PDI-Bench: Quantitative Video World Model Evaluation for Geometric Consistency

Jiaxin Wu[1], Yihao Pi[1], Yinling Zhang[2], Yuheng Li[3], and Xueyan Zou[1]

[1] Tsinghua University – IEI Lab
[2] University of Wisconsin–Madison
[3] Adobe Research

Abstract. Generative video models are increasingly studied as implicit world models, yet evaluating whether they produce physically plausible 3D structure and motion remains challenging. Most existing video evaluation pipelines rely heavily on human judgment or learned graders, which can be subjective and weakly diagnostic for geometric failures. We introduce **PDI-Bench** (**P**erspective **D**istortion **I**ndex), a quantitative framework for auditing *geometric coherence* in generated videos. Given a generated clip, we obtain object-centric observations via segmentation and point tracking (e.g., SAM 2, MegaSaM and CoTracker3), lift them to 3D cues with monocular reconstruction, and compute a set of projective-geometry residuals capturing three failure dimensions: *scale–depth alignment*, *3D motion consistency*, and *3D structural rigidity*. To support systematic evaluation, we build **PDI-Dataset**, covering diverse scenarios designed to stress these geometric constraints. Across state-of-the-art video generators, PDI reveals geometry-specific failure modes and provides a diagnostic signal for progress toward physically grounded video generation and physical world model. Our code and dataset can be found at `https://pdi-bench.github.io/`.

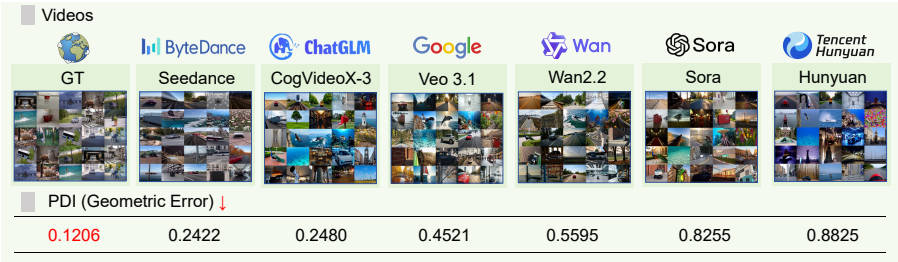


Fig. 1: Overview of the PDI-Bench Evaluation. (Top) Qualitative samples from our dataset, featuring real-world Ground Truth (GT) videos and generated sequences from state-of-the-art models. **(Bottom)** The corresponding **PDI-Scores** for GT and each model. Lower scores indicate better adherence to 3D physical laws (scale alignment, motion consistency, and structural rigidity).

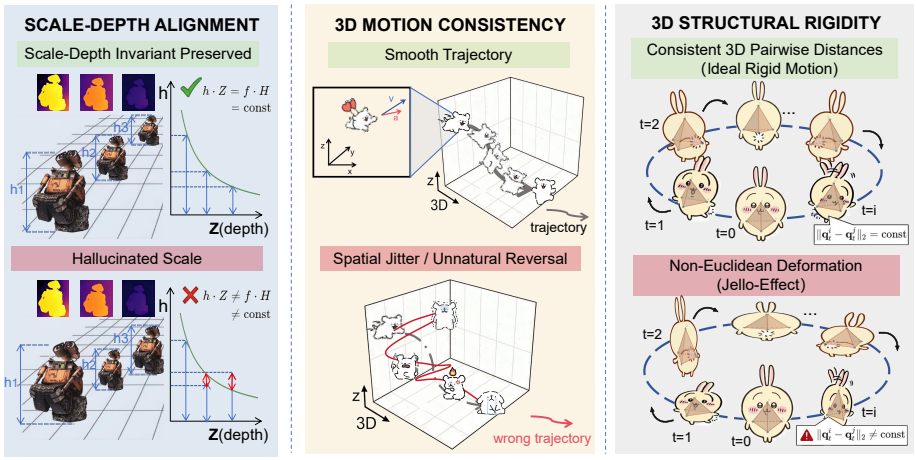


Fig. 2: The Three key perspectives that PDI-Bench is evaluating for geometric consistency. (Left) Scale-Depth Alignment: Based on the pinhole camera model, we verify if the product of projected height h and depth Z remains constant. **(Middle) 3D Motion Consistency:** We audit the smoothness of 3D world trajectories. **(Right) 3D Structural Rigidity:** By monitoring internal 3D pairwise distances between anchors (skeleton $\mathbf{q}^i, \mathbf{q}^j$), we detect non-Euclidean deformations.

1 Introduction

High-fidelity generative video models like Seedance [5], Veo 3.1 [10], and Sora [15, 17] have reshaped content creation with their unprecedented visual realism. This impressive quality has led many to view them as early "World Models". This terminology implies a critical shift from simple 2D pixel interpolation to a deeper, latent understanding of 3D structures and physical laws.

Despite their visual realism, a significant gap remains between visual plausibility and geometric rigor. Current state-of-the-art models frequently struggle with **spatial scale** and **perspective consistency**. While sometimes subtle, these artifacts clearly violate fundamental Euclidean properties. Common failure modes include “volume breathing,” where rigid objects unrealistically expand or contract, and “skating,” where an object’s motion is decoupled from the ground plane’s perspective. These geometric flaws stem from a lack of explicit structural constraints during generation, exacerbated by the absence of metrics capable of evaluating 3D properties from 2D videos. Existing metrics, such as Fréchet Video Distance (FVD) [22] or CLIP-based scores [18], rely on pixel distributions or semantic features, rendering them inherently “geometry-blind.” They fail to penalize physical errors, such as a train shrinking disproportionately to its velocity. To progress toward true physical simulation, the field needs a new evaluation standard that assesses 2D video motion using the strict rules of 3D projective geometry.

While evaluating physical common sense has emerged as a key research frontier, existing suites remain insufficient for *rigorous geometric verification*. Current pipelines typically depend on human judgment or rely on Large Multimodal

Models (LMMs) as proxy evaluators, which are prone to subjective bias. Automated benchmarks such as **PhysBench** [6] and **WorldBench** [23] improve scalability, yet primarily examine high-level physical phenomena or estimate scene-level physical properties. At a finer temporal level, **TRAJAN** [1] learns to reconstruct 2D point tracks and uses the resulting motion representation to detect and localize temporal inconsistencies. While effective for motion-quality assessment, it does not explicitly separate camera motion from object motion in a common 3D coordinate system or decompose failures into interpretable projective and structural residuals. PDI-Bench complements these approaches by lifting monocular video observations into world space and auditing three explicit geometric invariants: scale–depth alignment, 3D motion consistency, and 3D structural rigidity. This decomposition produces interpretable residuals that reveal not only whether a video is geometrically inconsistent, but also how the inconsistency arises.

As shown in Fig. 2, **PDI-Bench** evaluates physical realism via three orthogonal geometric metrics. *Scale-depth alignment* enforces the inverse correlation between an object's projected 2D height and its 3D depth to penalize unnatural scale hallucinations. *3D motion consistency* evaluates the object's 3D centroid motion in world coordinates, penalizing non-physical accelerations and abrupt angular shifts (e.g., spatial jitter or teleports) while fully decoupling camera movements. *3D structural rigidity* tracks internal 3D point-pair distances over time to penalize localized, non-Euclidean deformations (e.g., the "jello effect") and preserve rigid-body integrity.

To move beyond subjective visual assessment, PDI-Bench extracts 3D geometric evidence from 2D pixels via a "perception-to-reasoning" **Target-Uplift-Anchor** workflow. First, **Semantic Targeting** via SAM 2 [19] isolates the subject to establish precise 2D spatial boundaries and scale priors (h). Next, **3D Geometric Uplifting** via MegaSaM [14] reconstructs world-coordinate pointmaps and camera poses directly from the monocular sequence, lifting 2D observations into a unified 3D physical environment while fully decoupling camera ego-motion. Finally, **3D Structural Anchoring** via CoTracker3 [11] deploys dense point anchors within the subject mask. By utilizing the 2D trajectories as precise spatial indices into the aforementioned 3D pointmaps, we lift visual cues into structurally meaningful 3D trajectories ($\mathbf{q}_t^n$), enabling camera-motion-invariant evaluation of both Motion consistency and structural rigidity.

In summary, we claim the following contributions:

- We propose **PDI-Bench**, a quantitative framework that translates 2D pixel dynamics into 3D geometric reasoning to detect physical hallucinations.
- We introduce the **Perspective Distortion Index**, a metric system that utilizes 3D lifting to decouple kinematic errors from geometric alignment.
- We present **PDI-Dataset**, dedicated to geometric consistency, featuring real-world data and 6 SoTA open/closed-source video models.
- We demonstrate that PDI-Bench effectively identifies subtle spatial inconsistencies across five stress-test scenarios, offering critical insights for developing the next generation of space-aware generative systems.

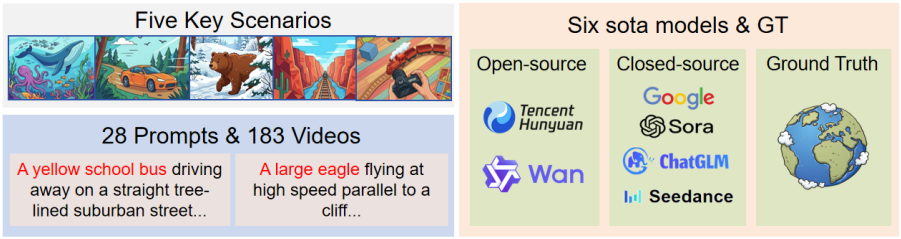


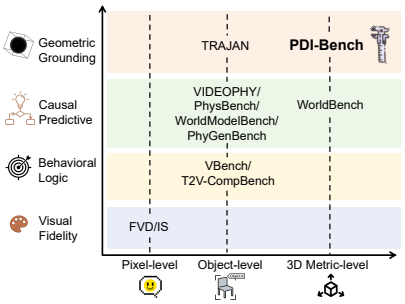
Fig. 3: Overview of PDI-Dataset and the experimental setup. The benchmark contains 183 real and generated videos spanning five geometric stress-test scenarios and evaluates six representative open- and closed-source video generators.

2 Related Work

Video Generation and World Model. Recent breakthroughs in diffusion-based generative models have significantly advanced the state of video synthesis. Early works focused on temporal extensions of 2D diffusion processes [3, 7]. More recently, large-scale models such as Sora [15], and Wan2.2 [25] have demonstrated a remarkable ability to generate high-fidelity, long-duration sequences with complex semantic alignment. The impressive visual quality has led to the conceptualization of these systems as "World Models", implying an internal representation of physical laws. However, whether these models truly simulate 3D space or simply replicate 2D statistical patterns remains a subject of intense debate. Our work aims to provide a geometric yardstick to quantify this distinction.

Video Quality Assessment and Prompt Following. The evaluation of generative videos has traditionally relied on distribution-based metrics such as Fréchet Video Distance (FVD) [22] and Inception Score (IS) [20], which prioritize frame-level textural and aesthetic quality over structural integrity. To assess text-video semantic alignment and prompt following, CLIP-based scores [18] have become the standard. Building on these foundations, comprehensive suites like VBench [9] and T2V-CompBench [21] have recently been proposed to provide multi-dimensional evaluations, encompassing dynamic quality, temporal consistency, and compositional prompt adherence.

Physical Consistency and Scene Understanding. The transition from video synthesis to world simulation requires evaluating true physical realism rather than mere perceptual quality. While recent benchmarks assess macroscopic physical laws—such as VIDEOPHY [2] and WorldModelBench [13] using VLM judges, PhysBench [6] querying interaction dynamics, PhyGenBench [16] testing dynamic constraints, and WorldBench [23] fitting physical constants—they primarily probe high-level semantic plausibility. Closer to our approach, **TRAJAN** [1] employs a trajectory autoencoder to detect anomalies in point tracks. However, it measures *implicit consistency* through learned representations. PDI-Bench fills a critical gap by enforcing *explicit physical laws* as quantitative constraints. Instead of evaluating high-level plausibility, our metric audits fundamental scale–depth alignment, 3D motion consistency, and structural rigidity to expose the spatial integrity of world models.



Visual Perception and 3D Reconstruction. The feasibility of 3D physical auditing is underpinned by the surge in foundation models for visual perception. Video segmentation has been revolutionized by memory-based propagation models like SAM 2 [19], while temporal correspondence has reached pixel-level precision through dense trackers such as CoTracker3 [11]. Furthermore, robust 3D reconstruction from dynamic monocular sequences has become accessible via semantic-aware SfM frameworks like MegaSaM [14]. These perceptual backbones serve as the foundational sensors in our pipeline, projecting 2D video dynamics into a verifiable 3D world coordinate system.

3 Method

The core objective of **PDI-Bench** is to bridge the gap between 2D pixel dynamics and 3D physical regularities. We operationalize this objective through a multi-stage *Target-Uplift-Anchor* pipeline, culminating in a three-dimensional geometric audit quantified by our proposed **Perspective Distortion Index(PDI)**.

3.1 Target-Uplift-Anchor: Perceptual Pipeline

As shown in Fig. 4, we orchestrate a hierarchical perception-to-reasoning stack through a collaborative *Target-Uplift-Anchor* workflow to extract the physical evidence required for geometric auditing.

Semantic Targeting (SAM 2). We initiate the pipeline by identifying the auditing subject using **Florence-2** [26] for automated text-to-box prompting. These prompts are then fed into **SAM 2** [19] to generate and propagate a temporal sequence of binary masks $\{M_t\}_{t=1}^{T}$. From these masks, we derive the instantaneous pixel height h_t and the 2D spatial boundaries.

3D Geometric Uplifting (MegaSaM). To recover the latent 3D physical environment, we employ **MegaSaM** [14] to obtain a coherent depth sequence $\{Z_t\}_{t=1}^{T}$, the estimated focal length f, and camera poses. More importantly, MegaSaM projects every pixel into a unified 3D world coordinate system, yielding world-space pointmaps $\mathbf{P}_{world} \in \mathbb{R}^{T\times H\times W\times 3}$. This critical step lifts 2D observations into pure 3D space, completely decoupling object kinematics from camera ego-motion.

3D Structural Anchoring (CoTracker3). With the 3D world-space constructed, we deploy **CoTracker3** [11] to monitor the subject's internal structural integrity. Within the region defined by the initial mask M_1, we seed anchor queries via a SIFT $\rightarrow$ Shi-Tomasi $\rightarrow$ uniform-grid cascade. After filtering by visibility and displacement-jump thresholds, we obtain a set of reliable 2D pixel-space trajectories $\{(u_t^n, v_t^n)\}$. Using CoTracker3's 2D pixel-space trajectories $\{(u_t^n, v_t^n)\}$ as spatial indices into the MegaSaM world-space pointmaps $\mathbf{P}_{world}$, we lift each

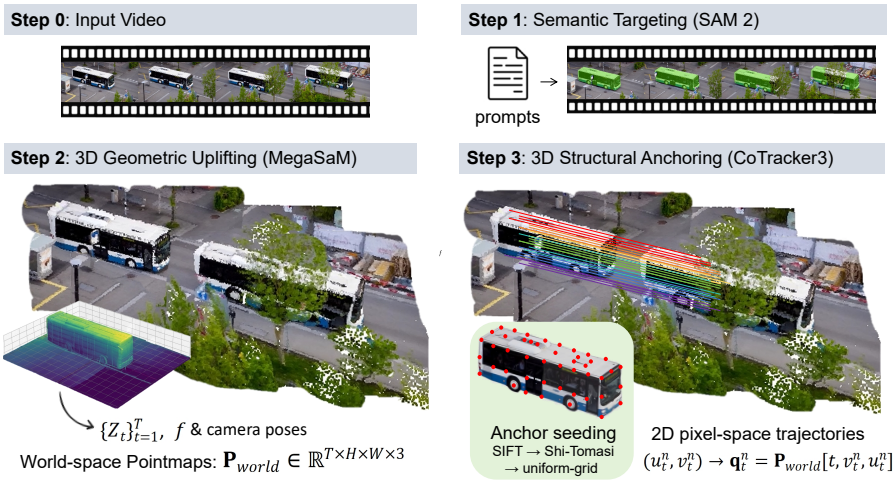


Fig. 4: Overview of the PDI-Bench Pipeline. Our *Target-Uplift-Anchor* pipeline initiates with **Semantic Targeting** (SAM 2) for object isolation. It then performs **Geometric Uplifting** via MegaSaM to construct 3D world-space pointmaps $\mathbf{P}_{world}$, followed by **Structural Anchoring** with CoTracker3 to lift 2D pixel-space trajectories into 3D coordinates $\mathbf{q}_t^n$. The resulting spatial-temporal data are synthesized into the PDI metric for consistency auditing.

tracked anchor to its 3D coordinate: $\mathbf{q}_t^n = \mathbf{P}_{world}[t, v_t^n, u_t^n]$. This operation effectively transforms 2D visual tracking into structurally meaningful 3D trajectories for subsequent rigidity auditing.

Finally, the extracted multi-dimensional features from this *Target-Uplift-Anchor* workflow are synthesized into the Perspective Distortion Index (PDI) via three orthogonal weighted residuals: ϵ_{scale}, ϵ_{traj}, and $\epsilon_{rigidity}$.

3.2 The Perspective Distortion Index (PDI)

We synthesize the multi-dimensional geometric evidence into the **Perspective Distortion Index (PDI)**, defined as a weighted sum of three orthogonal physical metrics:

$$\text{PDI} = w_1 \cdot \text{RMSE}(\epsilon_{scale}) + w_2 \cdot \text{RMSE}(\epsilon_{traj}) + w_3 \cdot \epsilon_{rigidity} \tag{1}$$

where $\sum_{i=1}^{3} w_i = 1$. Each term is designed to be scale-invariant and to capture a distinct failure mode. We apply the Root Mean Square Error (RMSE) to the scale and trajectory residuals to sensitize the index to catastrophic, high-magnitude physical hallucinations. Conversely, $\epsilon_{rigidity}$ is defined as the temporal mean of a robust dispersion statistic (MAD). We forgo RMSE for this term to avoid a redundant second-order penalty on an already aggregated measure of spatial incoherence, thereby preserving the direct physical interpretability of structural instability.

Scale-Depth Alignment (ϵ_{scale}) To establish a physical baseline, we model the generation process via pinhole camera geometry. For an object with physical height H and depth Z, its projected pixel height h satisfies $h = f \cdot H/Z$. Since f and H are constant for a rigid body, we derive the **Scale-Depth Invariant**:

$$h_t \cdot Z_t = f \cdot H = \text{Constant}, \quad \text{where } h_t \in \text{SAM 2}, Z_t \in \text{MegaSaM}. \tag{2}$$

Any fluctuation in this product indicates non-physical scaling (e.g., "volume breathing"). We quantify this via a log-space residual $\epsilon_{scale}^{(t)}$ to ensure symmetric penalties for expansions and contractions:

$$\epsilon_{scale}^{(t)} = \left|\ln(h_t \cdot Z_t) - \text{median}_{k\in[1,5]}\left(\ln(h_k \cdot Z_k)\right)\right| \tag{3}$$

where the median of the first five frames establishes a stable baseline against initialization noise. The final RMSE(ϵ_{scale}) serves as a robust measure of scaling severity, where h_t and Z_t are the per-frame pixel height and median object depth, respectively.

3D Motion Consistency (ϵ_{traj}) Rather than relying on 2D projective cues, we audit kinematic plausibility directly in the MegaSaM world coordinate system, which fully decouples object motion from camera ego-motion.

Centroid Extraction & Kinematics. The per-frame 3D foreground centroid $\mathbf{C}_t$ is computed as the coordinate-wise median of all foreground points in the world-space pointmaps $\mathbf{P}_{world}$ masked by M_t. After applying temporal median filtering ($k=3$) to suppress depth flickering noise, we compute the frame-rate-normalized 3D velocity and acceleration using the frame interval $\Delta t = 1/\text{fps}$:

$$\mathbf{v}_t = (\mathbf{C}_{t+1} - \mathbf{C}_t)/\Delta t, \qquad \mathbf{a}_t = (\mathbf{v}_{t+1} - \mathbf{v}_t)/\Delta t \tag{4}$$

To evaluate adherence to Newtonian inertia, we decompose kinematic anomalies into two complementary components: abnormal acceleration magnitude and unnatural directional shifts.

1. Acceleration Magnitude Penalty ($\tilde{a}_t$). To quantify spatial jitter without being biased by the absolute reconstruction scale, we define a dimensionless relative acceleration ratio r_t and its soft-saturated counterpart $\tilde{a}_t$ as

$$r_t = \frac{\|\mathbf{a}_t\|_2 \Delta t}{v_{\text{ref}}}, \qquad \tilde{a}_t = 2\tanh\left(\frac{r_t}{5}\right) \in [0, 2), \tag{5}$$

where

$$v_{\text{ref}} = \max\left(\underset{k}{\text{median}}\,\|\mathbf{v}_k\|_2,\, 2\Delta t\,\underset{k}{\text{median}}\,\|\mathbf{a}_k\|_2,\, \varepsilon_v\right) \tag{6}$$

is a robust reference speed, and $\varepsilon_v = 10^{-6}$ prevents numerical instability for near-stationary sequences. Multiplying the acceleration magnitude by the frame interval Δt converts it into a velocity-change magnitude, making r_t dimensionless. The tanh compression remains approximately linear for small deviations while bounding the influence of extreme outliers.

2. Directional Continuity Penalty (φ_t). Macroscopic objects cannot undergo instantaneous sharp turns without external forces. We penalize abrupt directional reversals via the cosine dissimilarity of consecutive velocity vectors. To prevent micro-tremors from triggering false penalties when the object is functionally stationary, this metric is activated only when the instantaneous speed exceeds a noise threshold ($\|\mathbf{v}\| > 0.1\, v_{\text{ref}}$):

$$\varphi_t = \begin{cases} 1 - \cos\angle(\mathbf{v}_{t-1},\, \mathbf{v}_t), & \text{if } \|\mathbf{v}_{t-1}\|, \|\mathbf{v}_t\| > 0.1\, v_{\text{ref}} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

Note that φ_t naturally falls within $[0, 2]$, where 0 indicates identical directions and 2 indicates a complete reversal, aligning perfectly with the scale of the magnitude penalty.

Final Residual. The magnitude and directional penalties are dimensionally aligned and combined with equal weights to form the holistic trajectory residual:

$$\epsilon_{traj}^{(t)} = 0.5 \cdot \tilde{a}_t + 0.5 \cdot \varphi_t \tag{8}$$

Crucially, this residual is absolutely orthogonal to ϵ_{scale}: it depends solely on the 3D centroid $\mathbf{C}_t$ sampled from world-space pointmaps, completely independent of the 2D bounding box height h_t.

Structural Rigidity ($\boldsymbol{\epsilon_{rigidity}}$) To quantify non-physical internal deformations (e.g., the "jello effect" or "volume breathing"), we audit the structural cohesion of the object across time based on 3D pairwise distance consistency.

We obtain the 3D world-space coordinates $\mathbf{q}_t^n$ by sampling the MegaSaM pointmaps at the specific pixel locations defined by CoTracker3's 2D trajectories. To guarantee high-fidelity correspondence and avoid boundary artifacts (such as depth bleeding), we select a set of optimal anchor pairs at the initial frame ($t=0$) through a rigorous triple-filtering strategy: 1) **Visibility:** Anchors must maintain a CoTracker3 tracking confidence > 0.5. 2) **Depth Smoothness:** We compute the Sobel gradient magnitude of the pointmap's Z-channel and discard points in the top-quartile (> 75th percentile) gradient regions to preclude depth discontinuities. 3) **Optimal Pair Scoring:** Valid points are paired by maximizing a joint heuristic score that balances the signal-to-noise ratio (large spatial separation) and inland reliability:$\mathcal{S}_{i,j} = \|\mathbf{q}_0^i - \mathbf{q}_0^j\|_2 \times \min(D_{mask}^i, D_{mask}^j)$, where D_{mask}^i denotes the pixel distance from anchor i to the nearest mask boundary.

In classical kinematics, a rigid body requires the 3D Euclidean distance between internal points to remain constant, i.e., $\|\mathbf{q}_t^i - \mathbf{q}_t^j\|_2 = \text{Constant}$. We define the distance ratio $r_{ij}(t)$ and the robust per-frame rigidity score as:

$$r_{ij}(t) = \frac{\|\mathbf{q}_t^i - \mathbf{q}_t^j\|_2}{\|\mathbf{q}_0^i - \mathbf{q}_0^j\|_2}, \qquad \text{score}(t) = \frac{\text{MAD}\big(\{r_{ij}(t)\}\big)}{\text{median}\big(\{r_{ij}(t)\}\big) + \varepsilon} \tag{9}$$

where MAD denotes the median absolute deviation, providing immunity against monocular global scale drift. Since $t = 0$ inherently yields a zero score, it is

excluded to prevent artificial suppression. The final rigidity metric is computed as the average over the active sequence:

$$\epsilon_{rigidity} = \frac{1}{T-1} \sum_{t=1}^{T-1} \text{score}(t) \tag{10}$$

4 Experiments

In this section, we conduct a systematic evaluation of state-of-the-art generative video models using the **PDI-Bench** framework to quantify the discrepancy between visual plausibility and geometric consistency. By auditing diverse scenarios, we expose vulnerabilities in latent 3D spatial representations and provide a diagnostic roadmap for physically grounded world simulators.

4.1 Experimental Setup

PDI-Dataset and Evaluation Scenarios. To systematically audit these generative systems, we curate a comprehensive benchmark containing 183 video sequences derived from 28 diverse textual prompts. This dataset comprises 15 high-quality, real-world Ground Truth (GT) videos serving as a reliable physical baseline for calibration, alongside 168 synthetic videos. To provide a holistic view of the field's current capabilities, we evaluate six representative state-of-the-art models categorized into two tiers: **Open-source** architectures (Wan 2.2 [25], HunyuanVideo [12]) and **Closed-source** systems (Sora (OpenAI) [17], Seedance 2.0 Fast (ByteDance via Doubao) [5], CogVideoX-3 (Zhipu AI via ChatGLM) [27], and Veo 3.1-Fast (Google via Flow) [10]). Furthermore, our evaluation is specifically designed to stress-test 3D spatial awareness by covering five critical geometric challenges: (1) *Longitudinal Convergence* (Longit. Conv.), (2) *Dynamic Tracking* (Dyn. Track.), (3) *Biological Motion* (Bio. Motion), (4) *Curved Motion* (Curved Mot.), and (5) *Partial Occlusion* (Part. Occl.). For the final PDI-score calculation, we empirically set the weights for scale-depth alignment, motion convergence, and structural rigidity to $(w_{scale}, w_{traj}, w_{rigidity}) = (0.4, 0.4, 0.2)$, respectively.

4.2 Perceptual and Geometric Fidelity Guard

To ensure the integrity of the PDI-Bench pipeline, we implement a multi-stage *Fidelity Guard* that independently validates the outputs of the underlying perception models before PDI synthesis.

Semantic Segmentation Audit (SAM 2): To verify that the segmentation masks generated by SAM 2 consistently adhere to the intended target, we utilize a Vision-Language Model (VLM), specifically **Doubao** [4]. We generate the RGB-mask overlays and the corresponding binary masks and prompt the VLM to evaluate whether the highlighted region accurately covers the object described by the *text_ query*.

Point Tracking Audit (CoTracker3): We audit tracking quality using *spatiotemporal trail maps* with cyan lines denoting historical trajectories. The VLM judges whether these trails represent the target object's motion trend.

3D Reconstruction Audit (MegaSaM): The fidelity of the 3D uplifted pointmaps is validated via a *Cross-frame Reprojection Consistency* check. For a sampled pair of frames $\{I_A, I_B\}$, we re-synthesize the view of frame B by projecting the world-space pointmaps of frame A onto the image plane of B using the estimated camera poses:

$$\hat{I}_B = \mathcal{R}(\mathbf{P}^A_{world}, R_B, \mathbf{T}_B, K) \tag{11}$$

where $\mathcal{R}$ denotes an off-screen renderer utilizing Z-buffering and point splatting. The reconstruction is deemed successful only if it satisfies predefined thresholds for *Coverage* (density), *MAE* (photometric error), and *L2 distance*.

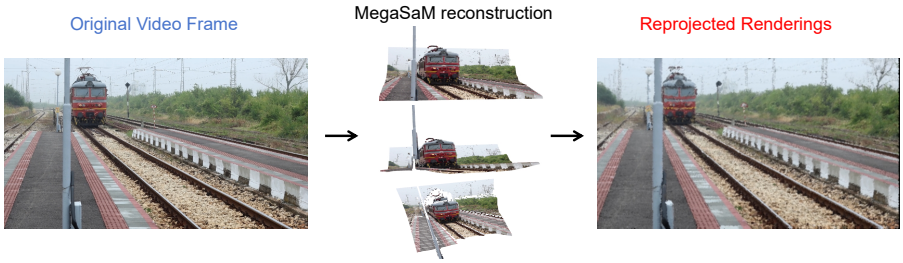


Fig. 5: Reconstruction Audit. We validate the fidelity of MegaSaM 3D pointmaps by reprojecting them onto target frames to ensure geometric consistency.

4.3 Quantitative Evaluation via PDI

Table 1: Quantitative comparison of geometric consistency on PDI-Bench. The PDI Score is the weighted aggregate of the scale–depth, trajectory, and rigidity residuals. Lower values indicate better geometric consistency. We additionally report the 95% confidence interval (CI), standard deviation, and outlier rate across evaluated samples. **GT** denotes the real-world reference and is excluded from the ranking. Best generative results are shown in **bold**.

Rank	Model	PDI Score ↓	95% CI	Scale ϵ_s ↓	Traj. ϵ_t ↓	Rigid. ϵ_r ↓	Std. ↓	Outlier ↓
–	*Ground Truth (GT)*	0.1206	[0.1018, 0.1386]	0.0660	0.1764	0.1182	0.0378	0.0%
1	Seedance 2.0	**0.2422**	[0.1954, 0.2920]	**0.2295**	0.2064	0.3392	**0.1315**	**0.0%**
2	CogVideoX-3	0.2480	[0.1656, 0.4093]	0.3135	**0.2033**	**0.2065**	0.3065	3.6%
3	Veo 3.1	0.4521	[0.2611, 0.7247]	0.7507	0.2271	0.3049	0.6980	7.1%
4	Wan 2.2	0.5595	[0.2572, 1.0766]	0.9317	0.2096	0.5150	1.2301	7.1%
5	Sora	0.8255	[0.2652, 1.4847]	1.6753	0.2711	0.2345	1.7312	14.3%
6	HunyuanVideo	0.8825	[0.3094, 1.6018]	1.8469	0.2515	0.2160	1.7730	14.3%

Table 1 summarizes the geometric consistency of the evaluated video generators using the PDI Score and its three components. The results reveal a clear gap

between real-world videos and generated videos under the proposed geometric audit.

Real-World Reference. The real-world GT videos obtain the lowest aggregate PDI Score (0.1206), together with the lowest scale–depth residual ($\epsilon_s = 0.0660$) and standard deviation (0.0378). Although the non-zero residuals reflect unavoidable errors introduced by monocular reconstruction, segmentation, and point tracking, the consistently lower GT values support the expected separation between real and generated videos. We therefore treat GT as an empirical reference rather than a theoretically perfect lower bound.

Top Generative Models Exhibit Complementary Strengths. Among the evaluated generators, Seedance 2.0 achieves the lowest sample-mean PDI Score (0.2422), while CogVideoX-3 follows closely with a score of 0.2480. Because their 95% confidence intervals substantially overlap, we regard their overall performance as comparable rather than statistically separated. Their component-wise behavior, however, differs. Seedance 2.0 obtains the lowest generative scale–depth residual ($\epsilon_s = 0.2295$), the lowest standard deviation (0.1315), and no detected outliers. In contrast, CogVideoX-3 achieves the lowest trajectory ($\epsilon_t = 0.2033$) and rigidity ($\epsilon_r = 0.2065$) residuals among the generative models, indicating comparatively stable 3D motion and internal structure.

Scale–Depth Errors Dominate Lower-Ranked Models. The performance degradation of Sora and HunyuanVideo is primarily associated with their scale–depth residuals, which reach 1.6753 and 1.8469, respectively. These values are more than 25× the GT scale residual. After applying the PDI weight $w_s = 0.4$, the scale component accounts for approximately 81% of Sora's aggregate PDI Score and 84% of HunyuanVideo's score. The results therefore suggest that scale–depth inconsistency is the dominant detected failure mode for these models under our evaluation protocol. Their larger standard deviations (> 1.7) and outlier rates (14.3%) further indicate substantial cross-sample variability in geometric consistency; they should not, however, be interpreted as evidence of stochastic instability without repeated generations from identical prompts and random seeds.

4.4 Category-Specific Physical Consistency Analysis

Building on Table 2 (a)–(e), we analyze how different motion patterns trigger specific physical hallucinations.

Longitudinal Convergence. This scenario audits scaling during axial motion. **HunyuanVideo** ($PDI = 0.10$) and **CogVideoX-3** ($PDI = 0.15$) approach the GT baseline, demonstrating a strong grasp of perspective laws. Conversely, **Wan2.2** and **Veo 3.1** exhibit high scale errors ($\epsilon_s > 0.32$), resulting in a "sliding" effect where object size mismatches relative 3D depth.

Biological Motion. Evaluating articulated dynamics reveals the difficulty of maintaining consistency in non-rigid entities. **Seedance 2.0** leads this category ($PDI = 0.25$), effectively preserving structural integrity. In contrast, **Veo 3.1** and **HunyuanVideo** suffer from "volumetric breathing" hallucinations, where body mass fluctuates inconsistently ($\epsilon_s > 1.97$) during gait cycles.

Table 2: Component-wise Physical Consistency and Human Alignment. (a)–(e) **Component-wise Error Analysis across Different Scenarios.** Each sub-table evaluates models on a specific challenge. GT (ref) denotes the real-world baseline. Lower values indicate higher physical consistency. (f) **Human Expert Evaluation Results.** Scores range from 1 (Best) to 10 (Worst). The human consensus ranking exhibits a perfect alignment ($\rho = 1.0$) with our automated PDI Score.

(a) Longitudinal Convergence

Model	PDI ↓	Scale ↓	Traj ↓	Rigid ↓
GT (ref)	**0.0715**	**0.0198**	**0.1406**	**0.0369**
HunyuanVideo	0.1004	0.0580	0.1533	0.0795
CogVideoX-3	0.1540	0.0814	0.1909	0.2252
Sora	0.2496	0.2656	0.2634	0.1903
Seedance 2.0	0.2623	0.2084	0.2351	0.4245
Veo 3.1	0.2958	0.4344	0.2115	0.1869
Wan2.2	0.3305	0.3278	0.2273	0.5424

(b) Dynamic Tracking

Model	PDI ↓	Scale ↓	Traj ↓	Rigid ↓
GT (ref)	**0.1167**	**0.0623**	**0.1734**	**0.1123**
CogVideoX-3	0.1620	0.1039	0.2222	0.1577
HunyuanVideo	0.1722	0.1839	0.2097	0.0738
Veo 3.1	0.2170	0.3233	0.1677	0.1029
Seedance 2.0	0.2334	0.2214	0.2039	0.3162
Wan2.2	0.2941	0.1661	0.2141	0.7098
Sora	1.2824	2.8387	0.2719	0.1910

(c) Biological Motion

Model	PDI ↓	Scale ↓	Traj ↓	Rigid ↓
GT (ref)	**0.1319**	**0.0887**	**0.1950**	**0.0921**
Seedance 2.0	0.2536	0.3401	0.1816	0.2246
CogVideoX-3	0.2773	0.4031	0.1781	0.2239
Wan2.2	0.2827	0.3632	0.1748	0.3377
Sora	0.3924	0.5968	0.2555	0.2571
HunyuanVideo	0.9760	2.1265	0.2310	0.1649
Veo 3.1	1.0230	1.9738	0.2268	0.7135

(d) Curved Motion

Model	PDI ↓	Scale ↓	Traj ↓	Rigid ↓
GT (ref)	**0.1550**	**0.1136**	**0.1366**	**0.2745**
Seedance 2.0	0.2558	0.3001	0.2122	0.2542
CogVideoX-3	0.2567	0.2944	0.2174	0.2600
Wan2.2	0.5223	0.8862	0.2042	0.4305
Veo 3.1	0.6037	1.0484	0.2728	0.3759
HunyuanVideo	0.7467	1.4705	0.2321	0.3282
Sora	2.1277	4.8660	0.3225	0.2617

(e) Partial Occlusion

Model	PDI ↓	Scale ↓	Traj ↓	Rigid ↓
GT (ref)	**0.1346**	**0.0626**	**0.2115**	**0.1250**
Seedance 2.0	0.2101	0.1075	0.1960	0.4433
Sora	0.2201	0.1617	0.2482	0.2807
Veo 3.1	0.2414	0.2270	0.2640	0.2251
CogVideoX-3	0.3964	0.6964	0.2058	0.1774
Wan2.2	1.3157	2.8131	0.2208	0.5109
HunyuanVideo	2.4104	5.3793	0.4248	0.4436

(f) Human Expert Study

Rank	Model	Mean Score ↓	Std. Dev.
1	**Ground Truth (GT)**	**1.57**	1.00
2	Seedance 2.0	2.96	1.86
3	CogVideoX-3	2.97	1.47
4	Veo 3.1	3.36	1.56
5	Wan2.2	3.37	2.26
6	Sora	3.60	1.99
7	HunyuanVideo	3.66	2.29

Curved Motion. Non-linear trajectories and rotations represent the most significant challenge. **Sora** exhibits catastrophic failure here ($PDI = 2.13$), driven by massive scale distortion ($\epsilon_s = 4.87$). This suggests that transformer-based generators often fail to preserve the $h \cdot Z$ invariant during rotational transformations. **CogVideoX-3** and **Seedance 2.0** remain the most robust in this scenario.

Partial Occlusion. This tests spatial memory and object permanence. **HunyuanVideo** shows severe degradation ($PDI = 2.41, \epsilon_s = 5.38$), indicating the model "forgets" physical dimensions while the object is obscured. Conversely, **Sora** and **Seedance 2.0** demonstrate superior resilience, maintaining structural cohesion ($\epsilon_r < 0.45$) upon the object's re-emergence.

Dynamic Tracking. This scenario assesses the decoupling of camera ego-motion from object kinematics. **CogVideoX-3** ($PDI = 0.16$) and **HunyuanVideo** ($PDI = 0.17$) excel, nearing GT fidelity. However, **Sora** struggles significantly with scale ($\epsilon_s = 2.84$), as its world model tends to conflate camera proximity with non-physical object growth, leading to a collapse of perspective coupling.

4.5 Human Expert Study and Perceptual Alignment

To validate the alignment between **PDI-Bench** and human physical intuition, we conducted a perceptual study involving seven computer vision experts. The evaluation set comprised 105 unique video clips: 15 real-world (GT) videos and 90 AI-generated clips (15 per model) sourced from six state-of-the-art generators. For each model, three clips were sampled for each of the five physical scenarios. Adhering to the PDI scoring protocol, experts assigned ratings ranging from **1**

(Physical Realism) to **10 (Catastrophic Failure)**. With each clip reviewed by all seven experts, we obtained $n = 105$ subjective ratings per model.

Results and Alignment Analysis. As summarized in Table 2 (f), the expert consensus yields a model ranking **identical** to our automated results. Real-world GT consistently anchors the high-fidelity bound with the lowest mean score (1.57 ± 1.00), confirming the experts' ability to distinguish natural physics from synthetic artifacts. Among the AI models, *Seedance 2.0* and *CogVideoX-3* emerge as the top performers with nearly identical scores (2.96 and 2.97, respectively). Their low mean scores suggest a higher degree of temporal consistency and fewer visible violations of basic physical laws. *Veo 3.1* and *Wan 2.2* occupy the middle tier (3.36–3.37). While competitive, they exhibit more frequent physical inaccuracies compared to the top-tier models. In contrast, *Sora* and *HunyuanVideo* were penalised for the severe "scale hallucinations" and "motion jitter" previously identified by our 3D geometric audit. Notably, the high standard deviations in the lower-ranked models (e.g., 2.29 for HunyuanVideo) suggest that their failure modes are often dramatic and highly visible to human observers. This consistent alignment underscores the validity of PDI-Bench as an objective, automated proxy for human perceptual assessment of physical laws in video generation.

5 Case Study: Diagnosing Autoregressive Extrapolation

To demonstrate the diagnostic power of PDI-Bench beyond standard evaluation, we conduct a stress test on autoregressive (AR) long-video generation. We evaluate the recently proposed **Self-Forcing** [8] paradigm, built upon the Wan2.1-T2V-1.3B architecture [24]. Although trained on 81-frame sequences, we extrapolate the model to 129 frames. As shown in Table 3, generation beyond the training context results in a substantial degradation of geometric consistency, dominated by **catastrophic scale drift**.

Table 3: Self-Forcing AR extrapolation.

Category	PDI ↓	Scale ↓	Traj. ↓	Rigid. ↓
Longit. Conv.	1.3063	2.8462	0.3078	0.2237
Dynamic Track.	0.1994	0.1515	0.2604	0.1730
Biological Mot.	2.0819	4.6326	0.3489	0.4464
Curved Motion	0.3111	0.3158	0.2530	0.4177
Partial Occl.	2.7570	6.2172	0.4097	0.5311
Overall Mean	**1.3407**	**2.8583**	**0.3170**	**0.3531**

As shown in Table 3, the scale residual increases to 2.8583, whereas the trajectory residual remains comparatively low at 0.3170. Scale distortion is particularly severe under biological motion and partial occlusion, reaching 4.6326 and 6.2172, respectively. This discrepancy indicates that long-horizon autoregressive extrapolation primarily disrupts perspective-consistent object scale rather than short-term motion smoothness.

Kinematic Success vs. Geometric Collapse. Table 3 reveals a striking dichotomy in the AR model's physical capabilities. Across all scenarios, the 3D Kinematic Trajectory error (ϵ_t) remains remarkably stable (overall mean: 0.3170). This indicates that the Self-Forcing training paradigm, combined with rolling

KV caching, successfully mitigates high-frequency spatial jitter and unnatural reversals often seen in long AR generation.

However, this kinematic smoothness masks a catastrophic failure in 3D projective geometry. The overall Scale-Depth Alignment error (ϵ_s) surges to 2.8583, particularly in *Longitudinal Convergence* and *Biological Motion*. This demonstrates severe *scale hallucination*: as the model extrapolates beyond its 81-frame training horizon, it loses the "spatial memory" of the object's original 3D volume, causing objects to expand or contract independently of their depth variations.

Vulnerability to Spatial Disconnects. The data further reveals that AR generation is highly sensitive to continuous visual context. In *Dynamic Tracking*, where the subject remains centrally focused, the scale error remains low (0.1515). Conversely, in *Partial Occlusion*, the model exhibits its worst performance ($PDI = 2.7570$). When the object is temporarily obscured, the AR mechanism loses its structural anchor within the context window, causing a complete failure in object permanence when the subject re-emerges. In *Curved Motion*, while trajectory is maintained, the complexity of rotation leads to a higher rigidity residual ($\epsilon_r = 0.4177$), indicating a mild "jello effect" during non-linear displacement. These insights highlight PDI-Bench's unique capability to provide fine-grained, frame-level diagnostic signals for emerging video generation architectures.

6 Conclusion

In this paper, we introduced **PDI-Bench**, a novel quantitative framework for auditing the perspective and scale consistency of generative video world models. By constructing a collaborative *Target-Uplift-Anchor* workflow, we successfully translated 2D pixel dynamics into verifiable 3D geometric reasoning. Our proposed **Perspective Distortion Index (PDI)** provides a multi-modal metric to identify subtle hallucinations such as "volume breathing" and "skating" effects. Through extensive benchmarking on our **PDI-Dataset**, we demonstrated that our framework offers a robust geometric yardstick that complements existing semantic-based metrics. We believe PDI-Bench will serve as a foundational tool for evaluating and improving the physical intelligence of future artificial world simulators.

Limitations. While PDI-Bench establishes a rigorous geometric yardstick, it presents three primary limitations. First, its accuracy relies on off-the-shelf perception tools; in degraded videos where 3D uplifting fails, the framework must fall back to 2D proxies, reducing depth-aware precision. Second, our geometric invariants rely on a rigid-body assumption, making the metric less theoretically suited for highly non-rigid or amorphous subjects. Finally, disentangling complex 3D rotation from axial translation using purely monocular cues is fundamentally ill-posed, occasionally introducing measurement noise despite our consensus-based mitigations.

References

1. Allen, K.R., Doersch, C., Zhou, G., Suhail, M., Driess, D., Rocco, I., Rubanova, Y., Kipf, T., Sajjadi, M.S.M., Murphy, K.P., Carreira, J., Steenkiste, S.V.: Direct motion models for assessing generated videos. In: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 1159–1183. PMLR (2025), https://proceedings.mlr.press/v267/allen25a.html
2. 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. In: Int. Conf. Learn. Represent. OpenReview.net (2025), https://openreview.net/forum?id=9D2QvO1uWj
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets (2023), https://arxiv.org/abs/2311.15127
4. ByteDance: Doubao large model. https://www.volcengine.com/product/doubao (2024), official product page; accessed August 12, 2026
5. ByteDance: Seedance 2.0 official launch. https://seed.bytedance.com/en/blog/official-launch-of-seedance-2-0 (2026), accessed: 2026-08-12
6. Chow, W., Mao, J., Li, B., Seita, D., Guizilini, V., Wang, Y.: Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In: Int. Conf. Learn. Represent. OpenReview.net (2025), https://openreview.net/forum?id=Q6a9W6kzv5, oral presentation
7. Ho, J., Jain, A.N., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33 (2020), https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html
8. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. In: Advances in Neural Information Processing Systems. vol. 38 (2025), https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html
9. 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. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21807–21818 (2024), https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html
10. Hume, T., Carey, M., Iljic, T.: Meet flow: Ai-powered filmmaking with veo 3. Google Blog (2025), https://blog.google/innovation-and-ai/products/google-flow-veo-ai-filmmaking-tool/, accessed: 2026-08-12
11. Karaev, N., Makarov, I., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In: Int. Conf. Comput. Vis. (2025), https://openaccess.thecvf.com/content/ICCV2025/html/Karaev_CoTracker3_Simpler_and_Better_Point_Tracking_by_Pseudo-Labelling_Real_Videos_ICCV_2025_paper.html
12. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang,

J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., Zhong, C.: Hunyuanvideo: A systematic framework for large video generative models (2024), https://arxiv.org/abs/2412.03603

13. Li, D., Fang, Y., Chen, Y., Yang, S., Cao, S., Wong, J., Luo, M., Wang, X., Yin, H., Gonzalez, J.E., Stoica, I., Han, S., Lu, Y.: Worldmodelbench: Judging video generation models as world models. In: Advances in Neural Information Processing Systems. vol. 38 (2025). https://doi.org/10.52202/085713-1834, https://proceedings.neurips.cc/paper_files/paper/2025/hash/4ec03ed08a3fcb59e1c815b5598beff1-Abstract-Datasets_and_Benchmarks_Track.html, datasets and Benchmarks Track
14. Li, Z., Tucker, R., Cole, F., Wang, Q., Jin, L., Ye, V., Kanazawa, A., Holynski, A., Snavely, N.: Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025), https://openaccess.thecvf.com/content/CVPR2025/html/Li_MegaSaM_Accurate_Fast_and_Robust_Structure_and_Motion_from_Casual_CVPR_2025_paper.html
15. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., He, L., Sun, L.: Sora: A review on background, technology, limitations, and opportunities of large vision models (2024), https://arxiv.org/abs/2402.17177
16. Meng, F., Liao, J., Tan, X., Lu, Q., Shao, W., Zhang, K., Cheng, Y., Li, D., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 43781–43806. PMLR (2025), https://proceedings.mlr.press/v267/meng25c.html
17. OpenAI: Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/ (2024), accessed: 2026-08-12
18. 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: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), https://proceedings.mlr.press/v139/radford21a.html
19. 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. In: Int. Conf. Learn. Represent. OpenReview.net (2025), https://openreview.net/forum?id=Ha6RTeWMd0
20. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in Neural Information Processing Systems. vol. 29 (2016), https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html
21. Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., Liu, X.: T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025), https://openaccess.thecvf.com/content/CVPR2025/html/Sun_T2V-CompBench_A_Comprehensive_Benchmark_for_Compositional_Text-to-video_Generation_CVPR_2025_paper.html

22. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges (2018), https://arxiv.org/abs/1812.01717
23. Upadhyay, R., Zhang, H., Solomon, J., Agrawal, A., Boreddy, P., Narayana, S.S., Ba, Y., Wong, A., de Melo, C.M., Kadambi, A.: Worldbench: Disambiguating physics for diagnostic evaluation of world models (2026), https://arxiv.org/abs/2601.21282
24. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models (2025), https://arxiv.org/abs/2503.20314
25. Wan-Video: Wan2.2: Wan: Open and advanced large-scale video generative models. https://github.com/Wan-Video/Wan2.2 (2025), gitHub repository
26. Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L.: Florence-2: Advancing a unified representation for a variety of vision tasks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4818–4829 (2024), https://openaccess.thecvf.com/content/CVPR2024/html/Xiao_Florence-2_Advancing_a_Unified_Representation_for_a_Variety_of_Vision_CVPR_2024_paper.html
27. Zhipu AI: Cogvideox-3. Official model documentation (2025), https://docs.bigmodel.cn/cn/guide/models/video-generation/cogvideox-3, released July 15, 2025; accessed August 12, 2026