WORLDOLYMPIAD: CAN YOUR WORLD MODEL SURVIVE A TRIATHLON?

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce WorldOlympiad, a benchmark for diagnosing video-based world models across physical faithfulness, geometric consistency, and interaction fidelity. While existing benchmarks often focus on visual quality, semantic alignment, or short-term temporal coherence, they provide limited insight into whether generated videos obey physical rules, preserve coherent 3D structure, and sustain controllable interactions over long horizons. To address this gap, WorldOlympiad decomposes world-model evaluation into three complementary dimensions. The physical track uses object segmentation and MLLM-as-judge to assess whether generated videos follow interpretable rules in mechanics, thermal phenomena, and material properties. The geometry track reconstructs generated videos with Gaussian splatting and evaluates structural consistency, cross-view coherence, and camera-trajectory alignment. The interaction track assesses whether generated rollouts follow complex action prompts and maintain smooth, coherent transitions across consecutive video chunks. WorldOlympiad further covers three major downstream scenarios, including gaming, robotics, and general real-world videos, capturing diverse challenges from interactive control and embodied manipulation to open-domain motion and camera dynamics. Together, these tracks and scenarios form a scalable and interpretable evaluation suite that exposes failure modes beyond generic video quality. Experiments on state-of-the-art models reveal substantial gaps in physical reasoning, 3D consistency, and long-horizon interaction, underscoring the need for more structured evaluation protocols for generative world models.

1 INTRODUCTION

Recent advances in video generation (Wan et al., 2025; Team et al., 2025; Kong et al., 2024; Yang et al., 2024; Blattmann et al., 2023; Brooks et al., 2024) have expanded its scope from passive content creation to video-based world modeling. A video-based world model is expected to predict future visual states from historical observations and control signals, which is crucial for game simulation (Che et al., 2025; Zhang et al., 2025a; He et al., 2025), robotic policy development (Agarwal et al., 2025; Ali et al., 2025; Chi et al., 2025), and real-world scene generation (Mao et al., 2025b; Team et al., 2026c; Wu et al., 2026; Sun et al., 2025; Yang et al., 2025). In these applications, high visual fidelity alone is insufficient: models must preserve state continuity, respect physical and geometric constraints, respond to user actions, and maintain plausible dynamics over long generation horizons. Together, these requirements make developing a comprehensive framework for evaluating video-based world models across multiple capability dimensions a key challenge.

To address this challenge, a straightforward solution is to reuse evaluation approaches originally designed for traditional video generation models, such as VBench (Huang et al., 2024), VBench-2.0 (Zheng et al., 2025), and CLIP-based metrics (Yang et al., 2025; Huang et al., 2025b; Liu et al., 2025). However, these approaches primarily measure perceptual quality or text–video alignment on short videos and miss core world-modeling properties. Although subsequent VBench++ (Huang et al., 2025c) extends evaluation to long videos, it still emphasizes visual appearance and temporal smoothness, leaving key capabilities underexplored. In particular, existing approaches pay limited attention to whether generated videos adhere to physical laws, preserve coherent 3D structure, and support controllable interactions over long horizons.

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.

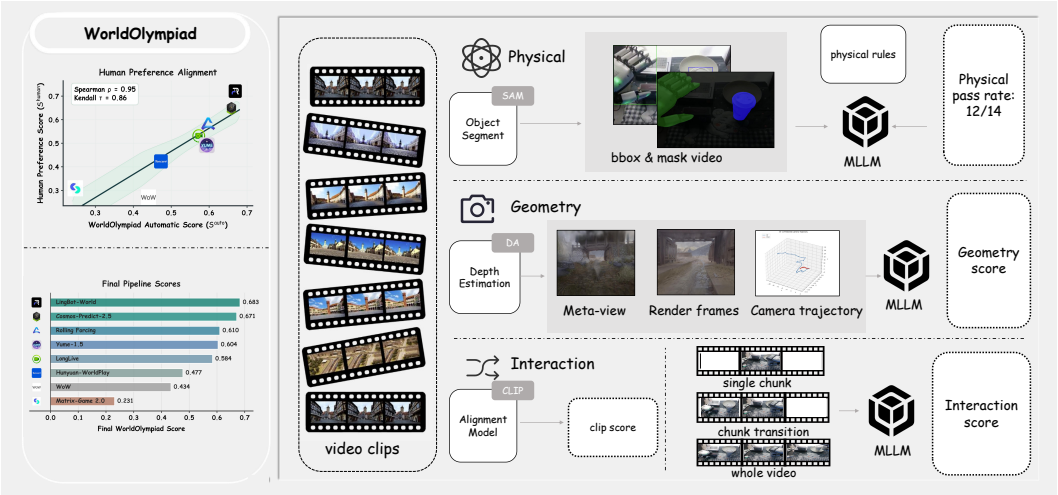


Figure 1: WorldOlympiad evaluates generated videos along three tracks: physical-rule compliance using object masks and multimodal judges, 3D consistency using reconstruction and camera trajectories, and interaction fidelity using chunk-level and full-video checks. The left panel summarizes benchmark-level statistics and human-preference alignment.

Moreover, recent benchmarks specifically designed for video-based world models often focus on a single downstream domain, such as gaming (Ye et al., 2026) or robotics (Shang et al., 2026a; Deng et al., 2026; Liu et al., 2026a), making it difficult to compare models under a unified protocol across gaming, robotics, and general real-world scenarios. As a result, current benchmarks still cannot fully answer a central question: Can existing video-based world models reliably simulate world dynamics across multiple domains, over long horizons, and in interactive settings?

Building on prior efforts, we identify two key challenges in evaluating video-based world models: developing capability-oriented metrics and ensuring task diversity. Evaluation should span heterogeneous domains, while the metrics should assess whether generated videos maintain long-horizon consistency, obey physical laws, preserve 3D geometry, and respond faithfully to control signals. Although existing benchmarks partially address these challenges, they do not comprehensively cover both aspects, as summarized in Table 1. To address this gap, we introduce **WorldOlympiad**, a unified benchmark for evaluating video-based world models across gaming, robotics, and real-world scenarios. WorldOlympiad is constructed from 1,000 high-quality long videos that cover diverse downstream requirements, including interactive control in games, embodied manipulation in robotics, and open-domain motion and camera dynamics in real-world scenes. Each video is processed into structured evaluation instances with scenario-specific prompts and chunk-level temporal descriptions, enabling controlled long-video generation and fine-grained diagnosis. Built on this benchmark, we evaluate generated videos from three complementary perspectives: physical faithfulness, geometric consistency, and interaction fidelity. With 8 representative long-video generation pipelines, WorldOlympiad reveals systematic limitations in current models and provides diagnostic evidence for developing more reliable video-based world models.

Our contributions are summarized as follows:

- We propose **WorldOlympiad**, a unified benchmark for evaluating interactive long-video world models across gaming, robotics, and real-world scenarios.
- We design multi-dimensional judge metrics that systematically assess physical-law adherence, 3D geometric consistency, and chunk-by-chunk interactive generation.
- We construct a dataset of 1,000 high-quality long videos and benchmark 8 long-video generation pipelines to assess reliability in downstream world-model applications.

Table 1: Comparison of representative video world-model benchmarks across long-horizon video, physical fidelity, 3D geometry, interaction, and application domains. A checkmark indicates that the benchmark explicitly evaluates the corresponding capability or setting.

Benchmark	Eval Metrics				Video Tasks		
	Long Video	Physical	Geometry	Interaction	Gaming	Robotics	Real-world
VBench (Huang et al., 2024)	✗	✗	✗	✗	✗	✗	✓
VBench++ (Huang et al., 2025c)	✓	✗	✗	✗	✗	✗	✓
VBench 2.0 (Zheng et al., 2025)	✗	✓	✗	✗	✗	✗	✓
MIND (Ye et al., 2026)	✓	✗	✗	✓	✓	✗	✗
EWMBench (Hu et al., 2025)	✗	✗	✗	✓	✗	✓	✗
WorldEval (Li et al., 2025c)	✗	✗	✗	✓	✗	✓	✗
WorldArena (Shang et al., 2026a)	✗	✓	✓	✓	✗	✓	✗
WorldOlympiad	✓	✓	✓	✓	✓	✓	✓

2 Related Work

2.1 Video Generation

Diffusion-based video generation models (Blattmann et al., 2023; Yang et al., 2024; Kong et al., 2024; Wan et al., 2025; Team et al., 2025) have demonstrated emergent physical consistency through large-scale training (Brooks et al., 2024; Wiedemer et al., 2025), including object permanence, 3D coherence, and plausible motion dynamics. Despite these properties, many early diffusion-based generators target 5–10-second clips, limiting their use as persistent world simulators. Recently, block diffusion has emerged as a promising paradigm for scalable long-horizon video synthesis. By performing iterative diffusion denoising within each block and conditioning on previously generated content via cross-block KV caching, this approach combines the high-quality parallel generation of diffusion models with the sequential consistency of autoregressive conditioning (Huang et al., 2025b; Yin et al., 2025; Zhu et al., 2026; Zhang et al., 2025b). Such a design preserves intra-block denoising quality while enabling scalable temporal extension, positioning block diffusion as a viable road toward video-based world models.

2.2 Video Generation Models as World Models

Video world models now target diverse applications, including interactive game generation (Team et al., 2026c) and robotics simulation (Agarwal et al., 2025; Ali et al., 2025). GameGen-X (Che et al., 2025) and Matrix Game (Zhang et al., 2025a) provide controllable game simulation, while robotics-oriented models support policy generation and data augmentation (Agarwal et al., 2025; Ali et al., 2025; Chi et al., 2025). Persistent state and real-time interaction nevertheless remain challenging. Existing work addresses this challenge through implicit memory, such as KV caching in LongLive (Yang et al., 2025); explicit 3D memory (Xiao et al., 2025; Huang et al., 2025a; Li et al., 2025b; Zhao et al., 2025; Yu et al., 2026; Wu et al., 2025b; Wang et al., 2026b); hybrid temporal-spatial memory (Yu et al., 2026; Team et al., 2026b); and controllable generation within block diffusion (Liu et al., 2026b; Shin et al., 2025). Systems such as LingBot-World (Team et al., 2026c) show that scaling can improve interactive generation, but persistent state, flexible control, and long-horizon consistency remain joint evaluation requirements.

2.3 World Model Benchmarks

Short-video benchmarks including VBench (Huang et al., 2024) and VBench 2.0 (Zheng et al., 2025) measure visual quality, motion, semantic consistency, and physical plausibility. World-model benchmarks assess physical-law adherence, simulation fidelity, and functional modeling (Duan et al., 2025; Qin et al., 2024; Li et al., 2025a; Ying et al., 2026), while robotics benchmarks additionally evaluate controllability, action conditioning, and closed-loop interaction (Shang et al., 2026a; Deng et al., 2026; Liu et al., 2026a). Recent benchmarks extend this coverage to physical-law and visually grounded tests (Inbasekar et al., 2026; Lin et al., 2026), action-conditioned or multi-turn interaction (Fang et al., 2026; Ying et al., 2026), and memory consistency (Zhang et al., 2026). Complementary work studies embodied multimodal capabilities and trustworthiness (Shang et al.,



Figure 2: The benchmark combines 400 robotics videos from RoboCOIN, 400 gaming videos from GameGen-X, and 200 real-world videos from LVD-2M. Together, the sources cover embodied manipulation, interactive control, and open-domain motion.

2026b; Li et al., 2026), as well as broader world reactivity (Yang et al., 2026). However, no benchmark jointly covers gaming, robotics, and general scenes while evaluating physical consistency, 3D geometry, interaction, and long-horizon behavior under one protocol. WorldOlympiad addresses this gap with 1,000 videos across these domains and unified perceptual and functional evaluation.

3 WorldOlympiad

3.1 Data Collection

Figure 2 summarizes the data collection process of WorldOlympiad. The benchmark contains 400 robotics videos, 400 gaming videos, and 200 real-world videos, covering complementary world-modeling requirements: robotics videos emphasize object manipulation and physical interaction, gaming videos emphasize interactive control and long-context state evolution, and real-world videos emphasize open-domain motion and camera dynamics. This diverse composition enables a comprehensive evaluation of video-based world models across their three most critical application domains.

3.1.1 Source Domains

Robotics domain. The robotics subset is built from RoboCOIN (Wu et al., 2025a), an open-source bimanual robotic manipulation data collection. We use this source because bimanual manipulation naturally contains object contact, gripper motion, state changes, and physically grounded interactions. RoboCOIN also includes multiple bimanual robot embodiments, giving the subset broad coverage for evaluating whether generated videos preserve action-consistent dynamics. From the downloaded RoboCOIN videos, we manually filter 400 videos as the robotics portion of the benchmark test set.

Gaming domain. The gaming subset is built from GameGen-X (Che et al., 2025), an interactive open-world game video dataset. We randomly sample videos from the official `OGameData_50K.csv` metadata file and download the corresponding videos. Since some gameplay videos are usually long and contain multiple interaction stages, we split long videos into shorter video chunks with 60 seconds before constructing the final 400-video gaming subset. This subset targets interactive world-modeling behavior such as camera movement, player navigation, combat events, skill execution, and game-state changes.

Real-world domain. The real-world subset is built from LVD-2M (Xiong et al., 2024), a long-take video dataset with temporally dense captions. We use the official `ytb_600k_720p.csv` subset and randomly select videos whose duration is longer than 60 seconds and whose motion score is greater

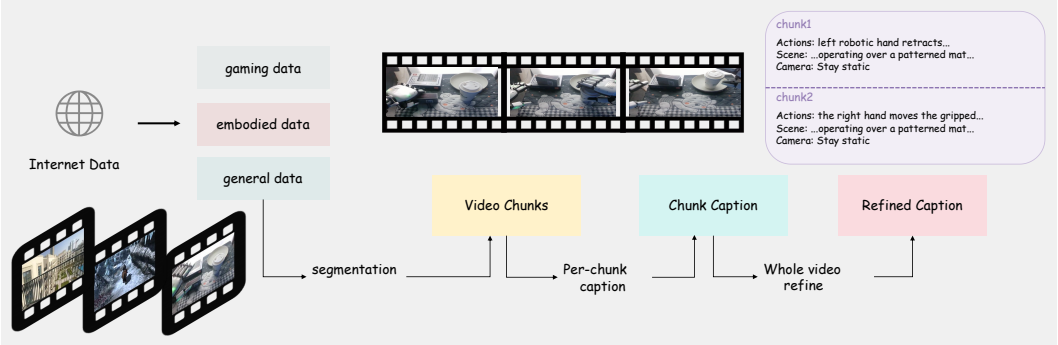


Figure 3: Three-stage annotation pipeline: source videos are segmented into temporal chunks, each chunk receives action and scene captions, and a full-video pass refines annotations for factual accuracy and cross-chunk consistency.

than 50. This filter favors long videos with visible motion, supporting evaluation of open-domain dynamics, camera movement, and geometric consistency in everyday scenes.

The resulting domain counts and selection rules are summarized in Appendix C.1.

3.1.2 Temporal Chunking and Captioning

Detailed video captions are essential for subsequent evaluation. Instead of relying on a single-pass MLLM, we design a three-stage chunk-caption-refine pipeline to ensure the resulting annotations are both accurate and comprehensive, as illustrated in Figure 3. We adopt Gemini-3-Pro-Preview (DeepMind, 2025) across all stages, owing to its superior performance in multimodal understanding.

Chunking, captioning, and refinement. The pipeline first identifies the main continuous execution interval and divides it into at most six contiguous, non-overlapping chunks. For gaming videos, chunking emphasizes combat, traversal, skill casting, and camera transitions; for real-world videos, it emphasizes continuous visual actions, object motion, interactions, and view transitions.

For each robotics, gaming, or real-world chunk, the captioning model produces an `action` field that maps camera motion to WASD-style controls (W/A/S/D denote forward/left/backward/right; `None` means the camera is static; see Appendix C.1) and an English `caption` describing the scene, visible entities, events, interactions, and outcomes. The action label uses camera motion only and ignores character animation, subject motion, visual effects, and UI changes. Finally, using the full video as context, we correct hallucinated details, standardize terminology across adjacent chunks, improve narrative continuity, and validate action labels. The Stage III outputs serve as the final captions for evaluation, and the active WorldOlympiad judge prompts are provided in Appendix B. Together, these annotations provide a shared temporal interface for local action following, transition checks, and long-range consistency across all three evaluation tracks.

Figure 4 summarizes these annotations: W is the most frequent action label, three-segment videos are most common, and the word cloud shows recurring terms in the captions.

3.2 Evaluation Metrics

3.2.1 Physical Evaluation

We evaluate physical faithfulness across mechanics, thermodynamics, and material properties. An MLLM identifies moving or deforming entities, and SAM3 (Carion et al., 2025) produces object masks for inspection. A relevance judge selects rules observable in each reference video; a compliance judge compares the generated video with its reference and returns a compliance label and confidence. Unobservable questions are excluded; each applicable question contributes its confidence if compliant and zero otherwise. Their mean is the physical score. Rule definitions, judge prompts, aggregation details, and per-question results are in Appendices B.2, B.3, and C.3.

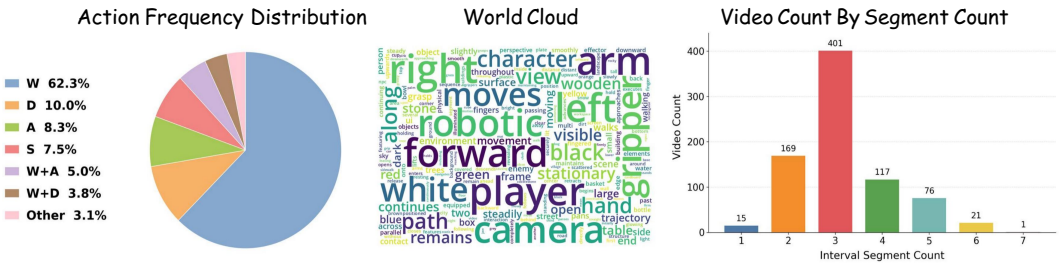


Figure 4: Annotation-pipeline statistics showing action-label frequencies, common caption terms, and the number of videos by temporal segment count. The segment-count distribution peaks at three segments per video, and W is the most frequent action label.

Mechanics. *Gravity* evaluates whether unsupported objects move downward under gravity, rather than floating upward or accelerating in physically implausible directions. *Buoyancy* focuses on whether objects in fluids remain near the surface or sink in accordance with their apparent density. *Compression* measures whether solids deform plausibly under load, instead of staying unrealistically rigid or buckling without sufficient cause. *Impact* examines whether collisions lead to reasonable post-impact dynamics, including momentum transfer, rebound, fracture, or eventual rest.

Thermodynamics. *Melting* assesses whether a heated solid gradually transitions into a liquid state. *Sublimation* captures direct solid-to-gas transitions without an intermediate liquid phase. *Vaporization* considers whether liquids turn into vapor through evaporation or boiling when heated or exposed over time. *Condensation* evaluates the formation of liquid droplets from cooled gas. *Deposition* describes the direct transformation from gas to solid without first becoming liquid. *Freezing* measures whether a cooled liquid solidifies into a stable solid state.

Material. *Color mixing* evaluates whether mixed colored liquids or paints yield the expected resultant color. *Solubility* focuses on whether soluble substances disperse and dissolve into the solvent, rather than remaining intact. *Hardness* distinguishes whether soft materials bend or tear easily while hard materials resist deformation or break sharply. *Combustibility* examines whether flammable materials ignite and produce physically consistent fire, smoke, or charring behavior.

3.2.2 Geometry Evaluation

We evaluate geometric consistency with three complementary signals (Wang et al., 2026a): S_{recon} scores a Gaussian-Splat render sequence, S_{meta} scores a single diagnostic image rendered from the recovered camera position farthest from the reconstruction origin, and S_{traj} scores agreement between recovered and reference camera trajectories. This *meta-view* is a still rendering of the reconstructed scene, not another model-generated frame; it provides a compact view of the reconstructed layout. The construction and judging procedure are detailed in Appendix B.5. Given a generated video $V = \{I_t\}_{t=1}^{T}$, we uniformly sample $\bar{V} = \{I_i\}_{i=1}^{N}$, with $N \leq 32$ in the implementation. When dynamic-object masks are available, foreground Gaussians are removed before rendering so that the 3D judge focuses on the static scene. Depth Anything 3 (Lin et al., 2025) estimates a Gaussian scene and camera parameters, and the Gaussian-Splat renderer produces two diagnostic artifacts:

$$\mathcal{F}_{\text{DA3}}(\bar{V}) \to \left(\mathcal{G}, \{E_i, K_i\}_{i=1}^{N}\right), \quad \hat{V}_{\text{GS}} = \mathcal{R}(\mathcal{G}, \{E_i, K_i\}_{i=1}^{N}), \quad \hat{I}_{\text{meta}} = \mathcal{R}(\mathcal{G}, E_{i^\star}, K_{i^\star}), \tag{1}$$

where $\mathcal{G}$ is the reconstructed Gaussian representation, E_i and K_i are recovered extrinsics and intrinsics, and $i^\star$ denotes the recovered camera pose farthest from the reconstruction origin.

The reconstruction and meta-view are scored separately by the calibrated MLLM judge for recognizable layout, coherent 3D structure, cross-view geometry, and prompt consistency. Camera trajectories are resampled and aligned, then scored for translation, rotation, path shape, motion extent, and jitter, with weights adapted to the reference motion profile. Judge rubrics, trajectory-score details, and submetric results are in Appendices B.5 and C.5.

The implementation records the raw 3D reward as the sum of the three bounded subscores, while all tables report the normalized geometry score:

$$S_{3D} = \frac{1}{3}\left(S_{\text{recon}} + S_{\text{meta}} + S_{\text{traj}}\right). \tag{2}$$

3.2.3 INTERACTION EVALUATION

We evaluate interaction fidelity under the chunk-by-chunk generation setting. Given a generated video divided into T chunks $\{v_i\}_{i=1}^{T}$ and their corresponding captions $\{p_i\}_{i=1}^{T}$, the interaction benchmark measures whether each chunk follows its local instruction, whether adjacent chunks transition coherently, and whether the full video remains temporally fluent. This design matches the way interactive video world models are typically rolled out: each new chunk is conditioned on the previous visual context and a new control or action caption, so a model must satisfy both local caption alignment and long-range continuity.

The first component is a CLIP-based semantic-adherence score. For each chunk, we uniformly sample eight frames, encode each frame and its caption with a CLIP model (Radford et al., 2021; Yang et al., 2025), and average their cosine similarities. We linearly calibrate the resulting video-level score to $[0, 1]$ and denote the calibrated value by $\widetilde{S}_{\text{clip}}$, using fixed thresholds $\tau_{\min} = 0.20$ and $\tau_{\max} = 0.40$. The same thresholds are used for every model, providing a lightweight auxiliary measure of whether generated chunks contain the semantic content requested by their captions.

The second component uses an MLLM as a structured rubric-based judge. We query the MLLM at three complementary levels, and all returned scores are clipped to the requested 0–5 range before being normalized to $[0, 1]$ for reporting. First, the MLLM receives each chunk v_i and its caption p_i, and scores *visual quality*, *text alignment*, and an *overall* chunk score a_i. Second, the MLLM receives each adjacent pair (v_i, v_{i+1}) together with their captions (p_i, p_{i+1}), and scores *transition smoothness* and an *overall* transition score b_i. Third, the MLLM receives the full generated video and scores *long-range consistency*, *global text alignment*, and a global *overall* score g. Let S_{chunk}, S_{trans}, and S_{global} denote the normalized averages of the corresponding overall judgments. We combine them as

$$S_{\text{mllm}} = \frac{1}{3}\left(S_{\text{chunk}} + S_{\text{trans}} + S_{\text{global}}\right). \tag{3}$$

The final interaction score uses the calibrated CLIP score as a lightweight semantic auxiliary term:

$$S_{\text{interact}} = (1 - \lambda) S_{\text{mllm}} + \lambda \widetilde{S}_{\text{clip}}, \quad \lambda = 0.1. \tag{4}$$

This design lets CLIP contribute frame-caption semantic grounding while keeping the interaction metric dominated by the structured MLLM judge, which evaluates chunk-level instruction following, boundary smoothness, state preservation, and full-video fluency. Judge rubrics and fine-grained results are in Appendices B.4 and C.4.

Finally, WorldOlympiad reports an overall score by averaging the three core evaluation tracks:

$$S_{\text{all}} = \frac{1}{3}\left(S_{\text{phys}} + S_{3D} + S_{\text{interact}}\right). \tag{5}$$

Equal weighting keeps the leaderboard aligned with the benchmark: physical faithfulness, geometric consistency, and interaction fidelity contribute equally to the final model ranking.

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

Evaluation models. We evaluate eight publicly available video-generation pipelines through OpenWorldLib (Team et al., 2026a). These pipelines cover three major families of video world models. The gaming-centric group includes Matrix-Game 2.0 (He et al., 2025) and LingBot-World (Team et al., 2026c); the robotics-centric group includes Cosmos-Predict-2.5 (Ali et al., 2025) and WoW (Chi et al., 2025); and the general long-video group includes Rolling Forcing (Liu et al., 2025), LongLive (Yang et al., 2025), Yume-1.5 (Mao et al., 2025a), and Hunyuan-WorldPlay (Sun et al., 2025). We evaluate these pipelines across gaming, robotics, and general real-world videos.

Implementation details. For fairness, we use each released pipeline with its official default generation configuration whenever possible. Since different pipelines may adopt different chunk sizes or segment-level generation settings, we dynamically map the temporal information in the chunk captions to each model's native generation configuration. This allows the temporal proportions of the

Table 2: Main benchmark results on **WorldOlympiad**. We evaluate eight representative video world models across gaming, robotics, and general long-video generation settings. **Physical** (S_{phys}): physical faithfulness; **3D Cons.** (S_{3D}): 3D spatial consistency; **Interact.** (S_{interact}): interaction fidelity with CLIP-based semantic grounding; **All** (S_{all}): overall composite score. Best and second-best results are marked in **bold** and underlined, respectively.

Category	Model	Evaluation Metrics				Rank
		Physical	3D Cons.	Interact.	All	
GAMING WORLD MODEL	Matrix-Game 2.0 (He et al., 2025)	0.325	0.255	0.113	0.231	8
	LingBot-World (Team et al., 2026c)	**0.942**	0.373	**0.734**	**0.683**	**1**
ROBOTICS WORLD MODEL	Cosmos-Predict-2.5 (Ali et al., 2025)	<u>0.906</u>	<u>0.399</u>	<u>0.707</u>	<u>0.671</u>	<u>2</u>
	WoW (Chi et al., 2025)	0.708	0.250	0.345	0.434	7
GENERAL WORLD MODEL	Rolling Forcing (Liu et al., 2025)	0.873	0.321	0.636	0.610	3
	LongLive (Yang et al., 2025)	0.863	0.363	0.526	0.584	5
	Yume-1.5 (Mao et al., 2025a)	0.863	0.301	0.649	0.604	4
	Hunyuan-WorldPlay (Sun et al., 2025)	0.692	**0.424**	0.316	0.477	6

All is the average of **Physical**, **3D Cons.**, and **Interact.**; overall ranks are computed by the unrounded **All** score. Displayed scores are rounded to three decimal places.

original chunk captions to be retained while respecting each generation pipeline's native training and inference configuration. For methods that include an explicit memory or long-context mechanism, such as Rolling Forcing, we preserve the official memory-management strategy during rollout. For pipelines without a dedicated long-horizon memory module, such as WoW, we generate long videos by conditioning each continuation segment on the preceding visual context.

All generated videos are evaluated by the same automatic WorldOlympiad evaluator. The evaluator reports physical faithfulness, 3D consistency, CLIP-augmented interaction fidelity, and an overall composite score. Judge-based component scores are reported after averaging their normalized subscores into the $[0, 1]$ range. Physical faithfulness aggregates rule-level judgments over mechanics, thermodynamics, and material behavior; 3D consistency combines reconstruction quality, meta-view quality, and camera trajectory consistency; and interaction fidelity measures chunk-level instruction following, CLIP-based semantic grounding, adjacent transition smoothness, and long-range coherence over the full generated video.

4.2 MAIN BENCHMARK RESULTS

Table 2 summarizes the video world models evaluated in OpenWorldLib, grouped by gaming, robotics, and general world-model categories. The table reports physical faithfulness, 3D consistency, CLIP-augmented interaction fidelity, and the overall score. Figure 5 further visualizes the score distribution across pipelines and evaluation dimensions.

From visual synthesis to stateful world simulation. The strongest models are distinguished not only by visual plausibility but also by their ability to preserve physical state and interaction semantics over long rollouts. LingBot-World achieves the highest overall score (0.683), with strong physical faithfulness (0.942) and interaction fidelity (0.734). Scale may help, but is not decisive: the 2B Cosmos-Predict-2.5 reaches 0.671, likely benefiting from physical-world training that generalizes beyond embodied manipulation. Targeted training and rollout design can therefore partly offset smaller model scale in stateful world simulation.

Physical regularity is emerging as a shared capability. Several pipelines already score highly on physical regularities: LingBot-World (0.942), Cosmos-Predict-2.5 (0.906), Rolling Forcing (0.873), LongLive (0.863), and Yume-1.5 (0.863). This suggests that models learn recurring motion, contact, support, and material patterns, but thermodynamics and material questions remain fragile, while weaker models still violate constraints during long rollouts.

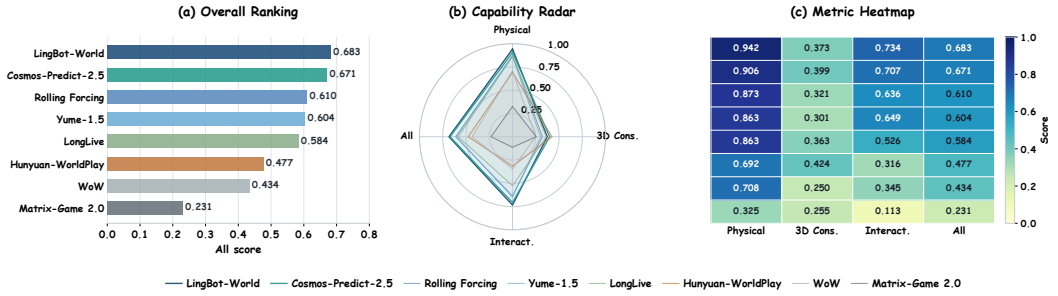


Figure 5: Results across the eight evaluated world-model pipelines. The panels show the overall ranking, each model's profile across physical, geometric, and interaction capabilities, and the metric values used to compute the composite score.

The geometry-simulation gap remains unresolved. Geometric consistency remains a major weakness: Hunyuan-WorldPlay reaches only (0.424), while most models remain in the (0.25)–(0.40) range. Its stronger 3D consistency may reflect reliance on camera or viewpoint control, which preserves layout under view changes but is more constrained than action-conditioned generation requiring object manipulation and multi-step state transitions. The result highlights a trade-off between cross-view structure and flexible interactive dynamics.

The specialization-generalization trade-off. Domain-specific training can transfer beyond its original domain: LingBot-World and Cosmos-Predict-2.5 rank highly, whereas WoW falls from 0.502 on embodied videos to 0.368 on gaming and 0.415 on general videos (Table 6). Specialization helps only when knowledge transfers broadly, so future models should combine domain training with transferable knowledge across domains.

Fine-grained diagnostics. WorldOlympiad is diagnostic rather than only leaderboard-driven. Beyond the aggregate scores in Table 2, we decompose behavior into domain results, physical questions, 3D submetrics, and interaction submetrics, identifying whether errors arise from a specific rule, unstable geometry, weak semantic grounding, or long-range drift. Detailed results are in Appendix C.

4.3 Qualitative Case Studies and Failure Modes

Qualitative cases complement aggregate scores by exposing three recurring failures: implausible physics (objects move against gravity, deform without contact, or change state abruptly); geometry that appears plausible in the original view but fails reconstruction, meta-view rendering, or camera-trajectory checks; and interaction drift, including state resets, object loss, or broken chunk continuity. Examples appear in Appendix D. Human and automatic rankings align (Spearman $\rho = 0.95$); preference results and protocol appear in Appendix E.

5 Conclusion

WorldOlympiad evaluates 1,000 videos across gaming, robotics, and real-world settings for physical faithfulness, 3D consistency, and interaction fidelity. Combining rule-based judging, 3D reconstruction, and chunk-level and long-horizon tests, it shows that plausible videos can still violate physics, distort geometry, or lose state. Track-level diagnostics expose these failures and their trade-offs, guiding improvements in data coverage, memory, and action conditioning.

Future work. Future work will compare KV-cache, 3D, linear-attention, and hybrid memory under matched data, model scale, and evaluation settings. We will assess their effect on long-horizon physical, geometric, and interaction consistency across domains.

AI USE STATEMENT

We used AI to assist with multimodal annotation of benchmark videos, including temporal chunking, action labeling, and caption generation and refinement. AI also assisted with writing and revising project scripts and designing and refining annotation and evaluation prompts. The authors drafted the manuscript; AI was used to polish author-written text for grammar, clarity, and readability, not to generate the paper's scientific claims or conclusions. The authors reviewed and checked the annotations against the source videos, inspected and tested the scripts, reviewed the prompts, and checked the language edits. The authors take responsibility for all content and artifacts in this work.

REPRODUCIBILITY STATEMENT

The dataset composition, processing and annotation pipeline, evaluation prompts, scoring rules, and experimental settings are specified in the paper and appendix. The fixed prompts, thresholds, formulas, and procedures support reproduction of the benchmark construction and reported results.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations*, volume 2025, pp. 37546–37593, 2025.

Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, et al. Wow: Towards a world omniscient world model through embodied interaction. *arXiv preprint arXiv:2509.22642*, 2025.

Google DeepMind. Gemini 3 pro model card, 2025. URL `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf`.

Yufan Deng, Zilin Pan, Hongyu Zhang, Xiaojie Li, Ruoqing Hu, Yufei Ding, Yiming Zou, Yan Zeng, and Daquan Zhou. Rethinking video generation model for the embodied world. *arXiv preprint arXiv:2601.15282*, 2026.

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

Jianjie Fang, Yingshan Lei, Qin Wan, Ziyou Wang, Yuchao Huang, Yongyan Xu, Baining Zhao, Weichen Zhang, Chen Gao, Xinlei Chen, et al. iworld-bench: A benchmark for interactive world models with a unified action generation framework. *arXiv preprint arXiv:2605.03941*, 2026.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Yue Hu, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198*, 2025a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025b.

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

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025c.

Karthik Inbasekar, Guy Rom, and Omer Shlomovits. Worldjen: An end-to-end multi-dimensional benchmark for generative video models. *arXiv preprint arXiv:2605.03475*, 2026.

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

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. *arXiv preprint arXiv:2502.20694*, 2025a.

Huiqiong Li, Jiayu Wang, Zhiting Mei, Anirudha Majumdar, Jingjing Chen, and Bin Zhu. Robotrust-bench: Benchmarking the trustworthiness of video world models for robotic manipulation. *arXiv preprint arXiv:2606.01600*, 2026.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25690–25699, 2025b.

Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Worldeval: World model as real-world robot policies evaluator, 2025c. URL https://arxiv.org/abs/2505.19017.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Juyi Lin, Arash Akbari, Yumei He, Lin Zhao, Haichao Zhang, Arman Akbari, Xingchen Xu, Zoe Y Lu, Enfu Nan, Hokin Deng, et al. Phyground: Benchmarking physical reasoning in generative world models. *arXiv preprint arXiv:2605.10806*, 2026.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Mingxin Liu, Shuran Ma, Shibei Meng, Xiangyu Zhao, Zicheng Zhang, Shaofeng Zhang, Zhihang Zhong, Peixian Chen, Haoyu Cao, Xing Sun, et al. Rise-video: Can video generators decode implicit world rules? *arXiv preprint arXiv:2602.05986*, 2026a.

Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026b.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025a.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025b.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026a.

Yu Shang, Yinzhou Tang, Yiding Ma, Zhuohang Li, Lei Jin, Weikang Su, Xin Jin, Zhaolu Wang, Ziyou Wang, Xin Zhang, et al. Worldarena 2.0: Extending embodied world model benchmarking on modality, functionality and platform. *arXiv preprint arXiv:2605.17912*, 2026b.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

DataFlow Team, Bohan Zeng, Daili Hua, Kaixin Zhu, Yifan Dai, Bozhou Li, Yuran Wang, Chengzhuo Tong, Yifan Yang, Mingkun Chang, et al. Openworldlib: A unified codebase and definition of advanced world models. *arXiv preprint arXiv:2604.04707*, 2026a.

InSpatio Team, Donghui Shen, Guofeng Zhang, Haomin Liu, Haoyu Ji, Hujun Bao, Hongjia Zhai, Jialin Liu, Jing Guo, Nan Wang, et al. Inspatio-world: A real-time 4d world simulator via spatiotemporal autoregressive modeling. *arXiv preprint arXiv:2604.07209*, 2026b.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026c.

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

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

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y. Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models. *arXiv preprint arXiv:2606.09828*, 2026b.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, et al. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026.

Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Bowen Yang, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025a.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025b.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. Lvd-2m: A long-take video dataset with temporally dense captions. *Advances in Neural Information Processing Systems*, 37:16623–16644, 2024.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Yuxue Yang, Shuyao Shang, Jiahe Wang, Zitong Zhou, Liang Tan, Junhan Zeng, Ruizhi Li, Junyan Li, Yu Liu, Xiao Yang, et al. Worldexam: Benchmarking world models from apparent appearance to inherent reactivity. *arXiv preprint arXiv:2608.02603*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. Mind: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. Wbench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026.

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, Animesh Garg, et al. Mosaicmem: Hybrid spatial memory for controllable video world models. *arXiv preprint arXiv:2603.17117*, 2026.

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Hanyang Wang, Chensheng Dai, Min Chen, Yifan Li, Yuxin Li, Yingjie Chen, et al. Mbench: A comprehensive benchmark on memory capability for video world models. *arXiv preprint arXiv:2606.00793*, 2026.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025a.

Zeyu Zhang, Shuning Chang, Yuanyu He, Yizeng Han, Jiasheng Tang, Fan Wang, and Bohan Zhuang. Blockvid: Block diffusion for high-quality and consistent minute-long video generation. *arXiv preprint arXiv:2511.22973*, 2025b.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. *arXiv preprint arXiv:2512.15716*, 2025.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

A Appendix

B WorldOlympiad Judge Prompt Templates

The prompt templates below cover dynamic-object extraction, physical consistency, interaction quality, and 3D reconstruction quality.

B.1 Dynamic-Object Extraction Prompt

Before physical and 3D scoring, WorldOlympiad uses a MLLM prompt to identify the primary dynamic or deforming objects for SAM-based visualization, masking, and background completion.

Dynamic-object extraction

System role: The model acts as an expert at spotting only the primary physical actors that visibly move or deform in a video.
Selection rules: Return the fewest distinct moving or deforming objects, with a maximum of three. Prefer dominant moving foreground objects and omit secondary or uncertain objects. Do not include static background, scenery, floors, tables, walls, tools, supports, containers, or objects that are merely visible. If an object does not visibly change position, orientation, or shape, do not return it. Merge duplicates and synonyms into one concise noun of one to three words.
User query: Watch the video and list only the main objects that visibly move or deform. The query includes {`video_prompt`} as the video description.
Output format: Return only a JSON array with one to three concise nouns, such as `["person", "ball"]`. No explanations are allowed.

B.2 Physical Judge Prompts

The physical pipeline first runs a relevance judge on the reference video to determine which physical rules are applicable. It then runs a compliance judge that compares the generated candidate against the reference.

Physical batch relevance judge

System role: `PhysicsFilterBatch`, an expert video physics relevance evaluator.
Input: One reference or ground-truth video, its textual prompt, and a list of physics questions. Each question contains a `question_id`, dimension, rule text, and success condition.
Decision rule: For every question, decide whether the reference video contains enough visual evidence to judge the physical rule. Use `related=true` when the rule can be judged from visible objects, materials, contacts, state changes, motion, support, heat or cold cues, liquid/gas/solid behavior, deformation, color or material behavior, burning, dissolving, or other relevant physical evidence. Use `related=false` only when the rule truly cannot be evaluated from the shown scene.
User query: The prompt provides {`video_prompt`} and {`question_list_json`}, and asks the judge to evaluate relevance for every `question_id`.
Output format: Return strict JSON with one result per input question: `question_id`, `related`, `confidence`, and a short evidence-based `reason`.

Table 3: Judge-related prompt families used by WorldOlympiad.

Component	Prompt family	Role in evaluation
Physical	Relevance and compliance judges	Select applicable physical rules and judge whether the generated video follows them against the reference.
Interaction	Chunk, transition, and global judges	Score local caption following, boundary smoothness, and long-range consistency.
3D	Static-scene rewrite and 3D MLLM scorers	Remove dynamic actors from the judging target and score Gaussian-splat reconstruction quality.
Preprocessing	Dynamic-object extraction	Select moving or deforming foreground actors for SAM-based masking and diagnostic videos.

Physical batch compliance judge

System role: `PhysicsJudgeBatch`, an expert video physics evaluator.
Input: Two videos are provided: the generated candidate video to judge and the ground-truth reference video, which supplies context for the intended event, scene, timing, and physical evidence. The prompt also provides {`video_prompt`} and the related subset of {`question_list_json`}.
Decision rule: For each related physics question, judge whether the generated candidate follows the physical rule and remains consistent with the reference. The judge does not require frame-exact matching, but it requires plausible physics, object and material identity consistency, correct temporal order, and visible evidence.
Output format: Return strict JSON with one result per input question: `question_id`, `compliant`, `confidence`, a 3–5 sentence `explanation` with specific visual evidence, and short concrete `observations`.

Physical question context. The `question_list_json` variable contains rule identifiers, dimensions, questions, and success conditions. These rules cover mechanics (`gravity`, `buoyancy`, `compression`, `impact`), thermodynamics (`melting`, `sublimation`, `vaporization`, `condensation`, `deposition`, `freezing`), and material behavior (`color_mixing`, `solubility`, `hardness`, `combustibility`).

Mechanics rule contexts

`gravity`
`rule`: Do free-moving objects move downward consistently with gravity?
`expected_behavior`: First judge whether objects are unsupported, airborne, falling, jumping, driving over uneven terrain, or staying grounded on a support surface. Unsupported objects should fall or arc downward under gravity; supported ground vehicles, people, and objects should remain in plausible contact with the ground unless a visible jump, ramp, collision, or lift explains vertical motion. Penalize floating, sinking through the ground, sudden vertical pops, or hovering without support.
`buoyancy`
`rule`: Do objects on or in a fluid behave consistently with buoyancy, such that floating items stay near the surface and sinking items submerge?
`expected_behavior`: Floating objects should remain on or near the surface; dense objects should descend.
`compression`
`rule`: When objects or support surfaces are stressed, loaded, squeezed, or pressed, do they deform or remain rigid in a plausible manner?
`expected_behavior`: Cans may dent when crushed, soft materials should compress smoothly under load, and rigid vehicles or metal bodies should mostly keep their shape unless there is collision or heavy force. In robot manipulation, vehicle, racing, or navigation scenes, grasping, pinching, pressing, tire-ground contact, suspension loading, soil displacement, body rigidity, or lack of impossible warping can make stress and deformation relevant. Deformation should start only after visible squeezing, support, load, contact, or other applied stress, and the same object, person, or vehicle should stay visually consistent instead of morphing into a different shape or identity.
`impact`
`rule`: Do contact, traction, collisions, impacts, and momentum changes produce reasonable motion transitions?
`expected_behavior`: Look for momentum transfer, bouncing, shattering, resting poses, ground contact, tire or foot contact, traction, braking, turning, or acceleration that matches visible forces and contacts. In robot manipulation, placing or releasing an object into contact also counts as an impact or contact event. In vehicle, racing, sports, or navigation scenes, tire-ground contact, acceleration from rest, sliding, dust kick-up, braking, steering, direction changes, speed changes, near-collisions, or collisions are relevant. Abrupt deformation, direction change, or speed change should happen after visible contact, impact, or control input rather than before.

Thermodynamics rule contexts

`melting`
`description`: A solid substance should gradually transition to liquid when heated above its melting point.
`expected_behavior`: The solid should decrease in size or volume as it transforms into liquid, and the process should be gradual and continuous.

violations: Solid increasing in size, liquid turning into solid, instantaneous disappearance without liquid formation, solid remaining unchanged despite heat, or no liquid formation.

sublimation

description: A solid should directly transition to gas without passing through the liquid phase.
expected_behavior: The solid should rapidly decrease or disappear while gas or vapor appears around it, with no visible liquid phase.
violations: Solid melting to liquid first, gas condensing to solid, solid remaining stable, or liquid forming as an intermediate phase.

vaporization

description: A liquid should transition to gas when heated or over time.
expected_behavior: The liquid should decrease in volume or disappear over time; bubbles, boiling, evaporation, or visible gas or vapor may appear.
violations: Liquid increasing in volume, gas condensing to liquid, liquid remaining constant despite heat or time, or liquid freezing.

condensation

description: A gas should transition to liquid when cooled below its condensation point.
expected_behavior: Liquid should form and increase in volume; small droplets may appear and merge, and gas may become less visible as it condenses.
violations: Liquid evaporating, gas remaining as gas, liquid freezing directly to solid, or no liquid formation.

deposition

description: A gas should directly transition to solid without passing through the liquid phase.
expected_behavior: Solid should form and increase in size or volume directly from gas, often as crystals or frost-like structures, with no visible liquid intermediate.
violations: Gas condensing to liquid first, solid melting, liquid freezing to solid, or no solid formation.

freezing

description: A liquid should transition to solid when cooled below its freezing point.
expected_behavior: The liquid should solidify and may expand slightly or contract; the surface may become rigid and the shape should become more defined and regular.
violations: Liquid remaining liquid, solid melting, liquid evaporating, or no solidification occurring.

Material rule contexts

color_mixing

question: When different colored liquids or paints mix, do they produce the correct resulting color?
success_condition: Colored liquids, paints, powders, smoke, dye, pigments, or other visibly colored substances should blend into plausible resulting colors when they contact and mix. Red and yellow should produce orange, blue and yellow should produce green, and red and blue should produce purple. If colored objects merely pass near each other without material transfer or blending, they should not change color.

solubility

question: Do soluble materials such as sugar or salt dissolve properly when placed in water or other solvents?
success_condition: Soluble or dispersible substances such as sugar, salt, powder, dye, tablets, or granular material should gradually disperse, dissolve, fade, or become suspended or invisible in a liquid solvent. Insoluble solids should remain visible or settle.

hardness

question: Do materials with different hardness levels behave correctly when grasped, pressed, cut, folded, or broken?
success_condition: Soft materials such as paper, cloth, foam, food, plants, soil, loose dirt, or powder should bend, fold, tear, compress, scatter, or deform when appropriate. Hard materials such as metal, stone, glass, rigid vehicle bodies, tools, or containers should resist deformation unless force or collision is strong enough. In robot, human, vehicle, racing, sports, or navigation scenes, grasping, pinching, pressing, stepping, tire-ground contact, placing, sliding, collision, or load-bearing can reveal rigidity versus softness. Shape change should follow visible contact or applied force, and the acted-on object, person, vehicle, or material should remain visually consistent.

combustibility

question: Do flammable materials burn correctly, producing fire, smoke, or char?
success_condition: Wood, paper, and fabric should ignite and produce flames or smoke; non-flammable materials should not.

B.3 Physical Score Aggregation

Let $\mathcal{R}$ be the set of physical questions marked relevant by the reference-video filter. For a related question q, let $c_q \in [0, 1]$ be the compliance judge's confidence. Its score is $s_q = c_q$ when the generated video is compliant, and $s_q = 0$ otherwise. The physical score is the mean of s_q over $\mathcal{R}$; unrelated questions are excluded rather than counted as failures. If no question is related, the implementation uses the documented default score of 0.5. Per-question and per-dimension results are reported in Appendix C.3.

B.4 Interaction Judge Prompts

The interaction pipeline evaluates chunk-generated long videos at three levels: individual chunks, adjacent chunk transitions, and the stitched full video.

Interaction judge shared instruction

System role: A strict evaluator for chunk-generated interactive videos.
Judgment scope: Judge whether each generated chunk follows its intended action and caption, whether adjacent chunks transition smoothly, and whether the full stitched video remains globally consistent. The reference video is used only as context for scene, style, camera, and intended interaction, not as a requirement for frame-exact matching.
Output rule: Always return strict JSON only. All scores must be numbers from 0 to 5.

Chunk-level interaction judge

Input metadata: {`chunk_index`}, {`source_interval`}, generated interval [{`generated_start_sec`}, {`generated_end_sec`}), {`action`}, and {`caption`}.
Primary evidence: Use the generated chunk frames as the primary evidence. Reference-video frames are used only for context about the intended scene and style.
Scoring criteria: Score `visual_quality` by clarity, realism, temporal stability, color and lighting consistency, and lack of artifacts. Score `text_alignment` by whether the visible content matches the intended action and caption.
Output format: Return strict JSON with `chunk_index`, `visual_quality`, `text_alignment`, `overall`, and a short evidence-based `reason`.

Transition-level interaction judge

Input metadata: The previous and next chunk indices, generated intervals, actions, and captions.
Decision rule: Judge whether the transition is smooth and continuous. Scene, lighting, style, and subject identity should remain coherent. Motion trajectory and camera movement should evolve naturally. Penalize abrupt jumps, object identity resets, impossible camera jumps, and visible stitching artifacts.
Output format: Return strict JSON with `from_chunk_index`, `to_chunk_index`, `transition_smoothness`, `overall`, and a short evidence-based `reason`.

Global interaction judge

Input: The whole stitched generated video and {`prompt_summary_json`}, which lists each chunk index, action, and caption.
Primary evidence: Use generated full-video frames as the primary evidence and reference frames only as context.
Scoring criteria: Judge whether subject, character, and object identity remain stable across the video; whether scene style, visual tone, lighting, and camera behavior remain coherent; and whether global semantics align with the combined intent of all chunk prompts.
Output format: Return strict JSON with `long_range_consistency`, `global_text_alignment`, `overall`, and a short evidence-based `reason`.

B.5 3D Judge Prompts

The 3D pipeline uniformly samples up to 32 frames and reconstructs a Gaussian scene and camera poses with Depth Anything 3 (DA3). When dynamic-object masks are available, the corresponding

foreground Gaussian components are removed from the reconstructed scene before rendering. The renderer produces a Gaussian-Splat sequence and one meta-view image; separate calibrated MLLM judgments provide S_{recon} and S_{meta}.

Meta-view construction. The meta-view is a single Gaussian-Splat rendering of the reconstructed scene, not a frame generated by the evaluated video model and not a newly sampled camera pose. From the DA3-recovered camera positions, the implementation selects the position with the greatest Euclidean distance from the reconstruction origin and renders the scene once from that pose. This compact view lets the judge inspect whether the recovered static layout remains recognizable and geometrically coherent from the selected camera position. It complements the render sequence; by itself, it does not establish consistency from every possible viewpoint.

Camera-trajectory score. The predicted DA3 trajectory and reference trajectory are resampled to a common number of poses and expressed relative to their first poses. When the reference has translation, Umeyama alignment is applied to the predicted poses before comparison. Let L_p^t, L_r^t be total predicted and reference translation lengths, and L_p^r, L_r^r the corresponding accumulated rotation angles. For a moving reference, the translation score combines path-direction similarity, motion-extent similarity, and mean position error:

$$S_t = 0.4S_{\text{path}} + 0.3S_{\text{extent}}^t + 0.3S_{\text{position}}.$$

The rotation score combines mean and final geodesic rotation errors with the rotation-extent similarity:

$$S_r = 0.5S_{\text{mean-rotation}} + 0.3S_{\text{final-rotation}} + 0.2S_{\text{extent}}^r.$$

Extent similarity is $\exp(-|\log((L_p+\epsilon)/(L_r+\epsilon))|)$ for the corresponding translation or rotation totals. If reference translation is below 10^{-4}, or reference rotation is below 10^{-3} degrees, the corresponding component instead penalizes predicted motion directly. For a near-static reference (less than 0.01 translation units and 0.5 degrees per step), both components use these direct penalties: $S_t = \exp(-L_p^t/3)$ and $S_r = \exp(-L_p^r/5)$. The final score is S_tS_r for a near-static reference, $0.2S_t + 0.8S_r$ when the reference has rotation only, $0.8S_t + 0.2S_r$ when it has translation only, and $0.5(S_t + S_r)$ when it has both. The result is clipped to $[0, 1]$.

For completeness, path-direction similarity is the flattened cosine similarity between predicted and reference camera centers, mapped from $[-1, 1]$ to $[0, 1]$; if both paths have zero norm it is set to 1, and if only one does, it is set to 0. The position term is $S_{\text{position}} = \exp[-\bar{e}_p/(2\max(L_r^t/(N-1), 1))]$, where $\bar{e}_p$ is mean aligned center error. The rotation terms are $S_{\text{mean-rotation}} = \exp(-\bar{e}_r/6)$ and $S_{\text{final-rotation}} = \exp(-e_r^{\text{last}}/6)$, with errors in degrees. These definitions follow the motion-profile implementation.

The raw 3D reward is the sum of the three bounded scores (range $[0, 3]$); the paper reports their mean, $S_{3D} = R_{3D}/3$, in $[0, 1]$.

Static-scene prompt rewrite

System role: Rewrite video generation prompts for static 3D reconstruction evaluation. Dynamic foreground Gaussian components are filtered with supplied masks before the diagnostic renders are scored.
Rewrite requirements: Keep static background, environment, layout, materials, lighting, weather, terrain, large structures, and camera/view behavior if present. Remove or explicitly ignore dynamic foreground actors and actions, including people, animals, vehicles, limbs, clothing motion, object manipulation, running, turning, and other moving subjects. If the original prompt has no camera information, set `camera_behavior` exactly to `"camera is static"`. Note that dynamic actors may be absent from the diagnostic renders when their Gaussian components were filtered by masks.
Input placeholder: {`prompt`}.
Output format: Return strict JSON with exactly `static_scene_description`, `camera_behavior`, and `ignore_for_3d`.

Gaussian-Splat video 3D judge

System role: A calibrated 3D reconstruction judge that follows the scoring rubric and returns strict JSON only.
Input: Several sampled frames from a Gaussian-Splat render reconstructed from a source video, with

dynamic foreground Gaussian components filtered by supplied masks when available, together with the static-scene prompt {prompt}.
Robustness instruction: Be robust to normal Gaussian-Splat and mask- filtering artifacts, blur, rain/fog/low-light appearance, and moderate texture noise. Dynamic foreground actors may be absent or incomplete because their Gaussian components were intentionally filtered; do not penalize their absence.
Scoring focus: Judge whether the reconstructed static background geometry is coherent across views or time, whether static structures and scene layout are spatially plausible, whether the render preserves recognizable camera-consistent organization, whether it is faithful to the static-scene description and expected camera behavior, and whether artifacts dominate the render. If the prompt says the camera is static, do not require camera motion or parallax.
Calibration: Scores in 0.8–1.0 indicate a clear coherent 3D scene; 0.6–0.8 indicate a recognizable mostly coherent scene with noticeable artifacts; 0.4–0.6 indicate partial recognition with significant artifacts; 0.2–0.4 indicate mostly failed reconstruction; and 0.0–0.2 indicate an unusable render. Do not assign a very low score solely because the render is blurry or noisy; if the static layout is recognizable, the score should usually be at least 0.5.
Output format: Return strict JSON with score in $[0, 1]$ and a short reason.

Meta-view image 3D judge

System role: A calibrated 3D reconstruction judge that follows the scoring rubric and returns strict JSON only.
Input: A single meta-view image rendered from a 3D reconstruction of a source video, with dynamic foreground Gaussian components filtered by supplied masks when available, together with the static-scene prompt {prompt}.
Robustness instruction: Be robust to normal Gaussian-Splat and mask- filtering artifacts, blur, rain/fog/low-light appearance, and moderate texture noise. Dynamic foreground actors may be absent or incomplete because their Gaussian components were intentionally filtered; do not penalize their absence.
Scoring focus: Judge whether the static background layout is recognizable, geometrically plausible, structurally coherent, and not catastrophically flat, floating, duplicated, or collapsed. If the prompt says the camera is static, do not require camera motion or parallax.
Calibration: Scores in 0.8–1.0 indicate a clear coherent static scene; 0.6–0.8 indicate a recognizable mostly coherent scene with noticeable artifacts; 0.4–0.6 indicate a partially recognizable scene; 0.2–0.4 indicate a mostly failed reconstruction; and 0.0–0.2 indicate an unusable image. Do not assign a very low score solely because the meta-view is blurry or noisy; if the static scene layout is recognizable, the score should usually be at least 0.5.
Output format: Return strict JSON with score in $[0, 1]$ and a short reason.

C Detailed Results

This section reports domain-wise scores, physical pass rates, interaction diagnostics, geometry diagnostics, and model-level submetrics.

C.1 Benchmark Data Composition

For reference, Table 4 summarizes the domain counts, data sources, and selection rules used to construct the WorldOlympiad test set.

Table 4: Data composition of the WorldOlympiad benchmark test set.

Domain	Count	Source	Selection rule
Robotics	400	RoboCOIN (Wu et al., 2025a)	Downloaded videos that are manually filtered.
Gaming	400	GameGen-X (Che et al., 2025)	Randomly sampled videos from the official OGameData_50K.csv; long videos are split into shorter evaluation chunks.
Real-world	200	LVD-2M (Xiong et al., 2024)	Videos selected from ytb_600k_720p.csv with duration longer than 60 seconds and motion score greater than 50.

Table 5: Detailed WorldOlympiad scores on the same-scene subset across gaming, robotics, and general domains. **All** is the equal-weight average of **Physical**, **3D Cons.**, and **Interact.**

Domain	Pipeline	Physical	3D Cons.	Interact.	CLIP Raw	CLIP Aux.	All
Gaming	Matrix-Game 2.0	0.332	0.189	0.111	0.230	0.150	0.211
	LingBot-World	0.884	0.366	0.778	0.315	0.575	0.676
	Cosmos-Predict-2.5	0.867	0.361	0.679	0.306	0.530	0.636
	WoW	0.633	0.223	0.249	0.247	0.235	0.368
	Rolling Forcing	0.853	0.289	0.675	0.332	0.660	0.606
	LongLive	0.851	0.292	0.554	0.322	0.610	0.566
	Yume-1.5	0.813	0.352	0.659	0.291	0.455	0.608
	Hunyuan-WorldPlay	0.852	0.348	0.471	0.296	0.480	0.557
Robotics	Matrix-Game 2.0	0.364	0.338	0.139	0.252	0.260	0.280
	LingBot-World	0.949	0.393	0.710	0.314	0.570	0.684
	Cosmos-Predict-2.5	0.937	0.479	0.721	0.321	0.605	0.712
	WoW	0.787	0.272	0.447	0.288	0.440	0.502
	Rolling Forcing	0.870	0.389	0.566	0.329	0.645	0.608
	LongLive	0.857	0.472	0.470	0.327	0.635	0.600
	Yume-1.5	0.851	0.288	0.624	0.312	0.560	0.588
	Hunyuan-WorldPlay	0.630	0.600	0.262	0.309	0.545	0.497
General	Matrix-Game 2.0	0.216	0.220	0.067	0.222	0.110	0.168
	LingBot-World	0.963	0.335	0.767	0.311	0.555	0.688
	Cosmos-Predict-2.5	0.939	0.317	0.736	0.313	0.565	0.664
	WoW	0.692	0.251	0.302	0.256	0.280	0.415
	Rolling Forcing	0.933	0.285	0.657	0.314	0.570	0.625
	LongLive	0.909	0.290	0.579	0.315	0.575	0.593
	Yume-1.5	0.925	0.302	0.694	0.302	0.510	0.640
	Hunyuan-WorldPlay	0.389	0.219	0.097	0.235	0.175	0.235

Action-label convention. The action field uses view-relative camera movement labels: `W` means forward, `A` left, `S` backward, and `D` right. Combined labels such as `W+A` indicate simultaneous movement in both directions. `None` denotes no discernible camera movement. These labels describe camera movement only; they do not encode character animation, subject motion, visual effects, or interface changes.

C.2 Domain-wise Results

Table 5 reports the detailed scores on the same-scene subset, grouped by evaluation domain. The table includes physical faithfulness, 3D consistency, CLIP-augmented interaction fidelity, raw and calibrated CLIP semantic alignment, and the overall score. The overall score is computed as the equal-weight average of physical faithfulness, 3D consistency, and interaction fidelity.

C.3 Fine-grained Physical Results

Table 6 reports physical pass rates aggregated by physical dimension. Table 7 further breaks these scores down into individual physical questions.

C.4 Fine-grained Interaction Results

Table 8 reports fine-grained interaction diagnostics. The chunk score measures local caption and action following, the transition score measures boundary smoothness between adjacent chunks, and the global score measures long-range consistency over the stitched video. The raw CLIP score is calibrated into a bounded auxiliary score with fixed thresholds, and the interaction score corresponds to the aggregate interaction metric reported in Table 5.

Table 6: Physical dimension pass rates on the same-scene subset.

Domain	Pipeline	Overall	Mechanics	Thermodynamics	Material
Gaming	Matrix-Game 2.0	0.332	0.433	0.172	0.184
	LingBot-World	0.884	0.983	0.450	0.969
	Cosmos-Predict-2.5	0.867	0.951	0.418	0.884
	WoW	0.633	0.806	0.226	0.446
	Rolling Forcing	0.853	0.941	0.418	0.854
	LongLive	0.851	0.941	0.377	0.865
	Yume-1.5	0.813	0.942	0.365	0.902
	Hunyuan-WorldPlay	0.852	0.944	0.426	0.843
Robotics	Matrix-Game 2.0	0.364	0.366	0.000	0.372
	LingBot-World	0.949	0.961	0.000	0.957
	Cosmos-Predict-2.5	0.937	0.939	0.000	0.968
	WoW	0.787	0.798	0.111	0.788
	Rolling Forcing	0.870	0.857	0.000	0.935
	LongLive	0.857	0.864	0.000	0.869
	Yume-1.5	0.851	0.857	0.000	0.872
	Hunyuan-WorldPlay	0.630	0.577	0.000	0.810
General	Matrix-Game 2.0	0.216	0.246	0.097	0.036
	LingBot-World	0.963	1.000	0.519	1.000
	Cosmos-Predict-2.5	0.939	0.977	0.613	0.875
	WoW	0.692	0.743	0.300	0.562
	Rolling Forcing	0.933	0.968	0.581	0.938
	LongLive	0.909	0.952	0.581	0.812
	Yume-1.5	0.925	0.979	0.370	0.906
	Hunyuan-WorldPlay	0.389	0.430	0.323	0.036

C.5 Fine-grained Geometry Results

Table 9 reports fine-grained geometry diagnostics. S_{recon} measures the quality of the Gaussian-splat reconstruction video, S_{meta} measures the quality of the rendered meta-view image, and S_{traj} measures camera-trajectory consistency. The 3D consistency score corresponds to the aggregate geometry metric reported in Table 5.

C.6 Model-level Fine-grained Results

Table 10 and Table 11 aggregate the fine-grained geometry and interaction diagnostics at the model-category level.

D Case Study

We provide representative qualitative cases that illustrate how WorldOlympiad diagnoses different failure modes beyond generic video quality. Each case uses the same source prompt or reference context across models, so the comparison focuses on model behavior rather than prompt variation.

Gaming case study. Figure 8 shows a gaming case study, where the main diagnostic signals come from geometry consistency and interaction fidelity. The geometry metric examines whether the generated video preserves a stable and spatially coherent game scene under camera movement. In particular, it checks whether the visual content remains consistent with the textual description of the scene, including the expected environment, objects, style, and spatial layout. When the camera moves, a strong model should maintain stable geometry and avoid sudden scene deformation, object disappearance, or inconsistent background structure. The interaction metric further evaluates whether the generated rollout follows the intended action sequence and preserves the game state across chunks. Failure cases include scene drift, unstable camera transitions, chunk-boundary resets, and actions that no longer match the captions.

Table 7: Physical question pass rates on the same-scene subset.

Domain	Pipeline	Grav.	Buoy.	Comp.	Impact	Melt	Sub.	Vap.	Cond.	Dep.	Freez.	Color	Sol.	Hard.	Comb.
Gaming	Matrix-Game 2.0	0.494	0.479	0.324	0.247	0.214	0.000	0.146	0.179	0.167	0.231	–	–	0.168	0.222
	LingBot-World	0.986	0.944	1.000	1.000	0.429	–	0.462	0.417	0.500	0.500	–	–	0.976	0.957
	Cosmos-Predict-2.5	0.958	0.986	0.986	0.868	0.357	0.000	0.292	0.513	0.833	0.538	–	–	0.901	0.843
	WoW	0.850	0.944	0.774	0.540	0.200	0.000	0.161	0.296	0.200	0.333	–	–	0.486	0.355
	Rolling Forcing	0.949	0.987	0.947	0.871	0.357	0.500	0.298	0.475	0.667	0.615	–	–	0.854	0.854
	LongLive	0.955	0.986	0.946	0.846	0.429	0.000	0.292	0.436	0.500	0.462	–	–	0.875	0.843
	Yume-1.5	0.955	1.000	0.900	0.786	0.600	0.000	0.286	0.333	0.500	0.600	–	–	0.902	0.900
	Hunyuan-WorldPlay	0.957	0.986	0.959	0.844	0.500	0.500	0.271	0.513	0.667	0.538	–	–	0.870	0.780
Robotics	Matrix-Game 2.0	0.427	0.467	0.290	0.239	–	–	0.000	–	–	–	–	0.000	0.374	–
	LingBot-World	0.965	1.000	0.985	0.935	–	–	0.000	–	–	–	1.000	0.500	0.962	–
	Cosmos-Predict-2.5	0.945	1.000	1.000	0.894	–	–	0.000	–	–	–	1.000	0.500	0.972	–
	WoW	0.840	0.929	0.845	0.662	–	–	0.111	–	–	–	0.000	0.000	0.800	–
	Rolling Forcing	0.889	1.000	0.889	0.766	–	–	0.000	–	–	–	–	0.000	0.941	–
	LongLive	0.879	0.933	0.938	0.791	–	–	0.000	–	–	–	–	0.000	0.873	–
	Yume-1.5	0.874	1.000	0.938	0.765	–	–	0.000	–	–	–	0.000	0.000	0.885	–
	Hunyuan-WorldPlay	0.644	0.812	0.615	0.373	–	–	0.000	–	–	–	0.000	0.000	0.822	–
General	Matrix-Game 2.0	0.310	0.267	0.111	0.130	0.125	–	0.000	0.000	1.000	0.000	–	–	0.037	0.000
	LingBot-World	1.000	1.000	1.000	1.000	0.600	–	0.400	0.500	1.000	0.667	–	–	1.000	1.000
	Cosmos-Predict-2.5	0.972	1.000	1.000	0.976	0.875	–	0.333	1.000	1.000	0.800	–	–	0.903	0.000
	WoW	0.767	0.806	0.889	0.634	0.500	–	0.133	–	0.500	0.400	–	–	0.581	0.000
	Rolling Forcing	0.978	0.935	1.000	0.951	0.875	–	0.267	1.000	1.000	0.800	–	–	0.935	1.000
	LongLive	0.956	1.000	1.000	0.915	0.875	–	0.400	0.000	0.500	0.800	–	–	0.839	0.000
	Yume-1.5	0.983	1.000	1.000	0.958	0.400	–	0.267	0.500	1.000	0.333	–	–	0.903	1.000
	Hunyuan-WorldPlay	0.494	0.467	0.500	0.260	0.500	–	0.067	0.000	1.000	0.600	–	–	0.037	0.000

Table 8: Fine-grained interaction diagnostics on the same-scene subset.

Domain	Pipeline	Chunk	Trans.	Global	Long Range	Global Text	CLIP Raw	CLIP Aux.	Interact.
Gaming	Matrix-Game 2.0	0.135	0.074	0.087	0.087	0.087	0.230	0.150	0.111
	LingBot-World	0.796	0.767	0.862	0.875	0.848	0.315	0.575	0.778
	Cosmos-Predict-2.5	0.704	0.677	0.700	0.718	0.680	0.306	0.530	0.679
	WoW	0.267	0.233	0.247	0.250	0.244	0.247	0.235	0.249
	Rolling Forcing	0.665	0.681	0.704	0.733	0.675	0.332	0.660	0.675
	LongLive	0.595	0.444	0.625	0.640	0.606	0.322	0.610	0.554
	Yume-1.5	0.645	0.727	0.668	0.702	0.632	0.291	0.455	0.659
	Hunyuan-WorldPlay	0.483	0.458	0.440	0.464	0.415	0.296	0.480	0.471
Robotics	Matrix-Game 2.0	0.136	0.167	0.041	0.042	0.041	0.252	0.260	0.139
	LingBot-World	0.670	0.714	0.881	0.893	0.869	0.314	0.570	0.710
	Cosmos-Predict-2.5	0.682	0.707	0.896	0.908	0.885	0.321	0.605	0.721
	WoW	0.413	0.501	0.472	0.493	0.451	0.288	0.440	0.447
	Rolling Forcing	0.498	0.600	0.632	0.661	0.603	0.329	0.645	0.566
	LongLive	0.484	0.288	0.587	0.619	0.556	0.327	0.635	0.470
	Yume-1.5	0.553	0.715	0.694	0.722	0.667	0.312	0.560	0.624
	Hunyuan-WorldPlay	0.245	0.254	0.134	0.140	0.129	0.309	0.545	0.262
General	Matrix-Game 2.0	0.069	0.064	0.031	0.031	0.029	0.222	0.110	0.067
	LingBot-World	0.752	0.819	0.829	0.838	0.812	0.311	0.555	0.767
	Cosmos-Predict-2.5	0.746	0.755	0.764	0.782	0.746	0.313	0.565	0.736
	WoW	0.314	0.294	0.292	0.293	0.285	0.256	0.280	0.302
	Rolling Forcing	0.620	0.727	0.661	0.684	0.629	0.314	0.570	0.657
	LongLive	0.598	0.520	0.639	0.661	0.603	0.315	0.575	0.579
	Yume-1.5	0.637	0.814	0.718	0.744	0.684	0.302	0.510	0.694
	Hunyuan-WorldPlay	0.130	0.030	0.073	0.073	0.073	0.235	0.175	0.097

Table 9: Fine-grained geometry diagnostics on the same-scene subset.

Domain	Pipeline	S_{recon}	S_{meta}	S_{traj}	3D Cons.
Gaming	Matrix-Game 2.0	0.160	0.159	0.247	0.189
	LingBot-World	0.389	0.372	0.337	0.366
	Cosmos-Predict-2.5	0.415	0.388	0.280	0.361
	WoW	0.232	0.205	0.231	0.223
	Rolling Forcing	0.324	0.292	0.250	0.289
	LongLive	0.328	0.292	0.256	0.292
	Yume-1.5	0.381	0.361	0.315	0.352
	Hunyuan-WorldPlay	0.397	0.363	0.284	0.348
Robotics	Matrix-Game 2.0	0.283	0.298	0.432	0.338
	LingBot-World	0.416	0.416	0.348	0.393
	Cosmos-Predict-2.5	0.451	0.464	0.523	0.479
	WoW	0.297	0.289	0.232	0.272
	Rolling Forcing	0.458	0.432	0.278	0.389
	LongLive	0.476	0.483	0.458	0.472
	Yume-1.5	0.337	0.340	0.185	0.288
	Hunyuan-WorldPlay	0.574	0.566	0.660	0.600
General	Matrix-Game 2.0	0.191	0.196	0.271	0.220
	LingBot-World	0.373	0.319	0.312	0.335
	Cosmos-Predict-2.5	0.341	0.322	0.288	0.317
	WoW	0.243	0.225	0.286	0.251
	Rolling Forcing	0.283	0.266	0.306	0.285
	LongLive	0.289	0.280	0.300	0.290
	Yume-1.5	0.318	0.306	0.282	0.302
	Hunyuan-WorldPlay	0.177	0.191	0.288	0.219

Table 10: Model-level 3D consistency submetrics.

Category	Model	GS	Meta	Camera Motion	3D Cons.
Gaming World Model	Matrix-Game 2.0	0.216	0.222	0.326	0.255
	LingBot-World	0.400	0.383	0.337	0.373
Robotics World Model	Cosmos-Predict-2.5	0.415	0.405	0.378	0.399
	WoW	0.262	0.245	0.244	0.250
General World Model	Rolling Forcing	0.359	0.332	0.272	0.321
	LongLive	0.379	0.365	0.345	0.363
	Yume-1.5	0.338	0.334	0.231	0.301
	Hunyuan-WorldPlay	0.426	0.412	0.436	0.424

Table 11: Model-level interaction submetrics.

Category	Model	Chunk	Trans.	Global	Long Range	Global Text	CLIP Raw	CLIP Aux.	Interact.
Gaming World Model	Matrix-Game 2.0	0.123	0.109	0.058	0.058	0.057	0.237	0.185	0.113
	LingBot-World	0.709	0.751	0.864	0.875	0.850	0.314	0.570	0.734
Robotics World Model	Cosmos-Predict-2.5	0.704	0.705	0.791	0.807	0.775	0.313	0.565	0.707
	WoW	0.339	0.359	0.352	0.362	0.341	0.267	0.335	0.345
General World Model	Rolling Forcing	0.600	0.666	0.671	0.699	0.641	0.327	0.635	0.636
	LongLive	0.552	0.398	0.613	0.636	0.585	0.323	0.615	0.526
	Yume-1.5	0.590	0.745	0.697	0.726	0.667	0.306	0.530	0.649
	Hunyuan-WorldPlay	0.320	0.294	0.247	0.259	0.235	0.290	0.450	0.316

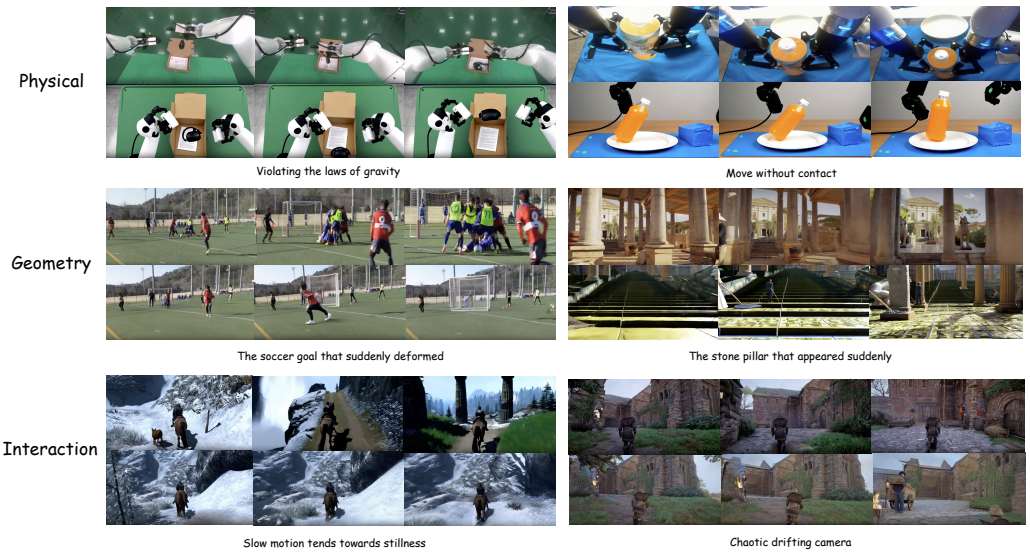


Figure 6: Representative WorldOlympiad case studies detected by the benchmark. The upper examples show high-quality generations that preserve the intended physical behavior, scene structure, or interaction state, while the lower examples show typical failure cases with visible rule violations, geometric inconsistency, or interaction drift.

Table 12: Representative case studies and the corresponding diagnostic signals.

Case	Evaluation focus	Typical success pattern	Typical failure pattern
Physical dynamics	Gravity, impact, deformation, or phase transition	The object follows the expected temporal order, preserves contact constraints, and changes state gradually when required.	The object floats, teleports, deforms without contact, changes phase instantaneously, or violates the expected direction of motion.
3D consistency	Gaussian-splat reconstruction and camera trajectory	The reconstructed scene remains stable under novel views, with consistent foreground objects and plausible camera motion.	The reconstruction contains stretched geometry, missing background structure, unstable object identity, or camera motion that disagrees with the reference trajectory.
Interactive rollout	Chunk-level instruction following and transition coherence	Each generated chunk follows its action caption, and the next chunk preserves scene state, agent pose, and object layout.	The model resets the scene at chunk boundaries, ignores control changes, changes object identity, or accumulates visual drift over long horizons.

Robotics case study. Figure 7 presents an robotics manipulation case, where WorldOlympiad jointly examines physical plausibility, scene-level geometric consistency, and instruction-following behavior. For physical evaluation, the case highlights failures such as an apple floating in mid-air despite the absence of visible support, indicating a violation of gravity and object-support constraints. For geometry evaluation, the benchmark further checks whether the scene layout remains coherent throughout the rollout. For example, a drawer may suddenly appear or disappear across frames, revealing inconsistent spatial structure and unstable background reconstruction. For interaction evaluation, the judge focuses on whether the robot follows the intended manipulation instruction, such as reaching toward the correct object, grasping the target item rather than a distractor, and maintaining a plausible object state after contact. This case shows that plausible robotics videos can fail when object dynamics, scene consistency, or robot-action alignment break.

General case study. Figure 9 presents a real-world case study, where all three evaluation dimensions are informative. For physical evaluation, the case checks whether the motion of a thrown frisbee follows a plausible trajectory, rather than floating, stopping unnaturally, or changing direction without visible cause. For geometry evaluation, the benchmark inspects whether the scene remains spatially and semantically consistent over time. For instance, a failure case may abruptly change an indoor scene into an outdoor scene, indicating severe scene-level inconsistency and poor long-range coherence. For interaction evaluation, the judge examines whether the generated video contains meaningful temporal evolution rather than becoming overly static. A strong sample should preserve realistic motion, maintain a coherent scene layout, and continue to reflect the intended event throughout the video. These cases expose failures in physics, geometry, and temporal interaction.

E Human Preference Study Details

Table 13 reports results from the protocol below.

Annotation protocol. For each selected evaluation prompt, annotators compare anonymized generated videos from the evaluated models under the same prompt or reference context. Five annotators participate in the study. We sample 20 prompts from the evaluation set and compare all $\binom{8}{2} = 28$ unordered model pairs under each prompt, resulting in 560 prompt-level pairwise comparisons. Each comparison is independently labeled by all five annotators, yielding 2,800 individual preference labels. Annotators are instructed to judge the overall preference using four criteria: visual quality, physical plausibility, temporal coherence, and interaction fidelity. Model names are hidden during annotation. Ties are allowed when neither video is clearly better.

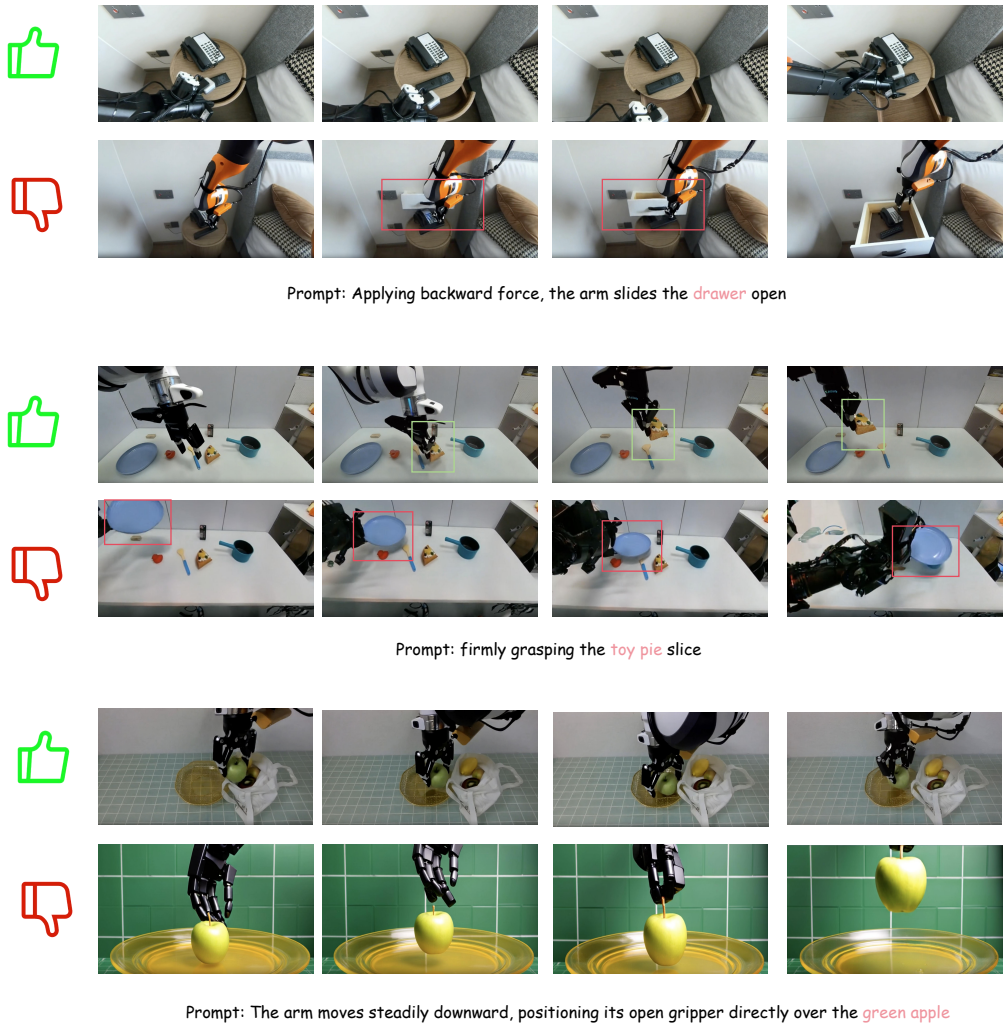


Figure 7: Robotics case study from WorldOlympiad. The example visualizes how the benchmark diagnoses physical interaction, object-state consistency, and temporal coherence in robotics world-model rollouts.

Table 13: Alignment between human preference rankings and WorldOlympiad automatic rankings. S^{human} denotes the pairwise human preference score, and S^{auto} denotes the WorldOlympiad **All** score. Rank gap is computed as human rank minus automatic rank.

Category	Model	S^{human}	S^{auto}	Human Rank	Auto Rank	Rank Gap
Gaming World Model	LingBot-World	0.721	0.683	1	1	0
Robotics World Model	Cosmos-Predict-2.5	0.648	0.671	2	2	0
General World Model	Rolling Forcing	0.579	0.610	3	3	0
General World Model	LongLive	0.532	0.584	4	5	-1
General World Model	Yume-1.5	0.491	0.604	5	4	1
General World Model	Hunyuan-WorldPlay	0.423	0.477	6	6	0
Gaming World Model	Matrix-Game 2.0	0.309	0.231	7	8	-1
Robotics World Model	WoW	0.271	0.434	8	7	1



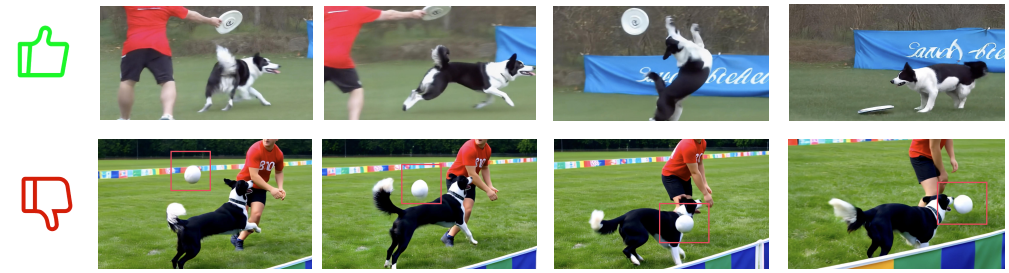
Figure 8: Gaming case study from WorldOlympiad. The example highlights how interactive game rollouts expose action-following, scene-state preservation, and cross-chunk transition failures.



Prompt: She expertly manipulates a long, colorful ribbon, creating large, flowing circles and spirals around her as she gracefully dances and leaps



Prompt: In an artificial wave pool, a young boy wearing a black long-sleeved swimsuit



Prompt: throws a white frisbee to the right. A black and white dog immediately sprints across the field to chase it

Figure 9: Real-world case study from WorldOlympiad. The example illustrates how open-domain videos reveal geometric consistency, camera-motion, and long-range visual-coherence issues.

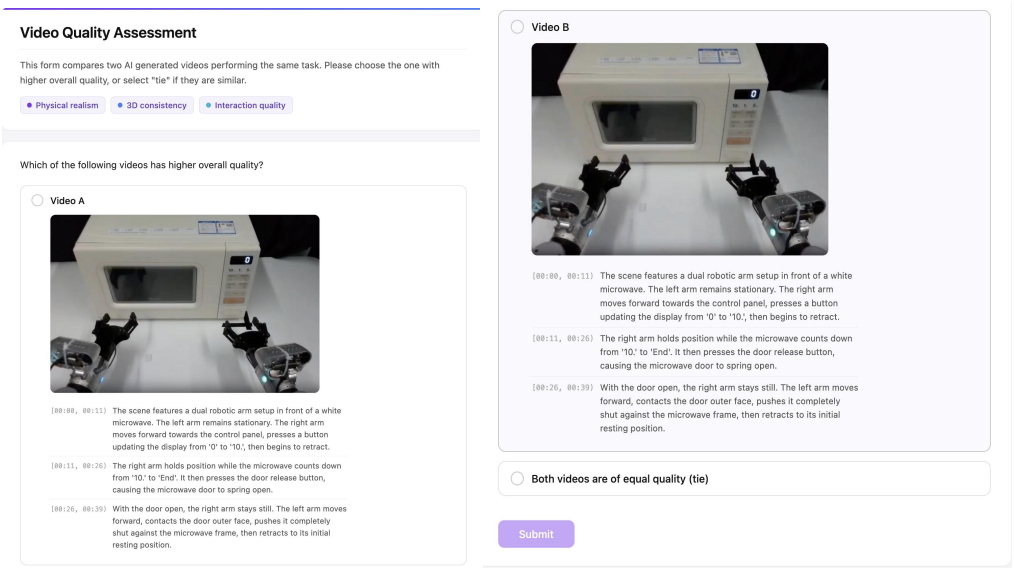


Figure 10: Human preference annotation interface used in the alignment study.

Human Preference Annotation Prompt

Input: Two anonymized generated videos, Video A and Video B, produced from the same prompt or reference context.
Instruction: Choose the video that better satisfies the prompt and shows more realistic world dynamics. Consider visual quality, physical plausibility, temporal coherence, object and scene consistency, and interaction fidelity. If neither video is clearly better, select Tie.
Output format: Return one label from {`A`, `B`, `Tie`} and a one-sentence rationale.

Score aggregation. Let $y_{m,n,a}$ denote the preference outcome assigned by annotator a for model m in comparison n. A win contributes 1, a tie contributes 0.5, and a loss contributes 0. We first average the five annotator labels for each pairwise comparison and then compute the model-level preference rate:

$$\bar{y}_{m,n} = \frac{1}{5}\sum_{a=1}^{5} y_{m,n,a}, \quad S_m^{\text{human}} = \frac{1}{N_m}\sum_{n=1}^{N_m} \bar{y}_{m,n},$$

where $N_m = 140$ is the number of aggregated valid comparisons involving each model. Human ranks are obtained by sorting S^{human} in descending order. WorldOlympiad ranks are obtained by sorting the automatic overall evaluation score in descending order, where S^{auto} is the same three-track average S_{all} used in the main benchmark table.

Rank correlation. We measure alignment using Spearman's rank correlation:

$$\rho = 1 - \frac{6\sum_{m=1}^{M} d_m^2}{M(M^2-1)}, \quad d_m = r_m^{\text{human}} - r_m^{\text{auto}},$$

where M is the number of evaluated models. For the eight models with human preference annotations, the correlation is 0.95, showing close agreement between human and automatic rankings.

Only two adjacent pairs swap ranks: LongLive with Yume-1.5 and Matrix-Game 2.0 with WoW. These swaps have little effect on the correlation while highlighting borderline cases where human and rubric-based automatic rankings differ.