code: https://github.com/tj12323/SGC
renrendwj@tongji.edu.cn; wenzhao.zheng@outlook.com;
lujiwen@tsinghua.edu.cn

Abstract. Recent generative models can produce high-fidelity videos, yet they often exhibit 3D spatial geometric inconsistencies. Existing evaluation methods fail to accurately characterize these inconsistencies: fidelity-centric metrics like FVD are insensitive to geometric distortions, while consistency-focused benchmarks often penalize valid foreground dynamics. To address this gap, we introduce SGC, a metric for evaluating 3D **S**patial **G**eometric **C**onsistency in dynamically generated videos. We quantify geometric consistency by measuring the divergence among multiple camera poses estimated from distinct local regions. Our approach first separates static from dynamic regions, then partitions the static background into spatially coherent sub-regions. We predict depth for each pixel, estimate a local camera pose for each subregion, and compute the divergence among these poses to quantify geometric consistency. Experiments on real and generative videos demonstrate that SGC robustly quantifies geometric inconsistencies, effectively identifying critical failures missed by existing metrics.

Keywords: Video Generation · Geometric Consistency

1 Introduction

Recent generative video synthesis models achieve high visual fidelity [3, 7, 32], approaching photorealism and enabling content creation and simulation applications. Despite strong performance on fidelity-centric metrics like FVD [40], current models often exhibit significant 3D spatial geometric inconsistencies. As illustrated in Fig. 1, these failures include (1) geometric warping—static backgrounds distorting during camera motion, (2) incoherent motion—static and dynamic objects illogically fusing, (3) object impermanence—structures flickering or morphing across frames, and (4) perspective failures—large-scale scene distortions violating 3D perspective principles.

A critical impediment to addressing these failures is the lack of robust metrics to measure 3D spatial geometric consistency in dynamically generated videos.

*Equal contributions. †Project leader. ✉Corresponding author.

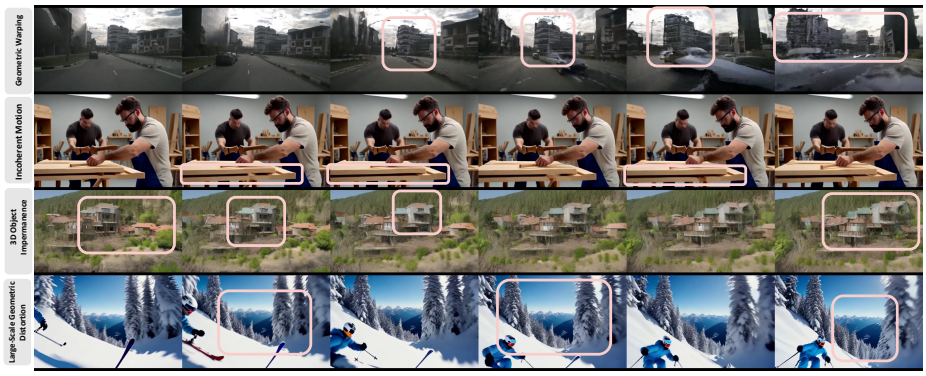


Fig. 1: Examples of 3D Spatial Geometric Inconsistencies in Generated Videos. Existing models often fail to maintain geometric consistency, exhibiting critical 3D spatial failures despite plausible per-frame visuals. (a) Geometric Warping: The rigid structure of the static buildings severely distorts as the camera moves. (b) Incoherent Motion: The static workbench illogically "sticks" to and moves with the dynamic piece of wood, violating physical separation. (c) Object Impermanence: A static structure on the mountain "flickers" and illogically changes its shape, failing to persist over time. (d) Perspective Failure: The distant mountains, which should remain stable, unnaturally warp and "narrow" as the skier moves forward, violating 3D perspective.

Existing metrics fall into two categories, each with fundamental limitations. Fidelity-centric metrics (e.g., FVD [40]) exhibit a content-over-motion bias, i.e., they are dominated by per-frame appearance and are largely insensitive to geometric distortions, thus rewarding plausible textures despite unstable geometry [4]. Conversely, existing consistency-focused metrics suffer from a fragility-to-motion bias, i.e., metrics from novel view synthesis [46, 48] are not robust to dynamic objects, while benchmarks like VBench [23] may penalize valid vigorous motion [30]. Other specialized tools like FVMD [30] focus on object kinematics while neglecting background stability. This fragmentation highlights a critical gap: no established metrics effectively isolate and diagnose 3D spatial geometric consistency amidst complex foreground dynamics.

To address this gap, we propose SGC (**S**patial **G**eometric **C**onsistency), a diagnostic metric to quantify this foundational 3D spatial geometric consistency, explicitly focusing on the stability of the static background. SGC is founded on the physical principle that in a coherent 3D world, the apparent motion of all static background points ($V_{P_i} = 0$) must be consistent with a single, shared camera transformation (T_{cam}). Our method first isolates the static background to analyze scene stability. Leveraging depth information, 3D points within these static regions are reconstructed and segmented into spatially coherent sub-regions. For each sub-region, the relative camera pose is estimated via Perspective-n-Point (PnP) using 2D feature tracks and their corresponding 3D points. In a physically coherent scene, these local pose estimates must agree, and their divergence signals geometric inconsistency. The SGC score quantifies this geometric instability by aggregating the divergence among these local pose estimates (measured

as inter-region variance), their agreement with a global camera trajectory, and additional criteria such as cross-frame depth alignment. We validate SGC on diverse videos from state-of-the-art generative models (spanning text-to-video, image-to-video, and video-to-video paradigms) and real-world sequences. Experiments confirm that SGC robustly quantifies these geometric inconsistencies, effectively identifying failures overlooked by existing metrics and establishing it as a necessary complementary diagnostic tool.

2 Related Work

Video Quality Assessment. Existing video quality metrics are either insensitive to geometric failures or overly sensitive to valid motion. Fidelity-centric metrics, including Fréchet Video Distance (FVD) [17,31,40], suffer from content-over-motion bias by rewarding plausible per-frame appearance while ignoring geometric distortions [4]. Similarly, prompt-alignment metrics like CLIPScore [21] assess frame-level semantics, neglecting dynamic properties like temporal consistency. These metrics are ill-suited for diagnosing foundational geometric failures. Conversely, consistency-focused metrics exhibit fragility-to-motion bias. Novel View Synthesis metrics (e.g., TSED [48], MEt3R [5]) fail when dynamic objects are present. Broad benchmarks like VBench [23] provide holistic assessments; however, using their consistency dimensions as geometric stability proxies risks confounding by valid high dynamics. Specialized tools fall short: FVMD [30] evaluates object kinematics while ignoring background stability. VideoPhy [6] assesses broad physical plausibility but cannot isolate background consistency from foreground dynamics. Furthermore, while a recent study [28] analyzes Sora's geometric consistency via 3D reconstruction quality, an automated, model-agnostic metric for arbitrary dynamic videos remains lacking. As existing metrics can hardly diagnose background geometric stability amidst complex foreground motion, we introduce a physically grounded metric that explicitly disentangles background stability from foreground dynamics.

Generative Video Models. Recent advancements in generative video models have led to distinct categories based on input modality, including Text-to-Video (T2V) [7, 9, 13, 18, 24, 26, 27, 32, 38, 41, 44, 45, 49, 53, 54], Image-to-Video (I2V) [8,16,24,35,36,47,50–52], and Video-to-Video (V2V) [3,20]. Architectural innovations like the Space-Time U-Net in Lumiere [7] target global temporal consistency by generating the entire video in one pass. To expand input flexibility, VideoPoet [26] leverages an LLM for zero-shot generation from text, images, and audio. Other works focus on high-fidelity synthesis and control [11,19,29,39,42]. I2VGen-XL [52] employs a cascaded diffusion strategy to separate semantic coherence from spatio-temporal refinement, while Diffusion as Shader [20] provides 3D-aware control via 3D tracking signals, enabling versatile manipulations. Models like Cosmos [3] scale to model complex 3D environmental dynamics.

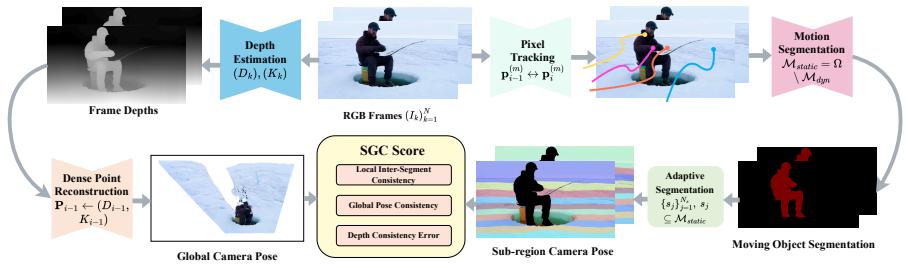


Fig. 2: Overview of the SGC computation pipeline. Input RGB frames undergo parallel processing: (i) depth estimation, leading to dense point reconstruction for global camera pose estimation; and (ii) pixel tracking followed by motion segmentation to isolate moving objects. The identified static background is then adaptively segmented. Local camera poses for these static sub-regions are subsequently estimated using information from pixel tracks and depth. Finally, the overall SGC score is computed by aggregating three key evaluations: local inter-segment consistency, global pose consistency, and cross-frame depth consistency.

3 Proposed Approach

We introduce the *SGC* score, a diagnostic metric designed to evaluate 3D spatial geometric consistency in dynamically generated videos. *SGC* is premised on the physical principle that all points belonging to the static background must adhere to a single, coherent camera transformation. The metric quantifies violations of this principle by measuring the variance among multiple local camera pose estimates derived exclusively from these static regions.

The implementation follows the feedforward pipeline shown in Fig. 2. To directly address the "fragility-to-motion" bias discussed earlier, the pipeline begins by isolating the static background $\mathcal{M}_{static}$ (Sec. 3.1). Subsequent steps then estimate global camera parameters and depth (Sec. 3.2), partition $\mathcal{M}_{static}$ into coherent sub-regions using depth (Sec. 3.3), estimate local relative poses $P_i^{loc,j}$ for these sub-regions via PnP, and finally aggregate metrics quantifying the divergence among these local estimates and their agreement with the global motion P_i^{glo} into the final *SGC* score (Sec. 3.4).

3.1 Static Background Isolation

A critical limitation of existing metrics is their "fragility-to-motion" bias, as they are often confounded by complex foreground dynamics. To establish a robust diagnostic for geometric consistency, our method first disentangles the static background from independently moving objects. Failure to do so introduces two key confounding factors: (1) a generator may be penalized for valid complex object motion, and (2) severe object-level inconsistencies can obscure the evaluation of an otherwise stable background. Therefore, our metric is explicitly designed to diagnose the stability of the static background, complementing kinematic metrics that evaluate dynamic elements. To achieve this isolation, we employ

SegAnyMo [22] as a preprocessing step. This method segments dynamic content by classifying long-range spatio-temporal point tracks, yielding precise per-frame moving-object masks. The union of these masks forms the dynamic region, denoted as $\mathcal{M}_{dyn}$. The static background region $\mathcal{M}_{static}$, which is the exclusive focus of our subsequent geometric analysis, is defined as the set complement:

$$\mathcal{M}_{static} = \Omega \setminus \mathcal{M}_{dyn}, \tag{1}$$

where Ω denotes the entire image spatial domain.

3.2 Dense Point Reconstruction

To accurately reconstruct the 3D scene, SGC relies on two complementary geometry sources: global camera geometry and per-frame scene depth.

Traditional Structure from Motion (SfM) pipelines (e.g., COLMAP [37]) often fail in reconstructing 3D scenes from generative videos. To address these limitations, we consider the Visual Geometry Grounded Transformer (VGGT) [43], a feed-forward neural network architected to directly infer a comprehensive suite of 3D scene attributes from a sequence of N input images $(I_k)_{k=1}^N$, where $I_k \in \mathbb{R}^{3\times H\times W}$. The model defines a function f that maps these images to their corresponding 3D annotations:

$$f\left((I_k)_{k=1}^N\right) = ((R_{wc,k}, t_{wc,k}, K_k), D_k, P_k, T_k)_{k=1}^N, \tag{2}$$

where $(R_{wc,k}, t_{wc,k})$ represents the global camera pose (with $R_{wc,k} \in SO(3)$ being the world-to-camera rotation matrix and $t_{wc,k} \in \mathbb{R}^3$ the translation vector), $K_k \in \mathbb{R}^{3\times 3}$ is the camera intrinsic matrix, D_k is the depth map, P_k is the viewpoint-invariant point map, and T_k are dense features.

Our system specifically exploits VGGT strictly for extracting robust global camera poses $(R_{wc,k}, t_{wc,k})$ and intrinsics K_k. Separately, to obtain high-quality and temporally consistent scene depth, we substitute VGGT depth output with Video Depth Anything [14] to estimate the per-frame dense depth map.

3.3 Local Relative Pose Estimation

We partition this region $\mathcal{M}_{static}$ into depth-consistent sub-regions. We apply clustering to the 1D depth values d_p associated with pixels (x, y) within $\mathcal{M}_{static}$. We robustly isolate distinct 3D surfaces by first partitioning $\mathcal{M}_{static}$ into globally coherent depth strata (e.g., robustly isolating a distant mountain range from a mid-ground building) and subsequently treating the resulting depth-based groups as sub-regions for local estimation.

The set of N_p valid depth values, $D = \{d_p \mid (x_p, y_p) \in \mathcal{M}_{static} \text{ and } d_p \text{ is valid}\}$, is processed by depth clustering into k_{seg} clusters of depth magnitudes, $\{S_j\}$. Each cluster S_j groups pixels from $\mathcal{M}_{static}$ that share similar depth values. These pixel groupings, after filtering by a minimum total count, constitute the N_s ($N_s \leq k_{seg}$) distinct subregions $s_j \subseteq \mathcal{M}_{static}$. Each sub-region s_j in frame

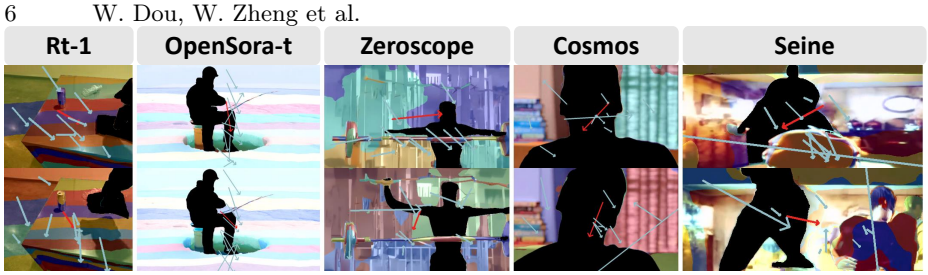


Fig. 3: Visualization of the Local Inter-Segment and Global Pose Consistency. Blue arrows depict image-plane projections of motion directions induced by local relative poses $P_i^{loc,j}$ of static sub-regions, while the red arrow denotes the projected global relative pose P_i^{glo}. Arrows are obtained by projecting a unit 3D direction vector under each relative transformation, length-normalized for visualization. Sequences from left to right exhibit decreasing consistency scores. (Left) High consistency: local poses are similar and align with global motion. (Right) Low consistency: local poses show high variance and/or deviate significantly from global motion.

f_i serves as an independent unit for subsequent pose estimation. For each such subregion s_j identified in frame f_i, we estimate its local relative camera pose with respect to the preceding frame f_{i-1}. This pose, denoted by the rotation $R_i^{loc,j} \in SO(3)$ and translation $\mathbf{t}_i^{loc,j} \in \mathbb{R}^3$, describes the transformation of points from the camera coordinate system of f_{i-1} to that of f_i.

Dense 2D tracking establishes the robust frame-to-frame correspondences required for local pose estimation. This yields per-pixel motion trajectories across consecutive frames, providing necessary 2D matching pairs. Specifically, for a set of 2D points $\{\mathbf{p}_{i-1}^{(m)} = (u_{i-1}^{(m)}, v_{i-1}^{(m)})\}$ observed in f_{i-1} and their corresponding tracked locations $\{\mathbf{p}_i^{(m)} = (u_i^{(m)}, v_i^{(m)})\}$ falling within subregion s_j in f_i, we first reconstruct their 3D coordinates in the camera frame of f_{i-1}. Given the depth map D_{i-1} and the camera intrinsic matrix K_{i-1} (obtained from the global geometry estimator), each 3D point $\mathbf{P}_{i-1}^{(m)}$ is reconstructed by unprojecting its 2D coordinates using the corresponding depth value and focal lengths.

With the set of 3D points $\{\mathbf{P}_{i-1}^{(m)}\}$ and their 2D projections $\{\mathbf{p}_i^{(m)}\}$ in subregion s_j, along with the intrinsic matrix K_i for frame f_i, the Perspective-n-Point (PnP) algorithm is employed to solve for $R_i^{loc,j}$ and $\mathbf{t}_i^{loc,j}$ that satisfy the projection equation for each correspondence m.

3.4 *SGC*: Spatial Geometric Consistency Score

To evaluate the 3D geometric consistency between consecutive frames (f_{i-1}, f_i), we define three component metrics that leverage the N_s local relative poses $P_i^{loc,j} = (R_i^{loc,j}, \mathbf{t}_i^{loc,j})$ estimated from static sub-regions and the global relative pose $P_i^{glo} = (R_i^{glo}, \mathbf{t}_i^{glo})$. High errors indicate failures to render a coherent 3D scene evolving consistently over time.

Local Inter-Segment Consistency. This metric measures motion estimation stability within the static background. Disagreement among local poses $\{P_i^{loc,j}\}$ from different static sub-regions indicates the generator's inability to maintain a rigid scene structure. We quantify this via two variance measures:

• Local Rotational Variance:

$$\sigma_{rot,loc}^2 = \frac{1}{N_s}\sum_{j=1}^{N_s}(d_\theta(R_i^{loc,j}, \overline{R}_i^{loc}))^2, \tag{3}$$

where $d_\theta(R_A, R_B) = \arccos\left(\frac{\text{Trace}(R_A^T R_B)-1}{2}\right)$ and $\bar{R}_i^{loc}$ is the mean local rotation.

• Local Translational Variance:

$$\sigma_{trans,loc}^2 = \frac{1}{N_s}\sum_{j=1}^{N_s}\left\|\mathbf{t}_i^{loc,j} - \bar{\mathbf{t}}_i^{loc}\right\|_2^2, \tag{4}$$

where $\bar{\mathbf{t}}_i^{loc}$ is the mean local translation.

Global Pose Consistency. This metric assesses the alignment between locally estimated motions and the global pose P_i^{glo}. Significant deviations suggest that local dynamics are erratic or disconnected from the plausible global camera movement, breaking large-scale spatio-temporal coherence (see Fig. 3). We evaluate this using:

• Global Rotational Variance:

$$\sigma_{rot,glob}^2 = \frac{1}{N_s}\sum_{j=1}^{N_s}(d_\theta(R_i^{loc,j}, R_i^{glo}))^2. \tag{5}$$

• Global Translational Variance:

$$\sigma_{trans,glob}^2 = \frac{1}{N_s}\sum_{j=1}^{N_s}\left\|\mathbf{t}_i^{loc,j} - \mathbf{t}_i^{glo}\right\|_2^2. \tag{6}$$

Depth Consistency Error (E_{depth}). This metric evaluates the stability of the 3D scene geometry over time using the global pose P_i^{glo} for alignment. It measures the discrepancy between the depth map of f_i (D_i) and the warped depth map from f_{i-1} ($D_{i-1\to i}$). A large error suggests the video fails to maintain a consistent 3D world.

$$E_{depth} = \frac{1}{|\mathcal{V}|}\sum_{\mathbf{u}\in\mathcal{V}}\left|D_{i-1\to i}(\mathbf{u}) - D_i(\mathbf{u})\right|. \tag{7}$$

where $\mathcal{V}$ is the set of valid overlapping pixels in the static scene mask of f_i.

Overall *SGC* Score. To aggregate these N_M component metrics (where N_M denotes the number of component metrics) into a single, robust *SGC* score, we employ an objective weighting procedure. First, all raw component scores $M_{d,k}$ are normalized to a common $[0, 1]$ scale (where 0 is best). This normalization, involving Z-score standardization followed by min-max scaling, is crucial to prevent metrics with high raw magnitudes (e.g., translational variance) from dominating the final score. Second, to avoid subjective manual tuning, we derive fixed weights w_k by performing Principal Component Analysis (PCA) on the

normalized score matrix from a diverse calibration dataset. The weights w_k are the normalized loadings of the first principal component (PC1). This process automatically assigns greater influence to metrics capturing the most significant shared variance related to geometric inconsistency. The final SGC score for a video d is the weighted sum of its normalized component scores $M''_{d,k}$:

$$SGC_d = \sum_{k=1}^{N_M} w_k M''_{d,k}. \tag{8}$$

This multi-component, variance-based design ensures SGC is robust to camera motion magnitude, penalizing geometric inconsistency rather than the motion.

4 Experiments

4.1 Experimental Setup

Datasets and Models. To rigorously evaluate spatial geometric consistency (SGC), we curate a comprehensive benchmark of 1,296 videos. This comprises 1,196 videos from the GenWorld dataset [15] alongside an additional 100 high-motion web videos sampled from OpenVidHD-0.4M. To establish robust ground-truth anchors for real-world geometry and dynamics, we utilize 300 real videos across three distinct domains: **nuScenes** [12] provides complex, multi-object driving scenarios; **RT-1** [10] offers highly dynamic robotic interactions; and **OpenVid** [33] serves as a crucial anchor that significantly enhances the diversity and representativeness of everyday web captures. The AI-generated portion consists of 996 videos sourced entirely from GenWorld, synthesized by 10 state-of-the-art generative models. To ensure comprehensive evaluation across paradigms, these models are grouped into: (1) **Text-to-Video (T2V)** [1,2,13,32,44,45,54]; (2) **Image-to-Video (I2V)** [16,54]; and (3) **Video-to-Video (V2V)** [3]. Detailed dataset statistics are deferred to the App.C.1.

Baselines. We compare our method against established metrics, categorized as follows: (1)**Feature/Consistency:** MEt3R [5] and its motion-segmented variant, MEt3R(+MOS); (2)**Benchmark Suites:** VBench [23] (specifically Background Consistency, Dynamic Degree, and a weighted composite); and (3) **Motion/Physics Proxies:** FVD [40], FVMD [30], and TRAJAN [4]. Moving object segmentation (MOS) is exclusively applied to MEt3R to create the MEt3R(+MOS) baseline, whereas all other baselines operate on unmasked videos. Further baseline implementation details are provided in the App.C.2.

Unified Evaluation Protocol. To ensure a strictly fair comparison, all evaluations adhere to a unified protocol. To isolate geometric stability, we define a static background mask, $\mathcal{M}_{static}$, where masked foreground pixels are explicitly ignored. For MEt3R(+MOS), this exact masking is applied by filtering the score map prior to final calculation. Furthermore, we standardize the underlying foundational estimators: VGGT [43] for global poses and intrinsics, Video Depth Anything [14] for depth, DELTA [34] for dense tracking, and SegAnyMo [22] for

Table 1: Quantitative comparison of generative video methods and real-world video datasets in terms of 3D spatial geometric consistency. The table presents an evaluation of the performance using MEt3R [5], FVD [40], the proposed SGC metric, and VBench [23] dimensions: background consistency (BC), dynamic degree (DD), weighted score (w), and TRAJAN [4] and FVMD [30]. The best and second-best scores are highlighted in **bold** and underlined, respectively.

	Methods	Task	SGC ↓	MEt3R [5] ↓	MEt3R [5] (+MOS) ↓	FVD [40] ↓	VBench [23] ↑ BC	VBench [23] ↑ DD	VBench [23] ↑ weighted	TRAJAN [4] ↑	FVMD [30] ↓
Gen.	Cosmos [3]	V2V	0.0722	0.0741	0.2350	760.3204	0.9461	0.6462	0.6037	0.6023	15389.47
	Hotshot [1]	T2V	0.1172	0.1371	0.3568	908.0547	0.9468	0.8190	0.6801	0.5518	-
	Latte [32]	T2V	0.3226	0.1707	0.2601	801.0069	0.9412	0.8559	0.4403	0.4418	19561.51
	Lavie [45]	T2V	0.1241	0.1122	0.2793	679.8203	<u>0.9576</u>	0.7559	0.7090	<u>0.6504</u>	<u>14424.44</u>
	Modelscope [44]	T2V	0.3129	0.1851	0.4208	784.6681	0.9196	0.7966	0.3313	0.5962	19511.57
	OpenSora-i [54]	I2V	0.1631	0.1030	0.3088	**751.2137**	0.8993	0.8333	0.3638	0.3208	**9250.18**
	OpenSora-t [54]	T2V	0.0831	0.0919	0.2513	<u>838.0382</u>	0.9259	0.8286	0.5370	0.2835	13169.15
	Seine [16]	I2V	0.2837	0.2613	0.6343	808.9575	0.8891	<u>0.8981</u>	0.1215	0.4669	23555.54
	VideoCrafter [13]	T2V	0.0973	0.1176	0.2205	699.7098	**0.9650**	0.6320	<u>0.7114</u>	0.6344	16125.86
	Zeroscope [2]	T2V	0.0912	0.1558	0.4602	823.1359	0.9326	0.5287	0.4974	0.5230	18714.92
Real	OpenVid [33]	–	**0.0530**	**0.0371**	**0.0486**	752.2600	0.9408	0.6300	**0.8330**	0.4869	-
	RT-1 [10]	–	0.0639	0.0799	<u>0.0698</u>	1352.8758	0.9229	**0.9800**	0.6499	**0.7362**	-
	nuScenes [12]	–	<u>0.0613</u>	<u>0.0485</u>	0.0753	1297.0720	0.9080	0.8800	0.5370	0.2320	-

MOS extraction. To evaluate local consistency, we robustly isolate distinct 3D surfaces by applying a GPU-accelerated k-means clustering to the 1D depth values d_p associated with valid pixels (x, y) within $\mathcal{M}_{static}$. Finally, to aggregate SGC components, PCA weights are learned exactly once on the full dataset. Additionally, the VBench weighted composite metrics are calculated using this identical weighting method on our dataset to ensure calibration parity. Detailed hyperparameter specifications are provided in the App.C.3.

4.2 Benchmarking Geometric Consistency in Generative Models

Unified Benchmark and Overall Performance. We present a comprehensive quantitative evaluation in Table 1, benchmarking 10 contemporary generative models against three diverse real-world anchors (RT-1, nuScenes, and OpenVid). To provide a rigorous and multi-faceted assessment, our unified benchmark compares the proposed SGC metric against established Feature/Consistency metrics (MEt3R [5], MEt3R(+MOS)), Benchmark Suites (VBench [23] BC, DD, and weighted composite), and Motion/Physics Proxies (FVD [40], FVMD [30], and TRAJAN [4]). As expected, the real-world datasets establish a strict lower bound for spatial error. Among the generative methods, Cosmos [3] achieves the most competitive SGC score (0.0722), though it still exhibits a noticeable gap when compared to real-world anchors like OpenVid (0.0530) or nuScenes (0.0613). Conversely, models such as Seine [16], Modelscope [44], and Latte [32] yield significantly higher SGC scores, indicating pronounced difficulties in maintaining spatial coherence across frames.

Isolating Background Dynamics. Comparing the standard MEt3R [5] metric with its motion-segmented counterpart, MEt3R(+MOS), yields a critical observation. Across all evaluated generative models, explicitly filtering the score map using the static background mask $\mathcal{M}_{static}$ induces a distinct and consistent upward shift in error scores. Rather than representing a baseline failure, this adjustment isolates the background by stripping away foreground motion priors. This shift demonstrates unmasked feature-matching metrics frequently conflate

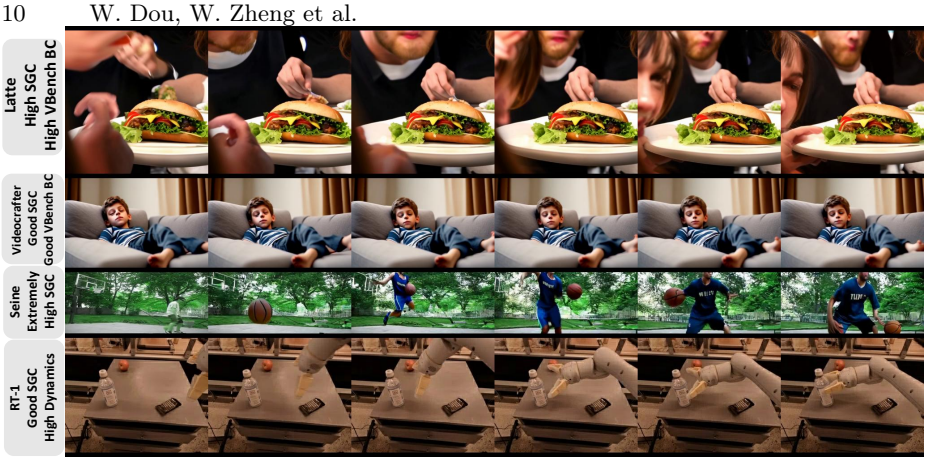


Fig. 4: Qualitative Validation: SGC Detects Geometric Failures Missed by Feature Metrics. Latte (R1) scores poorly on SGC (object instability), despite a high VBench-BC score from plausible textures. VideoCrafter (R2) is consistent, scoring well on both. Seine exhibits catastrophic geometric breakdown (e.g., subject distortion), reflected in its very high SGC score. RT-1, despite high dynamics (robot motion), scores excellently, demonstrating SGC motion robustness.

accurate tracking of foreground textures with true underlying environmental stability. While MEt3R(+MOS) successfully prevents dynamic foregrounds from artificially inflating perceived consistency, both MEt3R variants remain fundamentally constrained to 2D feature tracking. Compared to SGC, MEt3R(+MOS) lacks an explicit understanding of scene depth and camera ego-motion. By leveraging foundational estimators like VGGT [43] and Video Depth Anything [14], SGC evaluates true 3D background rigidity and multi-plane stability, providing a strictly more rigorous geometric constraint than 2D texture persistence alone. This distinction explains the performance gap for models like Cosmos [3]. Although Cosmos achieves a competitive SGC score, it exhibits significantly higher error under MEt3R(+MOS). A component breakdown reveals Cosmos performs well within real-world ranges for depth consistency and translation, struggling primarily with rotation. While MEt3R penalizes 2D feature persistence failures, SGC correctly rewards this underlying 3D structural preservation.

Diagnosing Foundational 3D Instabilities. The ranking derived from SGC diverges fundamentally from those produced by fidelity-centric and 2D physics metrics. For example, Latte and Modelscope achieve strong FVD [40] scores and high VBench [23] Background Consistency (BC) ratings, despite ranking poorly under SGC (as visualized in Fig. 4). This discrepancy illustrates that current generative models can synthesize highly plausible 2D textures (satisfying VBench-BC) and achieve high per-frame visual quality (satisfying FVD) while failing to preserve multi-plane stability and structural integrity. Ultimately, SGC is specifically sensitive to these foundational 3D geometric instabilities that remain hidden under texture-plausible generations, successfully exposing spatial inconsistencies that persist even after rigorous MOS adaptation. Furthermore, a blinded pairwise human study demonstrating that SGC aligns with human judgments of static-background 3D stability is detailed in the App.E.

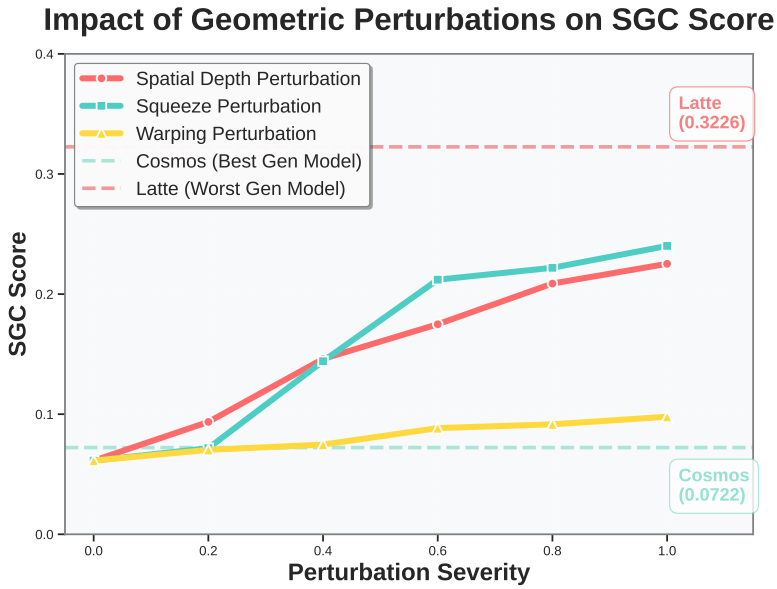


Fig. 5: Impact of Geometric Perturbations on SGC Score. This line plot presents a sensitivity and monotonicity analysis, illustrating SGC scores rising with geometric degradation severity (from 0.0 to 1.0) on the nuScenes dataset. Baseline (unperturbed) SGC scores for the best-performing (Cosmos, 0.0722) and worst-performing (Latte, 0.3226) generative models on the GenWorld dataset are included as dashed teal and red reference lines, respectively. This controlled test confirms that high SGC scores directly measure geometric failure and not semantics or texture shifts.

4.3 Controlled Validation of Geometric Sensitivity

Synthetic Perturbations on Real Videos. To address potential domain shift criticisms and verify that our metric isolates true geometric integrity, we conduct a causal-style controlled validation. We systematically inject three types of synthetic perturbations into the high-fidelity nuScenes dataset. To accurately simulate the progressive structural collapse often observed in generative models, the perturbation intensity scales linearly with time, compounding the distortions in later frames. We evaluate three distinct corruption paradigms: (1) **Warping:** A time-varying, non-rigid sinusoidal deformation applied jointly to the RGB image and 2D feature tracks, simulating continuous spatial drift while holding depth estimations constant. (2) **Perspective Squeeze:** A global anisotropic vertical compression applied exclusively to the 2D tracks. By keeping the RGB image static, this creates a severe cross-modal conflict between the apparent visual scale and the underlying multi-view geometry. (3) **Spatial Depth Warp:** A synchronized spatial remapping applied simultaneously across the depth map, RGB image, and 2D tracks. This aggressively disrupts temporal consistency by inducing continuous depth boundary drift and non-rigid trajectory shifts.

Sensitivity and Monotonicity. By strictly fixing the semantic content and texture quality, we evaluate SGC responsiveness purely to geometric degradation. As Fig. 5 illustrates, SGC exhibits a strict monotonic response to increasing perturbation severity (from 0.0 to 1.0) across all three corruption types. From a

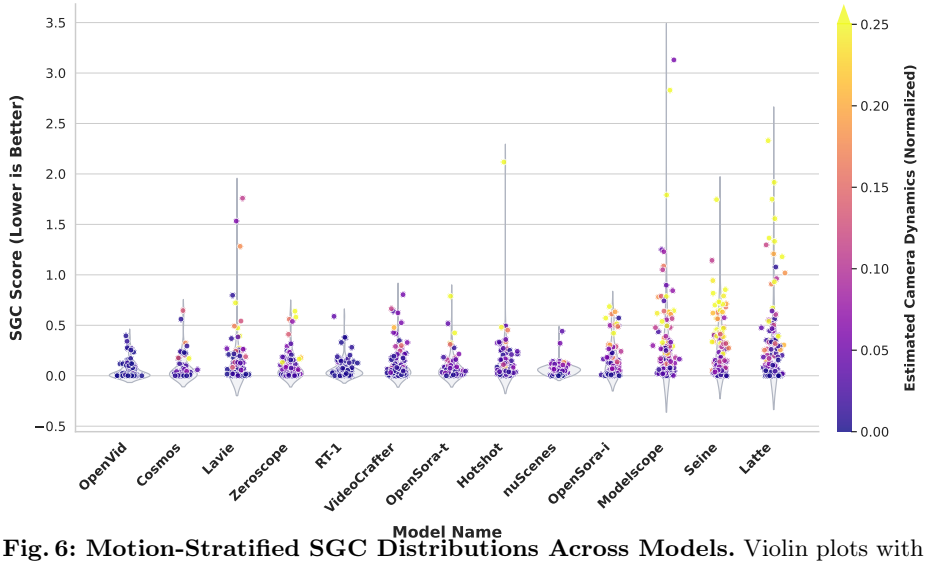


Fig. 6: Motion-Stratified SGC Distributions Across Models. Violin plots with jittered samples showing the distribution of SGC scores (lower is better) for each model. Individual points are colored by the normalized Estimated Camera Dynamics (ECD).

clean baseline score of 0.0613, the controlled injection of warping steadily inflates the SGC error to 0.0979. Spatial Depth Warp and Perspective Squeeze prove significantly more destructive to 3D consistency, precipitating steep degradation trajectories that culminate in SGC scores of 0.2252 and 0.2401, respectively.

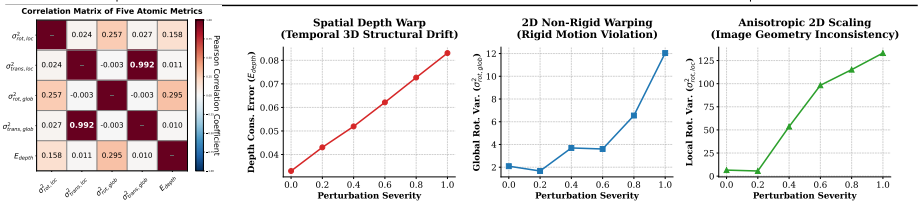
The clear monotonic trend and strong rank correlation with perturbation severity empirically validate our core design. Ultimately, these controlled tests demonstrate that SGC reliably tracks underlying 3D spatial instabilities independent of visual appearance. This significantly reduces the likelihood that observed performance disparities between models are merely artifacts of semantics or texture shifts, proving SGC strictly isolates and measures geometric failure.

4.4 Robustness, Disentanglement, and Diagnostic Analysis

Robustness to Camera Motion. A reliable 3D consistency metric must decouple true geometric instability from the natural artifacts of camera movement. Existing metrics, such as the VBench Dynamic Degree (DD), conflate object dynamics with camera dynamics. To explicitly isolate camera motion, we introduce Estimated Camera Dynamics (ECD), derived from the global relative pose. We compute the average translational and rotational magnitudes per frame, min-max normalize them across the dataset, and define ECD as the maximum of these scaled values. On real-world datasets, SGC exhibits only weak to moderate correlations with ECD (RT-1: $r = 0.44$, $p < 0.001$; nuScenes: $r = 0.52$, $p < 0.001$; OpenVid: $r = 0.17$, $p = 0.09$). The corresponding coefficients of determination (R^2) indicate that camera dynamics explain at most 27% of the SGC variance in RT-1 and nuScenes, and less than 3% in OpenVid. This confirms that

Table 2: Ablation study on SGC components across generative methods and real-world datasets. We report SGC scores (↓) for three variants: without moving object segmentation (w/o MOS), grid-based depth clustering, and the full model (SGC-full). Method names correspond to cited works.

Variant	Cosmos	Hotshot	Latte	Lavie	Modelscope	OpenSora-i	OpenSora-t	Seine	VideoCrafter	Zeroscope	OpenVid	RT-1	nuScenes
w/o MOS	0.052	0.104	0.286	0.133	0.222	0.157	0.080	0.259	0.086	0.075	0.014	0.163	0.064
Grid depth	0.067	0.109	0.353	0.170	0.314	0.190	0.107	0.306	0.101	0.071	0.062	0.093	0.163
Depth + Spatial	0.082	0.133	0.555	0.161	0.483	0.205	0.101	0.365	0.086	0.084	0.019	0.051	0.153
SGC-full	0.072	0.117	0.323	0.124	0.313	0.163	0.083	0.284	0.097	0.091	0.053	0.064	0.061



(a) **(b)**

Fig. 7: Validation of SGC components. Left: Pearson correlation matrix shows low inter-metric redundancy. Right: Response curves for component-specific perturbations.

SGC is inherently robust to natural camera motion, with motion explaining only a minor fraction of error variation in real data.

Disentangling Geometry from Motion. To verify that SGC penalizes structural inconsistency rather than motion intensity, we conduct a motion-stratified visual analysis (Fig. 6). While generated videos exhibit a stronger correlation between motion and SGC error (e.g., Latte: $r = 0.79$, Modelscope: $r = 0.47$), this sensitivity does not fully account for their elevated scores. Crucially, geometrically inconsistent models yield high SGC errors even under near-zero camera dynamics (dark-colored points), indicating a foundational deficit in maintaining 3D structure that is independent of motion. Conversely, real-world anchors securely maintain tightly bounded, low-SGC distributions (predominantly < 0.5) even at the upper bounds of estimated camera motion (bright-colored points). This demonstrates that structural geometric collapse, rather than motion intensity, is the dominant factor driving SGC degradation in generative models.

Complementarity to Fidelity Metrics. By successfully isolating spatial instability from motion artifacts, SGC serves as a critical complement to standard fidelity metrics such as Fréchet Video Distance (FVD). Conventional metrics are heavily biased toward 2D texture plausibility and frame-level visual quality, frequently overlooking severe underlying structural degradation. The consistent separation of SGC scores between real and generated videos—irrespective of camera dynamics—highlights its unique capability to diagnose foundational 3D geometric failures that traditional metrics fail to capture.

4.5 Ablation of SGC Design Choices

Full-Dataset Architectural Ablations. We validate SGC core design choices by ablating individual modules across 10 generative models and 3 real-world anchors (Tab. 2). First, we examine Moving Object Segmentation (MOS). Removing this module (w/o MOS) disrupts evaluating dynamic environments; on RT-1, the error increases from 0.064 to 0.163, as valid foreground articulation is erroneously penalized as background instability. Conversely, on largely static-camera

datasets like OpenVid, removing MOS yields a lower error (0.014) than the full pipeline (0.053). This mild full-model increase occurs because segmentation in largely fixed-view scenes may introduce boundary noise or mask leakage, culling otherwise reliable background points and slightly weakening spatial constraints for pose estimation. Nonetheless, for in-the-wild videos with salient dynamic foregrounds, MOS remains important to prevent object motion from corrupting geometric evaluation. Further sensitivity analysis on how MOS quality impacts the SGC score is detailed in App.D.2.

Next, we analyze the spatial partitioning strategy. Replacing depth-aware plane clustering with a naive spatial grid (Grid depth) degrades performance in complex scenes (e.g., nuScenes error increases from 0.061 to 0.163), indicating grid-based partitioning fails to isolate distinct geometric structures. To further interrogate our 1D depth clustering, we evaluate a Depth + Spatial variant to address the hypothesis that clustering by depth alone might group non-contiguous surfaces and yield spatially insufficient regions for PnP estimation. While 1D depth clustering can merge disjoint surfaces (such as forming thin horizontal bands across flat roads), this behavior can be beneficial in practice. Aggregating non-contiguous but co-planar patches often increases the effective image-plane span of correspondences, typically improving the conditioning of PnP estimates (especially for orientation components). Enforcing spatial contiguity can oversegment expansive surfaces into compact sub-regions; these spatially restricted regions tend to yield less stable local pose estimates, inflating the real-world nuScenes error to 0.153 and increasing instability penalties across generative models (e.g., Latte increases from 0.323 to 0.555). Overall, these results support allowing 1D depth clustering to merge disjoint regions when doing so preserves a broad spatial distribution, improving robust geometric verification.

Component Construct Validity. To verify SGC aggregated components capture distinct, interpretable failure modes, we analyze their empirical correlations and responses to controlled degradations. A Pearson correlation matrix across the dataset reveals most sub-metrics capture independent failure modes (Fig. 7a). While local ($\sigma^2_{trans,loc}$) and global ($\sigma^2_{trans,glob}$) translational variances exhibit high correlation ($r \approx 0.992$), this redundancy is an intentional architectural choice. This specific pair performs a hierarchical diagnostic check, comparing internal structural rigidity against global scene coherence, securely anchoring the PCA-weighted aggregation. To further validate this aggregation strategy, a cross-validation confirming the stability and generalizability across diverse video domains is detailed in the App.D.1. Crucially, targeted synthetic perturbations verify individual SGC components directly map to their intended physical interpretations (Fig. 7b). Each controlled degradation explicitly isolated its corresponding metric: continuous warping exclusively inflated global pose divergence, driving global rotation variance from a baseline of 2.090 to 12.042. Applying the Spatial Depth Warp predictably triggered depth inconsistency, raising depth error from 0.033 to 0.083. Finally, simulating perspective collapse via the Squeeze perturbation induced severe cross-plane rigidity violations, causing local rotation variance to spike from 6.522 to 133.091. This diagnostic alignment confirms SGC

Table 3: Ranking consistency across geometric estimators. NuScenes is excluded because Any4D [25] fails on this outdoor dataset. All correlations are significant at $p < 0.001$. Ranks per estimator are shown as superscripts.

Estimator	OpenVid	RT-1	Cosmos	OpenSora-t	Zeroscope	VideoCrafter	Hotshot	Lavie	OpenSora-i	Seine	Modelscope	Latte	ρ vs Ori.
Original	0.053^{1}	0.064^{2}	0.072^{3}	0.083^{4}	0.091^{5}	0.097^{6}	0.117^{7}	0.124^{8}	0.163^{9}	0.284^{10}	0.313^{11}	0.323^{12}	1.000
w/o VGGT	0.077^{1}	0.097^{2}	0.107^{3}	0.123^{5}	0.122^{4}	0.153^{6}	0.183^{7}	0.190^{8}	0.198^{9}	0.405^{10}	0.460^{11}	0.472^{12}	0.993
Any4D [25]	0.080^{2}	0.062^{1}	0.170^{3}	0.226^{5}	0.347^{9}	0.224^{4}	0.324^{8}	0.322^{7}	0.297^{6}	0.836^{11}	0.979^{12}	0.809^{10}	0.860
Page4D [55]	0.072^{2}	0.062^{1}	0.100^{5}	0.091^{4}	0.115^{6}	0.091^{3}	0.119^{7}	0.121^{8}	0.164^{9}	0.310^{11}	0.271^{10}	0.332^{12}	0.937

is not merely a generalized error scalar but a composite of necessary components rigorously isolating and identifying specific 3D geometric failures.

Estimator Robustness and Reliability. A potential concern regarding SGC is its reliance on foundational estimators like VGGT, which are primarily trained on static scenes, for dynamic video evaluation. To validate that SGC reliably captures intrinsic geometric failures rather than estimator artifacts, we conducted a robustness analysis replacing VGGT with state-of-the-art dynamic-scene estimators (Any4D [25] and Page4D [55]), alongside an ablation entirely removing the VGGT global pose component.

As shown in Tab. 3, substituting or removing the underlying estimator preserves remarkably strong rank correlations (Spearman $\rho \geq 0.860$). This confirms that the severe spatial inconsistencies detected in generative models are foundational structural failures, intrinsically captured by SGC within the evaluated estimators and datasets. Note that nuScenes was excluded from this specific subset analysis, as current dynamic-scene estimators like Any4D struggle to produce stable reconstructions in such complex, large-scale outdoor environments.

5 Conclusion

We presented SGC, a novel diagnostic metric for 3D spatial geometric consistency, addressing a gap where existing fidelity (e.g., FVD) and consistency (e.g., VBench-BC, MEt3R) metrics fail. These metrics are often insensitive to geometric warping or are confounded by valid foreground motion, respectively. SGC overcomes this by employing a physically grounded principle: all static background points must adhere to a single camera transformation. It enforces this by isolating the static background $\mathcal{M}_{static}$ (avoiding penalties for valid motion) and measuring the variance among local pose estimates to quantify geometric integrity. Ablations confirmed isolating motion is essential for fairness, and we proved SGC is robust to camera motion magnitude. Using this validated metric, we find many SOTA models suffer from significant geometric failures missed by other metrics, despite high FVD or VBench scores. SGC thus serves as an essential, complementary diagnostic to guide research towards generating truly geometrically coherent videos.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62441616, Grant 62321005, Grant 62125603, and Grant 62336004.

References

1. Hotshot. https://hotshot.co/ (2023)
2. Zeroscope-v2-xl. https://zeroscope.replicate.dev/ (2024)
3. 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)
4. Allen, K., Doersch, C., Zhou, G., Suhail, M., Driess, D., Rocco, I., Rubanova, Y., Kipf, T., Sajjadi, M.S., Murphy, K., et al.: Direct motion models for assessing generated videos. arXiv preprint arXiv:2505.00209 (2025)
5. Asim, M., Wewer, C., Wimmer, T., Schiele, B., Lenssen, J.E.: Met3r: Measuring multi-view consistency in generated images. In: CVPR (2024)
6. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520 (2024)
7. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al.: Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
8. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
9. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR. pp. 22563–22575 (2023)
10. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)
11. Burgert, R., Xu, Y., Xian, W., Pilarski, O., Clausen, P., He, M., Ma, L., Deng, Y., Li, L., Mousavi, M., et al.: Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In: CVPR. pp. 13–23 (2025)
12. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR. pp. 11621–11631 (2020)
13. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: CVPR. pp. 7310–7320 (2024)
14. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. arXiv:2501.12375 (2025)
15. Chen, W., Zheng, W., Zheng, Y., Chen, L., Zhou, J., Lu, J., Duan, Y.: Genworld: Towards detecting ai-generated real-world simulation videos (2025), https://arxiv.org/abs/2506.10975
16. Chen, X., Wang, Y., Zhang, L., Zhuang, S., Ma, X., Yu, J., Wang, Y., Lin, D., Qiao, Y., Liu, Z.: Seine: Short-to-long video diffusion model for generative transition and prediction. In: ICLR (2023)
17. Ge, S., Mahapatra, A., Parmar, G., Zhu, J.Y., Huang, J.B.: On the content bias in fréchet video distance. In: CVPR. pp. 7277–7288 (2024)
18. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: ICCV. pp. 22930–22941 (2023)

19. Geng, D., Herrmann, C., Hur, J., Cole, F., Zhang, S., Pfaff, T., Lopez-Guevara, T., Aytar, Y., Rubinstein, M., Sun, C., et al.: Motion prompting: Controlling video generation with motion trajectories. In: CVPR. pp. 1–12 (2025)
20. Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., Wang, W., Liu, Y.: Diffusion as shader: 3d-aware video diffusion for versatile video generation control. arXiv preprint arXiv:2501.03847 (2025)
21. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
22. Huang, N., Zheng, W., Xu, C., Keutzer, K., Zhang, S., Kanazawa, A., Wang, Q.: Segment any motion in videos (2025), https://arxiv.org/abs/2503.22268
23. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR. pp. 21807–21818 (2024)
24. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. arXiv preprint arXiv:2410.05954 (2024)
25. Karhade, J., Keetha, N., Zhang, Y., Gupta, T., Sharma, A., Scherer, S., Ramanan, D.: Any4d: Unified feed-forward metric 4d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14578–14589 (2026)
26. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.C., et al.: Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125 (2023)
27. 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)
28. Li, X., Zhou, D., Zhang, C., Wei, S., Hou, Q., Cheng, M.M.: Sora generates videos with stunning geometrical consistency (2024), https://arxiv.org/abs/2402.17403
29. Li, Y., Angel, M.C., Khan, S., Zhu, Y., Sun, J., Zhang, Y., Khan, F.S.: C-drag: Chain-of-thought driven motion controller for video generation (2025), https://arxiv.org/abs/2502.19868
30. Liu, J., Qu, Y., Yan, Q., Zeng, X., Wang, L., Liao, R.: Fr\'echet video motion distance: A metric for evaluating motion consistency in videos. arXiv preprint arXiv:2407.16124 (2024)
31. Luo, G.Y., Favero, G.M., Luo, Z.H., Jolicoeur-Martineau, A., Pal, C.: Beyond fvd: Enhanced evaluation metrics for video generation quality. arXiv preprint arXiv:2410.05203 (2024)
32. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
33. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:2407.02371 (2024)
34. Ngo, T.D., Zhuang, P., Gan, C., Kalogerakis, E., Tulyakov, S., Lee, H.Y., Wang, C.: Delta: Dense efficient long-range 3d tracking for any video. arXiv preprint arXiv:2410.24211 (2024)
35. Ni, H., Egger, B., Lohit, S., Cherian, A., Wang, Y., Koike-Akino, T., Huang, S.X., Marks, T.K.: Ti2v-zero: Zero-shot image conditioning for text-to-video diffusion models. In: CVPR. pp. 9015–9025 (2024)

36. Ren, W., Yang, H., Zhang, G., Wei, C., Du, X., Huang, W., Chen, W.: Consisti2v: Enhancing visual consistency for image-to-video generation. arXiv preprint arXiv:2402.04324 (2024)
37. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR. pp. 4104–4113 (2016)
38. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
39. Skorokhodov, I., Menapace, W., Siarohin, A., Tulyakov, S.: Hierarchical patch diffusion models for high-resolution video generation. In: CVPR. pp. 7569–7579 (2024)
40. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
41. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022)
42. Wang, H., Ma, C.Y., Liu, Y.C., Hou, J., Xu, T., Wang, J., Juefei-Xu, F., Luo, Y., Zhang, P., Hou, T., et al.: Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity. In: CVPR. pp. 2578–2588 (2025)
43. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025)
44. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
45. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. IJCV **133**(5), 3059–3078 (2025)
46. Watson, D., Chan, W., Martin-Brualla, R., Ho, J., Tagliasacchi, A., Norouzi, M.: Novel view synthesis with diffusion models. arXiv preprint arXiv:2210.04628 (2022)
47. Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., Wong, T.T.: Dynamicrafter: Animating open-domain images with video diffusion priors. In: ECCV. pp. 399–417. Springer (2024)
48. Yu, J.J., Forghani, F., Derpanis, K.G., Brubaker, M.A.: Long-term photometric consistent novel view synthesis with diffusion models. In: ICCV. pp. 7094–7104 (2023)
49. Yuan, X., Baek, J., Xu, K., Tov, O., Fei, H.: Inflation with diffusion: Efficient temporal adaptation for text-to-video super-resolution. pp. 489–496 (2024)
50. Zeng, Y., Wei, G., Zheng, J., Zou, J., Wei, Y., Zhang, Y., Li, H.: Make pixels dance: High-dynamic video generation. In: CVPR. pp. 8850–8860 (2024)
51. Zhang, D.J., Li, D., Le, H., Shou, M.Z., Xiong, C., Sahoo, D.: Moonshot: Towards controllable video generation and editing with multimodal conditions. arXiv preprint arXiv:2401.01827 (2024)
52. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 (2023)
53. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)

54. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
55. Zhou, K., Wang, Y., Chen, G., Beaudouin, G., Zhan, F., Liang, P.P., Wang, M.: PAGE-4d: VGGT-4d perception via disentangled pose and geometry estimation. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=Nfmzp5PBzr