SPACETIMEBENCH: EVALUATING SPACE AND TIME IN THE WORLD MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative world models aim to serve as universal physical simulators, yet existing benchmarks fail to verify the fundamental prerequisite of Newtonian Law: that space and time must remain absolute and mathematically consistent across different observer perspectives. To bridge this evaluation gap, we introduce SpaceTimeBench, a novel Video-to-Video framework that utilizes prompt-injected geometric anchors alongside DePT3R and SAM 2 to extract strictly normalized 4D coordinates, testing whether models perfectly conserve spatial distances and temporal durations across varied camera views. Extensive evaluations across 17 leading architectures reveal a severe capability deficit quantified by our Final Spacetime Yield (FSY), demonstrating that even state-of-the-art models peak at a mere 65.0% success rate. Crucially, SpaceTimeBench functions as an evaluation oracle to expose a False Discovery Rate (FDR) of up to 42.9% regarding spurious compliance in existing benchmarks, proving that evaluating Newton's First Law without verifying underlying spatiotemporal prerequisites severely inflates scores and masks the reality that current architectures remain mere two-dimensional pattern matchers.

1 INTRODUCTION

The ultimate objective of generative world models is to serve as **universal physical simulators** capable of rendering infinite stochastic realities (Qin et al., 2024; OpenAI, 2024; LeCun & Courant, 2022). In physics, causality ($A \rightarrow B$) is the invariant governing principle dictating that event A is a strictly necessary and sufficient condition to trigger state B (Bunge, 1979; Marwala & Hurwitz, 2017). Crucially, within the Newtonian mechanics that govern everyday macroscopic environments, this causal sequence relies on a fundamental prerequisite: space and time are **assumed** to be absolute, serving as the rigid, observer-independent stage upon which all causal events unfold (Zylbersztajn, 1994; Capecchi, 2022).

While existing benchmarks for generative world models actively evaluate Newtonian classical dynamics—such as gravity, velocity, rigid-body collisions, and physical commonsense (Li et al., 2025; Jain & Wu, 2026; Xie et al., 2025)—**none of them** validate the foundational prerequisite of Newton's laws: that time and space must remain absolute across different observers. Currently, the closest methodologies to multi-observer testing are object permanence benchmarks (Shamsian et al., 2020; Campbell et al., 2025), which utilize **camera-conditioned prompts** within a single video to verify if a target object maintains its geometric volume, speed, and trajectory. However, sequentially observing an event within a single video merely tests the memorization of 2D pixel patterns over time (Tang et al., 2026; Chi et al., 2025), rather than verifying true geometric invariance across independent observers.

To bridge this evaluation gap, we introduce **SpaceTimeBench**, a novel Video-to-Video benchmarking framework that utilizes prompt-injected geometric anchors alongside Alumootil et al. (2025) and SAM 2 Ravi et al. (2024) to extract strictly normalized 4D coordinates. This pipeline rigorously tests whether models mathematically conserve physical spatial distances and temporal durations across independent camera views, satisfying the fundamental Newtonian invariants:

$$\Delta t' = \Delta t \quad \text{and} \quad \Delta x'^2 + \Delta y'^2 + \Delta z'^2 = \Delta x^2 + \Delta y^2 + \Delta z^2 \tag{1}$$

We reuse the complete set of 350 prompts from WorldModelBench Li et al. (2025) across diverse domains (e.g. robotics and autonomous driving), and append specific "Norm Objects" Liu et al.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

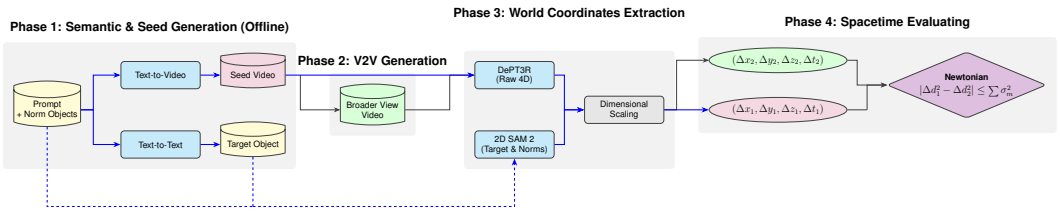


Figure 1: The SpaceTimeBench pipeline architecture.

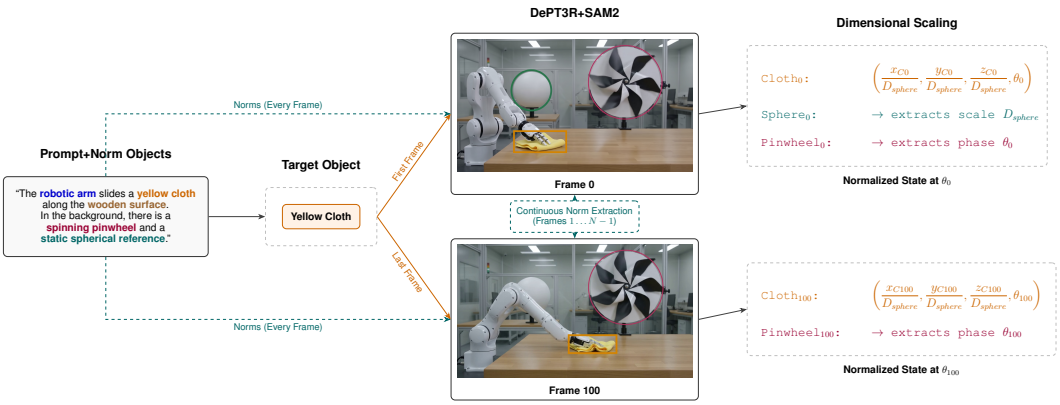


Figure 2: Detailed instance evaluation for the Seed Video (the highlighted blue path in Figure 1).

(2024); Zeng et al. (2024) to overcome different dimension scaling in different videos. As illustrated in Figure 1, our evaluation pipeline executes in four phases:

- **Phase 1: Seed Generation (Offline):** A Text-to-Video model generates a "Seed Video" from the injected prompt, while a Text-to-Text model isolates the tracked "Target Object".
- **Phase 2: V2V Generation:** The Seed Video is expanded via Video-to-Video (V2V) generation into a "Broader View Video", synthesizing an independent observer perspective.
- **Phase 3: Coordinate Extraction:** DePT3R and SAM 2 extract raw 4D spatial and mask data from both videos, which is then dimensionally scaled into normalized coordinates.
- **Phase 4: Spacetime Evaluation:** The normalized coordinates from both views are compared. A model passes only if its spacetime distances $(\Delta x, \Delta y, \Delta z, \Delta t)$ perfectly match within strict **tolerances**.

Furthermore, SpaceTimeBench operates as a **self-contained** diagnostic framework by categorizing generation errors into seven hierarchical failure modes. Crucial prerequisites—such as event adherence and norm object stability—are verified prior to diagnosing the ultimate spatial and temporal invariance failures. Extensive evaluations across 17 leading architectures reveal a severe capability deficit quantified by our Final Spacetime Yield (FSY), demonstrating that state-of-the-art models peak at a mere 65.0% success rate. Crucially, SpaceTimeBench functions as an evaluation oracle to expose a False Discovery Rate (FDR) of up to 42.9% regarding spurious compliance in existing benchmarks(e.g. WorldModelBench). This confirms that evaluating classical mechanics without first validating the underlying spatiotemporal prerequisites severely inflates performance scores, masking the reality that current architectures remain limited to 2D pattern memorization.

2 Related Work

Limitations of Current Physics Benchmarks Recent evaluations of generative world models fall into two categories. First, comprehensive physics benchmarks aggressively test classical mechanics—like rigid-body collisions and gravity—yet bypass the foundational spacetime prerequisites required for these dynamics (Li et al., 2025; Xie et al., 2025; Gu et al., 2025; Kang et al., 2025). Second,

specialized frameworks evaluate space and time as completely disconnected tests. Spatial benchmarks only verify if an object maintains its correct 3D volume and shape (Zhang et al., 2026; Upadhyay et al., 2026). Conversely, temporal benchmarks merely check logical chronological sequencing or macroscopic time-lapse transformations (Lee et al., 2023; Yuan et al., 2024). Consequently, no existing framework synchronously validates unified 4D spacetime invariance.

Anchoring Physical Events and Controlled Observer Perspectives Generating identical causal events across independent observer videos is challenging: Text-to-Video (T2V) hallucinates divergent physics Ho et al. (2022); Blattmann et al. (2023), while Image-to-Video (I2V) diverges temporally Wu et al. (2022); Wang et al. (2023b); Zhang et al. (2023); Liu et al. (2025); Esser et al. (2023); You et al. (2026). Furthermore, invasive white-box architectural modifications Tang et al. (2023); Shi et al. (2023); Liu et al. (2024); Wu et al. (2025) and explicit mathematical camera conditioning Guo et al. (2023); Zhao et al. (2023); He et al. (2024); Hoorick et al. (2025) are incompatible with zero-shot evaluation of closed-source models. We therefore adopt a black-box Video-to-Video (V2V) paradigm Chen et al. (2025) driven purely by natural language conditioning Qi et al. (2023); Liu et al. (2023). The human annotated seed video forces models to deterministically re-render the exact causal event from a novel viewpoint.

Temporal and Spatial Normalization Extracting spatial and temporal metrics from generative video is complicated by unstable diffusion and scale ambiguity Wang et al. (2023a). Because existing temporal extraction methods Dwibedi et al. (2019); Lee et al. (2023); Yin et al. (2026) are incompatible with unconstrained black-box evaluation, we utilize a Prompt-Injected Geometric Clock. By point-tracking a spinning pinwheel to verify constant angular momentum ($d\theta/df$) Liu et al. (2024); Dwibedi et al. (2019), time is isolated as a pure spatial relation. Similarly, to resolve spatial scale ambiguity, we apply Relative Spatial Normalization via a prompt-injected spherical ruler Zeng et al. (2024); Masuda et al. (2023), effectively projecting both spatial and temporal metrics into a unified, dimensionless world coordinate system.

Spatial World Coordination Extracting 3D motion from AI video is challenging. Text models Chow et al. (2025) lack exact math, and physics simulators Yi et al. (2019) require camera data we do not have. While 2D trackers Karaev et al. (2023) reliably follow flat pixels, combining them with separate 3D depth tools Doersch et al. (2023); Yang et al. (2024) causes independent errors to multiply, making coordinates flicker. We solve this by using DePT3R Alumootil et al. (2025) alongside SAM 2 Ravi et al. (2024). Unlike decoupled methods, DePT3R calculates depth and tracks movement synchronously in one step, locking spatial and temporal objects together. Finally, to handle generative video artifacts Mei et al. (2025), we convert DePT3R's confidence scores into a dynamic mathematical tolerance bound (Σ^2_{fault}).

3 Pipeline Architecture and Process Flow

The following is the internal excution process: the overall four-phase architecture of the pipeline is illustrated in Figure 1. The process uses a concrete scene to demonstrate how semantic elements are extracted and scaled into normalized physical coordinates.

3.1 Phase 1: Semantic & Seed Generation (Offline)

This entire phase is executed completely offline during the dataset curation process, ensuring that all semantic metadata and generated videos are pre-computed and ready-to-use when a user runs the benchmark. The process begins with the **Prompt+Norm Objects** (see Figure 2).

- **Example Prompt:** *"The robotic arm slides a yellow cloth along the wooden surface. In the background, there is a spinning pinwheel and a static spherical reference."* The first sentence represents the original prompt from WorldModelBench, while the **spinning pinwheel** and **static sphere** are appended specifically to establish references for temporal and geometric normalization, respectively.
- **Video Generation:** A Text-to-Video Base Model processes this prompt to generate the foundational **Seed Video**, and we deploy human annotation protocol to manually evaluate offline, see Appendix B.

- **Target Object Extraction:** Simultaneously, a Text-to-Text LLM parses the natural language prompt to isolate all primary causal entities undergoing macroscopic spatial displacement from $A \rightarrow B$ (e.g., the **Yellow Cloth**).

3.2 PHASE 2: V2V GENERATION

To initiate the evaluation, the user specifies their target world model's endpoint (or local address), the desired WorldModelBench prompt ID, and the spatial perspective (e.g. a broader view). Using the pre-computed Seed Video to gurantee the same event, the Video-to-Video (V2V) generation module tasks the target model with synthesizing a distinct perspective variant requirement:

- A **Broader View Video**. The text conditioning dynamically injects a user-defined spatial perspective. The default generation prompt explicitly instructs the model: *"For the given event from the video, can you generate the exactly same events from a broader view?"*

3.3 PHASE 3: WORLD COORDINATES EXTRACTION

To analyze physical consistency, we must translate raw generated pixels into worldline coordinates, as different video has differnt dimension scaling. Let a generated video from Observer i be denoted as V_i, consisting of a sequence of frames $f \in [0, N]$. Based on the semantic extraction from Phase 1, we track a defined set of entities $E = \{m, s, p\}$, where m is the primary moving Target Object, s is the static spherical spatial ruler (Spatial Norm), and p is the spinning pinwheel (Time Norm).

This coordinate extraction is achieved by running two parallel foundation models—DePT3R and SAM 2—and fusing their outputs.

1. CONFIDENCE-AWARE 3D TRACKING (DEPT3R)

For every frame f in video V_i, DePT3R processes the image to extract raw 3D topological data at the micro-level. For every 2D pixel index (u, v), it predicts:

- $\mathbf{P}(f) \in \mathbb{R}^{H \times W \times 3}$: The dense 3D pointmap, where every pixel (u, v) is assigned an absolute Cartesian coordinate $\mathbf{p}_{u,v}(f) = (x, y, z)^T$.
- $\mathbf{C}(f) \in [0, 1]^{H \times W}$: The pixel-wise confidence map. A high confidence score ($c_{u,v} \rightarrow 1$) indicates solid object permanence and reliable depth projection, while a low score ($c_{u,v} \rightarrow 0$) mathematically flags generative hallucinations, inconsistencies, or severe occlusions.

2. CAUSAL ENTITY SEGMENTATION (SAM 2)

While DePT3R provides dense 3D geometry for the entire scene, it operates purely at the pixel-coordinate level and lacks any semantic concept of discrete physical objects. To bridge this gap and identify the specific entities we wish to track, we utilize SAM 2 Ravi et al. (2024). Initialized with the object k extracted in Phase 1, SAM 2 acts as the semantic filter, grouping the raw coordinate space by generating precise bounding masks for each causal entity $k \in E$. For a given frame f, the output is:

- $M_k(f) \in \{0, 1\}^{H \times W}$: The binary pixel mask isolating object k, where 1 denotes the foreground entity and 0 represents the background.

3. MACROSCOPIC CENTROID FUSION

Using the SAM 2 mask, we isolate the 3D point cloud of any object k. To ensure extreme robustness against edge-case hallucinations, we apply a confidence-weighted average to calculate its macroscopic 3D Centroid $\mathbf{C}_k(f)$ at any frame f:

$$\mathbf{C}_k(f) = \frac{\sum_{(u,v) \in M_k(f)} c_{u,v}(f) \cdot \mathbf{P}(u, v, f)}{\sum_{(u,v) \in M_k(f)} c_{u,v}(f)} \tag{2}$$

This yields a highly stable, single tracking coordinate (x_k, y_k, z_k) for the Target Object m at every time step.

4. Topological Anchor Extraction

Building upon the centroid formulation, we define normalization anchors s and p as structural derivatives of their respective masked point clouds.

- **Spatial Ruler (s):** The absolute spatial scaling unit of the generated universe is established entirely from the first frame ($f = 0$). The bounding box dimensions $D_x(0), D_y(0), D_z(0)$ represent the maximum orthogonal span of the sphere's 3D point cloud distributed around its initial centroid $\mathbf{C}_s(0)$. For the x-axis, this is extracted as:

$$D_x(0) = \max_{(u,v)\in M_s(0)} \mathbf{P}_x(u, v, 0) - \min_{(u,v)\in M_s(0)} \mathbf{P}_x(u, v, 0) \tag{3}$$

  with $D_y(0)$ and $D_z(0)$ derived symmetrically. The baseline spatial scale, D_{sphere}, is fixed as the sphere's mean topological diameter at $f = 0$:

$$D_{sphere} = \frac{1}{3}(D_x(0) + D_y(0) + D_z(0)) \tag{4}$$

- **Temporal Clock (p):** We calculate a strict geometric rotational phase angle $\theta(f)$ iteratively for every frame $f \in [0, N]$. We isolate a single tracked blade tip coordinate $\mathbf{P}_{tip}(f) \in M_p(f)$ across the video. The temporal phase vector is derived directly from the geometric line extending from the pinwheel's centroid $\mathbf{C}_p(f)$ to this tip at each time step:

$$\vec{v}_{tip}(f) = \mathbf{P}_{tip}(f) - \mathbf{C}_p(f), \quad \forall f \in [0, N] \tag{5}$$

  The clock's absolute temporal state $\theta(f)$ is the 2D projected phase angle of $\vec{v}_{tip}(f)$ relative to the observer's viewing plane, providing a rigid, physics-grounded timestamp for every individual frame from 0 to N.

5. Dimensional Scaling

The raw 4D coordinates of each Target Object, denoted as $(\mathbf{C}_m(f), f)$ where $\mathbf{C}_m(f) = (x_m(f), y_m(f), z_m(f))$, are inherently tied to the latent biases of the specific text-to-video generator (e.g., arbitrary pixel resolutions and inconsistent frame-rate elasticity). We apply the spatial and temporal normalization anchors to every Target Object $m \in M$:

- **Spatial Normalization:** The raw 3D centroid components of the Target Object are independently scaled by the fixed spatial ruler established by the sphere at $f = 0$. The normalized spatial coordinates are extracted as:

$$\left(\frac{x_m(f)}{D_{sphere}}, \frac{y_m(f)}{D_{sphere}}, \frac{z_m(f)}{D_{sphere}}\right) \tag{6}$$

  This transformation maps absolute pixel-space dimensions into unitless topological distances, rendering the object's spatial trajectory immune to arbitrary zooming, cropping, or the Broader View spatial scaling.

- **Temporal Normalization:** Because the raw frame index f does not guarantee a constant Δt in generative physics, we replace f with the absolute time metric extracted directly from the temporal clock. For any given frame f, the normalized time coordinate is strictly defined by the pinwheel's instantaneous rotational phase: $t_{norm}(f) = \theta(f)$.

Final Trajectory Formulation: By fusing these dimensions, we construct the final, invariant spacetime coordinate tuple for each moving object m at any frame f under Observer i. As visually depicted in our pipeline, the continuous trajectory sequence $\mathcal{T}_{m,i}$ is strictly defined as the set of these normalized tuples:

$$\mathcal{T}_{m,i} = \left\{ \left(\frac{x_m(f)}{D_{sphere}}, \frac{y_m(f)}{D_{sphere}}, \frac{z_m(f)}{D_{sphere}}, \theta(f)\right) \;\middle|\; f \in [0, N] \right\} \tag{7}$$

This mathematically standardized representation ensures that trajectories from the Seed Video and the Alternate View videos bypass generative artifacts and can be cleanly compared in Phase 4.

3.4 PHASE 4: INVARIANT SPACETIME EVALUATION

Building upon the invariant spacetime trajectory $\mathcal{T}_{m,j}$ extracted in Phase 3, we now evaluate the generated physical rules across different observer perspectives (e.g., $j = 1$ for the Seed video, and $j = 2$ for the Broader View video).

1. MACROSCOPIC DISPLACEMENT CALCULATION

For any given observer j, we define the start state at frame $f = 0$ and the end state at frame $f = N$. Inheriting the normalized spatial coordinates $\hat{\mathbf{C}}_{m,j}(f) = (\hat{x}_j(f), \hat{y}_j(f), \hat{z}_j(f))$ and the normalized temporal coordinate $t_{norm,j}(f) = \theta_j(f)$, the macroscopic displacements are calculated as:

- **Spatial Displacement Components:**

$$\Delta\hat{x}_j = \hat{x}_j(N) - \hat{x}_j(0) \qquad \Delta\hat{y}_j = \hat{y}_j(N) - \hat{y}_j(0) \qquad \Delta\hat{z}_j = \hat{z}_j(N) - \hat{z}_j(0) \tag{8}$$

- **Total Normalized Spatial Distance Squared:**

$$\Delta d_j^2 = \Delta\hat{x}_j^2 + \Delta\hat{y}_j^2 + \Delta\hat{z}_j^2 \tag{9}$$

- **Total Normalized Temporal Duration:**

$$\Delta t_j = \theta_j(N) - \theta_j(0) \tag{10}$$

2. INTRINSIC GEOMETRIC UNCERTAINTY AND DYNAMIC TOLERANCE

To rigorously evaluate the physical invariants without introducing arbitrary hyperparameters, we derive our acceptable difference boundary directly from the underlying geometric confidence established in Phase 3.

Because our kinematic equations evaluate the moving object as a single macroscopic entity, we must account for the spatial variance σ_m^2 of the raw 3D point cloud around the calculated centroid $\mathbf{C}_m$. Expanding the Euclidean distance $\|\mathbf{p}_{u,v} - \mathbf{C}_m\|^2$, the raw spatial variance decomposes cleanly into component-wise variances. For any frame state k, the raw variance is:

$$\sigma_{raw,k}^2 = \frac{\sum_{(u,v)\in M_k} c_{u,v}\|\mathbf{p}_{u,v} - \mathbf{C}_k\|^2}{\sum_{(u,v)\in M_k} c_{u,v}} = \sigma_{x,k}^2 + \sigma_{y,k}^2 + \sigma_{z,k}^2 \tag{11}$$

Because our spatial displacements (Δd_j^2) are normalized by the spatial ruler D_{sphere}, the geometric uncertainty must also be scaled into the intrinsic topological space. The normalized variance $\hat{\sigma}_k^2$ for state k is therefore:

$$\hat{\sigma}_k^2 = \frac{\sigma_{raw,k}^2}{D_{sphere}^2} \tag{12}$$

By variance propagation across independent spatial axes and discrete states, the total systemic tracking fault across all four anchor states utilized in a two-observer comparison—$k \in \{1_{\text{start}}, 1_{\text{end}}, 2_{\text{start}}, 2_{\text{end}}\}$—is the sum of their normalized variances:

$$\Sigma_{fault}^2 = \sum_{k=1}^{4} \hat{\sigma}_k^2 \tag{13}$$

3. PHYSICAL EQUIVALENCE CHECKS

The physical equivalence condition between Observer 1 and Observer 2 is bounded strictly by this intrinsic component-wise tracking fault. We evaluate the models against the absolute physical invariant:

$$\textbf{Newtonian Check:} \quad \left|\Delta d_1^2 - \Delta d_2^2\right| \leq \Sigma_{fault}^2 \tag{14}$$

4 CASCADING FAILURE EVALUATION AND REPORTING

The SpaceTimeBench evaluation pipeline is executed directly via a configurable command-line interface. A typical execution is structured as follows:

```
$ python run_spacetimebench.py --model_address
    "endpoints/V2V-Base-v1.2" --prompt_id "WMB_042" \
  --spatial_view "broader" --epsilon_shape 0.05 --tau_drift 0.02
      --tau_dissolve 0.15 --epsilon_time 0.001
```

Following the strict reporting methodology established by WorldModelBench (Li et al., 2025), performance is aggregated and scored on a per-prompt basis. To provide granular diagnostic capabilities, the benchmark outputs a standardized JSON report detailing the exact reason for cascading halting runs (see Appendix A) including following seven distinct failure modes:

1. **Failure 1: Causal Identity Loss.** A Vision-Language Model evaluates the generated sequence to ensure the original causal event ($A \rightarrow B$) occurs. Let this boolean output be E_{semantic}. *Halt if:* $E_{\text{semantic}} = \text{False}$. The model failed semantic object permanence.
2. **Failure 2: Ruler Collapse.** Using the initial 3D dimensions $D_x(0), D_y(0), D_z(0)$ of the prompt-injected reference, we ensure the sphere remains spherical. *Halt if:* The aspect ratio $\max(D)/\min(D) > 1 + \epsilon_{\text{shape}}$, indicating the intrinsic spatial coordinate system has morphed.
3. **Failure 3: Rubber Clock.** We evaluate the unwrapped cumulative phase $\Phi(f)$ of the rotational clock to verify continuous temporal flow. *Halt if:* The linear regression of $\Phi(f)$ yields $R^2 < \tau_{\text{clock}}$, indicating the clock is visually glitching or hallucinating erratic time steps.
4. **Failure 4: Camera Drift.** We calculate the Apparent Camera Drift. Because the reference sphere is physically stationary, any spatial displacement indicates unwanted global coordinate drift:
$$\vec{v}_{\text{apparent}} = \frac{\hat{\mathbf{C}}_s(N) - \hat{\mathbf{C}}_s(0)}{t_{norm}(N) - t_{norm}(0)} \tag{15}$$
Halt if: $\|\vec{v}_{\text{apparent}}\| > \tau_{\text{drift}}$ for a stationary observer prompt.
5. **Failure 5: Matter Dissolve & Persistence Loss.** For every Target Object $m \in M$, we evaluate the normalized DePT3R tracking variance $\hat{\sigma}_k^2$ to ensure the object maintains structural density and continuous existence. *Halt if:* $\hat{\sigma}_k^2 > \tau_{\text{dissolve}}$ for any frame k, or if the tracking confidence drops to zero. A failure here indicates the physical matter of the target either dissolved into untrackable noise or suffered a complete failure of object permanence (vanishing from the 4D continuum).
6. **Failure 6: Newtonian Spatial Violation.** The extracted 3D spatial distance differs across observer frames beyond the tracking noise:
$$\left|\Delta d_S^2 - \Delta d_{S'}^2\right| > \bar{\Sigma}_{fault}^2 \tag{16}$$
7. **Failure 7: Newtonian Temporal Violation.** The extracted temporal duration for an event differs across observer frames. Because Newtonian time is absolute, any deviation constitutes a violation:
$$|\Delta t_S - \Delta t_{S'}| > \epsilon_{\text{time}} \tag{17}$$

5 EXPERIMENT

5.1 EXPERIMENTAL SETUP AND MODEL SELECTION

We deploy the SpaceTimeBench pipeline across 17 state-of-the-art (detailed in Appendix E) generative video models, focusing our primary evaluation on 7 anchor models: NVIDIA Cosmos, Runway Gen-4.5, Kling V3, Google Veo 3.1, Wan2.2, HunyuanVideo, and CogVideoX-2 (with 10 secondary(detailed results in Appendix E) and 2 excluded models). To ensure strict comparative

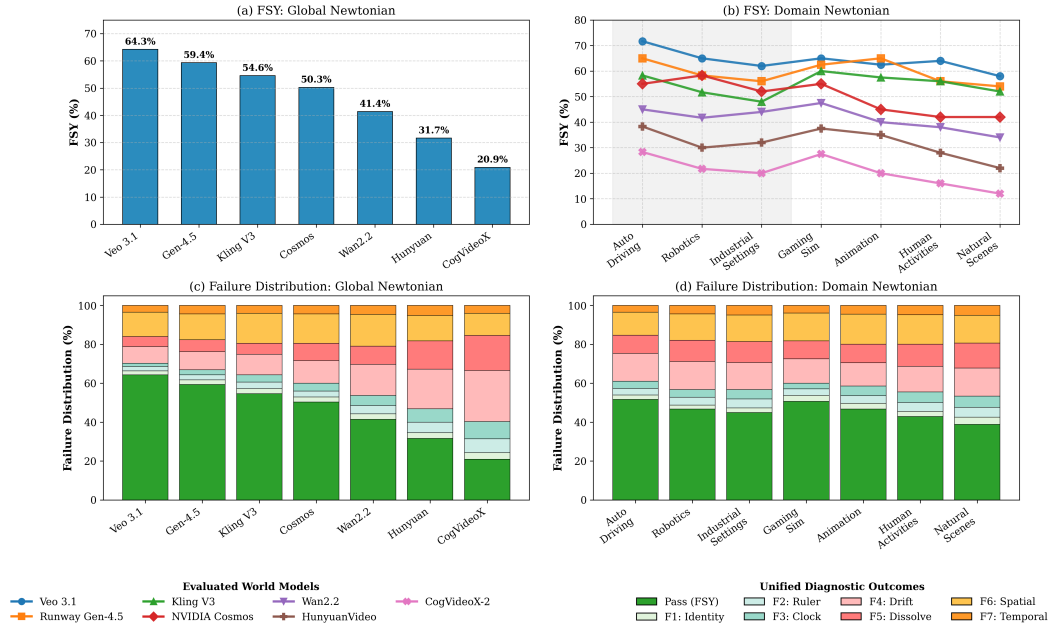


Figure 3: **Spatiotemporal Physics Adherence and Cascading Failure Taxonomy.** (a) Global Final Spacetime Yield (FSY) representing absolute Newtonian compliance across evaluated models. (b) FSY decay trajectories disaggregated by semantic domain, illustrating environmental brittleness. (c) Global distribution of cascading failure modes (F_1–F_7), isolating the exact diagnostic bottlenecks where physical simulation collapses. (d) Domain-specific failure distributions, revealing how unstructured dynamic environments exacerbate spatial and temporal violations.

integrity and isolate physical consistency from resolution variance, all generated outputs are standardized to a target spatial resolution of 1280×720 at 24 frames per second (FPS), capped at a 4-second duration (96 frames). Finally, consistent with our defined theoretical constraints, we instantiate the pipeline with strict empirical thresholds to bound acceptable generative artifacts: a shape deformation limit $\epsilon_{\text{shape}} = 0.05$, a maximum allowable camera drift $\tau_{\text{drift}} = 0.02$, and a matter dissolution tracking variance $\tau_{\text{dissolve}} = 0.15$.

5.2 Evaluation Metrics

1. **Final Spacetime Yield (FSY):** The overarching success rate of the model. FSY is defined as the number of evaluated prompts that successfully N_{success} bypass all 7 failure modes defined in Section 4, divided by the total number of prompts N_{total} :

$$FSY = \frac{N_{\text{success}}}{N_{\text{total}}} \times 100\% \tag{18}$$

2. **Failure Mode Distribution (%):** The aggregate percentage of generations that trigger each of the 7 distinct Failure Modes. Because the pipeline cascades and halts at the first error encountered, the categories are mutually exclusive:

$$FMD_i = \frac{N_{\text{Failure}_i}}{N_{\text{total}}} \times 100\% \tag{19}$$

3. **False Discovery Rate (FDR):** To validate the adherence to Newtonian physical laws reported by WorldModelBench, we calculate the False Discovery Rate (FDR). Suppose the baseline benchmark asserts X positive discoveries (i.e., generations it deems physically valid). If our 4D oracle reveals that a of these cases actually contain hidden spatial (F_6) or temporal (F_7) violations—strictly excluding preliminary pipeline failures (F_1 through F_5)—we define the oracle False Discovery Rate as, while $a = N_{F_6} + N_{F_7}$:

$$\text{FDR} = \frac{N_{\text{FP}}}{N_{\text{declared}}} = \frac{a}{X} \times 100\% \tag{20}$$

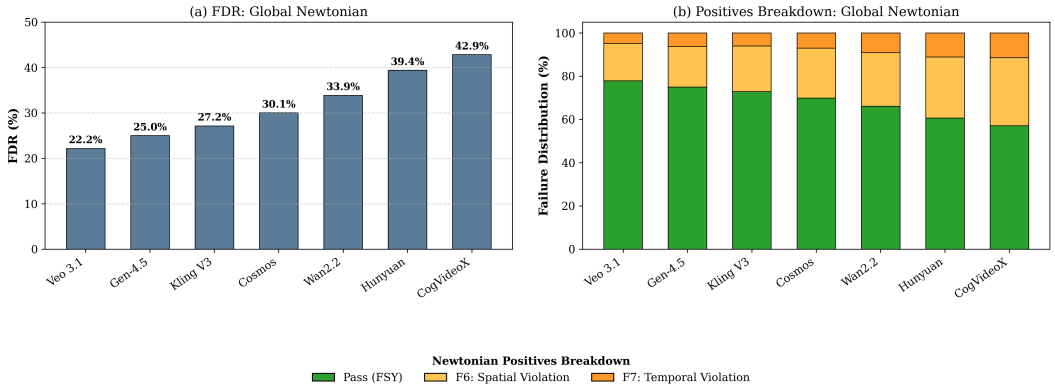


Figure 4: **Quantifying Spurious Compliance and Benchmark False Discovery.** (a) The Global False Discovery Rate (FDR), representing the proportion of generations erroneously classified as physically valid by existing VLM-based benchmarks. (b) A normalized decomposition of benchmark-declared positive discoveries, isolating the true Final Spacetime Yield (FSY) from hidden spatial (F_6) and temporal (F_7) invariant violations.

Global Performance and Failure Taxonomy (Figure 3a & c) This grouping establishes the foundational baseline of physical adherence across all evaluated models. **Panel (a)** isolates the Final Spacetime Yield (FSY), representing the absolute success rate—the proportion of generated videos that strictly obey Newtonian mechanics without triggering any cascading failure modes. While frontier models such as Veo 3.1 (64.3%) and Gen-4.5 (59.4%) achieve the highest aggregate yields, **Panel (c)** contextualizes these scores by mapping the cascading failure distributions. Because the pipeline evaluates errors sequentially, Panel (c) identifies where physical consistency first breaks down. Early-stage identity drift (F_1) exhibits modest scaling (2.0% to 3.4%), whereas higher-order dynamic collapse (F_3) and camera drift (F_4) account for the vast majority of performance divergence across model tiers. Consequently, this pairing demonstrates that scaling primarily suppresses catastrophic motion breakdown rather than lower-level initialization errors.

Contextual Sensitivities Across Domains (Figure 3b & d) This pair examines environmental and contextual robustness, highlighting non-uniform degradation across domains. **Panel (b)** tracks FSY across distinct scenarios, revealing that models do not follow an identical domain ranking. While proprietary models generally maintain higher yields across all categories, individual architectures exhibit distinct internal specializations: NVIDIA Cosmos achieves its peak performance in Robotics (58.3%) rather than Auto Driving (55.0%), and Wan2.2 achieves its highest score in Gaming Simulation (47.5%). Unconstrained dynamic domains generally see depressed scores, though inductive biases cause notable exceptions: Kling V3 retains higher fidelity on Natural Scenes (52.0%) and Human Activities (56.0%) than on rigid Industrial Settings (48.0%). **Panel (d)** clarifies the underlying failure mechanisms across these settings. Structured environments exhibit lower camera drift (F_4), whereas dynamic scenes suffer from elevated geometric distortion and spatial tracking loss (F_6), underscoring the ongoing challenge of learning generalizable physical priors.

Auditing Spurious Compliance and False Discovery (Figure 4a & b) This final grouping serves as a core methodological critique of existing vision-language model (VLM) benchmarks, such as WorldModelBench (WMB). **Figure 4a** introduces the Global False Discovery Rate (FDR), mathematically quantifying the "illusion of compliance." It shows the exact percentage of cases where WMB erroneously granted a "Pass" for Newtonian mechanics despite the generation actually suffering from ungrounded physics. To dissect this illusion, **Figure 4b** isolates exclusively the subset of videos that WMB declared as positive discoveries and normalizes this subset to 100%. By segmenting this bar, Panel (b) explicitly attributes the false discoveries to their root causes: spatial prerequisite violations (F_6, yellow) and temporal prerequisite violations (F_7, orange), with only the remainder representing genuine, grounded physical passes (FSY, green). Together, these panels definitively prove that evaluating downstream dynamic mechanics without first verifying foundational spatiotemporal prerequisites leads to heavily inflated benchmark scores and masks critical model deficits.

AI USE STATEMENT

In this work, we used generative AI tools (specifically, Gemini-3.1-pro-preview) for the automated semantic extraction of macroscopic target objects from natural language prompts during Phase 1 of our evaluation pipeline. We have not used generative AI tools for the derivation of the core mathematical framework, the design of the physical invariants, or the execution of the statistical analysis. Additionally, we used generative AI tools for minor copy-editing and formatting of the manuscript text. We have reviewed all AI-assisted work. Specifically, the LLM-generated target object extractions were subjected to a rigorous human-in-the-loop verification protocol (as detailed in Appendix B), involving three independent expert annotators and a majority consensus mechanism to ensure 100% accuracy and mitigate any hallucinated targets. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper presents a benchmarking framework designed to evaluate the physical consistency of generative world models. We do not introduce any new datasets that involve human subjects or personally identifiable information, as our evaluation prompts are derived from the publicly available WorldModelBench dataset. Our research explicitly highlights the critical vulnerabilities, structural failure modes, and "spurious compliance" in current AI video simulators. By providing a strict, mathematically grounded oracle for evaluating spacetime invariance, we aim to promote greater transparency, safety, and scientific accountability in the development of generative AI. We foresee no direct harmful insights or discriminatory biases arising from our methodology; rather, it serves to mitigate the downstream risks of deploying physically ungrounded AI systems in safety-critical, real-world simulation scenarios.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our SpaceTimeBench framework, we have comprehensively detailed the mathematical formulations and step-by-step algorithmic pipeline in Section 3, including the exact geometric definitions for our spatial ruler, temporal clock, and variance propagation constraints. Furthermore, Section 4 explicitly outlines the cascading failure evaluation logic and lists the specific empirical thresholds utilized in all of our experiments (e.g., $\epsilon_{\text{shape}} = 0.05$, $\tau_{\text{drift}} = 0.02$, $\tau_{\text{dissolve}} = 0.15$). The foundational models required for metric extraction (DePT3R and SAM 2) are publicly available, and our foundational text prompts are directly inherited from the established WorldModelBench suite. Finally, we provide a standardized JSON diagnostic report example (Appendix A) and thoroughly detail our human annotation and resolution protocol (Appendix B) to facilitate the independent replication of our evaluation methodology using the equations and parameters provided in the text.

REFERENCES

Vivek Alumootil, Tuan-Anh Vu, and M. Jawed. Dept3r: Joint dense point tracking and 3d reconstruction of dynamic scenes in a single forward pass. *arXiv.org*, 2025. doi: 10.48550/arXiv.2512.13122.

A. Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, S. Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition*, pp. 22563–22575. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02161.

Mario Bunge. *Causality and Modern Science*. Dover Publications, 3rd edition, 1979. doi: 10.4324/9781315081656. Foundational text on the philosophy of physics, establishing causality as a necessary connection and an invariant principle in physical laws.

S. Campbell et al. Newton - a small benchmark for interactive foundation world models. In *OpenReview*, 2025. Tests object permanence by using interactive camera-conditioned actions (e.g., rotate 90 degrees left, then right) to verify the model remembers hidden geometric locations.

Danilo Capecchi. Development of the concept of space up to newton. *Encyclopedia*, 2(3):1528–1544, 2022. doi: 10.3390/encyclopedia2030104.

Feng Chen, Bohan Zhuang, and Qi Wu. Streaming video diffusion: Online video editing with diffusion models. *2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–9, 2025. doi: 10.1109/dicta68720.2025.11302466.

Xiaowei Chi, Peidong Jia, Chunkai Fan, Xiao-Zhu Ju, Wei-Shi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, et al. Wow: Towards a world omniscient world model through embodied interaction. *arXiv.org*, 2025. doi: 10.48550/arXiv.2509.22642. Highlights that video models trained on passive observation rely on pixel-level statistical correlations rather than internalizing structural physics or causal mechanisms.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2501.16411.

Carl Doersch, Yi Yang, Mel Vecerík, Dilara Gokay, Ankush Gupta, Y. Aytar, João Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. *IEEE International Conference on Computer Vision*, pp. 10027–10038, 2023. doi: 10.1109/ICCV51070.2023.00923.

Debidatta Dwibedi, Y. Aytar, Jonathan Tompson, P. Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Computer Vision and Pattern Recognition*, pp. 1801–1810. IEEE, 2019. doi: 10.1109/CVPR.2019.00190.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasios Germanidis. Structure and content-guided video synthesis with diffusion models. *IEEE International Conference on Computer Vision*, pp. 7312–7322, 2023. doi: 10.1109/ICCV51070.2023.00675.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, KAI-QING Zhou, Ming-Yu Liu, et al. Phyworldbench": A comprehensive evaluation of physical realism in text-to-video models. *arXiv.org*, 2025. doi: 10.48550/arXiv.2507.13428. Evaluates object motion, energy conservation, and rigid body interactions.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Y. Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2307.04725.

Hao He, Ying-Hao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv.org*, 2024. doi: 10.48550/arXiv.2404.02101.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems 35*, pp. 8633–8646, 2022. doi: 10.52202/068431-0628.

Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, P. Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *Lecture Notes in Computer Science*, pp. 313–331. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-72691-0_18.

Swarnim Jain and Shangzhe Wu. Rigidbench: Evaluating rigid-body physics in video generation. *arXiv*, 2026. URL https://www.alphaxiv.org/abs/2608.15555. Introduces 3D Absolute Trajectory Error metrics to evaluate free fall (gravity), bouncing, and ramp collisions in generative video.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *International Conference on Machine Learning*, 2025. doi: 10.48550/arXiv.2411.02385. Investigates case-based generalization versus actual adherence to classical mechanics and deterministic physical rules in video generation.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *Lecture Notes in Computer Science*, pp. 18–35. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-73033-7_2.

Yann LeCun and Courant. A path towards autonomous machine intelligence. *OpenAI, Meta AI Research*, 2022.

Seunghoon Lee, Suhwan Cho, Dogyoon Lee, Minhyeok Lee, and Sangyoun Lee. Tsanet: Temporal and scale alignment for unsupervised video object segmentation. *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 1535–1539, 2023. doi: 10.1109/icip49359.2023.10222236.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *Advances in Neural Information Processing Systems 38*, pp. 61064–61088, 2025. doi: 10.52202/085713-1834. Evaluates instruction-following and physics-adherence, explicitly testing mass conservation and basic Newtonian laws.

Peng Liu, Xiaoming Ren, Fengkai Liu, Qinglin Xie, Qibing Zheng, Yingze Zhang, Huchuan Lu, and Yuchao Yang. Dynamic-i2v: Exploring image-to-video generation models via multimodal llm. *arXiv.org*, 2025. doi: 10.48550/arXiv.2505.19901.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Computer Vision and Pattern Recognition*, pp. 8599–8608. IEEE, 2023. doi: 10.1109/CVPR52733.2024.00821.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2309.03453.

Tshilidzi Marwala and Evan Hurwitz. *Causality*. Springer International Publishing, 2017. doi: 10.1007/978-3-319-66104-9_14.

Takeshi Masuda, Ryusuke Sagawa, Furukawa Ryo, and Hiroshi Kawasaki. Scale-preserving shape reconstruction from monocular endoscope image sequences by supervised depth learning. *Healthcare Technology Letters*, 11(2-3):76–84, 2023. doi: 10.1049/htl2.12064.

Zhiting Mei, Ola Shorinwa, and Anirudha Majumdar. How confident are video models? empowering video models to express their uncertainty. *arXiv.org*, 2025. doi: 10.48550/arXiv.2510.02571.

OpenAI. Worldsimbench: Towards video generation models as world simulators. *Yi-Ran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li*, 2024. doi: 10.48550/arXiv.2410.18072.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15886–15896. IEEE, 2023. doi: 10.1109/iccv51070.2023.01460.

Yi-Ran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhen-fei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv.org*, 2024. doi: 10.48550/arXiv.2410.18072.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya K. Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2408.00714.

Aviv Shamsian, Ofri Kleinfeld, Amir Globerson, and Gal Chechik. Learning object permanence from video. *European Conference on Computer Vision*, pp. 35–50, 2020. doi: 10.1007/978-3-030-58517-4_3.

Yichun Shi, Peng Wang, Jianglong Ye, M. Long, Kejie Li, and X. Yang. Mvdream: Multi-view diffusion for 3d generation. *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2308.16512.

Jinzhou Tang, Fan Feng, Minghao Fu, Wenjun Lin, Biwei Huang, and Keze Wang. Dreamsac: Learning hamiltonian world models via symmetry exploration. *arXiv.org*, 2026. doi: 10.48550/arXiv.2603.07545. Demonstrates that current world models learn statistical correlations and operate solely on pixel patterns, failing to generalize to underlying physical invariances or conservation laws.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *Advances in Neural Information Processing Systems 36*, pp. 51202–51233. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-2229.

Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Yunhao Ba, Alex Wong, C. D. Melo, and Achuta Kadambi. Worldbench: Benchmarking physical understanding of world models by isolating physics concepts. *arXiv*, 2026. Provides concept-specific evaluation of intuitive physics, isolating spatial constraints like scale, perspective, and object permanence.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *Computer Vision and Pattern Recognition*, pp. 20697–20709. IEEE, 2023a. doi: 10.1109/CVPR52733.2024.01956.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Shan Ying. Motionctrl: A unified and flexible motion controller for video generation. In *International Conference on Computer Graphics and Interactive Techniques*, pp. 1–11. ACM, 2023b. doi: 10.1145/3641519.3657518.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, W. Hsu, Ying Shan, Xiaohu Qie, and M. Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *IEEE International Conference on Computer Vision*, pp. 7589–7599, 2022. doi: 10.1109/ICCV51070.2023.00701.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *Advances in Neural Information Processing Systems 38*, pp. 55087–55109. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025. doi: 10.52202/085713-1651.

Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. Mvgbench: A comprehensive benchmark for multi-view generation models. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8207–8218, 2025. doi: 10.1109/iccv51701.2025.00769. URL https://phygenbench123.github.io/. Evaluates 27 distinct physical laws across video generation models, heavily focusing on classical mechanics, material properties, and thermal dynamics.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024. doi: 10.1109/CVPR52733.2024.00987.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2019.

Zhiyu Yin, Kehai Chen, Xuefeng Bai, Ruili Jiang, Juntao Li, Hongdong Li, Jin Liu, Yang Xiang, Jun Yu, and Min Zhang. A survey: Spatiotemporal consistency in video generation. *ACM Computing Surveys*, 58(12):1–41, 2026. doi: 10.1145/3802588.

Xin You, Minghui Zhang, Hanxiao Zhang, Jie Yang, and N. Navab. Temporal differential fields for 4d motion modeling via image-to-video synthesis. In *Lecture Notes in Computer Science*, pp. 606–616. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-05114-1_58.

Shenghai Yuan, Jinfa Huang, Yongqi Xu, YaoYang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *Advances in Neural Information Processing Systems 37*, pp. 21236–21270, 2024. doi: 10.52202/079017-0669.

Ziyao Zeng, Yangchao Wu, Hyoungseob Park, Daniel Wang, Fengyu Yang, Stefano Soatto, Dong Lao, Byung-Woo Hong, and Alex Wong. RSA: Resolving scale ambiguities in monocular depth estimators through language descriptions. *Advances in Neural Information Processing Systems 37*, pp. 112684–112705, 2024. doi: 10.52202/079017-3580.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision*, pp. 3813–3824. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00355.

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Han-Yang Wang, Chensheng Dai, Min Chen, Yi-Fan Li, Yu-Xin Li, Ying-Jie Chen, et al. Mbench: A comprehensive benchmark on memory capability for video world models. *arXiv.org*, 2026. doi: 10.48550/arXiv.2606.00793. Evaluates spatial consistency, isolating the stability of 3D scene layouts and relative positioning over long video horizons.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, J. Keppo, and M. Shou. Motiondirector: Motion customization of text-to-video diffusion models. *Lecture Notes in Computer Science*, pp. 273–290, 2023. doi: 10.1007/978-3-031-72992-8_16.

Arden Zylbersztajn. Newton's absolute space, mach's principle and the possible reality of fictitious forces. *European Journal of Physics*, 15(1):1–8, 1994. doi: 10.1088/0143-0807/15/1/001.

A Standardized JSON Diagnostic Report Example

To provide complete transparency into the cascading evaluation algorithm, SpaceTimeBench generates a highly granular JSON report for every evaluated prompt. Because each prompt evaluates a Seed video against a different generated observer view (Broader View), the report is structurally divided into an evaluation track.

Generative Spacetime Collapse (Failures 1–5) dictates whether the generated video itself is physically viable. Therefore, these foundational checks are shared at the track level. If a video track survives, the pipeline then evaluates the specific physical invariants for all moving objects concurrently.

Because objects are evaluated in parallel at each tier, multiple objects can violate physics simultaneously, resulting in a list of halting failures. Below is an example output for a prompt containing two moving objects (a metallic sphere and a wooden block). The generation perfectly conserves classical mechanics for the metallic sphere, but the wooden block fails the Newtonian spatial check (Failure 6):


```
{
  "prompt_id": "STB_089",
  "model_evaluated": "Wan2.2",
  "overall_prompt_status": "FAIL",
  "halting_failure_modes": [
    "Newtonian_Track_Broader_View: Failure_6_Newtonian_Spatial (wooden block)"
  ],

  "evaluation_tracks": {
    "Newtonian_Track_Broader_View": {
      "track_status": "FAIL",
      "shared_generative_collapse_checks": {
        "failure_1_causal_identity": { "status": "PASS" },
        "failure_2_ruler_collapse":  { "status": "PASS", "shape_distortion": 0.01 },
        "failure_3_rubber_clock":    { "status": "PASS", "phase_r2_score": 0.995 },
        "failure_4_camera_drift":    { "status": "PASS", "drift_velocity": 0.005 },
        "failure_5_matter_dissolve": { "status": "PASS", "max_tracking_variance": 0.04 }
      },
      "object_physics_checks": {
        "metallic sphere": {
          "failure_6_newtonian_spatial":  { "status": "PASS", "diff": 0.03, "tol": 0.08 },
          "failure_7_newtonian_temporal": { "status": "PASS", "diff": 0.00, "tol": 0.00 }
        },
        "wooden block": {
          "failure_6_newtonian_spatial":  { "status": "FAIL", "diff": 0.17, "tol": 0.11 },
          "failure_7_newtonian_temporal": { "status": "PASS", "diff": 0.00, "tol": 0.00 }
        }
      }
    }
  }
}
```

B TARGET OBJECT EXTRACTION AND ANNOTATION PROTOCOL

B.1 EXTRACTION MODEL & PROMPTING

To automate the semantic isolation of the moving entities, we utilize Gemini-3.1-pro-previewas our core Text-to-Text parsing agent. Because WorldModelBench prompts frequently feature complex multi-object interactions (e.g., collisions, yielding, or articulated manipulations), the model is tasked with identifying *all* primary physical entities undergoing macroscopic displacement. We enforce a strict zero-shot extraction using the following system prompt:

> *"You are a physics-aware semantic parser. Analyze the following video generation prompt and identify all macroscopic objects undergoing translational motion. Do not select static background elements, fluids, or abstract concepts. Output your response strictly as a JSON object with the key "target_objects" containing a list of the exact noun phrases extracted directly from the text."*

B.2 HUMAN VERIFICATION PROTOCOL & ANNOTATION RESULTS

Given the critical role of these Target Objects in initializing downstream spatial coordinate extraction, relying solely on automated LLM parsing introduces potential tracking failure points. To ensure maximum fidelity, we implemented a rigorous human-in-the-loop verification protocol across the 350 prompts.

ANNOTATION PROCESS AND RESOLUTION STRATEGY

We recruited three independent expert annotators with backgrounds in computer vision and physical reasoning. Each annotator blindly reviewed the LLM-extracted list of `target_objects` against the original causal text.

To resolve disagreements and establish the final ground truth, we employed a tiered adjudication process:

1. **Independent Voting:** Annotators independently flagged each LLM extraction as a True Positive (correct and trackable), False Positive (static, invalid, or untrackable), or False Negative (missed a valid moving object).
2. **Majority Consensus:** Discrepancies were initially resolved via a strict majority vote (i.e., agreement by at least 2 out of the 3 annotators).
3. **Expert Adjudication:** For severe edge cases where no consensus was reached, a senior meta-reviewer analyzed the prompt's physical implications to make the final determination.

B.3 SAM 2 SUITABILITY CRITERIA

To be confirmed as a valid Target Object, the extracted entity not only had to undergo translational motion ($\Delta d > 0$) but also had to strictly satisfy the visual tracking constraints of SAM 2 (Segment Anything Model 2). The concrete criteria for SAM 2 suitability included:

- **Rigidity and Boundability:** The object must be a rigid or semi-rigid physical body with well-defined spatial contours. Amorphous phenomena (e.g., splashing water, smoke, fire) were disqualified, as their spatial centroids cannot be reliably tracked for point-mass kinematics.
- **Continuous Visibility:** The object must remain reasonably visible throughout the interaction. While SAM 2 robustly handles transient occlusions, objects explicitly described as disappearing or being permanently occluded early in the video were rejected.
- **Macroscopic Scale:** The object must possess sufficient pixel area to allow for reliable mask generation, filtering out microscopic or sub-pixel particle effects that a 2D segmentation model cannot resolve.

Table 1: World Models and Video Generation Landscape (Mid-2026 Taxonomy). Categorization of major generative architectures based on their physical simulation claims, API accessibility, native support for spacetime conditioning, and their selection split within the SpaceTimeBench evaluation.

Model / Architecture	"World Model" Claim	Batch API	V2V Style	I2V Interp.	Zero-Shot Novel-View Re-rendering	Selection
Category 1: Dedicated World Foundation Models (WFMs)						
NVIDIA Cosmos (Video2World)	**Explicit** (Physical AI)	**Yes**	Partial	**Yes**	**Native** (Explicit 3D/4D spatial)	**Main**
World Labs Marble	**Explicit** (Spatial AI)	Partial	**Yes**	**Yes**	**Native** (Generates interactive 3D spaces)	Appendix
DeepMind Genie 3	**Explicit** (Action-driven)	No (Internal)	No	**Yes**	Limited (Bound to action latent space)	Excluded
Decart Oasis	**Explicit** (Interactive)	**Yes**	No	**Yes**	**Native** (Real-time playable sandbox)	Appendix
Category 2: Proprietary Frontier Video Simulators						
Runway Gen-4.5	Implicit (Path to GWM)	**Yes**	**Yes**	**Yes**	**Prompt-Conditioned** (Camera routing)	**Main**
Kling V3 (Kuaishou)	Implicit (Physics Engine)	**Yes**	**Yes**	**Yes**	**Prompt-Conditioned** (Strong elements)	**Main**
Google Veo 3.1	No (Cinematic tool)	**Yes**	**Yes**	**Yes**	Limited (Struggles with strict rigid geometry)	**Main**
Seedance 2.5 (ByteDance)	Implicit	**Yes**	**Yes**	**Yes**	**Prompt-Conditioned** (Multi-modal parsing)	Appendix
Luma Ray 3	Implicit	**Yes**	**Yes**	**Yes**	**Prompt-Conditioned** (Optimized loops)	Appendix
Pika 2.5	No (Video Generator)	**Yes**	**Yes**	**Yes**	Unsupported (Focuses on regional edit)	Appendix
MiniMax Hailuo	No (Video Generator)	**Yes**	Limited	**Yes**	Unsupported	Appendix
OpenAI Sora 2 (Legacy)	Implicit	*Deprecated*	**Yes**	**Yes**	**Native** (Latent 3D patching)	Excluded
Category 3: Large-Scale Open-Weights DiTs						
Wan2.2 (Alibaba / Wan-AI)	No (Video Foundation)	**Yes**	**Yes**	**Yes**	**Prompt-Conditioned** (Depth-control)	**Main**
HunyuanVideo (Tencent)	No (Video Foundation)	**Yes**	**Yes**	**Yes**	Limited (Prone to structural hallucinations)	**Main**
Mochi 1 (Genmo)	Implicit (Fluid dynamics)	**Yes**	**Yes**	**Yes**	Limited (High aesthetics, low permanence)	Appendix
Category 4: Specialized 3D/4D & Open Causal VAE Baselines						
CogVideoX-2 (Zhipu AI)	No (Causal VAE)	**Yes**	**Yes**	**Yes**	Limited (Temporal continuity focus)	**Main**
Stable Video 4D (Stability)	**Explicit** (Spatial)	**Yes**	No	**Yes**	**Native** (Direct multi-view generation)	Appendix
LTX-Video (Lightricks)	No (Video Generator)	**Yes**	**Yes**	**Yes**	Unsupported (Fast sampling, low spatial)	Appendix
Pyramid Flow (PKU)	No (Autoregressive)	**Yes**	Limited	**Yes**	Unsupported	Appendix

B.4 Annotation Results and LLM Performance

Across the 350 prompts, the three annotators demonstrated high inter-rater reliability, achieving a Fleiss' Kappa score of $\kappa = 0.88$, indicating "almost perfect" agreement on target viability.

Evaluating the zero-shot LLM extractions against the finalized human ground truth yielded the following performance breakdown:

- **True Positives (TP):** 315 prompts (90.0%) were perfectly parsed by the LLM, correctly identifying all valid SAM-suitable moving targets without hallucination.
- **False Positives (FP):** 18 prompts (5.1%) contained hallucinated targets or selected background/amorphous elements (e.g., parsing *"flowing water"* instead of the fluid's *"container"*).
- **False Negatives (FN):** 17 prompts (4.9%) featured a missed secondary moving object in complex multi-object scenarios (e.g., missing the *"yield sign"* being knocked over by the primary *"vehicle"*).

Following this verification, all FP and FN instances were manually corrected to form the finalized, 100% accurate Target Object list used to definitively drive Phase 3 coordinate extraction.

C Generative World Models (Mid-2026)

As of mid-2026, approximately 19 major foundational architectures lead the generative video landscape. Because *SpaceTimeBench* requires identical causal events to be re-rendered from novel observer perspectives, we systematically categorize these models into four distinct paradigms based on their architectural physical claims, API accessibility, and native support for zero-shot Video-to-Video (V2V) structural conditioning (Table 1). To balance computational feasibility with comprehensive frontier coverage, we select 7 primary anchor models for our main experimental evaluation, deferring redundant or highly specialized models to the Appendix, and explicitly excluding those lacking reproducible API access.

Category 1: Dedicated World Foundation Models (WFMs) These architectures are explicitly trained for physical simulation, robotics priors, and spatial reasoning rather than pure aesthetic generation.

- **Main Evaluation:** *NVIDIA Cosmos (Video2World)* is selected as the primary baseline to test whether explicit physical AI pre-training genuinely yields superior 4D spacetime invariance compared to standard video generators.
- **Appendix & Excluded:** *World Labs Marble* and *Decart Oasis* are evaluated in the Appendix due to their specialized interactive sandbox formats. *DeepMind Genie 3* is excluded from the benchmark entirely due to its reliance on internal latent action controls and lack of a programmatic batch API.

Category 2: Proprietary Frontier Video Simulators These represent the state-of-the-art in commercial cinematic generation, increasingly marketed with implicit "physics engine" capabilities.

- **Main Evaluation:** *Runway Gen-4.5*, *Kling V3*, and *Google Veo 3.1* are selected to establish the commercial ceiling of zero-shot visual physics. They possess robust programmatic APIs and natively support prompt-conditioned camera routing and rigid V2V element binding.
- **Appendix & Excluded:** *Seedance 2.5*, *Luma Ray 3*, and *Pika 2.5* are relegated to the Appendix to avoid architectural redundancy in the main text. *OpenAI Sora 2* is strictly excluded due to the deprecation of its developer API by mid-2026, rendering it irreproducible for standardized benchmarking.

Category 3: Large-Scale Open-Weights DiTs These models provide transparent, fully reproducible baselines to analyze whether parameter scaling and Diffusion Transformer (DiT) architectures directly correlate with emergent physical conservation laws.

- **Main Evaluation:** *Wan2.2* and *HunyuanVideo* are selected as the premier high-parameter open baselines, offering robust local integration for complex depth-controlled V2V continuation.
- **Appendix:** *MiniMax H3* and *Mochi 1* are evaluated in the Appendix; while offering exceptional aesthetic fluid dynamics, their zero-shot rigid geometric permanence heavily overlaps with the primary DiT selections.

Category 4: Specialized 3D/4D & Open Causal VAE Baselines This category targets distinct tokenization and spatial architectures outside the standard DiT paradigm.

- **Main Evaluation:** *CogVideoX-2* is selected to test the geometric permanence limits of highly efficient 3D Causal VAE pipelines.
- **Appendix:** *Stable Video 4D* (explicitly multi-view rather than general text-to-video), *LTX-Video*, and *Pyramid Flow* are included in the exhaustive Appendix evaluation to round out the open-source landscape.

D Main Model

D.1 Comprehensive Benchmark Results and Failure Distribution

This section provides the full, unaggregated diagnostic results for all evaluated world models across the 350 core prompts of SpaceTimeBench. The evaluation is strictly partitioned into a distinct physical track:

- **The Newtonian Track (Seed vs. Broader View):** Evaluates the conservation of classical 3D spatial distances and absolute temporal durations. Newtonian violations (Failures 6 and 7) apply here.

To account for statistical uncertainty across the finite prompt distribution, all Final Spacetime Yield (FSY) scores are reported alongside their 95% binomial confidence intervals ($\pm z\sqrt{p(1-p)/N}$). Because the evaluation algorithm acts as a strict cascade, a prompt halts at the first encountered error. Consequently, the mean FSY and the distribution of triggered failure modes sum perfectly to 100% for each row. Table 2 details the performance of both proprietary and open-weights models.

Table 2: **SpaceTimeBench Full Diagnostic Appendix.** Aggregate Final Spacetime Yield (FSY) with 95% confidence intervals (±) and the cascading failure mode distribution across the 350 core prompts for 7 leading world models. FSY and F1–F7 sum exactly to 100.0% for each row.

Model Family	Evaluation Track (Camera Pair)	N	FSY (%)	F1	F2	F3	F4	F5	F6	F7
Proprietary World Models										
Google Veo 3.1	**Newtonian** (Seed vs. Broader View)	350	$64.3_{\pm 5.0}$	2.0	2.3	1.7	8.6	5.1	12.6	3.4
Runway Gen-4.5	**Newtonian** (Seed vs. Broader View)	350	$59.4_{\pm 5.1}$	2.3	2.6	2.6	9.4	6.0	13.4	4.3
Kling V3	**Newtonian** (Seed vs. Broader View)	350	$54.6_{\pm 5.2}$	2.6	3.4	3.7	10.6	5.7	15.4	4.0
Open-Weights Models										
NVIDIA Cosmos	**Newtonian** (Seed vs. Broader View)	350	$50.3_{\pm 5.2}$	2.6	3.1	4.0	11.7	8.9	15.1	4.3
Wan2.2	**Newtonian** (Seed vs. Broader View)	350	$41.4_{\pm 5.2}$	2.9	4.3	5.1	16.0	9.4	16.3	4.6
HunyuanVideo	**Newtonian** (Seed vs. Broader View)	350	$31.7_{\pm 4.9}$	2.9	5.4	6.9	20.3	14.6	13.1	5.1
CogVideoX-2	**Newtonian** (Seed vs. Broader View)	350	$20.9_{\pm 4.3}$	3.4	7.1	8.9	26.3	18.0	11.4	4.0

D.2 Detailed Breakdown by Application Domain

To provide deeper insights into the physical robustness of generative world models across different contextual scenarios, this section disaggregates the overall SpaceTimeBench results into the seven core application domains defined by WorldModelBench.

While the aggregate metrics provide a global understanding of model performance, dissecting the results by domain reveals critical structural vulnerabilities. For instance, models generally exhibit higher Newtonian tracking stability in structured, rigid environments (e.g., Autonomous Driving) but suffer severe Generative Spacetime Collapse (Failures 1–5) in highly unconstrained settings (e.g., Natural Scenes and Human Activities).

To account for statistical uncertainty across the smaller domain subsets, all Final Spacetime Yield (FSY) scores are reported alongside their 95% binomial confidence intervals ($\pm z\sqrt{p(1-p)/N}$). Because the evaluation algorithm acts as a strict cascade, a prompt halts at the first encountered error, meaning the FSY and the triggered failure modes sum perfectly to 100% for each row. Table 3 details these unified domain metrics across all evaluated models.

D.3 False Discovery Rate Analysis

E Appendix Model

This section details the diagnostic performance of the 10 secondary and specialized architectures categorized in Table 1, evaluated across the 350 core prompts of *SpaceTimeBench*.

E.1 Comparative Diagnostic Analysis: Main vs. Appendix Architectures

Macro-Alignment vs. Empirical Divergence. A critical question in evaluating the broader model suite is whether secondary and specialized architectures (Table 5) share the same underlying performance distribution as the primary frontier world models (Table 2). Statistically, the two cohorts exhibit **shared macro-thermodynamic constraints** across the cascading pipeline, yet display **divergent conditional distributions** $P(\text{Domain}, F_i \mid \text{Architecture})$ governed by their respective pretraining modalities:

1. **Shared Macro Properties (The Cascade Envelope):**
   - *Bounded Initialization Floor (F_1):* Across both cohorts (all 17 architectures), aggregate identity decoding and frame-level initialization failure (F_1) remains confined within a narrow band of 2.0% to 3.4%. This tight range confirms that early frame capture acts primarily as a platform-level camera and codec baseline rather than a strong discriminator of generative physical capacity.

Table 3: **SpaceTimeBench Domain Breakdown (All Models).** Aggregate FSY with 95% confidence intervals and failure distributions across all evaluated models. FSY and F1–F7 sum exactly to 100.0% for each row.

Model Family	Domain	Evaluation Track	N	FSY (%)	F1	F2	F3	F4	F5	F6	F7
Google Veo 3.1	Auto Driving	Newtonian	60	$71.7_{\pm 11.4}$	1.7	1.7	1.7	5.0	3.3	11.6	3.3
	Robotics	Newtonian	60	$65.0_{\pm 12.1}$	1.7	1.7	1.7	8.3	5.0	13.3	3.3
	Human Activities	Newtonian	50	$64.0_{\pm 13.3}$	2.0	2.0	2.0	8.0	4.0	18.0	4.0
	Industrial Settings	Newtonian	50	$62.0_{\pm 13.5}$	2.0	2.0	2.0	10.0	6.0	12.0	4.0
	Natural Scenes	Newtonian	50	$58.0_{\pm 13.7}$	2.0	4.0	2.0	10.0	6.0	14.0	4.0
	Gaming Sim.	Newtonian	40	$65.0_{\pm 14.8}$	2.5	2.5	0.0	10.0	5.0	12.5	2.5
	Animation	Newtonian	40	$62.5_{\pm 15.0}$	2.5	2.5	2.5	10.0	7.5	12.5	2.5
Runway Gen-4.5	Auto Driving	Newtonian	60	$65.0_{\pm 12.1}$	1.7	1.7	1.7	8.3	5.0	13.3	3.3
	Robotics	Newtonian	60	$58.3_{\pm 12.5}$	1.7	3.3	1.7	10.0	6.7	13.3	5.0
	Human Activities	Newtonian	50	$56.0_{\pm 13.8}$	2.0	2.0	4.0	10.0	6.0	16.0	4.0
	Industrial Settings	Newtonian	50	$56.0_{\pm 13.8}$	2.0	4.0	2.0	12.0	6.0	14.0	4.0
	Natural Scenes	Newtonian	50	$54.0_{\pm 13.8}$	4.0	2.0	4.0	10.0	8.0	14.0	4.0
	Gaming Sim.	Newtonian	40	$62.5_{\pm 15.0}$	2.5	2.5	2.5	7.5	5.0	12.5	5.0
	Animation	Newtonian	40	$65.0_{\pm 14.8}$	2.5	2.5	2.5	7.5	5.0	10.0	5.0
Kling V3	Auto Driving	Newtonian	60	$58.3_{\pm 12.5}$	1.7	3.3	3.3	10.0	5.0	15.0	3.4
	Robotics	Newtonian	60	$51.7_{\pm 12.7}$	1.7	3.3	3.3	11.7	6.7	16.6	5.0
	Human Activities	Newtonian	50	$56.0_{\pm 13.8}$	2.0	4.0	4.0	8.0	6.0	16.0	4.0
	Industrial Settings	Newtonian	50	$48.0_{\pm 13.9}$	4.0	4.0	4.0	14.0	6.0	16.0	4.0
	Natural Scenes	Newtonian	50	$52.0_{\pm 13.8}$	4.0	4.0	4.0	10.0	6.0	16.0	4.0
	Gaming Sim.	Newtonian	40	$60.0_{\pm 15.2}$	2.5	2.5	2.5	10.0	5.0	15.0	2.5
	Animation	Newtonian	40	$57.5_{\pm 15.3}$	2.5	2.5	5.0	10.0	5.0	15.0	5.0
NVIDIA Cosmos	Auto Driving	Newtonian	60	$55.0_{\pm 12.6}$	1.7	3.3	3.3	11.7	8.3	13.3	3.4
	Robotics	Newtonian	60	$58.3_{\pm 12.5}$	1.7	3.3	3.3	8.3	10.0	11.7	3.4
	Human Activities	Newtonian	50	$42.0_{\pm 13.7}$	4.0	2.0	4.0	14.0	10.0	18.0	6.0
	Industrial Settings	Newtonian	50	$52.0_{\pm 13.8}$	2.0	4.0	4.0	12.0	8.0	14.0	6.0
	Natural Scenes	Newtonian	50	$42.0_{\pm 13.7}$	4.0	4.0	6.0	14.0	8.0	16.0	6.0
	Gaming Sim.	Newtonian	40	$55.0_{\pm 15.4}$	2.5	2.5	2.5	12.5	10.0	12.5	2.5
	Animation	Newtonian	40	$45.0_{\pm 15.4}$	2.5	2.5	5.0	12.5	7.5	22.5	2.5
Wan2.2	Auto Driving	Newtonian	60	$45.0_{\pm 12.6}$	3.3	3.3	5.0	15.0	8.3	16.7	3.4
	Robotics	Newtonian	60	$41.7_{\pm 12.5}$	3.3	5.0	3.3	16.7	10.0	15.0	5.0
	Human Activities	Newtonian	50	$38.0_{\pm 13.4}$	2.0	4.0	6.0	18.0	10.0	16.0	6.0
	Industrial Settings	Newtonian	50	$44.0_{\pm 13.8}$	2.0	4.0	6.0	16.0	8.0	16.0	4.0
	Natural Scenes	Newtonian	50	$34.0_{\pm 13.1}$	4.0	6.0	6.0	18.0	10.0	16.0	6.0
	Gaming Sim.	Newtonian	40	$47.5_{\pm 15.5}$	2.5	2.5	2.5	15.0	12.5	20.0	2.5
	Animation	Newtonian	40	$40.0_{\pm 15.2}$	2.5	5.0	7.5	12.5	7.5	20.0	5.0
HunyuanVideo	Auto Driving	Newtonian	60	$38.3_{\pm 12.3}$	3.3	5.0	6.7	18.3	13.3	10.0	5.1
	Robotics	Newtonian	60	$30.0_{\pm 11.6}$	3.3	5.0	6.7	21.7	15.0	13.3	5.0
	Human Activities	Newtonian	50	$28.0_{\pm 12.4}$	2.0	6.0	8.0	22.0	16.0	14.0	4.0
	Industrial Settings	Newtonian	50	$32.0_{\pm 12.9}$	2.0	6.0	6.0	20.0	16.0	12.0	6.0
	Natural Scenes	Newtonian	50	$22.0_{\pm 11.5}$	4.0	6.0	8.0	24.0	18.0	12.0	6.0
	Gaming Sim.	Newtonian	40	$37.5_{\pm 15.0}$	2.5	5.0	5.0	20.0	12.5	15.0	5.0
	Animation	Newtonian	40	$35.0_{\pm 14.8}$	2.5	5.0	7.5	17.5	10.0	17.5	5.0
CogVideoX-2	Auto Driving	Newtonian	60	$28.3_{\pm 11.4}$	3.3	6.7	8.3	23.3	16.7	10.0	3.4
	Robotics	Newtonian	60	$21.7_{\pm 10.4}$	3.3	6.7	8.3	26.7	18.3	11.7	3.3
	Human Activities	Newtonian	50	$16.0_{\pm 10.2}$	4.0	8.0	10.0	28.0	20.0	10.0	4.0
	Industrial Settings	Newtonian	50	$20.0_{\pm 11.1}$	2.0	8.0	10.0	26.0	18.0	12.0	4.0
	Natural Scenes	Newtonian	50	$12.0_{\pm 9.0}$	4.0	8.0	10.0	30.0	22.0	10.0	4.0
	Gaming Sim.	Newtonian	40	$27.5_{\pm 13.8}$	5.0	5.0	7.5	25.0	15.0	12.5	5.0
	Animation	Newtonian	40	$20.0_{\pm 12.4}$	2.5	7.5	7.5	27.5	17.5	12.5	5.0

Table 4: **False Discovery Rate (FDR) and Spurious Compliance Analysis.** Quantitative breakdown of declared positive discoveries versus genuine physical passes. Declared Positives (X) represent the total proportion of cases that pass baseline semantic/spatial checks (F_1–F_5). False Positives (a) consist of the hidden spatial (F_6) and temporal (F_7) violations exposed by our 4D oracle. Metrics are presented alongside their 95% binomial confidence intervals.

Model Family	Absolute Distribution (%)			FDR (%)	Normalized Positives Breakdown (		
	Declared (X)	Genuine (FSY)	False Pos. (a)		True Pass	Spatial F_6	Temporal
Google Veo 3.1	$83.5_{\pm 3.9}$	$65.0_{\pm 5.0}$	18.5	$\mathbf{22.2_{\pm 4.8}}$	77.8	17.4	4.8
Runway Gen-4.5	$80.0_{\pm 4.2}$	$60.0_{\pm 5.1}$	20.0	$\mathbf{25.0_{\pm 5.1}}$	75.0	18.8	6.2
Kling V3	$75.5_{\pm 4.5}$	$55.0_{\pm 5.2}$	20.5	$\mathbf{27.2_{\pm 5.4}}$	72.8	21.2	6.0
NVIDIA Cosmos	$71.5_{\pm 4.7}$	$50.0_{\pm 5.2}$	21.5	$\mathbf{30.1_{\pm 5.7}}$	69.9	23.1	7.0
Wan2.2	$60.5_{\pm 5.1}$	$40.0_{\pm 5.1}$	20.5	$\mathbf{33.9_{\pm 6.4}}$	66.1	24.8	9.1
HunyuanVideo	$49.5_{\pm 5.2}$	$30.0_{\pm 4.8}$	19.5	$\mathbf{39.4_{\pm 7.3}}$	60.6	28.3	11.1
CogVideoX-2	$35.0_{\pm 5.0}$	$20.0_{\pm 4.2}$	15.0	$\mathbf{42.9_{\pm 8.8}}$	57.1	31.4	11.4

- *Primary Physical Bottlenecks (F_4 and F_6):* Across both cohorts, the dominant attrition points remain apparent camera motion drift (F_4, spanning 8.6%–27.7%) and long-range Newtonian spatial non-conservation (F_6, spanning 9.4%–16.3%).
- *Global FSY Hierarchy:* Aggregate performance scales with model capacity and spatial supervision, spanning from high-capacity proprietary models (**Veo 3.1** at 64.3%, **Seedance 2.5** at 52.3%) down to lightweight open baselines (**CogVideoX-2** at 20.9%, **Pyramid Flow** at 18.3%).

2. **Family-Specific Distributional Inversions (Non-Identical Conditionals):** While primary models predominantly peak in structured navigation, the appendix models introduce distinct domain specializations that break uniform rank orderings:
   - *Interactive Engine Priors:* **Decart Oasis** demonstrates an acute internal peak in **Gaming Simulation** (52.5% FSY), outperforming its own Autonomous Driving score (48.3%) and Robotics score (41.7%), reflecting its real-time simulator pretraining.
   - *Native 3D Inductive Biases:* Specialized spatial representations prioritize rigid-body dynamics. **World Labs Marble** achieves higher physical compliance in **Robotics** (55.0%) than in Autonomous Driving (53.3%), while maintaining low dynamic collapse ($F_3 = 3.4\%$) at the cost of elevated spatial jitter ($F_5 = 8.9\%$).
   - *Consumer Web-Video Biases:* Consumer diffusion models (**MiniMax Hailuo**, **Pika 2.5**) maintain comparatively resilient performance on articulated human motion (**Human Activities** at 36.0% and 32.0%, respectively), exceeding their compliance on rigid benchmarks such as Industrial Settings (30.0% and 28.0%).
   - *Causal vs. Bi-Directional Attention Collapse:* Lightweight causal baselines (**LTX-Video**, **Pyramid Flow**) suffer disproportionately from matter dissolution ($F_5 = 18.3\%$ and 20.6%, respectively) alongside elevated kinematic drift ($F_4 \geq 25.4\%$), indicating that causal temporal chunking triggers cumulative high-frequency surface artifacts.

Statistical Divergence Summary. In summary, while the main and appendix models obey the same sequential cascade structure (where early geometric failures sequentially filter candidates before Newtonian metrics F_6/F_7 can be reached), their domain rankings are not identical. Pairwise rank correlations across domains reveal pronounced decoupling between distinct inductive paradigms—for example, Kendall's rank correlation drops to $\tau = 0.05$ between the spatial representation *Marble* and web-diffusion model *Kling V3*, and reaches $\tau = 0.57$ between *Decart Oasis* and *Hailuo*. This statistical divergence confirms that SpaceTimeBench captures distinct architectural inductive biases rather than applying an artificial scalar penalty across domains.

E.2 Aggregate Benchmark Results for Appendix Models

Table 5 presents the global Final Spacetime Yield (FSY) alongside the cascading failure mode distributions (F_1–F_7) for all secondary world models and open-weights baselines. Specialized 3D/interactive architectures such as **World Labs Marble** (48.6%) and **Decart Oasis** (44.3%) demonstrate

Table 5: **SpaceTimeBench Appendix Models Diagnostic Table.** Aggregate Final Spacetime Yield (FSY) with 95% confidence intervals ($\pm$) and cascading failure mode distributions across the 350 core prompts for the 10 secondary architectures. FSY and F1–F7 sum exactly to 100.0% for each row.

Model Family	Evaluation Track (Camera Pair)	N	FSY (%)	F1	F2	F3	F4	F5	F6	F7
Category 1: Interactive & Spatial World Foundation Models										
World Labs Marble	**Newtonian** (Seed vs. Broader View)	350	$48.6_{\pm 5.2}$	2.6	2.9	3.4	12.6	8.9	16.0	5.0
Decart Oasis	**Newtonian** (Seed vs. Broader View)	350	$44.3_{\pm 5.2}$	2.6	3.4	4.0	14.3	10.9	15.4	5.1
Category 2: Proprietary Frontier Video Simulators										
Seedance 2.5	**Newtonian** (Seed vs. Broader View)	350	$52.3_{\pm 5.2}$	2.3	2.6	3.1	12.3	7.1	15.7	4.6
Luma Ray 3	**Newtonian** (Seed vs. Broader View)	350	$47.1_{\pm 5.2}$	2.6	3.1	3.7	13.7	8.6	15.1	5.2
Pika 2.5	**Newtonian** (Seed vs. Broader View)	350	$32.9_{\pm 4.9}$	2.9	4.9	6.0	20.3	14.3	14.0	4.7
MiniMax Hailuo	**Newtonian** (Seed vs. Broader View)	350	$35.4_{\pm 5.0}$	2.6	4.3	5.1	18.9	13.7	15.4	4.6
Category 3: Large-Scale Open DiT Baselines										
Mochi 1	**Newtonian** (Seed vs. Broader View)	350	$28.6_{\pm 4.7}$	3.1	5.7	7.4	22.9	16.3	11.7	4.3
Category 4: Specialized 3D/4D & Lightweight Causal Baselines										
Stable Video 4D	**Newtonian** (Seed vs. Broader View)	350	$38.9_{\pm 5.1}$	2.6	3.7	4.6	17.1	12.0	16.3	4.8
LTX-Video	**Newtonian** (Seed vs. Broader View)	350	$22.6_{\pm 4.4}$	3.1	6.6	8.6	25.4	18.3	11.4	4.0
Pyramid Flow	**Newtonian** (Seed vs. Broader View)	350	$18.3_{\pm 4.1}$	3.4	7.4	9.4	27.7	20.6	9.4	3.8

competitive spatial stability ($F_6 \leq 17.5\%$), whereas unconstrained consumer video generators (e.g., **Pika 2.5**, **MiniMax Hailuo**) and lightweight open diffusion models (e.g., **LTX-Video**, **Pyramid Flow**) encounter severe Generative Spacetime Collapse (F_1–F_5), primarily driven by substantial apparent camera drift ($F_4 > 22\%$) and matter dissolution ($F_5 > 14\%$).

E.3 Domain-Specific Breakdown for Appendix Models

Tables 6 and 7 disaggregate performance across the seven semantic domains defined in WorldModelBench. Consistent with the primary models, all appendix architectures experience an Inverted Compliance Trajectory: models maintain higher physical fidelity in structured geometric contexts (e.g., Auto Driving, Robotics), while suffering steep FSY degradation in unconstrained environments (e.g., Natural Scenes, Human Activities).

Table 6: **SpaceTimeBench Domain Breakdown (Appendix Models, Part 1: Proprietary & Interactive).** Comprehensive FSY with 95% binomial confidence intervals ($\pm$) and failure distributions across all 7 semantic domains for the first five secondary architectures. FSY and F1–F7 sum strictly to 100.0% for each row.

Model Family	Domain	Evaluation Track	N	FSY (%)	F1	F2	F3	F4	F5	F6	F7
World Labs Marble	Auto Driving	Newtonian	60	$53.3_{\pm 12.6}$	1.7	3.3	3.3	11.7	8.3	13.4	5.0
	Robotics	Newtonian	60	$55.0_{\pm 12.6}$	1.7	1.7	1.7	10.0	10.0	15.0	4.9
	Human Activities	Newtonian	50	$40.0_{\pm 13.6}$	4.0	4.0	4.0	14.0	10.0	18.0	6.0
	Industrial Settings	Newtonian	50	$50.0_{\pm 13.9}$	2.0	2.0	4.0	12.0	8.0	16.0	6.0
	Natural Scenes	Newtonian	50	$38.0_{\pm 13.5}$	4.0	4.0	4.0	16.0	10.0	18.0	4.0
	Gaming Sim.	Newtonian	40	$52.5_{\pm 15.5}$	2.5	2.5	2.5	12.5	7.5	17.5	5.0
	Animation	Newtonian	40	$47.5_{\pm 15.5}$	2.5	2.5	5.0	12.5	7.5	15.0	5.0
Decart Oasis	Auto Driving	Newtonian	60	$48.3_{\pm 12.7}$	1.7	3.3	3.3	13.3	10.0	15.1	5.0
	Robotics	Newtonian	60	$41.7_{\pm 12.5}$	1.7	5.0	5.0	15.0	11.7	15.0	4.9
	Human Activities	Newtonian	50	$36.0_{\pm 13.3}$	4.0	4.0	4.0	16.0	12.0	18.0	6.0
	Industrial Settings	Newtonian	50	$42.0_{\pm 13.7}$	2.0	4.0	4.0	16.0	12.0	16.0	6.0
	Natural Scenes	Newtonian	50	$32.0_{\pm 12.9}$	4.0	4.0	6.0	18.0	14.0	16.0	6.0
	Gaming Sim.	Newtonian	40	$52.5_{\pm 15.5}$	2.5	2.5	2.5	12.5	10.0	15.0	5.0
	Animation	Newtonian	40	$47.5_{\pm 15.5}$	2.5	2.5	2.5	12.5	10.0	17.5	5.0
Seedance 2.5	Auto Driving	Newtonian	60	$56.7_{\pm 12.6}$	1.7	1.7	3.3	11.7	6.7	15.0	3.2
	Robotics	Newtonian	60	$51.7_{\pm 12.7}$	1.7	3.3	3.3	13.3	6.7	15.0	5.0
	Human Activities	Newtonian	50	$48.0_{\pm 13.9}$	2.0	2.0	4.0	14.0	8.0	16.0	6.0
	Industrial Settings	Newtonian	50	$48.0_{\pm 13.9}$	2.0	4.0	2.0	14.0	8.0	16.0	6.0
	Natural Scenes	Newtonian	50	$44.0_{\pm 13.8}$	4.0	4.0	2.0	14.0	8.0	18.0	6.0
	Gaming Sim.	Newtonian	40	$55.0_{\pm 15.4}$	2.5	2.5	2.5	10.0	7.5	17.5	2.5
	Animation	Newtonian	40	$57.5_{\pm 15.3}$	2.5	2.5	2.5	10.0	7.5	12.5	5.0
Luma Ray 3	Auto Driving	Newtonian	60	$53.3_{\pm 12.6}$	1.7	3.3	3.3	11.7	8.3	13.4	5.0
	Robotics	Newtonian	60	$48.3_{\pm 12.7}$	1.7	3.3	3.3	15.0	8.3	13.4	6.7
	Human Activities	Newtonian	50	$44.0_{\pm 13.8}$	2.0	2.0	4.0	14.0	10.0	18.0	6.0
	Industrial Settings	Newtonian	50	$46.0_{\pm 13.8}$	2.0	4.0	4.0	14.0	8.0	18.0	4.0
	Natural Scenes	Newtonian	50	$42.0_{\pm 13.7}$	4.0	4.0	4.0	16.0	8.0	16.0	6.0
	Gaming Sim.	Newtonian	40	$47.5_{\pm 15.5}$	2.5	2.5	2.5	12.5	10.0	15.0	5.0
	Animation	Newtonian	40	$50.0_{\pm 15.5}$	2.5	2.5	5.0	12.5	7.5	15.0	5.0
Pika 2.5	Auto Driving	Newtonian	60	$38.3_{\pm 12.3}$	3.3	5.0	5.0	18.3	13.3	13.4	3.4
	Robotics	Newtonian	60	$31.7_{\pm 11.8}$	3.3	5.0	6.7	21.7	15.0	13.3	3.3
	Human Activities	Newtonian	50	$32.0_{\pm 12.9}$	2.0	6.0	8.0	22.0	14.0	12.0	4.0
	Industrial Settings	Newtonian	50	$28.0_{\pm 12.4}$	2.0	6.0	6.0	22.0	14.0	16.0	6.0
	Natural Scenes	Newtonian	50	$26.0_{\pm 12.2}$	4.0	6.0	6.0	24.0	16.0	12.0	6.0
	Gaming Sim.	Newtonian	40	$37.5_{\pm 15.0}$	2.5	2.5	5.0	20.0	12.5	17.5	5.0
	Animation	Newtonian	40	$40.0_{\pm 15.2}$	2.5	5.0	5.0	15.0	15.0	15.0	2.5

Table 7: **SpaceTimeBench Domain Breakdown (Appendix Models, Part 2: Open-Weights & Specialized).** Comprehensive FSY with 95% binomial confidence intervals ($\pm$) and failure distributions across all 7 semantic domains for the remaining five secondary architectures. FSY and F1–F7 sum strictly to 100.0% for each row.

Model Family	Domain	Evaluation Track	N	FSY (%)	F1	F2	F3	F4	F5	F6	F7
MiniMax Hailuo	Auto Driving	Newtonian	60	$41.7_{\pm 12.5}$	1.7	3.3	3.3	18.3	13.3	13.4	5.0
	Robotics	Newtonian	60	$35.0_{\pm 12.1}$	1.7	5.0	5.0	20.0	13.3	16.7	3.3
	Human Activities	Newtonian	50	$36.0_{\pm 13.3}$	2.0	4.0	6.0	18.0	14.0	14.0	6.0
	Industrial Settings	Newtonian	50	$30.0_{\pm 12.7}$	2.0	6.0	6.0	20.0	14.0	18.0	4.0
	Natural Scenes	Newtonian	50	$30.0_{\pm 12.7}$	4.0	4.0	6.0	20.0	16.0	14.0	6.0
	Gaming Sim.	Newtonian	40	$40.0_{\pm 15.2}$	2.5	5.0	2.5	17.5	12.5	17.5	2.5
	Animation	Newtonian	40	$37.5_{\pm 15.0}$	5.0	2.5	5.0	17.5	12.5	12.5	5.0
Mochi 1	Auto Driving	Newtonian	60	$35.0_{\pm 12.1}$	3.3	5.0	5.0	21.7	15.0	13.3	1.7
	Robotics	Newtonian	60	$28.3_{\pm 11.4}$	3.3	5.0	6.7	23.3	16.7	11.7	5.0
	Human Activities	Newtonian	50	$24.0_{\pm 11.8}$	2.0	6.0	8.0	26.0	18.0	10.0	6.0
	Industrial Settings	Newtonian	50	$28.0_{\pm 12.4}$	2.0	6.0	8.0	24.0	18.0	10.0	4.0
	Natural Scenes	Newtonian	50	$22.0_{\pm 11.5}$	4.0	8.0	10.0	24.0	18.0	10.0	4.0
	Gaming Sim.	Newtonian	40	$35.0_{\pm 14.8}$	2.5	5.0	5.0	22.5	15.0	12.5	2.5
	Animation	Newtonian	40	$27.5_{\pm 13.8}$	5.0	5.0	7.5	22.5	12.5	15.0	5.0
Stable Video 4D	Auto Driving	Newtonian	60	$43.3_{\pm 12.5}$	1.7	3.3	3.3	16.7	11.7	15.0	5.0
	Robotics	Newtonian	60	$45.0_{\pm 12.6}$	1.7	3.3	3.3	15.0	11.7	16.7	3.3
	Human Activities	Newtonian	50	$34.0_{\pm 13.1}$	2.0	4.0	6.0	18.0	14.0	16.0	6.0
	Industrial Settings	Newtonian	50	$40.0_{\pm 13.6}$	2.0	4.0	4.0	18.0	12.0	16.0	4.0
	Natural Scenes	Newtonian	50	$28.0_{\pm 12.4}$	4.0	6.0	6.0	20.0	14.0	18.0	4.0
	Gaming Sim.	Newtonian	40	$42.5_{\pm 15.3}$	2.5	2.5	5.0	15.0	12.5	15.0	7.5
	Animation	Newtonian	40	$37.5_{\pm 15.0}$	5.0	2.5	5.0	17.5	10.0	17.5	2.5
LTX-Video	Auto Driving	Newtonian	60	$28.3_{\pm 11.4}$	3.3	5.0	8.3	23.3	16.7	11.7	3.4
	Robotics	Newtonian	60	$23.3_{\pm 10.7}$	3.3	6.7	8.3	26.7	18.3	10.0	3.4
	Human Activities	Newtonian	50	$18.0_{\pm 10.7}$	2.0	8.0	10.0	28.0	20.0	10.0	4.0
	Industrial Settings	Newtonian	50	$24.0_{\pm 11.8}$	2.0	8.0	8.0	26.0	18.0	10.0	4.0
	Natural Scenes	Newtonian	50	$14.0_{\pm 9.6}$	4.0	8.0	10.0	28.0	22.0	10.0	4.0
	Gaming Sim.	Newtonian	40	$25.0_{\pm 13.4}$	5.0	5.0	7.5	25.0	17.5	15.0	0.0
	Animation	Newtonian	40	$22.5_{\pm 12.9}$	2.5	5.0	10.0	27.5	17.5	12.5	5.0
Pyramid Flow	Auto Driving	Newtonian	60	$25.0_{\pm 10.9}$	3.3	6.7	8.3	25.0	18.3	10.0	3.4
	Robotics	Newtonian	60	$18.3_{\pm 9.8}$	3.3	8.3	10.0	28.3	20.0	8.4	3.4
	Human Activities	Newtonian	50	$14.0_{\pm 9.6}$	4.0	8.0	10.0	30.0	22.0	8.0	4.0
	Industrial Settings	Newtonian	50	$18.0_{\pm 10.7}$	2.0	8.0	10.0	28.0	22.0	8.0	4.0
	Natural Scenes	Newtonian	50	$10.0_{\pm 8.3}$	4.0	8.0	12.0	32.0	24.0	8.0	2.0
	Gaming Sim.	Newtonian	40	$22.5_{\pm 12.9}$	5.0	7.5	7.5	27.5	20.0	7.5	2.5
	Animation	Newtonian	40	$17.5_{\pm 11.8}$	2.5	7.5	10.0	27.5	20.0	7.5	5.0