WBENCH: A COMPREHENSIVE MULTI-TURN BENCHMARK FOR INTERACTIVE VIDEO WORLD MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive world models are advancing rapidly, yet existing benchmarks cover only part of the required competencies, leaving no unified standard for systematic evaluation. To fill this gap, we introduce **WBench**, a comprehensive multi-turn benchmark for interactive world model evaluation along five dimensions, namely video quality, setting adherence, interaction adherence, consistency, and physics compliance. WBench contains 289 test cases and 1,058 interaction turns, where each case specifies a world setting and a multi-turn interaction sequence, covering diverse scenes, styles, subjects, and both first- and third-person perspectives, together with four interaction types, including navigation, subject action, event editing, and perspective switching. For navigation, WBench unifies text, 6-DoF pose, and discrete-action control, enabling evaluation of models with different native input interfaces. Evaluation uses 22 automatic sub-metrics that combine specialist vision models with vision-language models. Across 20 state-of-the-art models, we find that no single model performs strongly across all dimensions. We hope these diagnostic insights into model strengths, weaknesses, and open challenges will help advance research on interactive world models.

1 INTRODUCTION

Recent advances in video generation (Ho et al., 2020; Kong et al., 2024; Wan et al., 2025; Polyak et al., 2024; Brooks et al., 2024) have enabled interactive world models with controllable generation across games (Bruce et al., 2024; Parker-Holder et al., 2024; Ball et al., 2025; He et al., 2025; Wang et al., 2026; Li et al., 2025b; Tang et al., 2025), autonomous driving (Hu et al., 2023; Gao et al., 2024), embodied interaction (Yang et al., 2023; Zhu et al., 2024), and open-domain scenarios (HunyuanWorld, 2025; Mao et al., 2025b;a; Team et al., 2026c). However, evaluation remains fragmented, with many works relying on selected demos or task-specific protocols, making fair comparison and failure diagnosis difficult across visual quality, controllability, memory, and physics.

A capable interactive world model must fulfill five complementary roles, analogous to the subsystems of a game engine: a *Renderer* for visually convincing video, a *Director* for correct world initialization, a *Controller* for faithful interaction execution, a *Memory* for preserving world state across turns, and an *Engine* for physically compliant world evolution. Existing benchmarks cover these roles only partially (Table 1). Video-generation benchmarks such as VBench (Huang et al., 2024; Zheng et al., 2025) focus on perceptual quality without interactive control. World-model benchmarks evaluate more dimensions but remain limited in scope: WorldMark (Xu et al., 2026) and MIND (Ye et al., 2026) cover navigation and memory but lack semantic interactions, Omni-WorldBench (Wu et al., 2026a) adds causal interaction but supports only first-person perspective, and WorldLens (Liang et al., 2025) evaluates multiple dimensions but is restricted to autonomous driving. None provides a unified protocol spanning open-domain scenes, both perspectives, and all four interaction types.

To address this gap, we introduce WBench, a comprehensive multi-turn benchmark for interactive world model evaluation. As shown in Fig. 1, each test case is defined by a *world setting* (scene, subject, style, and perspective) together with a multi-turn *interaction* sequence. The top row illustrates a concrete case: a realistic snowy mountain scene with a human subject in third-person perspective, followed by forward navigation, a jump, the appearance of a helicopter, and a perspective switch

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.

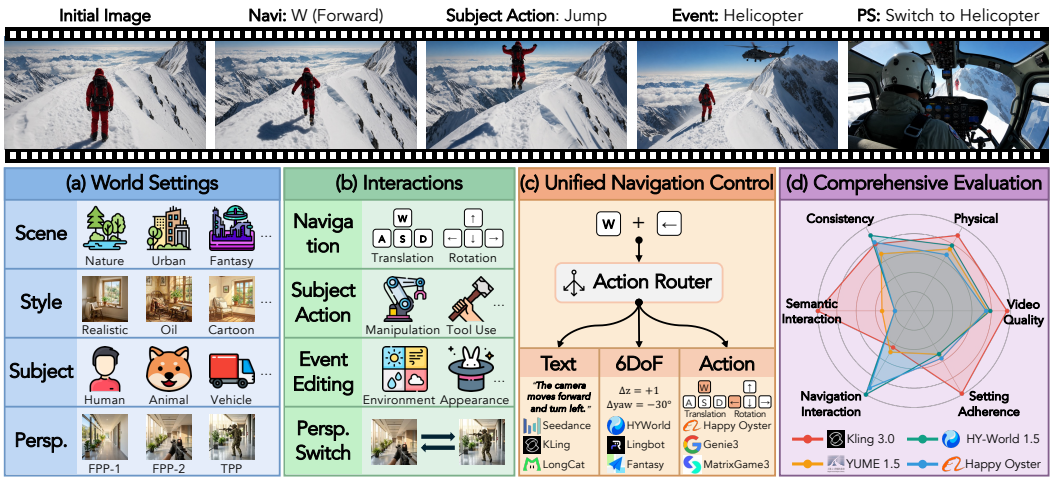


Figure 1: Overview of **WBench**. **Top:** a multi-turn case with navigation, subject action, event editing, and perspective switching. **Bottom:** the benchmark design, including world settings, interaction taxonomy, unified navigation control, and evaluation over video quality, setting adherence, interaction adherence (navigation and semantic interactions), consistency, and physics compliance.

to the cockpit. More broadly, the benchmark spans diverse open-domain scenes, rendering styles, subject categories, and both first- and third-person perspectives (Fig. 1 (a)), with four interaction types shown in Fig. 1 (b): navigation, subject action, event editing, and perspective switching. This design separates what the world *is* from what the user *requests*, making failure modes easier to locate: a model may render the initial scene well but ignore later actions, or follow a single instruction correctly but lose identity and spatial consistency over multiple turns.

WBench also supports fair comparison across different control paradigms. As shown in Fig. 1 (c), navigation interactions are represented in three aligned forms, namely text, camera pose, and discrete action, so that models can be evaluated through their native interfaces. Accordingly, we adopt a dual-track evaluation protocol: all 20 models are compared on a shared navigation subset of 158 cases, while text-driven models are further evaluated on the full benchmark (289 cases, 1,058 turns). The 22 sub-metrics combine specialist vision models and vision-language models (VLMs).

Experiments on 20 models reveal that: 1) no model dominates all five dimensions, 2) navigation correlates weakly with other dimensions, 3) navigation and perspective consistency rankings differ, 4) physical scores correlate more with rendering quality than navigation, 5) benchmark difficulty is structured by perspective, scene type, and subject category, and 6) four interaction types degrade unevenly over turns, with navigation most fragile.

Our contributions are 1) an interaction taxonomy combining four interaction types with explicit world settings and both perspectives in multi-turn cases, 2) a unified navigation interface for text, camera-pose, and discrete-action control, and 3) an evaluation pipeline with 22 sub-metrics across five dimensions, establishing diagnostic baselines for 20 models.

2 Related Work

Video Generation Models. Video generation has evolved rapidly, from early U-Net-based diffusion models (Ho et al., 2020; 2022; Blattmann et al., 2023) to scalable Diffusion Transformers (Yang et al., 2024; Polyak et al., 2024; Kong et al., 2024) trained with flow-matching objectives on large-scale data, yielding longer, higher-resolution, and temporally coherent outputs. Building on this foundation, the current frontier like Sora 2 (OpenAI, 2025b), Kling 3.0 (Kuaishou Technology, 2025), Veo 3 (Google DeepMind, 2025), Wan 2.7 (Alibaba Cloud, 2026), and others (Seedance et al., 2026; Shengshu Technology, 2025; Kong et al., 2024; HaCohen et al., 2024; Gu, 2025; Team et al., 2025), collectively advance cinematic quality, prompt adherence, efficient inference, physical grounding, and long-horizon continuation. Despite these advances, evaluation still centers

Table 1: Comparison with representative benchmarks. Input: T = text, camera, action. FPP /TPP: first-/third-person perspective. Navi, SA, EE, PS: navigation, subject action, event editing, and perspective switching. Qual, Adh, Inter, Cons, Phys: video quality, setting adherence, interaction adherence, consistency, and physics compliance. Full comparison is provided in Table 3.

Benchmark	Input	Persp.		Interactions				Dimensions					Scale	
		FPP	TPP	Navi	SA	EE	PS	Qual	Adh	Inter	Cons	Phys	Cases	Turns
VBench (Huang et al., 2024)	T	–	–	✗	✗	✗	✗	✓	✗	✗	✗	✗	946	946
WorldScore (Duan et al., 2025)	T,	✓	✗	✓	✗	✓	✗	✓	✓	✗	✗	✗	3,000	3,000
WorldModelBench (Li et al., 2025a)	T	✓	–	✓	✓	✗	✗	✗	✓	✓	✗	✓	350	350
MIND (Ye et al., 2026)	,	✓	–	✓	✗	✗	✗	✓	✗	✓	✓	✗	250	–
WorldArena (Shang et al., 2026)	T,	✓	✗	✓	✓	✗	✗	✓	✗	✓	✓	✓	500	500
Omni-WorldBench (Wu et al., 2026a)	T,	✓	–	✓	✓	✓	✗	✓	✗	✓	✓	✓	1,068	1,068
WorldLens (Liang et al., 2025)	,	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	✓	–	–
WorldMark (Xu et al., 2026)	T,	✓	✓	✓	✗	✗	✗	✓	✗	✓	✓	✗	500	–
WBench (Ours)	T,	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	289	1,058

on distributional metrics (FID (Heusel et al., 2017), FVD (Unterthiner et al., 2018)), text-alignment scores, or multi-dimensional quality suites (Huang et al., 2024), which do not assess interactive control.

Interactive Video World Models. World models (Ha & Schmidhuber, 2018; LeCun et al., 2022) predict environment evolution in response to actions. While traditionally realized as latent state-space models, recent video generators have enabled a new paradigm: *interactive video world models* that directly synthesize next frames from the current observation and an action signal, enabling closed-loop simulation. Although early systems also appeared in robotic manipulation and autonomous driving, such as UniSim (Yang et al., 2023), IRASim (Zhu et al., 2024), GAIA-1 (Hu et al., 2023), and Vista (Gao et al., 2024), we focus on the open-domain branch most relevant to WBench. Among world models evaluated in this work, YUME 1.5 (Mao et al., 2025a) represents language-driven interaction, using natural-language actions for multi-turn world evolution, while HY-World 1.5 (HunyuanWorld, 2025) and LingBot-World (Team et al., 2026c) represent camera-controlled generation with an emphasis on navigation and geometric consistency. Action-conditioned systems such as Hunyuan-GameCraft (Li et al., 2025b; Tang et al., 2025), Matrix-Game 2.0 (He et al., 2025), and Matrix-Game 3.0 (Wang et al., 2026) push real-time keyboard-and-mouse control, with Matrix-Game 3.0 further improving long-horizon consistency through explicit memory. Closed-source systems such as Genie 3 (Ball et al., 2025), Happy Oyster (Alibaba Token Hub, 2026), and Marble (World Labs, 2025) further highlight the momentum of this area.

World Model Evaluation. As shown in Table 1, existing benchmarks fall into two broad groups. Non-interactive suites such as VBench (Huang et al., 2024; 2025; Zheng et al., 2025), EvalCrafter (Liu et al., 2024), and VideoPhy (Bansal et al., 2024; Meng et al., 2024) assess video quality, text alignment, or physical commonsense, but do not take action inputs or evaluate multi-turn interaction. Among world model benchmarks, WorldScore (Duan et al., 2025) evaluates camera-trajectory-conditioned generation, WorldModelBench (Li et al., 2025a) studies decision-oriented world-model quality, WorldArena (Shang et al., 2026) targets embodied agents in closed domains, MIND (Ye et al., 2026) probes closed-loop memory consistency, Omni-WorldBench (Wu et al., 2026a) focuses on causal interaction, WorldLens (Liang et al., 2025) targets autonomous driving, and WorldMark (Xu et al., 2026) measures navigation consistency. Additional efforts examine complementary aspects (Han et al., 2025; Liu et al., 2023; Gu et al., 2026; Cai et al., 2025; Upadhyay et al., 2026; Lu et al., 2025; Team et al., 2026b; Qin et al., 2024; Yue et al., 2025; Zhang et al., 2025; Tang et al., 2025; Zhou et al., 2026). Despite this rapid progress, no existing benchmark jointly covers (i) diverse open-domain scenes, (ii) both first- and third-person perspectives with perspective-dependent action semantics, (iii) a comprehensive interaction taxonomy spanning navigation, subject action, event editing, and perspective switching, and (iv) multi-turn closed-loop evaluation targeting long-horizon consistency and physics compliance. WBench fills this gap with a unified framework across all four axes and 22 automatic sub-metrics. The comparison highlights interaction coverage, multi-turn structure, and control interfaces. Case and turn counts indicate scale, while WBench prioritizes diagnostic breadth.

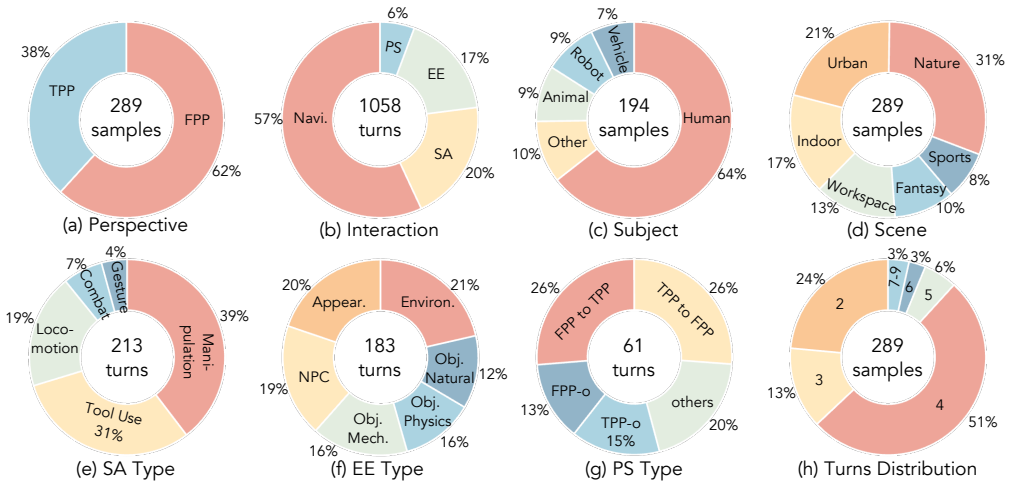


Figure 2: Dataset composition of WBench across eight axes. We discuss these in Section 3.2.

3 WBench Dataset

An interactive world model (Ha & Schmidhuber, 2018; LeCun et al., 2022) predicts the next observation o_{t+1} from observation history $o_{\leq t} = (o_0, \ldots, o_t)$ and actions $a_{\leq t} = (a_0, \ldots, a_t)$,

$$o_{t+1} \sim f_\theta(o_{t+1} | o_{\leq t}, a_{\leq t}). \tag{1}$$

Each WBench case specifies a **World Setting** $\mathcal{W}$ defining the initial state o_0 and an **Interaction sequence** $\mathcal{I} = (a_0, a_1, \ldots, a_{T-1})$ defining user control over T consecutive turns.

3.1 Dataset Construction

World Settings. A world setting $\mathcal{W}$ is defined by four attributes: **1) Scene**, the environment type, spatial layout, and inherent dynamics, including both elements visible in the initial frame (*e.g.*, terrain, buildings) and offscreen elements expected to appear during interaction (*e.g.*, a river behind the camera); **2) Style**, the rendering appearance, such as realistic, cartoon, anime, cinematic, CG, or oil painting; **3) Perspective**, either first- or third-person; and **4) Subject**, the primary entity in the scene, such as a human, animal, vehicle, or robot. The subject attribute applies to all third-person cases and first-person cases where the viewer holds or controls a visible entity (*e.g.*, a tool or an ego robot arm), while environment-only first-person scenes have no associated subject. These four attributes are composed into an **environment prompt** (Scene + Style) and a **subject prompt** (Perspective + Subject), which together with an initial frame form the input to each evaluated model. Initial frames are generated by Nano Banana 2 (Google, 2025) and GPT-Image-1.5 (OpenAI, 2025a), supplemented by web-collected and manually captured images, with all frames manually checked for quality.

Interactions. Each case specifies a multi-turn interaction sequence drawn from four complementary types that can be freely composed within a single case, as shown in Fig. 1 (top). **1) Navigation** governs camera or ego-agent motion through four *translational* controls `W`/`S`/`A`/`D` and four *rotational* controls ←/→/↑/↓, composable into compound actions such as `W`+←. The same key drives the camera in first-person mode and the subject in third-person mode. Trajectories span six path topologies for motion diversity (Appendix A.3). **2) Subject Action** covers actions performed by the primary subject, including manipulation, locomotion, tool use, combat, and gestural interaction. **3) Event Editing** covers externally imposed changes to the environment, such as weather transitions, time-of-day shifts, object appearances, *etc.* **4) Perspective Switching** covers transitions between first- and third-person perspectives, including same-subject switches, multi-subject switches, and scope mode transitions.

Case Construction. Annotators first design world settings, then derive physically executable interaction sequences with coherent semantics and causal ordering. Examples include manipulation in a kitchen, outdoor weather changes, and navigation trajectories consistent with the scene layout. Sampling is stratified by scene, style, perspective, subject, and interaction type. Following Appendix D.5, all selected cases are manually reviewed for prompt-frame consistency and inter-turn coherence.

3.2 Dataset Statistics

The navigation split contains 158 cases with at least one navigation turn, including mixed-interaction cases. The remaining 131 cases contain no navigation turns and form the non-navigation split.

WBench contains 289 cases and 1,058 turns, with 62% first-person and 38% third-person cases. Navigation accounts for 57% of turns, followed by subject action (20%), event editing (17%), and perspective switching (6%). Figure 2 summarizes coverage across world settings and interactions.

Scene, Subject, and Style Diversity. Scenes span six categories, led by nature (31%) and urban environments (21%), with indoor (17%), workspace (13%), fantasy (10%), and sports (8%) settings completing the spectrum, see Fig. 2 (d). Across the 194 cases with an explicit subject, humans dominate (64%), followed by animals (9%), robots (9%), vehicles (7%), and 10% miscellaneous objects (Fig. 2 (c)). Photorealistic rendering covers 52% of the cases, while the remaining 48% span styles including anime, cartoon, CG, oil painting, ink wash, pencil sketch, and flat or abstract styles.

Interaction Sub-type Taxonomy. Subject action covers manipulation (39%), tool use (31%), locomotion, combat, and gestures. Event editing includes environment changes (21%), appearance-state changes (20%), NPC motion (19%), and mechanical, physical, or natural object-state transitions. Of the 61 perspective-switch turns, each cross-perspective direction accounts for 26%, intra-perspective switches for 28%, and other transitions include TPP-to-scope, as shown in Fig. 2 (e–g).

Multi-turn Interaction Depth. Cases span 2–9 turns with an average of 3.7. Four-turn cases (51%) mostly follow navigation trajectories, while longer 5–9-turn cases (12%) typically interleave subject action and event editing. These sequences probe temporal consistency beyond single-turn evaluation. Appendix A details navigation coverage, metric activation, and lexical diversity.

4 WBench Evaluation Suite

WBench evaluates five complementary dimensions through 22 sub-metrics, all reported on a 0–100 scale with higher values indicating better performance. Appendix C provides metric definitions, score transformations, and implementation details.

Video Quality. Video quality measures the perceptual quality of the generated video irrespective of the conditioning signal. We adopt five sub-metrics from VBench (Huang et al., 2024): V.1 Aesthetic Quality, V.2 Imaging Quality, V.3 Temporal Flickering, V.4 Dynamic Degree and V.5 Motion Smoothness, plus the percentile-normalized preference score V.6 HPSv3-Norm (Ma et al., 2025).

Setting Adherence. Setting adherence measures agreement with the world setting $\mathcal{W}$. For S.1 **Scene Adherence**, a VLM[1] checks whether initially visible elements remain consistent and specified offscreen elements appear later. For S.2 **Subject Adherence**, it compares the subject's appearance and movement with the subject prompt.

Interaction Adherence. Interaction adherence evaluates execution of $\mathcal{I}$ through geometric pose estimation for navigation and turn-level VLM checks for semantic interactions.

I.1 Navigation Score. We compare MegaSaM-estimated camera poses (Li et al., 2025c) with a synthetic trajectory derived from the action sequence. First-person rotations change heading, while third-person rotations orbit the subject. After alignment and arc-length resampling, normalized Absolute Trajectory Error (nATE) measures accuracy. For symmetric action pairs, cross-turn consistency compares trajectory shapes after mirror alignment. NavScore averages accuracy and consistency when both are available. Model-level aggregation first averages each component over its own valid cases and then combines the two means.

I.2 Event Editing and I.3 Subject Action Adherence. A shared VLM protocol applies five binary checks per turn, covering change detection, event occurrence, completion, detail accuracy, and anomaly absence. Each satisfied check contributes one point. The resulting $[0, 5]$ grades are averaged across turns and scaled to 0–100, with prompts and criteria in Appendices C.4.2 and C.4.3.

I.4 Perspective Switching Adherence. The VLM checks early and late frames for transition visibility, target-type consistency, and structural compliance, as defined in Appendix C.4.4. We report the percentage of turns satisfying all three criteria for the requested perspective.

[1] Unless otherwise noted, VLM scoring uses `doubao-seed-2-0-lite`.

Consistency. We assess geometric, visual, and perspective stability across successive interactions.

C.1 Spatial Consistency and C.2 Gated Spatial Consistency. For roundtrip trajectories (Ye et al., 2026) (*e.g.* left$\times 2 \rightarrow$ right$\times 2$), we use MegaSaM-estimated camera poses to locate the return frame best matching the initial viewpoint, then compute DreamSim (Fu et al., 2023) perceptual similarity with the first frame. The *gated* variant additionally samples intermediate frames and computes their minimum similarity to the first frame, suppressing the score when the video barely moves.

C.3 Segment Continuity. The score is the fraction of videos without unexpected hard cuts detected by TransNetV2 (Soucek & Lokoc, 2024).

C.4 Perspective Consistency. We measure the stability of the SAM2-tracked subject centroid (Ravi et al., 2024), weighted by the fraction of frames with a visible subject.

C.5 Geometric Consistency and C.6 Photometric Consistency. We use Depth Anything 3 (Lin et al., 2025) to estimate per-frame depth and camera poses, then reproject pixels across views. Geometric consistency measures agreement between reprojected and predicted depths, while photometric consistency measures appearance stability via PSNR between reprojected colors and observed frames.

C.7 Subject Consistency. We use SAM2 masks to crop the subject, falling back to the full frame when a valid mask is unavailable. We average DINOv2 (Oquab et al., 2023) adjacent-frame cosine similarity and CLIP first-frame anchored similarity to capture local continuity and global drift.

C.8 Background Consistency. We follow VBench (Huang et al., 2024) and average CLIP similarity to preceding and initial frames, measuring local and accumulated background changes.

Physical. The physical dimension covers high-level causal fidelity and low-level visual plausibility.

P.1 Causal Fidelity. The VLM assesses frames sampled across all turns in two stages, each using integer scores from 0 to 3. **Stage 1** checks global rendering-physics and causal violations, including motion discontinuities, failures of object permanence, implausible character physics, and effects without causes. It also checks causes that fail to produce effects and unrelated objects that appear unexpectedly. Expected changes caused by instructed actions are not treated as violations. **Stage 2** checks fluid and smoke, collision, surface tracks, deformation, wind, reflection, and human motion. Only applicable sub-dimensions are scored, using a manually verified set held fixed across models. We average Stage 1 with the mean Stage 2 score and scale to 0–100. Appendix C.6.1 gives the applicability rules and prompts.

P.2 Visual Plausibility. We fine-tune Qwen3-VL-30B-A3B (Bai et al., 2025) on expert-annotated data to score geometric distortion, object penetration, and unnatural deformation. Its continuous $[1, 5]$ score is multiplied by 20. This complements causal fidelity with artifact detection that requires no case-specific questions, as detailed in Appendix C.6.2.

5 Experiments

5.1 Evaluated Models and Protocol

As shown in Table 2, we evaluate 20 models spanning three paradigms: (1) **text-driven** (9 models, *e.g.* Seedance 1.5 (Seedance et al., 2025), Kling 3.0 (Kuaishou Technology, 2025), Wan 2.7 (Alibaba Cloud, 2026)), which accept all four interaction types and are evaluated on the full 289-case test set via iterative last-frame forwarding, (2) **camera-controlled** (5 models, *e.g.* HY-World 1.5 (HunyuanWorld, 2025), LingBot-World (Team et al., 2026c)), and (3) **action-conditioned** (6 models, *e.g.* Genie 3 (Ball et al., 2025), Matrix-Game 3.0 (Wang et al., 2026)), both restricted to the navigation subset of 158 cases. Each navigation action is defined as a canonical camera movement and is mapped to natural-language prompts, relative 6-DoF poses, or discrete keyboard commands depending on the paradigm, allowing fair cross-paradigm comparison. Semantic interactions are text-only and restricted to text-driven models, with configurations in Appendix B.1. For each case and model, we generate two samples and evaluate each twice, averaging the four scores. Appendices D.4 and D.7 report repeatability and confidence intervals from archived per-case scores.

5.2 Per-Dimension Results

Table 2 presents a unified view across all five dimensions. We analyze each dimension below:

Table 2: Main results on the WBench navigation split (158 navigation cases). Columns from left to right: text-driven, camera-controlled, and action-conditioned models. Event editing, subject action, and perspective switching are evaluated on cases containing the corresponding interactions and are reported only for text-driven models. All scores $\in [0, 100]$ and higher is better. Within each paradigm group, bold and underline mark the best and second best model, and *Average* is the group mean. "–" marks metrics that are not applicable.

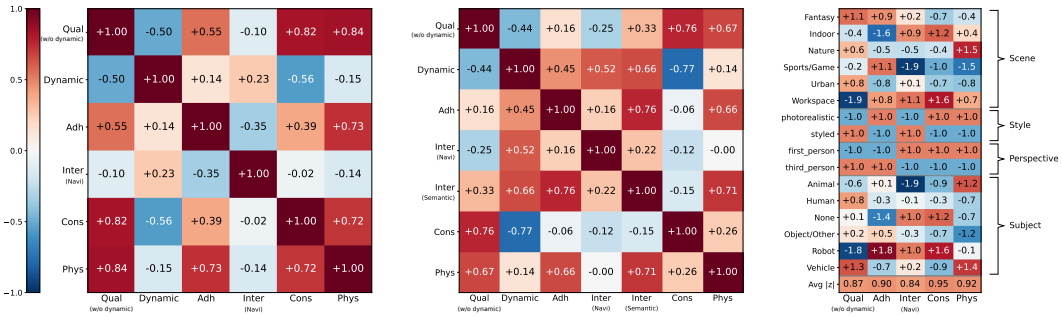
	Metrics	Seedance 1.5	Wan 2.7	Kling 3.0	YUME 1.5	HY-Video 1.5	LTX 2.3	LongCat-Video	Kairos 3.0	Cosmos 2.5	*Average*	LingBot-World	HY-World 1.5	Fantasy-World	InSpatio-World	Astra	*Average*	Happy Oyster	Matrix-Game 3.0	Genie 3	Matrix-Game 2.0	HY-GameCraft	Infinite-World	*Average*
Video Quality	Aesthetic	61.0	61.4	63.0	58.7	63.4	57.9	**66.5**	59.9	61.8	61.5	**66.9**	60.1	63.0	64.4	48.6	60.6	56.6	46.4	51.6	54.0	52.6	**58.7**	53.3
	Imaging	69.3	68.0	68.1	63.3	67.4	61.0	**69.6**	62.7	66.9	66.3	**67.9**	65.4	62.8	67.6	52.5	63.2	63.9	**70.0**	59.3	60.3	58.7	66.1	63.1
	Flickering	92.4	92.2	93.2	93.0	94.2	93.2	94.8	**95.4**	94.8	93.7	94.1	93.5	95.8	**96.0**	**96.0**	95.1	94.0	86.3	**95.0**	94.6	93.7	94.1	93.0
	Dynamic	99.4	**100.0**	97.5	96.8	73.9	98.1	45.9	70.1	49.0	81.2	66.2	**91.1**	49.0	26.1	79.6	62.4	94.2	**97.5**	92.4	94.9	96.8	82.8	93.1
	Smoothness	97.5	96.3	97.6	97.0	**98.7**	96.4	97.9	97.5	98.2	97.5	96.9	98.1	97.9	**98.8**	97.7	97.9	97.0	95.4	97.8	**98.2**	97.6	98.0	97.3
	HPSv3-Norm	73.0	71.1	69.1	57.0	68.0	56.1	**77.6**	58.5	66.5	66.3	**81.4**	60.5	65.8	76.1	28.0	62.4	58.3	57.1	55.2	41.0	38.3	**62.3**	52.0
	Average	**82.1**	81.5	81.4	77.6	77.6	77.1	75.4	74.0	72.9	77.7	**78.9**	78.1	72.4	71.5	67.1	73.6	**77.3**	75.5	75.2	73.8	73.0	77.0	75.3
Setting	Scene	71.6	88.3	**89.0**	53.1	77.5	81.3	53.1	52.2	72.4	70.9	51.6	**53.5**	52.4	51.7	43.4	50.5	57.4	48.9	**61.1**	49.4	50.6	54.0	53.6
	Subject	94.2	**94.6**	92.9	91.7	93.6	89.2	91.5	88.5	94.2	92.3	**93.6**	90.8	90.1	91.1	75.9	88.3	**91.1**	78.4	83.8	84.9	82.5	84.5	84.2
	Average	82.9	**91.4**	91.0	72.4	85.5	85.2	72.3	70.3	83.3	81.6	**72.6**	72.2	71.2	71.4	59.7	69.4	**74.2**	63.7	72.5	67.2	66.5	69.2	68.9
Interaction	Navigation	66.5	64.7	69.6	**72.0**	71.8	66.0	62.7	64.0	62.9	66.7	80.2	**87.5**	72.0	73.1	66.7	75.9	**84.7**	83.6	74.1	80.3	66.3	75.6	77.4
	Event Editing	80.4	**84.0**	81.4	57.8	63.8	53.0	50.4	46.8	48.2	62.9	–	–	–	–	–	–	–	–	–	–	–	–	–
	Subject Action	80.0	83.4	**85.6**	47.0	55.6	51.8	48.4	41.4	41.6	59.4	–	–	–	–	–	–	–	–	–	–	–	–	–
	Persp. Switching	45.0	**55.0**	**55.0**	16.7	27.6	25.0	18.3	13.3	20.0	30.7	–	–	–	–	–	–	–	–	–	–	–	–	–
	Average	68.0	71.8	**72.9**	48.4	54.7	49.0	44.9	41.4	43.2	54.9	–	–	–	–	–	–	–	–	–	–	–	–	–
Consistency	Background	89.6	89.4	92.3	90.3	92.1	88.3	**95.1**	91.1	92.3	91.2	**96.9**	92.7	94.2	95.0	85.3	92.8	**91.4**	85.7	90.7	86.9	86.5	88.8	88.3
	Spatial	72.7	71.0	75.2	71.5	79.2	70.2	**83.3**	76.8	78.1	75.3	92.7	90.6	80.6	**93.8**	64.7	84.5	77.7	**81.0**	79.9	64.5	60.5	74.9	73.1
	Gated Spatial	72.4	71.0	**75.1**	71.4	**75.1**	70.2	66.2	62.0	74.3	70.9	67.1	**84.9**	64.2	66.5	63.3	69.2	75.8	**80.4**	78.4	64.5	60.5	74.4	72.3
	Segment	96.2	92.4	93.0	**99.4**	**99.4**	75.8	**99.4**	94.3	94.3	93.8	99.4	**100.0**	**100.0**	**100.0**	86.6	97.2	96.2	89.8	93.6	21.0	99.4	**100.0**	83.3
	Perspective	70.5	78.2	76.8	48.0	**86.6**	69.8	81.5	76.3	84.3	74.7	**90.9**	62.5	79.8	72.5	30.0	67.1	**75.0**	13.3	54.5	29.2	17.9	33.8	37.3
	Subject	90.1	90.7	88.5	88.8	91.6	87.2	**93.4**	90.8	92.3	90.4	93.5	89.1	92.5	**94.4**	83.5	90.6	**91.5**	83.0	90.4	87.2	82.6	88.4	87.2
	Geometric	82.4	83.7	88.9	88.0	94.6	76.9	**95.4**	89.0	94.6	88.2	95.4	92.0	95.3	**97.3**	85.6	93.1	87.2	87.6	88.6	86.1	88.3	**94.3**	88.7
	Photometric	76.8	76.4	79.9	**83.3**	80.3	79.2	82.2	80.8	81.6	80.1	83.3	83.1	84.8	87.4	**87.5**	85.2	79.8	75.3	84.5	81.3	85.0	**85.1**	81.8
	Average	81.3	81.6	83.7	80.1	**87.4**	77.2	87.1	82.6	86.5	83.1	**89.9**	86.9	86.4	88.4	73.3	85.0	**84.3**	74.5	82.6	65.1	72.6	79.9	76.5
Physical	Causal Fidelity	76.0	**83.3**	78.0	72.7	75.0	74.0	76.0	62.7	74.7	74.7	**77.7**	74.0	74.0	67.3	48.3	68.3	69.3	64.7	**71.7**	59.3	68.3	67.0	66.7
	Visual Plausibility	60.7	60.3	60.7	57.7	59.7	55.7	**61.8**	58.0	60.1	59.4	**64.8**	58.6	59.7	63.1	54.6	60.2	57.6	54.0	**59.7**	55.0	56.5	57.2	56.7
	Average	68.3	**71.8**	69.3	65.2	67.3	64.8	68.9	60.4	67.4	67.0	**71.2**	66.3	66.8	65.2	51.5	64.2	63.5	59.4	**65.7**	57.1	62.4	62.1	61.7

Video Quality. Flickering and smoothness scores approach saturation across paradigms. Text-driven models lead narrowly, with Seedance 1.5 at 82.1 and Wan 2.7 at 81.5. LingBot-World (78.9) and Happy Oyster (77.3) also achieve competitive quality alongside their native control capabilities.

Setting Adherence. Wan 2.7 (91.4) and Kling 3.0 (91.0) substantially exceed Happy Oyster (74.2) and LingBot-World (72.6). The gap is concentrated in scene adherence rather than subject adherence, indicating uneven following of the specified world settings.

Interaction. Navigation favors native control interfaces. Action-conditioned (77.4) and camera-controlled (75.9) models exceed text-driven models (66.7) by 9 to 11 points. YUME 1.5 leads the text-driven group at 72.0. For semantic interactions, Kling 3.0 and Wan 2.7 lead event editing and subject action, while perspective switching remains hardest, averaging 30.7 across text-driven models.

Consistency. LingBot-World achieves the highest overall consistency (89.9), but consistency is multi-faceted. Camera-controlled models lead geometric consistency (93.1 *vs.* 88.2 for text-driven) thanks to explicit pose supervision, yet average only 67.1 on perspective consistency, underperforming text-driven models (74.7). As shown in Fig. 3 (a), dynamic degree is negatively correlated with consistency (r=−0.56), and camera-controlled models drop 15.3 points from spatial to gated spatial



(a) 6-Dim Correlation (20 models) (b) 7-Dim Correlation (9 text-driven models) (c) World Setting Z-score Deviation

Figure 3: Cross-dimension correlation and per-setting deviation analysis. **(a)** Pearson correlation among six dimensions (n=20 models, navi split). **(b)** seven-dimension correlation (n=9 text-conditioned models), with Interaction split into Navigation and Semantic. **(c)** per-setting Z-score deviation across five dimensions. Positive (red) = easier, negative (blue) = harder. Qual is the five-metric video-quality average excluding Dynamic Degree, whereas Table 2 reports all six metrics.

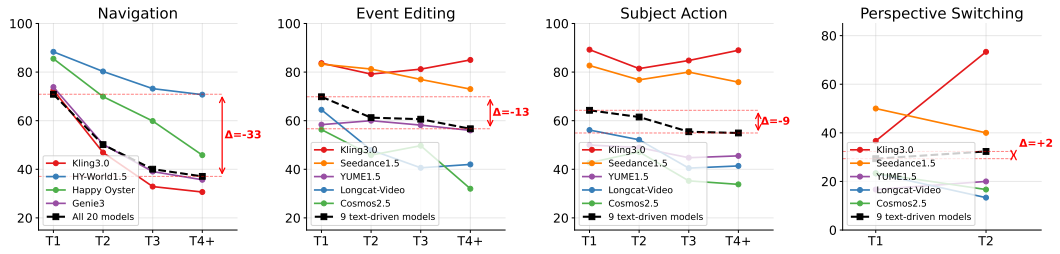


Figure 4: Per-turn performance degradation. T4+ aggregates all turns from the fourth turn onward.

consistency (*vs.* 4.5 for text-driven), confirming that some high scores are inflated by scene stasis rather than maintaining genuine consistency through active motion.

Physical. Text-driven models average 67.0 on the physical dimension, compared with 64.2 for camera-controlled and 61.7 for action-conditioned models. Wan 2.7 leads at 71.8, with a causal fidelity score of 83.3. LingBot-World leads camera-controlled models at 71.2 and also achieves high consistency at 89.9. These scores describe differences in the evaluated cohort without isolating the effects of training data or control-specific training.

5.3 Cross-Dimension Analysis

No model or paradigm dominates uniformly. Figure 3 examines associations across 20 models, including nine text-driven models. These exploratory correlations may reflect control paradigms or shared evaluator backbones and do not establish causality or capability independence.

Navigation Exhibits Distinct Cross-Dimension Patterns. Across the evaluated models, navigation has weak correlations with video quality (r=−0.10), consistency (r=−0.02), and physical scores (r=−0.14). Physical scores correlate more strongly with video quality (r=0.84) and consistency (r=0.72). Within text-driven models, semantic interaction scores correlate more strongly with setting adherence (r=0.76) than with navigation (r=0.22). These patterns suggest that rankings differ across dimensions in this cohort, without identifying the training or architectural factors responsible.

Navigation and Perspective Consistency Rankings Differ. Navigation also correlates weakly with perspective consistency. Among the 11 camera-controlled and action-conditioned models, HY-World 1.5 ranks first in navigation and fifth in perspective consistency, while Matrix-Game 3.0 ranks third and eleventh. The differing rankings motivate measuring both capabilities.

Open-Source Models Are Competitive. HY-World 1.5 leads navigation (87.5), and LingBot-World leads consistency (89.9). Both are open source, while closed-source Happy Oyster leads action-conditioned navigation (84.7). The leaders span release categories and control interfaces.

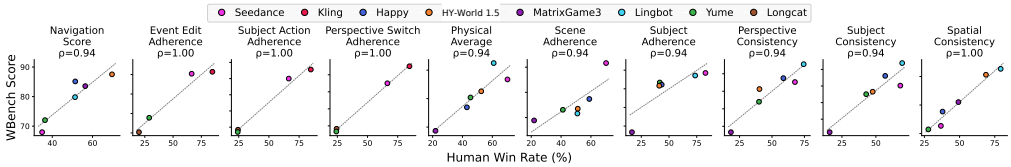


Figure 5: Spearman ρ between per-model human win rates (x-axis) and automated WBench scores (y-axis) across ten evaluation aspects. All aspects achieve $\rho \geq 0.94$, with four reaching $\rho = 1.00$.

World Settings Induce Structured Difficulty. As shown in Fig. 3 (c), first-person perspective makes navigation easier (z=+1.0) thanks to the direct action-to-camera mapping, while third-person adds geometric complexity from joint subject-camera control. Sports/game scenes are hardest for navigation (z=−1.9) and animal subjects likewise (z=−1.9) due to fast dynamics and complex non-rigid motion, while workspace scenes (z=+1.1) and robot subjects (z=+1.0) are easiest given their static geometry and rigid-body motion. Benchmark difficulty is thus governed by control-mapping complexity, scene dynamism, and subject rigidity.

Navigation Breaks Down Over Turns. In Fig. 4, navigation drops 33 points from turn 1 to turn 4+, compared with 13 for event editing and 9 for subject action. HY-World 1.5 degrades less than Kling 3.0 on navigation. On event editing and subject action, Kling 3.0 remains relatively stable, while Seedance 1.5 shows smaller declines than the text-driven average. Perspective switching changes little (+2), against a low average score of 30.7. These patterns indicate uneven multi-turn robustness. Appendix D.3 further evaluates two camera-controlled models over 2 to 20 turns in two navigation scenarios.

5.4 Human Preference Alignment

Following the human alignment protocol of VBench (Huang et al., 2024), we validate automated metrics by computing the Spearman rank correlation between per-model human win rates and WBench scores. We recruit 400 crowdsourced annotators to perform blind pairwise comparisons across ten evaluation aspects, using model subsets appropriate to each aspect. For each comparison, annotators select *A wins*, *B wins*, or *Tie* based on dimension-specific criteria. We aggregate preferences into per-model win rates (Tie counts as 0.5 for each side) and correlate against the corresponding automated score. As shown in Fig. 5, all ten aspects achieve Spearman $\rho \geq 0.94$, with four (event editing, subject action, perspective switching, and spatial consistency) reaching $\rho = 1.00$. Details are provided in Appendix D.2. We report case-level human–metric agreement in Appendix D.6.

6 Conclusion

We introduce WBench, a benchmark for evaluating interactive world models across five complementary dimensions through explicit world settings and multi-turn interactions. Experiments on 20 models reveal complementary strengths, differing navigation and consistency rankings, and uneven multi-turn robustness. Physical scores correlate more strongly with rendering quality than with navigation. These findings motivate joint evaluation across all five dimensions. By separating world settings from interaction sequences, WBench makes these differences explicit across scenes, perspectives, and control interfaces. Its fine-grained profiles help identify which capabilities fail under a given interaction, rather than treating visual quality or a single aggregate score as a substitute for world-model competence. We therefore encourage reporting dimension-level results alongside overall comparisons when evaluating progress in interactive generation.

Limitations. **(1) Scale and coverage.** WBench remains limited in scale and coverage, motivating expansion to more test cases and more diverse world settings and interactions. **(2) Interaction duration.** Current evaluations cover relatively short interaction sequences. Extending evaluation to sustained interactions lasting minutes or hours is an important direction for future work. **(3) Evaluator reliability.** VLM-based physical assessment may overlook subtle violations, motivating further human validation across styles and control interfaces. **(4) Interaction latency.** We report generation efficiency but do not evaluate the delay between a control input and its visible response. Measuring this latency is an important future direction because responsiveness is essential to real-time interaction. **(5) Generalization.** The observed correlations describe the evaluated model cohort and do not establish causality or generalization to unseen model families.

AI USE STATEMENT

We used AI tools to assist with information retrieval, drafting and polishing manuscript sections, interpreting reported results, and LaTeX formatting. Nano Banana 2 and GPT-Image-1.5 generated part of the benchmark's initial imagery, which underwent manual review as described in Section 3.1. Vision-language models also support automated evaluation, as detailed in Appendix C. The authors take responsibility for the final manuscript and all AI-assisted content.

ETHICS STATEMENT

WBench evaluates interactive video generation, and its scores should not be interpreted as evidence of safe deployment or reliable real-world physical behavior. The human-preference study involves a pool of 400 annotators, with task instructions, training, and quality-control procedures described in Appendix D.2. Human preferences and model-based judgments can reflect subjective or systematic biases, and the sampled scenarios do not represent all populations or deployment contexts. Initial imagery combines generated, web-collected, and manually captured images, so source-specific privacy and usage rights require consideration when reusing the data. Appendix D.8 documents the release scope, including annotations that remain internal. Improved video generation can support simulation and research but can also enable misleading synthetic media, which remains a relevant downstream risk for work using this benchmark.

REPRODUCIBILITY STATEMENT

We document dataset construction and case selection in Section 3.1 and Appendix D.5, model configurations and inference settings in Appendix B.1, and metric definitions and scoring procedures in Appendix C. Human evaluation protocols, repeated-evaluation results, and confidence-interval analyses are provided in Appendices D.2, D.4 and D.7. Appendix D.8 summarizes resource availability and release limitations. In particular, the physical scorer's training and held-out annotations and the human-preference annotations are not released. The available scorer supports inference, but does not enable independent reproduction of its training or held-out validation. Reproductions should record the evaluated checkpoint or API version, interface, sampling settings, and evaluation configuration. Changes in hosted models or scoring tools may affect results, so comparisons with archived scores should distinguish implementation differences from changes in model capability.

REFERENCES

ACE Robotics. Kairos 3.0-4b: Real-time generative world model for embodied intelligence. https://github.com/kairos-agi/kairos-sensenova/tree/main, 2026.

Alibaba Cloud. Wan 2.7: Image-to-video api. Alibaba Cloud Model Studio documentation, 2026. URL https://www.alibabacloud.com/help/en/model-studio/image-to-video-general-api-reference. Accessed September 20, 2026.

Alibaba Token Hub. Happy Oyster: An open-ended world model for real-time world creation and interaction. https://happyoyster.cn/, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang,

Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/, 2025.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 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.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Zefan Cai, Haoyi Qiu, Tianyi Ma, Haozhe Zhao, Gengze Zhou, Kung-Hsiang Huang, Parisa Kordjamshidi, Minjia Zhang, Wen Xiao, Jiuxiang Gu, et al. Mmgr: Multi-modal generative reasoning. *arXiv preprint arXiv:2512.14691*, 2025.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. *arXiv preprint arXiv:2509.21657*, 2025.

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.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024.

Google. Nano banana 2. https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/, 2025.

Google DeepMind. Veo 3: State-of-the-art video generation with audio. https://deepmind.google/models/veo/, 2025.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. "phyworldbench": A comprehensive evaluation of physical realism in text-to-video models, 2026. URL https://arxiv.org/abs/2507.13428.

Jinwei Gu. Cosmos world foundation models for physical ai. In *Proceedings of the 3rd International Workshop on Rich Media With Generative AI*, pp. 39–39, 2025.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, et al. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18858–18868, 2025.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

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*, 2025.

Team HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. *arXiv preprint*, 2025.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

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.

Kuaishou Technology. Kling 3.0 pro. https://klingai.com, 2025.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

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.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2(3):6, 2025b.

Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9801–9810, 2023.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10486–10496, 2025c.

Ao Liang, Lingdong Kong, Tianyi Yan, Hongsi Liu, Wesley Yang, Ziqi Huang, Wei Yin, Jialong Zuo, Yixuan Hu, Dekai Zhu, et al. Worldlens: Full-spectrum evaluations of driving world models in real world. *arXiv preprint arXiv:2512.10958*, 2025.

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.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22139–22149, 2024.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36:62352–62387, 2023.

Yiting Lu, Wei Luo, Peiyan Tu, Haoran Li, Hanxin Zhu, Zihao Yu, Xingrui Wang, Xinyi Chen, Xinge Peng, Xin Li, et al. 4dworldbench: A comprehensive evaluation framework for 3d/4d world generation models. *arXiv preprint arXiv:2511.19836*, 2025.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score, 2025. URL https://arxiv.org/abs/2508.03789.

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.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

OpenAI. GPT-Image-1.5. https://openai.com/zh-Hans-CN/index/new-chatgpt-images-is-here/, 2025a.

OpenAI. Sora 2. https://openai.com/zh-Hans-CN/index/sora-2/, 2025b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, et al. Genie 2: A large-scale foundation world model. https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/, 2024.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

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.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

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*, 2026.

Shengshu Technology. Vidu q3 pro. https://www.vidu.com, 2025.

Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11218–11221, 2024.

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, et al. Hunyuan-gamecraft-2: Instruction-following interactive game world model. *arXiv preprint arXiv:2511.23429*, 2025.

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*, 2026a.

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.

PAN Team, Qiyue Gao, Kun Zhou, Jiannan Xiang, Zihan Liu, Dequan Yang, Junrong Chen, Arif Ahmad, Cong Zeng, Ganesh Bannur, Xinqi Huang, Zheqi Liu, Yi Gu, Yichi Yang, Guangyi Liu, Zhiting Hu, Zhengzhong Liu, and Eric Xing. World reasoning arena, 2026b.

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.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Pranay Boreddy, Shruti Satya Narayana, Yunhao Ba, Alex Wong, Celso M de Melo, and Achuta Kadambi. Worldbench: Disambiguating physics for diagnostic evaluation of world models. *arXiv preprint arXiv:2601.21282*, 2026.

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.

Zile Wang, Zexiang Liu, Jaixing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

World Labs. Marble: A multimodal world model. https://www.worldlabs.ai/blog/marble-world-model, 2025.

Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, et al. Omni-worldbench: Towards a comprehensive interaction-centric evaluation for world models. *arXiv preprint arXiv:2603.22212*, 2026a.

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*, 2026b.

Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Kaipeng Zhang, and Yongtao Ge. Worldmark: A unified benchmark suite for interactive video world models. *arXiv preprint arXiv:2604.21686*, 2026.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

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.

Hu Yue, 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.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M Patel, Paul Pu Liang, et al. World-in-world: World models in a closed-loop world. *arXiv preprint arXiv:2510.18135*, 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.

Yang Zhou, Hao Shao, Letian Wang, Zhuofan Zong, Hongsheng Li, and Steven L Waslander. Drivinggen: A comprehensive benchmark for generative video world models in autonomous driving. *arXiv preprint arXiv:2601.01528*, 2026.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: Learning interactive real-robot action simulators. *arXiv preprint arXiv:2406.14540*, 1(2):3, 2024.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Astra: General interactive world model with autoregressive denoising. *arXiv preprint arXiv:2512.08931*, 2025.

A Additional Dataset Statistics and Analysis

A.1 Extended Benchmark Comparison

Table 3: Extended comparison of WBench with all surveyed benchmarks. FPP/TPP denotes first-/third-person perspective. Navi, SA, EE, and PS denote navigation, subject action, event editing, and perspective switching. Qual, Adh, Inter, Cons, and Phys denote video quality, setting adherence, interaction adherence, consistency, and physics compliance.

Benchmark	Venue	Persp.		Interaction Type				Evaluation Dimension					Scale	
		FPP	TPP	Navi	SA	EE	PS	Qual	Adh	Inter	Cons	Phys	Cases	Turns
VBench (Huang et al., 2024)	[CVPR'24]	-	-	✗	✗	✗	✗	✓	✗	✗	✗	✗	946	946
EvalCrafter (Liu et al., 2024)	[CVPR'24]	-	-	✗	✗	✗	✗	✓	✗	✗	✗	✗	700	700
WorldSimBench (Qin et al., 2024)	[arXiv'24]	✓	✗	✓	✓	✗	✗	✗	✗	✓	✓	✗	35k	35k
PhyGenBench (Meng et al., 2024)	[ICML'25]	-	-	✗	✗	✗	✗	✓	✗	✗	✗	✓	160	160
VBench++ (Huang et al., 2025)	[TPAMI'25]	-	-	✗	✗	✗	✗	✓	✗	✗	✗	✓	1,260	1,260
Video-Bench (Han et al., 2025)	[arXiv'25]	-	-	✗	✗	✗	✗	✓	✗	✗	✗	✗	419	419
VideoPhy-2 (Bansal et al., 2025)	[arXiv'25]	-	-	✗	✗	✗	✗	✓	✗	✗	✗	✓	590	590
WorldScore (Duan et al., 2025)	[ICCV'25]	✓	✗	✓	✗	✓	✗	✓	✓	✗	✗	✗	3,000	3,000
WorldModelBench (Li et al., 2025a)	[NeurIPS'25]	✓	-	✓	✓	✗	✗	✗	✓	✓	✗	✓	350	350
4DWorldBench (Lu et al., 2025)	[arXiv'25]	-	-	✗	✗	✗	✗	✓	✗	✗	✗	✓	333	333
InterBench (Tang et al., 2025)	[arXiv'25]	✓	✓	✗	✓	✓	✗	✓	✗	✓	✓	✓	920	920
MIND (Ye et al., 2026)	[arXiv'26]	✓	-	✓	✗	✗	✗	✓	✗	✓	✓	✗	250	–
WorldArena (Shang et al., 2026)	[arXiv'26]	✓	✗	✓	✓	✗	✗	✓	✗	✓	✓	✓	500	500
Omni-WorldBench (Wu et al., 2026a)	[arXiv'26]	✓	-	✓	✓	✓	✗	✓	✗	✓	✓	✓	1,068	1,068
WR-Arena (Team et al., 2026b)	[arXiv'26]	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	✓	62	$\leq$558
WorldLens (Liang et al., 2025)	[CVPR'26]	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	✓	–	–
World-in-World (Zhang et al., 2025)	[ICLR'26]	✓	✗	✓	✓	✗	✗	✓	✗	✓	✗	✗	1,079	–
WorldMark (Xu et al., 2026)	[arXiv'26]	✓	✓	✓	✗	✗	✗	✓	✗	✓	✓	✗	500	–
WBench (Ours)	–	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	289	1,058

Table 3 summarizes the full set of surveyed benchmarks. Representative examples span quality-oriented suites such as VBench (Huang et al., 2024), which dissects text-to-video quality into 16 hierarchical dimensions with tailored prompts, and EvalCrafter (Liu et al., 2024), which scores around 700 prompts with 17 objective metrics calibrated against user preferences. Physics-focused work includes PhyGenBench (Meng et al., 2024), with 160 prompts across 27 physical laws in four domains, and VideoPhy-2 (Bansal et al., 2025), which targets action-centric physics over 200 real-world actions. WorldScore (Duan et al., 2025) casts world generation as a sequence of next-scene tasks along explicit camera trajectories on 3,000 cases, while WorldModelBench (Li et al., 2025a) evaluates instruction following and physics adherence using 67K human labels in robotics and driving settings. Among interactive efforts, WorldSimBench (Qin et al., 2024) pairs perceptual evaluation with a manipulative evaluation that checks whether generated videos translate into correct control signals, World-in-World (Zhang et al., 2025) builds a closed-loop platform with a unified planner and standardized action API that reports task success as the primary metric, and WorldMark (Xu et al., 2026) introduces a WASD-style action-mapping layer to enable apples-to-apples comparison across interactive video world models under identical scenes and trajectories.

A.2 Dataset Gallery

This section provides qualitative showcases of WBench along the taxonomy axes introduced in Section 3.1. We show the thumbnails of all the initial frames of the WBench in Fig. 6.



Figure 6: Thumbnail gallery of all cases in WBench.

A.2.1 SCENE AND STYLE GALLERY

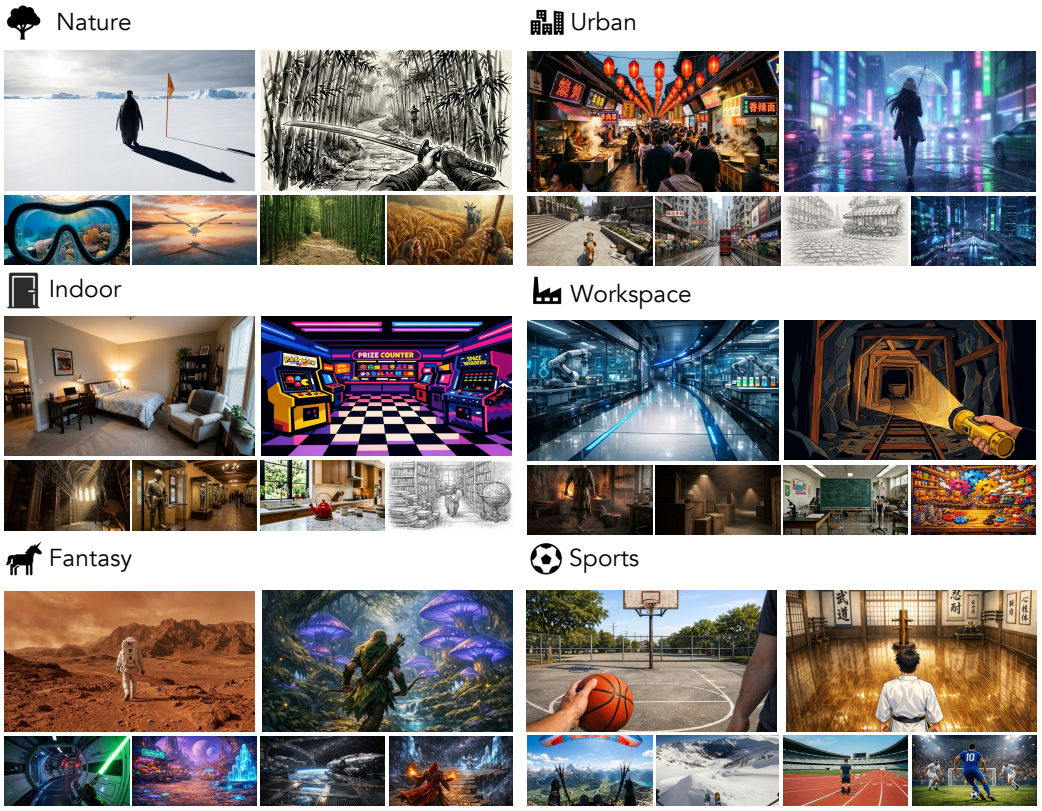


Figure 7: Scene and style coverage. Two categories per row are presented, each shown as a photorealistic/stylized pair that shares the same underlying scene specification. The six categories cover nature, urban, indoor, workspace, fantasy, and sports/game.

Fig. 7 displays two photorealistic and stylized pairs per row, covering all six scene categories in a compact three-row layout. The six categories, nature, urban, indoor, workspace, fantasy, and sports or game, span both naturalistic settings (outdoor landscapes, cityscapes, living spaces) and more constructed or speculative environments (office or workshop, surreal or magical worlds, and curated activity spaces), and each category is shown through a photorealistic and a stylized rendering of the same underlying scene specification, so that readers can compare how the rendering pipeline preserves the scene semantics while varying the visual appearance. Fig. 8 further enlarges the stylized rendering to show the full range of visual styles supported by WBench, including realistic, anime, cartoon, oil painting, ink wash, flat, and pencil sketch, which together cover photographic, illustrative, painterly, and sketch-like appearances used by practical content-creation workflows.

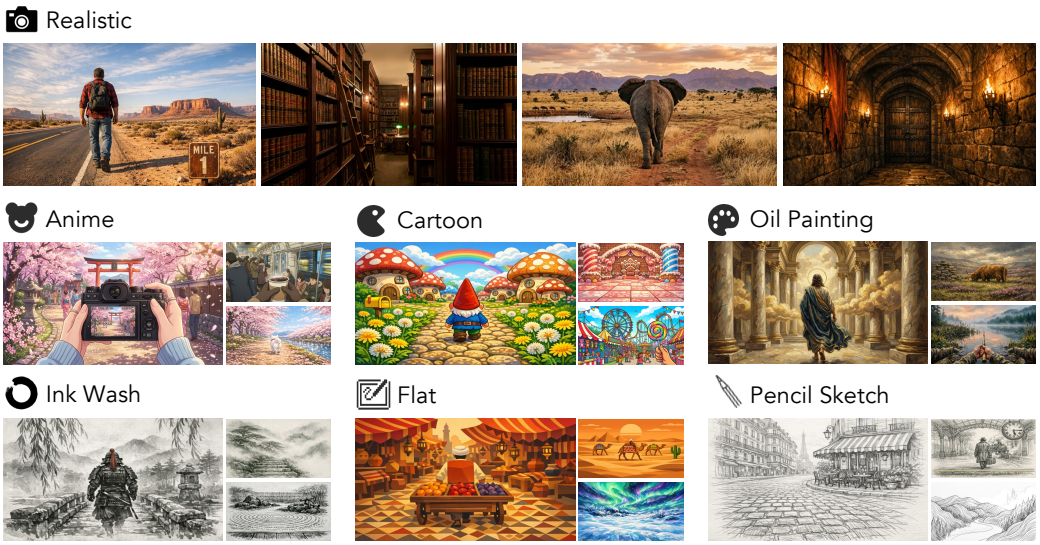


Figure 8: Style gallery. Representative initial frames spanning the rendering styles covered by WBench: realistic, anime, cartoon, oil painting, ink wash, flat, and pencil sketch.

A.2.2 Perspective and Subject Gallery

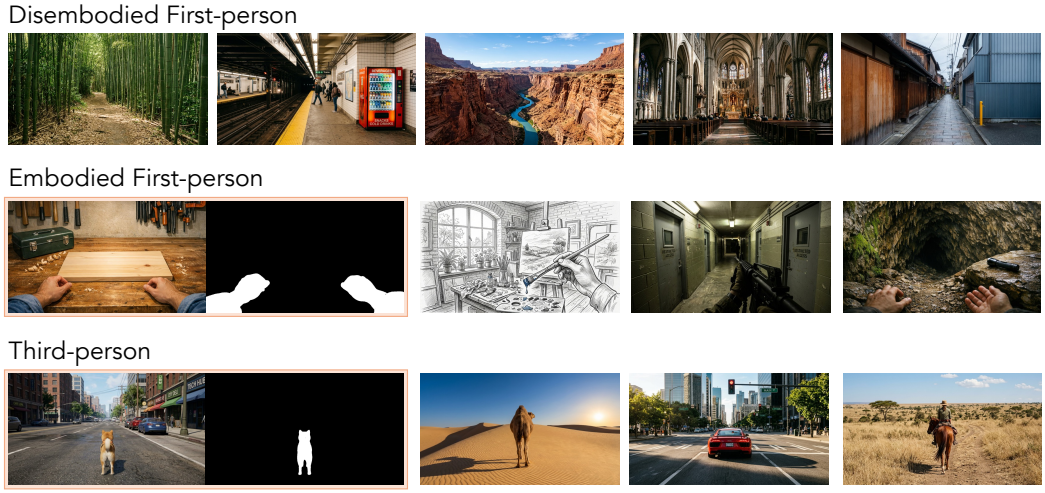


Figure 9: Perspective gallery. Cases are grouped into three rows by perspective type: disembodied first-person with no visible agent, embodied first-person with a visible body part such as hands or weapon, and third-person with the controlled subject in view.

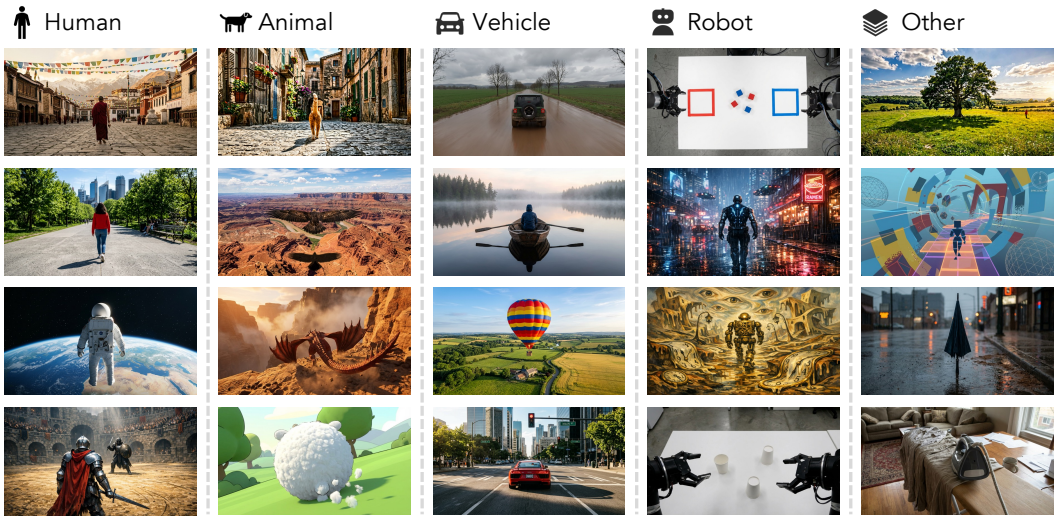


Figure 10: Subject gallery. One column per subject category, covering human, animal, vehicle, robot, and other objects. Each column lists multiple representative cases that span different scenes and styles.

Fig. 9 separates the perspective axis into three rows. The top row shows *disembodied first-person* cases, where the camera is attached to an implicit agent and no body part is visible, so that perspective is defined purely by the camera's position and motion. The middle row shows *embodied first-person* cases, where a body part (typically hands, held tools, or a mounted weapon) is visible at the bottom of the frame and grounds the perspective in a specific agent. The bottom row shows *third-person* cases, where the controlled subject is visible inside the frame and the camera follows it externally. This three-way split matters because embodiment changes how navigation and subject-action prompts should be interpreted: in disembodied first-person, instructions like "walk forward" collapse to camera translation, whereas in embodied first- and third-person settings the same instruction must be grounded in a specific agent's motion. Fig. 10 further enumerates the five controllable-subject categories on the third-person axis, namely human, animal, vehicle, robot, and other objects, each illustrated by multiple representative cases that vary scene and style, so that subject diversity is shown to be orthogonal to scene and style diversity rather than confounded with them.

A.2.3 Subject Action and Event Editing Examples

Unlike scenes, styles, perspectives, and subjects, which are naturally illustrated by static initial frames, subject action and event editing are fundamentally temporal phenomena and are therefore hard to render compactly as image galleries. In place of such galleries, Table 4 lists condensed example phrases abstracted from the released WBench cases, so that reviewers can directly read off the semantic breadth each sub-type aims to cover. On the Subject Action side, the five sub-types range from fine-grained hand-object interaction (manipulation, tool use) to whole-body movement and agent–environment physics (locomotion, combat), and further to socially grounded behaviors that carry communicative intent rather than physical effect (gestural interaction). On the Event Editing side, the six sub-types progressively move outward from the subject: appearance or state changes that attach to existing objects, environmental changes and natural phenomena that reshape global scene attributes, NPC motion that introduces new dynamic agents, mechanical transitions that reconfigure scene geometry, and physical effects that trigger irreversible state changes such as explosions or collapses. Together the eleven sub-types are designed to stress-test different aspects of world-model behavior, from local controllability over a subject to global responsiveness of the surrounding world.

A.2.4 Perspective Switching Prompts Gallery

Fig. 11 illustrates the three perspective-switching sub-types in a single row. Each panel shows the source perspective as the initial frame, with the released switching prompt baked in to describe the intended target perspective, so that the interaction shown is exactly the one the model is asked to

Table 4: Condensed example phrases for each Subject Action and Event Editing sub-type, abstracted from released WBench cases. Each cell lists three semicolon-separated phrases that cover the typical semantic range of the subtype.

Sub-type	**Example phrases**
Subject Action	
Manipulation	open a box lid; knead dough; sweep leaves into a pile.
Tool use	saw a plank along a line; light a candle with a match; trim branches with shears.
Locomotion	glide on ice; dive and descend underwater; squeeze through a narrow cave.
Combat	downward sword slash; straight punch at a dummy; axe strike on a door.
Gestural interaction	expressive hand-gesture speech; kneel before a herald; pat a horse.
Event Editing	
Environment change	rain begins; sunset turns to night; fog rolls in.
Appearance state change	runes glow cyan; armor visor lights up; window fogs from steam.
NPC motion	a passer-by walks by; a fish school drifts past; a dog dashes out.
Mechanical transition	drawbridge lowers; gate slides up; elevator doors open.
Physical effect	canister explodes; building collapses with dust; ice cracks and water floods in.
Natural phenomenon	volcano erupts; sandstorm approaches; hot-air balloons rise at sunset.

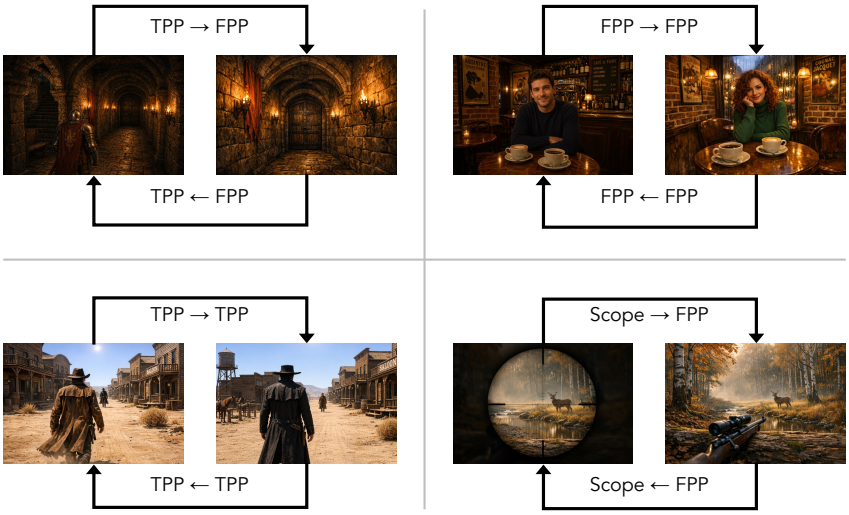


Figure 11: Perspective-switching taxonomy showcase. One representative case per sub-type, covering same-subject switches, multi-subject switches, and scope-mode transitions. The source perspective and the switching prompt are rendered inside each frame.

execute at evaluation time. *Same-subject* switches change the camera relative to the same controlled subject, for example from a third-person follow shot to an embodied first-person perspective of the same character, probing whether the model can re-anchor the camera without losing identity or pose. *Multi-subject* switches hand control over from one subject to another within the same scene, so that the new perspective is attached to a different agent while the surrounding scene must remain consistent. *Scope mode* transitions change the spatial scope of the perspective itself, for example zooming from a narrow first-person perspective out to a broader third-person shot or vice versa, testing whether the model maintains a coherent world across scale changes. The three sub-types together cover the principal ways a user can reconfigure perspective at run time without regenerating the world from scratch.

A.3 Navigation Design and Distribution

Navigation is the most common interaction type in WBench and receives the most systematic treatment. This section details the action definition, distribution statistics, and trajectory design.

Action definition. WBench adopts a WASD plus arrow-key scheme for discrete navigation control. As shown in Fig. 12 and Table 5, the same key triggers different physical motions depending on the perspective. Under first-person perspective, W/S/A/D translate the camera forward, backward, left, and right, while arrow keys rotate the viewpoint. Under third-person perspective, W/S/A/D move the subject, while arrow keys orbit the camera around the subject. This mode-dependent mapping mirrors how game engines handle first- and third-person control, and ensures that each key carries a clear, unambiguous spatial semantics for evaluation.

Atomic action distribution. Fig. 13 shows the distribution of 644 atomic actions across 601 navigation turns, with compound actions counted once for each constituent action. Translational actions (W/S/A/D) account for 63.2% and rotational actions (arrow keys) for 36.8%. Forward motion (W, 30%) is the most frequent translational action, followed by backward (S, 12%), leftward (A, 12%), and rightward (D, 10%) motion. Rotational actions are distributed across left (12%), right (15%), up (5%), and down (5%). Percentages for individual actions are rounded to the nearest integer.

Trajectory type design. We categorize navigation sequences into six trajectory types, each targeting different aspects of spatial understanding. Table 6 lists representative action sequences for each type. **Round-trip** trajectories move forward and then retrace the path, directly testing whether the model can maintain spatial consistency when revisiting previously seen regions. **Progressive** trajectories combine multiple distinct directions without returning, requiring the model to coherently extend the scene into new areas. **Repeat** trajectories apply the same action consecutively, testing whether the model sustains smooth and consistent motion over extended sequences. **L-shape** trajectories execute a single sharp turn, probing the model's ability to handle abrupt direction changes while preserving scene geometry. **Loop** trajectories traverse a closed path that returns to the starting area via a different route, demanding global spatial coherence across the entire sequence. **Zigzag** trajectories alternate between opposing directions, challenging the model to maintain stable scene structure under rapid, repeated directional switches. Among these, round-trip trajectories are the most frequent (32%) as they directly enable spatial consistency evaluation through revisitation, followed by Progressive (22%), Repeat (15%), and Loop (13%). L-Shape and Zigzag make up the remaining.

Overall navigation distribution. Figure 13 summarizes navigation coverage through atomic action frequencies, trajectory types, and the proportions of pure translation, pure rotation, and mixed navigation.

Table 5: Navigation action semantics under different perspectives. The same key triggers different physical motions depending on whether the case is first-person or third-person.

Type	Key	First-Person	Third-Person
Translation	W	Camera pushes forward	Subject walks forward
	S	Camera pulls backward	Subject steps backward
	A	Camera strafes left	Subject moves left
	D	Camera strafes right	Subject moves right
Rotation	←	View turns left	Camera orbits left
	→	View turns right	Camera orbits right
	↑	View tilts up	Camera elevates
	↓	View tilts down	Camera descends

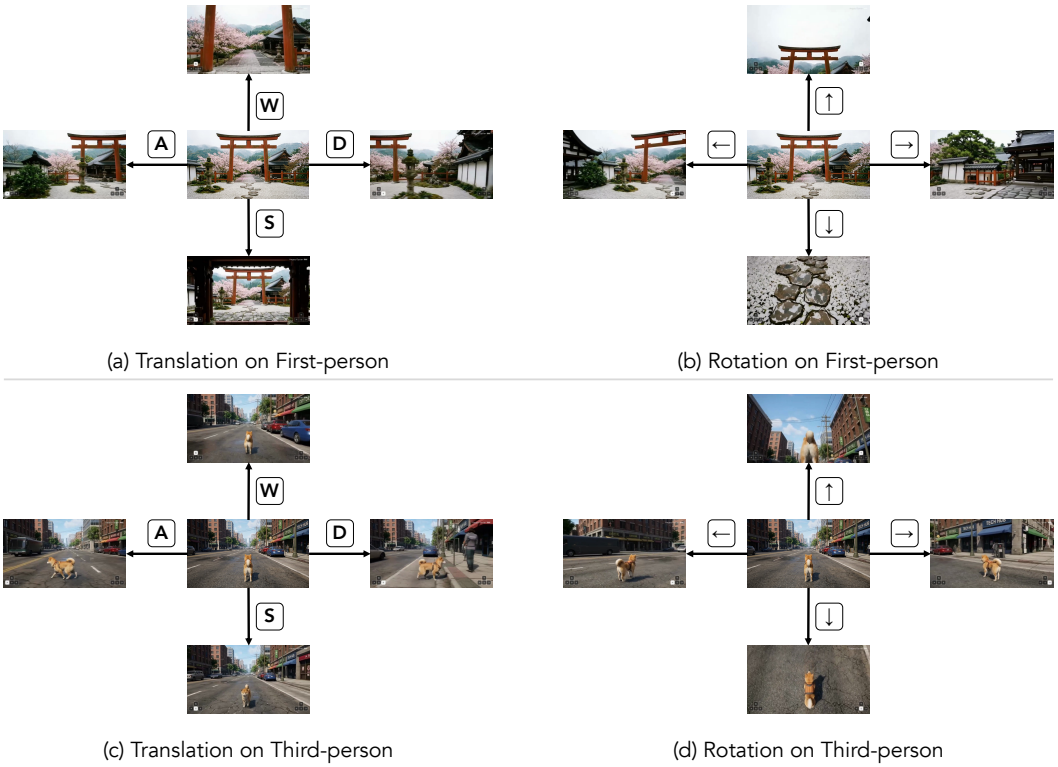


Figure 12: Navigation action definition. Illustration of how each WASD and arrow key maps to a physical motion under first-person and third-person perspectives. Frames following the navigation definition panel are from Happy Oyster (Alibaba Token Hub, 2026).

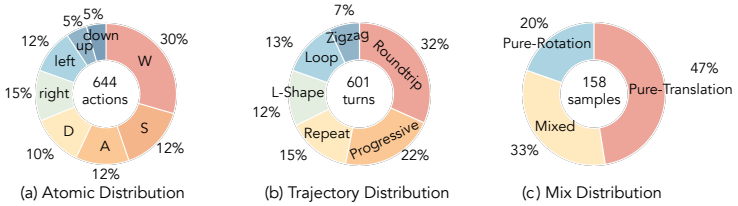


Figure 13: Navigation distributions. **(a)** Atomic action distribution over 644 actions. **(b)** Trajectory type distribution over 601 navigation turns. **(c)** Distribution of pure translation, pure rotation, and mixed navigation across 158 samples.

Table 6: Trajectory type examples in navigation test cases.

Trajectory Type	Examples
Round-trip	[A, A, D, D], [left, left, right, right], [W, W, W, S, S, S]
Progressive	[W, right, right, left], [W, A, W, D], [W, W, left, left, right, right]
Repeat	[W, W, W, W], [W, W, W]
L-shape	[W, W, A, A]
Loop	[W, A, S, D], [A, A, W, S, D, D]
Zigzag	[right, W, right, W], [W+A, W+D, W+A, W+D], [up, down, up, down]

B Evaluated Models

B.1 Per-Model Configuration

Table 7: Detailed overview of all 20 evaluated models. "Params" = parameter count (— if undisclosed). "Frames" = number of frames generated per interaction turn (RT = real-time streaming). "Inference : Video" is the ratio of wall-clock generation time to the duration of the produced video, lower is faster.

Model	Access	Architecture	Params	Resolution	Frames	Inference : Video	Hardware / Note
Text-driven models							
Seedance 1.5 (Seedance et al., 2025)	API	-	—	1280×720	120	—	API
Wan 2.7 (Alibaba Cloud, 2026)	API	-	—	1920×1080	80	—	API
Kling 3.0 (Kuaishou Technology, 2025)	API	-	—	1280×720	120	—	API
YUME 1.5 (Mao et al., 2025a)	Open	DiT	5B	1280×704	58	1.7	A100 80G
HY-Video 1.5 (Kong et al., 2024)	Open	DiT	8B	848×480	120	—	A100 80G
LTX 2.3 (HaCohen et al., 2024)	Open	DiT	22B	1536×1024	120	—	A100 80G
LongCat (Team et al., 2025)	Open	DiT	14B	832×480	48	621.8	A100 80G
Kairos 3.0 (ACE Robotics, 2026)	Open	DiT	4B	832×480	96	—	A100 80G
Cosmos 2.5 (Gu, 2025)	Open	DiT	2B	1280×704	93	—	A100 80G
Camera-controlled models							
LingBot-World (Team et al., 2026c)	Open	Multi-stage DiT	14B	1280×704	32	268.8	A100 80G
HY-World 1.5 (HunyuanWorld, 2025)	Open	Streaming DiT	8B	832×480	96	18.8	A100 80G
Fantasy-World (Dai et al., 2025)	Open	DiT + 3D head	14B	592×336	80	2320	A100 80G
InSpatio-World (Team et al., 2026a)	Open	V2V AR DiT	1.3B	832×480	96	—	A100 80G
Astra (Zhu et al., 2025)	Open	AR DiT	2B	832×480	80	30.0	A100 80G
Action-conditioned models							
Happy Oyster (Alibaba Token Hub, 2026)	Closed	—	—	1280×720	RT	—	Web interface
Matrix-Game 3.0 (Wang et al., 2026)	Open	AR Diffusion	6B	1280×704	70	0.5	A100 80G
Genie 3 (Ball et al., 2025)	Closed	—	—	1280×720	RT	—	Web interface
Matrix-Game 2.0 (He et al., 2025)	Open	AR Diffusion	2B	640×352	48	5.6	A100 80G
HY-GameCraft (Li et al., 2025b)	Open	Hist-Cond DiT	13B	832×480	132	2.5	A100 80G
Infinite-World (Wu et al., 2026b)	Open	Hierarchical DiT	1.5B	896×448	160	49.2	A100 80G

Table 7 covers the 20 models of Table 2.

B.1.1 Text-driven Models

Since these models do not natively accept action signals, we adopt an iterative image-to-video (I2V) protocol: for each interaction turn, we extract the last frame of the previous clip, construct a text prompt that combines the world-setting description with the current interaction instruction, and generate the next video segment.

Seedance 1.5 (Seedance et al., 2025). Seedance 1.5 is ByteDance's commercial video generation model, accessed through a closed-source API. It supports image-conditioned generation at up to 1280×720 resolution and 120 frames per clip. In WBench, it is evaluated under the iterative I2V protocol, with interaction instructions injected as text prompts.

Wan 2.7 (Alibaba Cloud, 2026). Wan 2.7 is Alibaba's closed-source video generation model, accessed through an API. In WBench, we evaluate it with 80 frames per clip under the iterative I2V protocol.

Kling 3.0 (Kuaishou Technology, 2025). Kling 3.0 is Kuaishou's commercial video generation model, accessed through a closed-source API. It supports image-conditioned generation with interaction instructions provided as natural-language prompts. In WBench, it is evaluated under the iterative I2V protocol.

YUME 1.5 (Mao et al., 2025a). YUME 1.5 is the first fully open-source interactive world model to support the combined capabilities of text-to-world generation, image-to-world initialization, and text-driven event editing, all within a single Diffusion Transformer. For WBench, we use the released `Yume-5B-720P` checkpoint, which contains approximately 5.2B parameters. Unlike keyboard-driven world models, YUME 1.5 uses natural language to specify both the scene content and the desired interaction, making it the only IWM in this evaluation that shares the same text-based control interface as the general I2V models. Its training combines large-scale text-video paired data with

curated interactive gameplay footage, enabling the model to ground language descriptions in plausible world dynamics. Each turn is assembled by concatenating fixed-size 29-frame iterations, with the number of iterations per turn chosen to best approximate the target turn duration.

HunyuanVideo 1.5 (Kong et al., 2024). HunyuanVideo 1.5 is Tencent's open-source video generation model based on a causal 3D full-attention DiT, pre-trained on a large-scale video and image corpus. The released 480p I2V checkpoint contains roughly 8.3B transformer parameters. Its architecture jointly models spatial and temporal attention without factorized approximations, enabling high visual fidelity and strong temporal coherence across long clips. The model supports image-conditioned generation at 848×480 and 120 frames, making it well-suited for multi-turn evaluation where each turn requires consistent subject appearance and background layout.

LTX 2.3 (HaCohen et al., 2024). LTX 2.3 is a 22B-parameter Diffusion Transformer developed by Lightricks for joint audio-video generation. We evaluate the 22B model as an open-weight I2V baseline to assess its interaction adherence under text-prompted control.

LongCat-Video (Team et al., 2025). LongCat-Video is a DiT model developed by Meituan. Its released DiT transformer has approximately 13.6B parameters.

Kairos 3.0 (ACE Robotics, 2026). Kairos is a DiT-based video generation model from ACE Robotics, originally developed to support embodied intelligence applications that require egocentric video synthesis conditioned on navigation commands. For WBench, we use the officially released `kairos-common-4B-720P-16fps` checkpoint, which contains roughly 4.05B parameters. Its training pipeline incorporates large volumes of first-person footage collected from both simulated and real-world environments, endowing the model with a strong prior over forward-facing camera dynamics. We evaluate Kairos under the standard iterative I2V protocol using text prompts, treating it as a general I2V baseline that benefits from egocentric pre-training.

Cosmos-Predict 2.5 (Gu, 2025). Cosmos 2.5 is NVIDIA's world foundation model, pre-trained on a large-scale mixture of real-world video, physical simulation output, and synthetic robotics demonstrations. For WBench, we evaluate the Cosmos-Predict2.5 2B post-trained variant (`2B/post-trained`) in image-to-world mode, which generates 93 frames per turn at 1280×704 resolution. The Cosmos platform is designed as a general-purpose foundation for physical AI, with a Diffusion Transformer backbone that supports world-consistent video generation conditioned on text, images, or structured action specifications. Although Cosmos is primarily positioned for robotics and autonomous-driving applications, we include it as a general-purpose I2V baseline given its large-scale pre-training and publicly available weights.

B.1.2 Camera-controlled Models

These models natively accept camera-control inputs in the form of 6DoF camera parameter matrices. For each interaction turn, we provide the model with the last generated frame and a camera trajectory, represented as a sequence of 6DoF camera matrices derived from the current navigation instruction. The model then generates the next video segment conditioned on this prescribed viewpoint change.

LingBot-World (Team et al., 2026c). LingBot-World is an interactive world model targeting minute-level generation horizons with sub-second latency at 8 fps, designed for long-context robotic simulation. It is built on a Wan2.2 14B I2V backbone and consumes camera-pose sequences as the control signal: the given navigation action is first converted into a deterministic pose sequence by a geometric mapping and then fed to the backbone for frame generation, with no learned planner involved. For WBench, we use the officially released `lingbot-world-base-cam` checkpoint. Training data is a hybrid mixture of real-world footage, game-engine recordings, and synthetic renderings, providing broad coverage of camera dynamics and scene diversity.

HY-World 1.5 (HunyuanWorld, 2025). HY-World 1.5 is Tencent Hunyuan's streaming Diffusion Transformer for real-time interactive world generation, built on the 480p I2V branch of HunyuanVideo 1.5. The deployed transformer contains roughly 8.3B parameters. It introduces a dual action representation that encodes both discrete keyboard tokens and continuous camera deltas, together with a reconstituted context memory module that compresses long-horizon history into a fixed-size token buffer to maintain geometric consistency across hundreds of generation steps. This design allows the model to sustain coherent spatial layouts and stable subject appearances over extended multi-turn trajectories, which is directly reflected in its leading navigation adherence scores on WBench.

Fantasy-World (Dai et al., 2025). Fantasy-World achieves geometry-consistent interactive world modeling by augmenting a standard video Diffusion Transformer with a parallel geometric prediction branch that jointly estimates depth, point, and camera trajectory maps for each generated frame. For WBench, we use the officially released `acvlab/FantasyWorld-Wan2.1-I2V-14B-480P` adapter. Within each generation segment, the estimated geometric representations are fed back as conditioning signals for subsequent frames via IRG blocks, creating an implicit 3D consistency loop that constrains the generated video to respect plausible spatial structure. Across segments, only the last RGB frame is carried forward as the next conditioning input, so the geometric feedback operates at the segment level rather than globally across the whole video. The model is trained on a curated dataset of game-engine renderings with dense geometric annotations, and it is evaluated on navigation cases in WBench where the 3D geometric consistency signal is most informative.

InSpatio-World (Team et al., 2026a). InSpatio-World introduces a state-anchored world modeling paradigm for 4D exploration, in which the world state is represented as a set of spatiotemporal anchor tokens that persist across steps and are updated incrementally as the user navigates. For WBench, we use the officially released 1.3B distilled configuration, `InSpatio-World-1.3B`. Generation is performed via a causal block-wise autoregressive pipeline with KV caching, where each new latent block is conditioned on the current action and the cached history of previously generated blocks, rather than in a single-shot V2V pass. Since the V2V interface expects a fixed-length video input, we construct it by replicating the initial conditioning frame up to the required number of frames.

Astra (Zhu et al., 2025). Astra is a general interactive world model based on an autoregressive denoising framework that jointly predicts the next-frame latent and its denoising trajectory conditioned on an action token sequence. It is built on a Wan2.1-T2V-1.3B base with an action-conditioned adapter, and the combined model contains roughly 2B parameters. Operating at 832×480 resolution, Astra prioritizes temporal coherence and action responsiveness over raw visual fidelity, and its open-source architecture makes it accessible for ablation studies and downstream fine-tuning. The model is evaluated at its native resolution, and scores are reported without upsampling to ensure fair comparison.

B.1.3 Action-conditioned Models

These models natively accept discrete or continuous action signals and generate subsequent observations conditioned on the full action history.

Happy Oyster (Alibaba Token Hub, 2026). Happy Oyster is Alibaba's real-time interactive world creation system, supporting two distinct interaction paradigms: a *directing* mode in which the user issues discrete navigation commands, such as W/A/S/D and rotation keys, and a *wandering* mode in which the model autonomously explores an initialized world. The model is built on a streaming generative architecture that renders video frames at native resolution with low latency, allowing real-time human-in-the-loop interaction through its official web interface. Like Genie 3, it is evaluated following the standardized web-based protocol in Appendix B.2.

Matrix-Game 2.0 (He et al., 2025) and Matrix-Game 3.0 (Wang et al., 2026). Matrix-Game is a series of autoregressive diffusion models for interactive world generation developed by the Skywork team. Matrix-Game 2.0 introduces a few-step AR diffusion framework that decouples temporal prediction from per-frame quality refinement, enabling real-time streaming generation with competitive visual quality. Matrix-Game 3.0 scales the architecture and adds a camera-aware memory retrieval module that indexes previously generated keyframes for spatial consistency, and achieves 40 fps at 720p. For WBench, Matrix-Game 3.0 is deployed locally on a single NVIDIA A100 80GB GPU and evaluated using predefined action sequences. We evaluate both versions to trace the progression of the Matrix-Game design across generations.

Genie 3 (Ball et al., 2025). Genie 3 is Google DeepMind's third-generation interactive world model, generating diverse and visually rich playable environments at 24 fps conditioned on keyboard and mouse inputs. The model is built on a large-scale autoregressive Transformer trained on internet-scale gameplay video, enabling multi-minute visual memory and zero-shot generalization to novel scene descriptions from a single conditioning image. Because Genie 3 does not expose a public API, evaluation on WBench follows the automated web-interface protocol described in Appendix B.2. A browser script executes predefined keyboard commands, and the recorded video is downloaded for offline metric computation.

Hunyuan-GameCraft (Li et al., 2025b). Hunyuan-GameCraft is Tencent's interactive game video generation model, employing a history-conditioned DiT architecture that conditions each new frame on a rolling window of past frames and control tokens. The deployed model is built on the full HunyuanVideo dual-stream Transformer, and the distilled checkpoint we evaluate contains roughly 13B parameters. The model is trained on a large-scale dataset of over one million gameplay recordings across more than 100 AAA games with paired keyboard and mouse annotations, specializing it for high-dynamic game-engine-style interactions where rapid camera movement, character animation, and environmental changes must be rendered consistently. It supports real-time generation via model distillation, taking keyboard and mouse as native input modalities. Each turn is assembled by concatenating fixed-size 33-frame segments, with the number of segments per turn chosen as $\lceil T_{\text{turn}}/T_{\text{seg}} \rceil$ to best approximate the target turn duration T_{turn}.

Infinite-World (Wu et al., 2026b). Infinite-World is an interactive world model developed at Nankai University and Meituan, designed to scale generation horizons to 1000+ frames without explicit geometric priors. It is built on a Wan2.1-T2V-1.3B base with an additional action encoder, and the combined checkpoint contains roughly 1.5B parameters. The model generates video autoregressively, conditioning each new chunk on the latent representation of previously generated frames together with the current discrete keyboard action, which allows it to sustain coherent spatial layouts over very long rollouts. Each turn is assembled by concatenating fixed-size 80-frame iterations, with the number of iterations per turn chosen to best approximate the target turn duration.

B.1.4 Inference Speed Analysis

Table 7 also reports the generation efficiency of some evaluated models, measured by the ratio between wall-clock generation time and the duration of the produced video. This value can be interpreted as seconds of computation per second of output video. Equivalently, for frame-based generation, it corresponds to $\text{fps} \times s_{\text{frame}}$, where s_{frame} denotes the average generation time per frame. Lower values indicate faster generation, and values below 1 indicate faster-than-real-time generation.

For text-driven models, the reported value corresponds to the time required for a single API call or local inference pass to generate one video clip. For camera-controlled and action-conditioned models, we measure the wall-clock time required to generate the number of frames used for one evaluation turn and divide it by the corresponding video duration. All locally deployed models are evaluated on a single NVIDIA A100 80GB GPU to ensure comparable efficiency measurements, while API- or web-only models are marked separately when their underlying hardware or runtime is not accessible.

B.2 Web-Based Evaluation Protocol (Genie 3 & Happy Oyster)

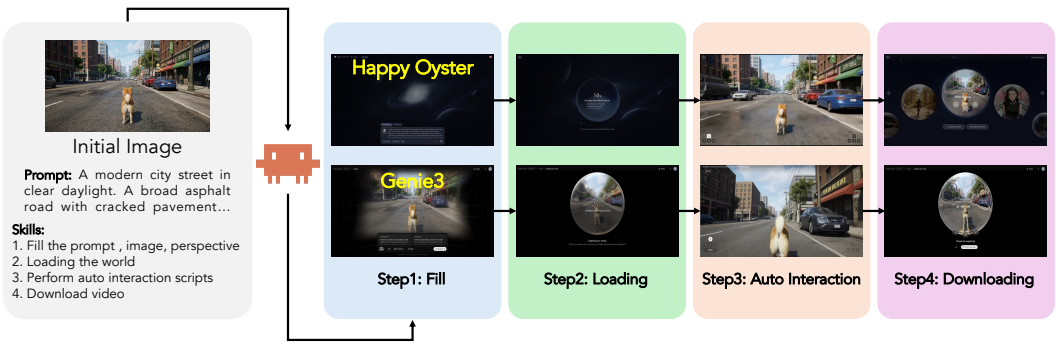


Figure 14: Automated web-based evaluation pipeline for Genie 3 and Happy Oyster. Given an initial image and a text prompt, the browser-use agent follows four steps: **Step 1** fills in the prompt, image, and perspective; **Step 2** waits for world loading to complete; **Step 3** executes the interaction script with each turn lasting 5 seconds; **Step 4** downloads the recorded video.

Genie 3 (Ball et al., 2025) and Happy Oyster (Alibaba Token Hub, 2026) are only accessible through preview web interfaces and do not expose public APIs or downloadable weights. To enable programmatic evaluation at scale, we automate the entire interaction pipeline using Claude Code's

browser-use capability, which controls Chrome via structured commands. As illustrated in Fig. 14, the process is decomposed into four steps: **(1) Fill**: the script fills in the text prompt, uploads the initial image, and configures the perspective setting on the model's web page; **(2) Loading**: the script waits for the model to finish loading the world; **(3) Auto Interaction**: once world loading is detected as complete, the script automatically executes the pre-defined interaction sequence, holding each navigation key for 5 seconds per turn; **(4) Downloading**: the generated video is downloaded for offline evaluation.

C Per-Metric Evaluation Details

C.1 Metric Quick Reference

Table 8 provides a unified index of all 22 evaluation sub-metrics. Entries in the **Detail** column are hyperlinked to the corresponding methodology description.

Table 8: Metric quick-reference table. "EM" = Expert Model; "VLM" = vision-language model judge.

Dimension	Sub-Metric	Method	Tool / Model	Detail
Video Quality	Aesthetic Quality	EM	CLIP + LAION head	Appendix C.2.1
	Imaging Quality	EM	MUSIQ	Appendix C.2.2
	Temporal Flickering	EM	Pixel MAE	Appendix C.2.3
	Dynamic Degree	EM	RAFT optical flow	Appendix C.2.4
	Motion Smoothness	EM	AMT-S interpolation	Appendix C.2.5
	HPSv3-Norm	EM	HPSv3	Appendix C.2.6
Setting Adh.	Scene Adherence	VLM	doubao-seed-2-0-lite	Appendix C.3.1
	Subject Adherence	VLM	doubao-seed-2-0-lite	Appendix C.3.2
Interaction Adh.	NavScore	EM	MegaSaM pose estimation	Appendix C.4.1
	Event Editing Adh.	VLM	doubao-seed-2-0-lite	Appendix C.4.2
	Subject Action Adh.	VLM	doubao-seed-2-0-lite	Appendix C.4.3
	Persp. Switching Adh.	VLM	doubao-seed-2-0-lite	Appendix C.4.4
Consistency	Subject Consistency	EM	DINOv2 + CLIP + SAM2	Appendix C.5.1
	Background Cons.	EM	CLIP ViT-B/32	Appendix C.5.2
	Spatial Consistency	EM	MegaSaM + DreamSim	Appendix C.5.3
	Gated Spatial Cons.	EM	MegaSaM + DreamSim	Appendix C.5.3
	Segment Continuity	EM	TransNetV2	Appendix C.5.4
	Perspective Cons.	EM	SAM2	Appendix C.5.5
	Geometric Cons.	EM	Depth Anything 3 reproj.	Appendix C.5.6
	Photometric Cons.	EM	Depth Anything 3 reproj.	Appendix C.5.6
Physical	Causal Fidelity	VLM	doubao-seed-2-0-lite	Appendix C.6.1
	Visual Plausibility	EM	Qwen3-VL-30B-A3B (ft.)	Appendix C.6.2

This appendix provides per-sub-metric evaluation details organized by dimension. Each sub-metric includes its definition and evaluation examples (VLM prompts or qualitative cases where applicable). The full list of all 22 sub-metrics is summarized in the unified metric quick-reference table (Table 8 in Appendix C.1).

C.2 Video Quality

All six Video Quality sub-metrics are computed by expert models and produce scores in $[0, 100]$ (higher is better). For a target sampling rate f, the implementation selects every $\max(1, \mathrm{round}(f_{\mathrm{video}}/f))$-th frame, so the effective rate depends on the source frame rate.

C.2.1 Aesthetic Quality

Definition. Frames are sampled at 2 FPS. Each frame is encoded by CLIP ViT-L/14, and a LAION Aesthetic linear head (Schuhmann et al., 2022) predicts an aesthetic score in $[0, 10]$. The final score is the mean over all sampled frames, rescaled to $[0, 100]$:

$$S_{\mathrm{aesth}} = \frac{1}{N} \sum_{i=1}^{N} f_{\mathrm{LAION}}(\mathrm{CLIP}(x_i)) \times 10. \tag{2}$$

C.2.2 IMAGING QUALITY

Definition. Frames are sampled at 2 FPS and resized such that the longer edge ≤ 512. MUSIQ (Ke et al., 2021) predicts a per-frame quality score in $[0, 100]$:

$$S_{\text{imag}} = \frac{1}{N} \sum_{i=1}^{N} \text{MUSIQ}(x_i). \tag{3}$$

C.2.3 TEMPORAL FLICKERING

Definition. Frames are sampled at a target rate of 10 FPS, and consecutive sampled pairs are compared via pixel-level mean absolute error (MAE). The score is inverted so that lower flickering yields a higher score:

$$S_{\text{flick}} = \frac{255 - \frac{1}{N-1} \sum_{i=1}^{N-1} \text{MAE}(x_i, x_{i+1})}{255} \times 100. \tag{4}$$

C.2.4 DYNAMIC DEGREE

Definition. Frames are sampled at 8 FPS. RAFT optical flow is computed for each consecutive pair, and the top 5% magnitude values are averaged per pair:

$$m_i = \text{mean}\big(\text{top}_{5\%} \|\mathbf{f}_i\|\big), \quad S_{\text{dyn}} = \begin{cases} 100 & \text{if } \sum_i \mathbf{1}[m_i > \tau] \geq N_{\min}, \\ 0 & \text{otherwise}, \end{cases} \tag{5}$$

where $\mathbf{f}_i$ is the optical flow field for frame pair i. Frames are resized to a short edge of at most 512 pixels. With short edge s and N sampled frames, the implementation uses $\tau = 6s/256$ and $N_{\min} = \max(1, \text{round}(N/4))$. The released code falls back to OpenCV Farneback flow if RAFT cannot be loaded. This alternative changes the flow estimator and should be distinguished from the RAFT configuration when reporting results.

C.2.5 MOTION SMOOTHNESS

Definition. Following VBench, we sample frames at a target rate of 10 FPS and resize them to a short edge of at most 512 pixels. We use AMT-S (Li et al., 2023) to interpolate even-indexed frames from odd-indexed neighbors, then compute pixel MAE between predicted and actual even frames:

$$S_{\text{app}} = \frac{255 - \frac{1}{M} \sum_{j=1}^{M} \text{MAE}(\hat{x}_{2j}, x_{2j})}{255} \times 100. \tag{6}$$

C.2.6 HPSV3-NORM

Definition. We uniformly sample up to 20 frames per video, compute HPSv3 rewards with an empty text prompt, and average them before normalization. We use fixed reference constants p_1 and p_{99} for the 1st and 99th percentiles of per-video mean rewards:

$$S_{\text{HPS}} = \text{clip}\left(\frac{\bar{r} - p_1}{p_{99} - p_1} \times 100,\ 0,\ 100 \right), \tag{7}$$

where $\bar{r}$ is the per-video mean raw reward. In our evaluation, $p_1 = -5.21$ and $p_{99} = 8.61$. These constants remain fixed when evaluating additional models.

C.3 SETTING ADHERENCE

Setting adherence evaluates whether the generated video faithfully realizes the declared world settings. It comprises two VLM-based sub-metrics: *scene adherence* (environment and off-screen elements) and *subject adherence* (appearance and motion style). All prompts are issued to `doubao-seed-2-0-lite` using the first interaction turn. The released API path submits this video segment directly, with frame extraction handled by the service.

Pre-processing: world-setting decomposition. Before evaluation, a VLM automatically decomposes each case's world-setting description into a *visible part* (elements present in the initial frame)

and an *offscreen part* (elements expected to appear only after camera movement). For subject adherence, the subject description is further decomposed into an *appearance part* (visual attributes) and an *action part* (expected movement pattern). The decomposition is precomputed once per case and shared across all models.

C.3.1 Scene Adherence

Definition. Scene adherence combines two complementary signals: (i) *environment maintenance* ($s_{\text{maint}} \in [1, 5]$): how well the initially visible scene elements and visual style are preserved across the generated segment; (ii) *offscreen content appearance* ($s_{\text{off}} \in \{0, 1\}$): whether elements described as offscreen in the initial frame become visible after the prescribed camera movement. The per-case scene adherence score is $(s_{\text{maint}}/5 + s_{\text{off}})/2$.

Prompt. Both aspects are evaluated in a single VLM call with structured JSON output:

Scene Adherence Prompt

```
You are a video analysis expert.  You will see video frames from Turn 1 of a generated
video (sampled at 2fps).

Evaluate the video on two aspects:

**1.  Environment Maintenance (1-5)**
How well does the video maintain these scene elements that were visible in the initial
frame?  Also consider whether the visual style remains consistent.
Scene description:  [VISIBLE_PART]
Visual style:  [STYLE]

Scoring:
- 5:  All elements perfectly maintained, style fully consistent
- 4:  Most elements maintained, style mostly consistent, minor issues
- 3:  Some elements maintained or style partially drifts
- 2:  Few elements maintained, significant changes or style mismatch
- 1:  Scene barely matches the description or style completely wrong

**2.  Offscreen Content Appearance (0 or 1)**
The following elements were described as existing OFF-SCREEN (not visible in the initial
frame).  Did ANY of them appear in the video as the camera moved?
Description:  [OFFSCREEN_PART]

- 1:  At least one offscreen element appeared
- 0:  None of the offscreen elements appeared

Answer in JSON format ONLY. Each reason must be within 20 words.
{"maintenance":  <1-5>, "maintenance_reason":  "...", "offscreen":  <0 or 1>,
"offscreen_reason":  "..."}
```

C.3.2 Subject Adherence

Definition. Subject adherence (third-person cases only) combines: (i) *appearance maintenance* ($s_{\text{app}} \in [1, 5]$): how well the subject's visual appearance (shape, color, clothing, equipment) is maintained throughout the segment; (ii) *action match* ($s_{\text{act}} \in \{0, 1\}$): whether the subject exhibits the described movement or action pattern. The per-case subject adherence score is $(s_{\text{app}}/5 + s_{\text{act}})/2$.

Prompt. Both aspects are evaluated in a single VLM call with structured JSON output:

Subject Adherence Prompt

```
You are a video analysis expert.  You will see video frames from Turn 1 of a generated
video (sampled at 2fps).

Evaluate the video on two aspects:

**1.  Subject Appearance Maintenance (1-5)**
How well does the video maintain the subject's visual appearance (shape, color, clothing,
equipment, category)?
Subject appearance:  [APPEARANCE_PART]

Scoring:
- 5:  Subject appearance perfectly maintained throughout
- 4:  Mostly maintained, minor visual changes
- 3:  Some features maintained, noticeable changes
- 2:  Few features maintained, significant appearance drift
```

```
- 1:  Subject barely recognizable or completely changed

**2.  Subject Action (0 or 1)**
Does the subject exhibit the described movement/action pattern in the video?
Expected action:  [ACTION_PART]

- 1:  The described action/movement pattern is clearly visible
- 0:  The action/movement pattern is not visible or completely different

Answer in JSON format ONLY. Each reason must be within 20 words.
{"maintenance":  <1-5>, "maintenance_reason":  "...", "action":  <0 or 1>, "action_reason":
"..."}
```

C.4 Interaction Adherence

Interaction adherence comprises navigation trajectory evaluation (expert model) and three VLM-based sub-metrics for event editing, subject action, and perspective switching.

C.4.1 NavScore

Definition. NavScore evaluates navigation accuracy by comparing predicted camera trajectories against synthetic ground-truth (GT) trajectories constructed from the action sequence.

Camera pose estimation. We extract per-frame camera-to-world poses using MegaSaM (Li et al., 2025c), a monocular structure-from-motion method that produces dense 4×4 pose matrices from video. Each multi-turn video yields a single globally consistent pose sequence, which is then split at turn boundaries for per-turn and global analysis.

Ground-truth trajectory construction. GT trajectories are built adaptively per turn based on the action type:

- *Pure translation* (W/S/A/D): a straight-line trajectory in the action direction under the initial camera orientation, with length matched to the predicted displacement. If the predicted displacement is below a minimum threshold (0.1 units), a fallback length of 1.0 is used to penalize non-responsive models.
- *Rotation* (left/right/up/down): an adaptive trajectory uses a radius estimate $R = \text{chord}/(2\sin(\theta_{\text{pred}}/2))$, where θ_{pred} is the total predicted rotation. When this estimate is available and $\theta_{\text{pred}} \geq 3^\circ$, the GT angle is $\max(\theta_{\text{pred}}, 30^\circ)$, with $R \geq 1.0$ for third-person orbits. Otherwise the construction uses fallback parameters $\theta = 30^\circ$ and $R = 1.0$. First-person rotations change heading, while third-person rotations trace an orbit.

The GT is then aligned to the predicted coordinate frame (translation offset + rotation alignment), while the predicted trajectory remains unchanged.

Scope of adaptive ground truth. The evaluated interfaces do not share a calibrated displacement for one navigation command. Matching GT extent to predicted motion reduces dependence on interface-specific step sizes, so NavScore primarily assesses commanded direction, trajectory shape, and consistency across turns. The minimum-motion thresholds and fallback trajectories above penalize insufficient responses, but do not make the score a measure of absolute distance or rotation-magnitude accuracy. A high NavScore therefore does not establish that a model executed a prescribed distance or angle. Evaluating that capability would require calibrated target magnitudes shared across interfaces.

Arc-length resampling. To ensure equal weighting across turns regardless of frame count or motion speed, both GT and predicted trajectories are resampled to a fixed number of points ($K = 20$) uniformly along their respective arc lengths. Position is linearly interpolated and rotation is interpolated via Slerp along the arc-length parameter.

Evaluation metrics. We evaluate navigation accuracy using normalized Absolute Trajectory Error (ATE), which measures the global shape agreement between predicted and ground-truth trajectories after per-turn arc-length resampling and concatenation:

$$\text{nATE}_t = \min\left(\frac{\text{ATE}_t}{\max(L_{\text{pred}},\ 0.5)},\ 1\right), \tag{8}$$

$$\text{nATE}_r = \min\left(\frac{\text{ATE}_r}{\max(\Theta_{\text{pred}},\ 10^\circ)},\ 1\right), \tag{9}$$

where L_{pred} and Θ_{pred} are the predicted path length and total rotation. The minimum denominators (0.5 for translation, 10° for rotation) prevent instability when the model produces near-zero motion. We also compute Relative Pose Error (RPE) and report it as an auxiliary diagnostic, but it is excluded from the final score because action-conditioned world models often exhibit non-uniform motion profiles, making frame-level velocity-matched GT trajectories unreliable. ATE therefore serves as the primary signal for trajectory-shape correctness.

Trajectory consistency. For repeated or symmetric actions, we compare all valid same-group turn pairs, including W↔S, A↔D, left↔right, and up↔down. Each turn is first normalized to its own start pose, and symmetric pairs are mirrored into a shared canonical frame when needed. We then compute pairwise normalized ATE between the two turns and average over all valid pairs:

$$\text{cnATE}_t = \frac{1}{P}\sum_{p=1}^{P} \text{nATE}_t^{(p)}, \tag{10}$$

$$\text{cnATE}_r = \frac{1}{P}\sum_{p=1}^{P} \text{nATE}_r^{(p)}, \tag{11}$$

where P is the number of valid same-group turn pairs. The trajectory consistency term is then defined as

$$\text{Cons} = 1 - \frac{\text{cnATE}_t + \text{cnATE}_r}{2}. \tag{12}$$

If no valid same-group pair exists, both consistency and the per-case NavScore are treated as null for that case.

NavScore aggregation. The per-case NavScore combines accuracy and consistency when both are available:

$$\text{Acc} = 1 - \frac{\text{nATE}_t + \text{nATE}_r}{2}, \tag{13}$$

$$\text{NavScore}^{(\text{case})} = \frac{\text{Acc} + \text{Cons}}{2}. \tag{14}$$

When a case has no valid same-group pair, only the accuracy term is defined, and the per-case NavScore is marked as null. At the model level, accuracy and consistency are each averaged over their own valid case sets, and the model-level NavScore is the average of the two means, rescaled to $[0, 100]$. NavScore jointly considers trajectory accuracy and cross-turn consistency, as both following the commanded motion and maintaining consistent responses across turns contribute to a model's navigation capability.

Pose-fidelity interpretation. NavScore evaluates the complete pose trajectory rather than masking pose components that are not directly commanded. For a pure translation turn, the target trajectory moves in the commanded direction while preserving orientation, such that unintended rotation is penalized. For a first-person rotation turn, the target changes orientation while preserving camera position, such that unintended translational drift is penalized. Third-person rotation commands induce orbital motion and therefore involve both translation and rotation. Consequently, both components are retained when computing pose fidelity. This design intentionally measures whether the generated trajectory follows the commanded motion without introducing unintended motion in the remaining pose components. Because the two components are averaged, cases with one correct and one incorrect component may receive partial credit; we therefore report and analyze the translation and rotation errors separately as auxiliary diagnostics.

Limitation of adaptive normalization. NavScore adapts the target motion magnitude to the predicted trajectory, which can reduce sensitivity to under- or over-rotation relative to a prescribed camera trajectory. In a preliminary diagnostic on 46 first-person HY-World 1.5 cases containing 136 rotation turns, reconstructed target and MegaSAM-estimated endpoint rotation magnitudes average 69.30° and 54.98° across turns, respectively. The mean geodesic rotation error is 9.08° after anchoring each turn to its initial orientation, averaging frames within turns, turns within cases, and then cases equally. These measurements include pose-estimation error, and their interpretation depends on the reconstructed controls matching the original generation inputs. A high NavScore therefore does not necessarily imply accurate reproduction of the requested motion magnitude, motivating complementary evaluation against the original camera controls when available.

Qualitative Results. Fig. 15 shows a four-turn navigation case (T1: W, T2: →, T3: →, T4: ←) evaluated on three models. Happy Oyster and HY-World 1.5 produce trajectories that consistently follow the instructed directions across all turns, while HY-Video 1.5 reverses the rotation direction in T2 and T3, resulting in clear trajectory deviation. Fig. 16 illustrates how the adaptive ground-truth is constructed per model. Because different models accept different input interfaces and exhibit varying motion magnitudes, the GT trajectory adapts its scale to the predicted path length. Both Happy Oyster and HY-World 1.5 navigate correctly but with different amplitudes, so their adaptive GTs differ in scale while both yielding high NavScores. In contrast, HY-Video 1.5 moves in the wrong direction, and the adaptive GT faithfully reflects this directional error, penalizing it appropriately.

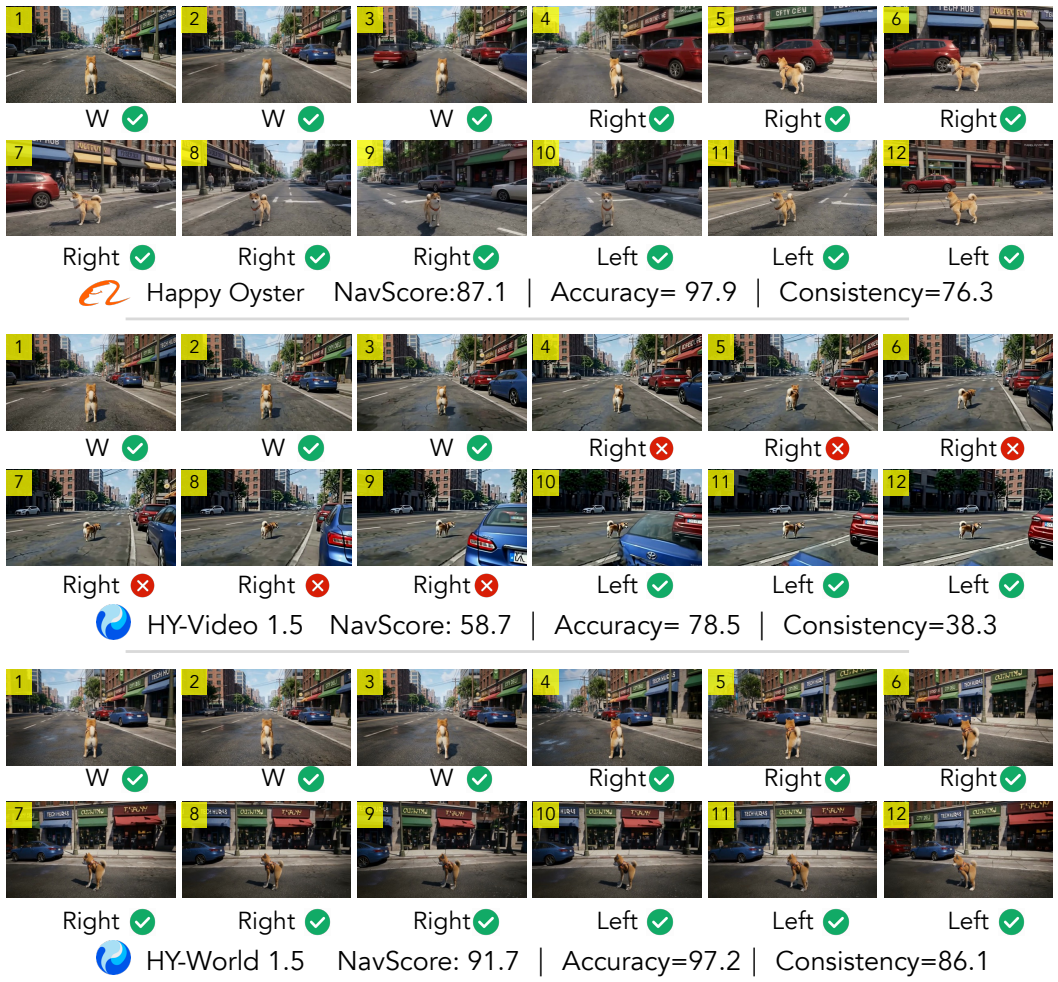


Figure 15: A four-turn navigation case (T1: W, T2: →, T3: →, T4: ←) across three models. Frames 1–3 correspond to T1, 4–6 to T2, 7–9 to T3, and 10–12 to T4. Happy Oyster and HY-World 1.5 follow the instructed directions correctly, while HY-Video 1.5 reverses the rotation in T2 and T3.

C.4.2 Event Editing Adherence

Definition. Event editing adherence uses `doubao-seed-2-0-lite` with video frames sampled at 3 fps. Five progressive binary questions are asked per turn, each with an *expected* answer. The turn receives one point per question whose response matches the expected answer, yielding a raw score in $[0, 5]$. The case score is the mean over all event-edit turns, rescaled to $[0, 100]$ by multiplying by 20.

Prompt. Event editing and subject action share the following system prompt:

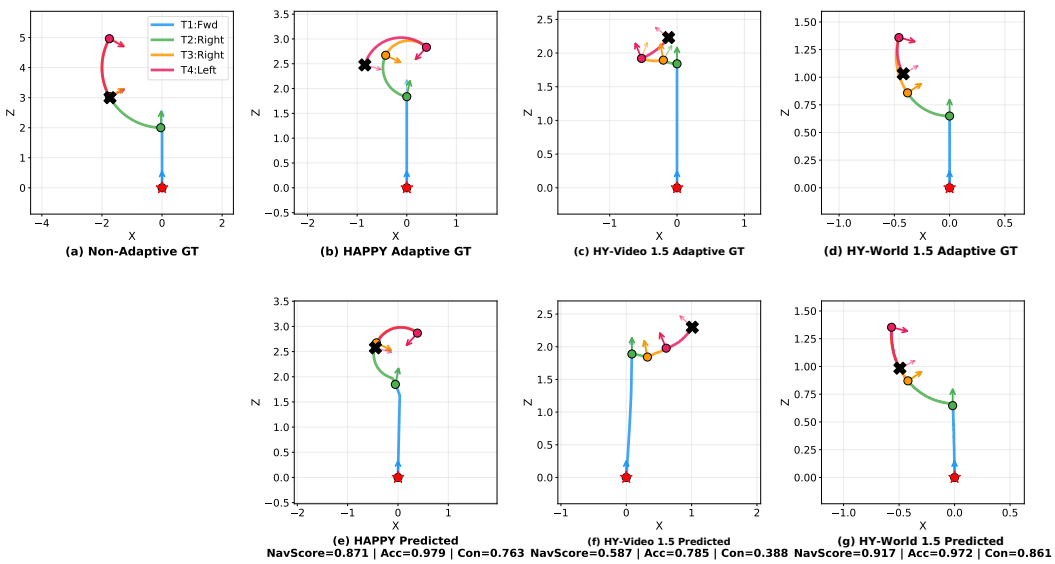


Figure 16: Adaptive GT construction for the same case. The GT trajectory adapts to each model's predicted motion magnitude, so correctly navigating models with different amplitudes (Happy Oyster vs. HY-World 1.5) both receive fair evaluation. For HY-Video 1.5, the wrong-direction motion is captured as trajectory error rather than a scale mismatch.

Shared System Prompt (event editing & subject action)

System Prompt

```
You are evaluating AI-generated video segments from interactive world models.
Each video segment corresponds to one interaction turn where a specific event was
instructed to happen.

Your task: answer with ONLY 'Yes' or 'No' based on what you observe in the provided
frames. Follow these guidelines:
- These are AI-generated videos, not real footage. Minor visual artifacts, slight
inconsistencies, or imperfect rendering should NOT cause a 'No' if the overall intent
is recognizable.
- Focus on whether the CORE semantic meaning of the described event is visually
present, not whether every literal detail matches perfectly.
- A long or detailed event description may only be partially depicted. Judge based on
whether the main action or change is recognizable.
- When in doubt between Yes and No, lean toward the interpretation that better matches
what you actually see, not what you expect to see.
```

Each response is followed by a one-line `Reason:` field (at most 50 words) that names the specific issue whenever the point is not awarded.

Event Editing Adherence (Q1–Q5)

```
World setting: [SCENE_DESCRIPTION]
Interaction instruction for this turn: [ACTION]

The following frames are uniformly sampled from the generated video segment for this
turn.
[FRAME_SEQUENCE]

Q1 (Static Scene Check): Does the scene remain largely unchanged, with no clear
indication of '[ACTION]' occurring? Answer No if there is ANY recognizable change related
to this event. Expected: No.
Q2 (Event Occurrence): Does the video show something resembling or consistent with the
described event '[ACTION]'? Partial depiction counts as Yes. Expected: Yes.
Q3 (Event Completion): By the end of this segment, has the described event '[ACTION]'
reached a clear conclusion or outcome? Ongoing events without a visible end state receive
No. Expected: Yes.
Q4 (Detail Accuracy): Are the key details of '[ACTION]' accurately depicted, such as the
```

```
correct objects, agents, directions, and quantities?  Expected:  Yes.
Q5 (Anomaly Absence):  Does any unexpected object, entity, or visual anomaly unrelated
to '[ACTION]' appear in this segment?  Natural consequences of the event and minor AI
rendering artifacts do not count.  Expected:  No.
```

Qualitative Results. Fig. 17 contrasts representative outputs across models on two event-edit turns. High-scoring cases exhibit a clear state transition aligned with the instructed event (matching Q2–Q3), while low-scoring cases either leave the scene untouched (Q1 fires) or introduce unrelated entities that trigger the anomaly check (Q5).

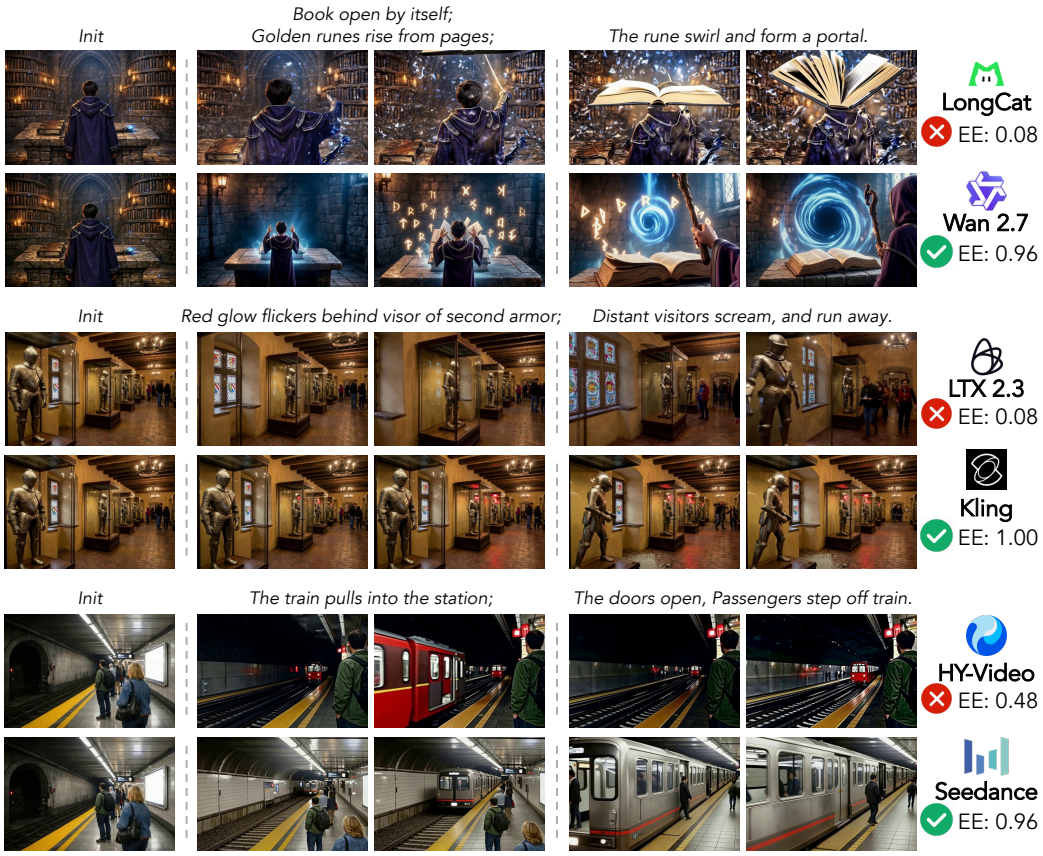


Figure 17: Qualitative comparisons on event editing.

C.4.3 Subject Action Adherence

Definition. Subject action adherence uses the same shared system prompt and five-question scoring rule as event editing adherence, but with question templates tuned to whether the subject performs the instructed action. The case score is the mean over all subject-action turns.

Prompt. We show the prompt below.

Subject Action Adherence (Q1–Q5)

```
World setting:  [SCENE_DESCRIPTION]
Interaction instruction for this turn:  [ACTION]

The following frames are uniformly sampled from the generated video segment for this
turn.
[FRAME_SEQUENCE]

Q1 (Idle Subject Check):  Does the subject remain idle or stationary, with no attempt
to perform '[ACTION]'? Answer No if the subject shows ANY movement or gesture related to
this action.  Expected:  No.
Q2 (Action Occurrence):  Does the video show the subject performing something resembling
```

or consistent with '[ACTION]'? *Partial execution counts as Yes.* Expected: Yes.
Q3 (Action Completion): By the end of this segment, has the subject's action '[ACTION]' reached a clear conclusion or outcome? Expected: Yes.
Q4 (Detail Accuracy): Are the key details of '[ACTION]' accurately depicted, such as the correct subject, target objects, and manner of the action? Expected: Yes.
Q5 (Unnatural Motion Check): Does the subject exhibit any unnatural body movement, impossible pose, or physically implausible interaction while performing '[ACTION]'? *Only clearly unnatural movements or impossible poses count; slight jitter or minor hand artifacts do not.* Expected: No.

Qualitative Results. Fig. 18 illustrates how subject-action adherence differentiates models on the same instruction. Successful generations show the subject initiating and completing the prescribed action successfully, whereas failure cases leave the subject idle or execute only a partial gesture that never reaches a clear end state.

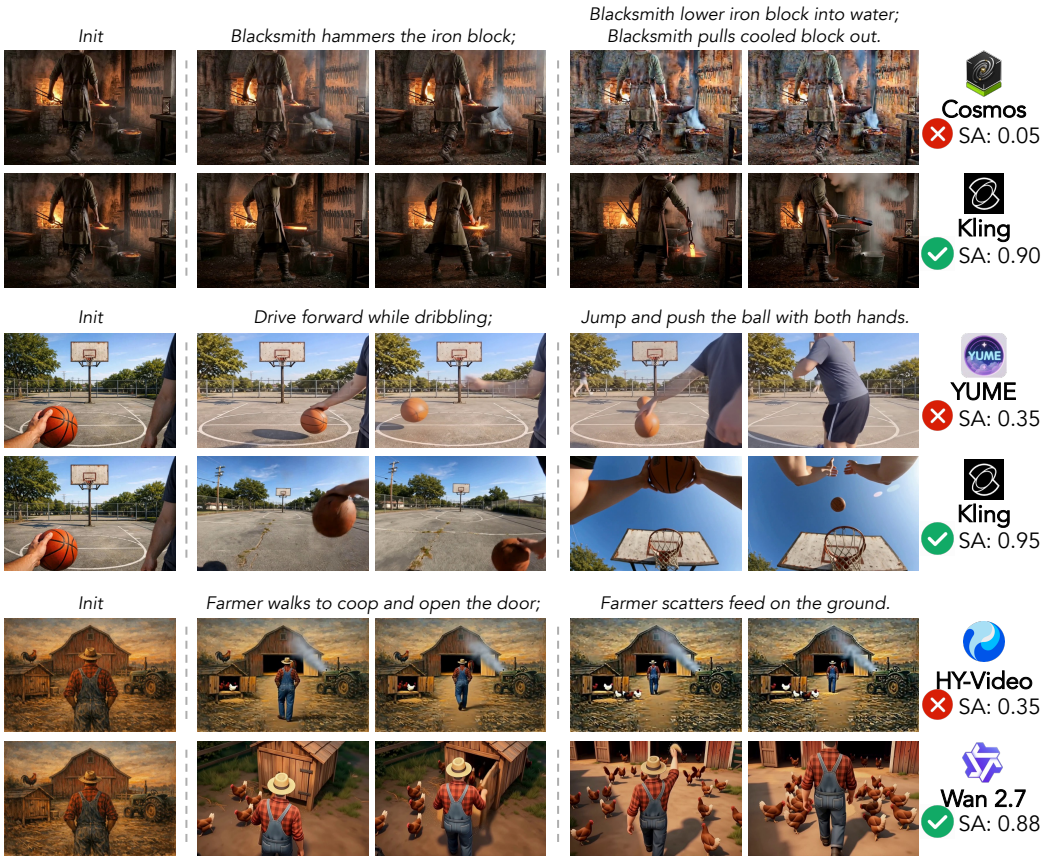


Figure 18: Qualitative comparisons on subject action.

C.4.4 Perspective Switching Adherence

Definition. Perspective switching adherence evaluates whether the model executes a requested change of perspective. Three binary checks determine the score, labeled Q1, Q3, and Q4 in the code. Q1 receives three early and three late frames, while Q3 and Q4 receive the three late frames. A turn scores 1 only if all three answers are `Yes`, and 0 otherwise. An additional camera-motion question uses frames sampled across the turn but does not enter the score.

Prompt. The following summary presents the three scoring criteria. The released prompts instantiate the requested source and target perspectives separately for each question.

Perspective Switching Adherence (Q1, Q3, Q4)

System:

System Prompt

```
You are a precise video evaluator.  You will be given frames from the beginning and
end of a generated video turn and a perspective-switching instruction.  Answer each
question with Yes or No only.
```

User:

User Prompt

```
Perspective switching instruction:  [SWITCH_INSTRUCTION]
(e.g., "Switch from first-person to third-person view following the character")
Source perspective:  [SOURCE_PERSPECTIVE] (first-person / third-person)
Target perspective:  [TARGET_PERSPECTIVE] (first-person / third-person)

Early frames (beginning of the turn segment):
[EARLY_FRAME_SEQUENCE]
Late frames (end of the turn segment):
[LATE_FRAME_SEQUENCE]

Answer each question with Yes or No only.

Q1 (Transition Visibility):  Can a clear change in perspective be observed between
the early and late frames?
Q3 (Target Consistency):  Does the late-frame perspective match the target perspective
([TARGET_PERSPECTIVE])?
Q4 (Quality Compliance):  Does the new perspective satisfy the target framing
requirements?  For third-person, is the character roughly centered, viewed from behind,
and shown at an appropriate distance?  For first-person, are impossible views of the
character's own face, back, or full body absent?
```

Qualitative Results. Fig. 19 shows some text-driven model cases on the same switching instruction. Successful cases satisfy all three checks with a clear perspective change, a scene matching, and a structurally clean framing, while failures either preserve the source perspective.

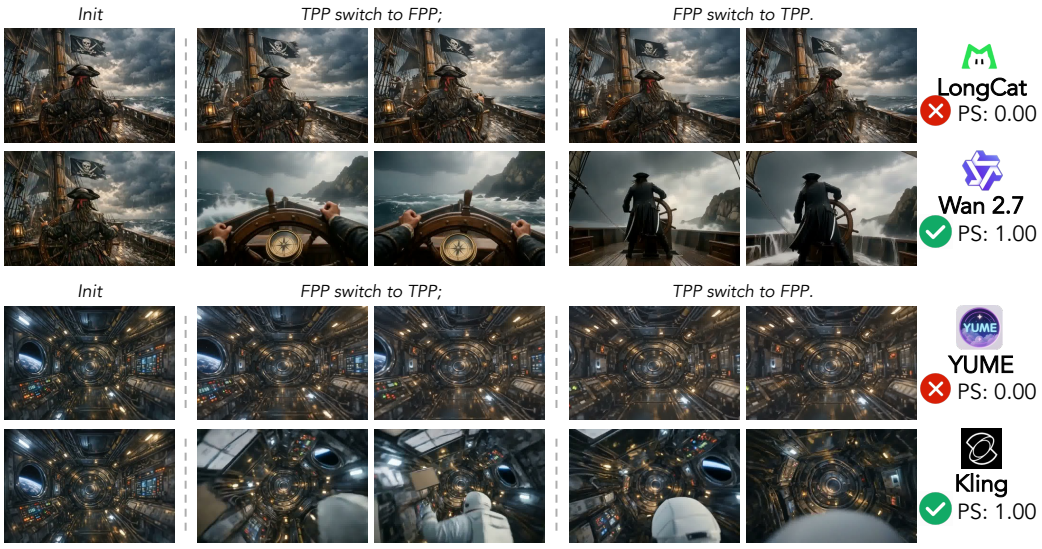


Figure 19: Qualitative comparisons on perspective switching.

C.5 Consistency

The eight consistency sub-metrics measure complementary aspects of temporal coherence within each generated clip.

C.5.1 SUBJECT CONSISTENCY

Definition. For each frame with a valid SAM2 mask (area ≥ 10 px), we crop the subject region with 10-pixel padding and fill the background with gray. When a mask is absent or too small, the implementation retains the full frame. We extract DINOv2 ViT-B/14 and CLIP ViT-B/16 features from these inputs. We compute two similarity signals: (i) DINOv2 adjacent-frame cosine similarity $s_i^{\text{dino}} = \max(0, \cos(\mathbf{f}_{i-1}, \mathbf{f}_i))$, and (ii) CLIP first-frame anchored similarity $s_i^{\text{clip}} = \max(0, \cos(\mathbf{g}_0, \mathbf{g}_i))$, where $\mathbf{f}$ and $\mathbf{g}$ denote the respective feature sequences. The score is $(s_i^{\text{dino}} + s_i^{\text{clip}})/2$, averaged over frames after the first. The full-frame fallback can introduce background information when subject tracking fails.

C.5.2 BACKGROUND CONSISTENCY

Definition. Following VBench, we extract CLIP ViT-B/32 features at a target rate of 2 FPS. Each frame after the first is compared with both its predecessor and the first frame, with negative cosine similarities clipped to zero:

$$s_{\text{bg}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{\max(0, \cos(\mathbf{f}_{i-1}, \mathbf{f}_i)) + \max(0, \cos(\mathbf{f}_0, \mathbf{f}_i))}{2}. \tag{15}$$

C.5.3 SPATIAL CONSISTENCY

Definition. For roundtrip trajectories, we use MegaSaM-estimated camera poses to identify the *return frame* in the final turn, i.e., the frame whose rotation matrix R_k minimizes $\arccos\big((\text{tr}(R_0^\top R_k) - 1)/2\big)$ relative to the first frame. Let s_{ret} be the DreamSim similarity between the first frame and this return frame, computed as $1/(1+d)$ where d is the DreamSim distance. We uniformly sample 10 intermediate frames and compute the minimum similarity $s_{\min}$ to the first frame. A motion gate prevents trivially high scores from near-static videos:

$$\text{score} = s_{\text{ret}} \cdot \min\Big(1,\ \frac{1 - s_{\min}}{\tau}\Big), \quad \tau = 0.15. \tag{16}$$

Return-frame selection. Spatial Consistency is designed as a viewpoint-centric visual revisitation metric rather than an exact SE(3) loop-closure metric. For round-trip trajectories, we identify the frame in the final return segment whose orientation is closest to the initial view, and evaluate whether the visual content is recovered at that viewpoint. We intentionally use orientation for return-frame selection because monocular pose estimation is generally more sensitive to translational scale and accumulated drift, whereas orientation provides a more stable cue for identifying a revisited viewing direction. Translational and geometric fidelity are evaluated separately by NavScore and the reconstruction-based geometric/photometric consistency metrics. The motion gate further prevents near-static videos from receiving trivially high revisitation scores.

Qualitative Results. Fig. 20 illustrates the motion gate at work. The near-static video generated by LongCat-Video achieves a high return-frame similarity s_{ret} but is penalised by the motion gate. A video that returns to the initial view after meaningful motion receives a higher gated spatial consistency score.

C.5.4 SEGMENT CONTINUITY

Definition. TransNetV2 predicts per-frame scene-boundary probabilities. Frames exceeding a confidence threshold of 0.5 are flagged as cut candidates, with a minimum scene length of 10 frames to suppress spurious detections. Each video receives a binary score: 1 if no cut is detected, 0 otherwise. The model-level score is the fraction of cut-free videos: $\text{score} = 1 - n_{\text{cut}}/N$.

C.5.5 PERSPECTIVE CONSISTENCY

Definition. For each frame where the SAM2-tracked target is present (mask area ≥ 10 px), we record the normalized centroid (c_x, c_y) within the mask region. Let $p = n_{\text{valid}}/n_{\text{total}}$ denote the target presence rate. The score measures how stably the subject remains positioned across frames:

$$s_{\text{centroid}} = \max\Big(0,\ 1 - \frac{\sqrt{\sigma_{c_x}^2 + \sigma_{c_y}^2}}{0.3}\Big) \times p, \tag{17}$$

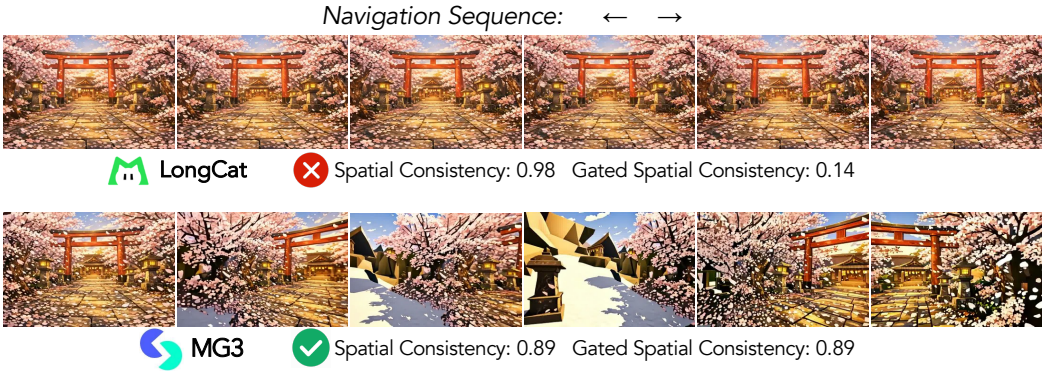


Figure 20: Qualitative comparisons on spatial consistency. The gated score penalises the static scene.

where σ_{c_x} and σ_{c_y} are the standard deviations of the normalized centroid coordinates over valid frames. The presence weighting ensures that models where the target frequently disappears receive lower scores, while the threshold constant (0.3) is set so that moderate drift already incurs significant penalty.

C.5.6 Reconstruction Consistency

Definition. Depth Anything 3 (Lin et al., 2025) jointly estimates per-frame depth maps and camera poses. We evaluate 3D coherence via two complementary signals that share the same depth-based reprojection pipeline.

For a pixel $\mathbf{u}_i = (u_i, v_i, 1)^\top$ in frame i with depth $d_i(\mathbf{u}_i)$, we first back-project it to 3D in camera i,

$$\mathbf{X}_i = d_i(\mathbf{u}_i) K^{-1} \mathbf{u}_i, \tag{18}$$

where K is the camera intrinsic matrix. Using the relative pose from frame i to frame j, we transform the point into camera j,

$$\mathbf{X}_j = R_{ji}\mathbf{X}_i + \mathbf{t}_{ji}, \tag{19}$$

and project it onto frame j,

$$\hat{\mathbf{u}}_j = \pi(K\mathbf{X}_j), \tag{20}$$

where $\pi([x, y, z]^\top) = (x/z, y/z)$ denotes perspective projection.

Geometric consistency compares the projected depth $z_{i\to j}$ with the depth d_j predicted at the rounded target pixel. For each target frame j, points from the other frames are projected into its view. Among points with positive projected and target depths inside the image, the relative depth error is

$$e_{i\to j} = \frac{|z_{i\to j} - d_j(\hat{\mathbf{u}}_j)|}{d_j(\hat{\mathbf{u}}_j)}. \tag{21}$$

The implementation takes the median error e_j over valid projected points in each target frame, then averages these medians over valid frames $\mathcal{J}$. The score is

$$\bar{e} = \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} e_j, \qquad s_{\text{geo}} = \frac{1}{1+\bar{e}}. \tag{22}$$

Frames without valid depth comparisons are excluded, and a video with no valid frame receives zero.

Photometric consistency renders colors from the other frames into each target view using the same reprojection mapping. Let $\hat{I}_j$ denote the rendered image and I_j the observed frame. PSNR is evaluated over rendered pixels, and views with fewer than 100 rendered pixels are excluded. The implementation averages positive per-view PSNR values and converts the mean to the reported score:

$$p_j = -10\log_{10}\left(\text{MSE}(\hat{I}_j/255, I_j/255) + 10^{-10}\right), \qquad S_{\text{photo}} = 100\left(1 - 10^{-\bar{p}/20}\right), \tag{23}$$

where $\bar{p}$ is the mean of valid positive p_j values, or zero if none are available. This nonlinear transformation is used by the released report generator. The reconstruction module also exposes a legacy linear PSNR diagnostic, which is not used in the final report.

The two signals measure depth agreement and appearance agreement under the same estimated geometry. Errors in depth or pose can affect both scores, so they should be interpreted as reconstruction-based diagnostics.

C.6 PHYSICAL

The physical dimension is evaluated via two sub-metrics: causal fidelity (VLM-based) and visual plausibility (fine-tuned Qwen3VL).

C.6.1 CAUSAL FIDELITY

Definition. Causal fidelity uses `doubao-seed-2-0-lite` with video frames sampled at 3 fps. It is scored along two complementary tracks, both using integer grades from 0 to 3. Both tracks receive frames from the entire video. Track 1 assigns a global score for rendering-level physics and causal consistency. Track 2 averages scores over a precomputed, manually verified subset of seven physics dimensions that is fixed across models. When Track 2 has valid scores, the case score is $(s_{\text{track1}} + \bar{s}_{\text{track2}})/2$. Otherwise it falls back to s_{track1}. The result is divided by 3 for a unit-scale score and multiplied by 100 for reporting.

Prompt. The used prompots are shown as below.

Causal Fidelity — Track 1: Rendering Physics & Causal Consistency (0–3)

System:

System Prompt

```
You are evaluating the RENDERING-LEVEL visual plausibility and causal consistency of
an AI-generated video segment.  You will be shown a sequence of frames in temporal
order.

IMPORTANT RULES:
- Do NOT judge whether depicted actions are possible in the real world.  Fantasy /
sci-fi elements are acceptable if the setting calls for them.
- Do NOT penalize camera behaviour (tilting, shaking, panning, zooming, rotation).
Camera motion is a cinematography choice, NOT a physics issue of the world.
- ONLY evaluate the physics of OBJECTS and CHARACTERS within the scene.

Before scoring, mentally scan the frames for notable events, unexplained appearances
/ disappearances, and whether every visible effect has a plausible cause and every
depicted action produces its expected consequences.

Evaluate the following aspects:
A. Rendering physics.  (1) solid-body integrity, (2) consistent gravity, (3) motion
continuity, (4) object permanence, (5) natural deformation, (6) character physics.
B. Causal consistency.  (7) no effect without cause (for non-instructed changes), (8)
no cause without effect (the instructed action must produce its consequences), (9) no
unrelated objects or effects appearing from nothing.  The instructed action and its
direct consequences are EXPECTED changes and must NOT be penalised.
```

User:

User Prompt

```
Scene context: [SCENE_DESC]
The instructed action for this segment is: '[ACTION]'.
Special physics rule: [RULE_DESC]

[FRAME_SEQUENCE]

First, note any notable events, state changes, or objects appearing / disappearing
across the frames. Then rate the overall rendering physics AND causal consistency of
this segment from 0 to 3:
3 - Good: physics and causality are natural and consistent, no significant artifacts
or causal errors.
2 - Fair: noticeable problems such as some clipping, jerky motion, or 1-2 causal
errors, but main actions still make sense.
1 - Poor: multiple physics violations and / or causal breakdowns, objects appear from
nowhere, frequent clipping, missing consequences.
0 - Failure: physics and causality are mostly or completely broken.
```

Causal Fidelity — Track 2: Scene-Aware Physics Dimensions (0–3, averaged)

Stage 2A (dimension selection, text-only; precomputed once per case):

System Prompt

```
You are a physics evaluation assistant for AI-generated videos. Given scene context,
environment details, and an action description, identify which specific physics
phenomena are relevant and should be evaluated. Consider BOTH the action AND the
environment, for example rain implies fluid / reflection, snow implies surface tracks,
wind implies environmental forces. Select only dimensions that are clearly relevant,
but do not miss obvious ones.
```

User Prompt

```
Scene description: [SCENE_DESCRIPTION]
Interaction sequence: [FULL_INTERACTION_SEQUENCE]
Environment details: [ENV_DETAILS]

Select ALL applicable dimensions from:
(1) Fluid & Smoke: liquids, smoke, fire, steam, and underwater behaviour (buoyancy,
bubbles, drifting).
(2) Collision & Clipping: objects or characters physically colliding, striking,
pushing, or overlapping. Do NOT select for scenes without physical contact.
(3) Surface Tracks & Imprints: movement over soft, deformable surfaces (snow, sand,
mud, wet ground, dust). Do NOT select for hard surfaces.
(4) Deformation & Destruction: breaking, shattering, crumbling, tearing, or bending,
including explosion debris.
(5) Wind & Environmental Forces: wind acting on hair, cloth, leaves, particles;
currents and air resistance.
(6) Reflection & Lighting: water / mirror reflections, explicit point lights,
directional shadows, metallic / glass reflections, or fire / lava illumination.
(7) Human Motion & Expression: clearly visible human or humanoid figures whose body
motion and face coherence can be judged across frames.
```

Stage 2B (shared system prompt and user template for per-dimension scoring):

System Prompt

```
You are evaluating a specific physics phenomenon in an AI-generated video segment.
Score based ONLY on the specified physics dimension, not on overall video quality.
These are AI-generated videos; minor rendering artifacts are acceptable. Focus on
whether the specific physical behaviour is plausible.
```

User Prompt

```
World setting: [SCENE_DESCRIPTION]

[FRAME_SEQUENCE]

Evaluate dimension [DIM_ID]: [DIM_NAME].
Description: [DIM_DESCRIPTION]
[DIM_SCORING_PROMPT]
```

The seven dimension-specific `[DIM_SCORING_PROMPT]` templates are reproduced verbatim below.

Track 2 Rubric — Fluid & Smoke

```
Rate fluid, smoke, and related effects (0-3).
3 = Realistic flow, splash, smoke rise, underwater drift.
2 = Minor artifacts but behaviour is recognisable.
1 = Notable issues: liquid defies gravity, smoke static, no bubbles.
0 = Fluid / smoke behaviour completely unrealistic or absent, or completely unrealistic.
```

Track 2 Rubric — Collision & Clipping

```
Rate collision and solid-body integrity (0-3). ONLY evaluate collisions / contacts that
actually occur.
3 = All contacts are natural, no clipping or pass-through.
2 = Minor clipping but mostly plausible contacts.
1 = Obvious clipping: limbs pass through objects or surfaces.
0 = Severe clipping, objects completely ignore each other.
```

Track 2 Rubric — Surface Tracks & Imprints

```
Rate surface interaction and imprint accuracy (0-3).
3 = Clear, appropriate marks visible (footprints in snow, tracks in sand).
2 = Some marks visible but incomplete or inconsistent.
1 = No marks visible on a surface that should clearly show them.
0 = Surface shows no response at all to contact.
```

Track 2 Rubric — Deformation & Destruction

```
Rate deformation and destruction effects (0-3).
3 = Realistic shattering, crumbling, bending, debris scatter.
2 = Minor issues but deformation is recognisable.
1 = Deformation clearly unnatural or inconsistent.
0 = No deformation when expected, or completely unrealistic.
```

Track 2 Rubric — Wind & Environmental Forces

```
Rate wind and environmental force effects (0-3).
3 = Natural swaying, drifting, fluttering of affected objects.
2 = Minor issues but environmental effects are recognisable.
1 = Effects are stiff, static, or inconsistent with the scene.
0 = No environmental effects visible when expected.
```

Track 2 Rubric — Reflection & Lighting

```
Rate the accuracy of SPECIAL reflection and lighting effects (0-3). Evaluate ONLY optical
effects that are actually VISIBLE in the video; do NOT penalise for effects that are
absent from the scene.
Check for:
- Water / mirror reflections: do they match the shape and motion of the source?
- Point light sources: is the illumination range and falloff plausible?
- Shadows: do they point in the correct direction relative to the light?
- Metallic / glass surfaces: do they show plausible environment reflections?

3 = All visible optical effects are accurate and physically consistent.
2 = Minor errors in one effect (e.g. reflection slightly misaligned).
1 = Clear errors: reflections show wrong content, shadows point wrong way, or lighting
does not respond to scene changes.
0 = Optical effects are completely wrong or absent when clearly expected.
```

Track 2 Rubric — Human Motion & Expression

```
Rate human motion naturalness and facial expression quality (0-3). Focus on MOVEMENT
ACROSS FRAMES, not just single-frame anatomy:
- Does the character move fluidly between poses, or jerk / teleport?
- Are body movements biomechanically plausible (natural gait, realistic reach, believable
weight shifts)?
- Do limbs move in coordinated, purposeful ways, or flail randomly?
- Is the face coherent (no melting, distortion, or uncanny shifts)?
Minor structural issues (slight hand distortion, brief extra finger) should NOT heavily
impact the score if overall motion is natural.

3 = Smooth, natural human motion throughout; face is coherent.
2 = Mostly natural motion with minor stiffness or brief glitches.
1 = Clearly unnatural: jerky limbs, puppet-like movement, frozen poses, or obvious face
distortion.
0 = Severely broken: limbs flail impossibly, body contorts unnaturally, face melts or
deforms grotesquely.
```

Track 2 score $\bar{s}_{track2}$ is the arithmetic mean of valid 0–3 scores over the selected dimensions. Dimension selection is precomputed once per case, manually verified, and shared across models. Both tracks score the entire video. With valid scores from both tracks, $s = (s_{track1} + \bar{s}_{track2})/2$. Otherwise a valid Track 1 score is used alone. Cases without a valid Track 1 score are excluded. The case score is divided by 3 for normalization.

Qualitative Results. Fig. 21 shows the comparison of good and bad cases, generated by different models on each sub-dimension evaluated in Track 2.

Results. Table 9 lists every one of the **50** cases that receive Track 2 scoring, together with a short scene description and the seven dimension flags. A ✓ indicates that the dimension is included in that case's Stage 2B scoring pool, and a ✗ indicates that it is skipped.

Figure 22 further decomposes Track 2 causal fidelity across the seven physics sub-dimensions. The results reveal heterogeneous failure modes across models. Reflection & Lighting is nearly saturated for most models, while Deformation & Destruction shows much larger variation despite being activated in only 4 of the 50 Track 2 cases, suggesting that destructive event dynamics remain challenging. Tier-wise gaps are also axis-dependent: strong models such as Wan 2.7 lead on most dimensions, but their advantage is much larger on Deformation and Human Motion than on Reflection, indicating that aggregate scores can obscure the underlying failure axes. Moreover, models with similar overall performance exhibit distinct trade-offs. Flagship text-driven systems are stronger on Fluid and Collision, LingBot-World performs best on Human Motion and Reflection but lags on Fluid and Surface, and YUME 1.5 stands out on Wind, consistent with its outdoor navigation-oriented fine-tuning. These patterns show that causal fidelity is a multi-dimensional capability, where similar overall scores may reflect qualitatively different physics profiles.

Table 9: Per-case Track 2 dimension activation for all **50** cases that receive Track 2 scoring. Columns: F = Fluid & Smoke, C = Collision & Clipping, S = Surface Tracks, D = Deformation & Destruction, W = Wind & Environmental Forces, R = Reflection & Lighting, H = Human Motion & Expression. The final row aggregates the activation count of each dimension over all 50 cases, summing to 153 dimension-case pairs (mean ≈ 3.06 dimensions per case).

ID	Scene description	F	C	S	D	W	R	H	#Dims
1	blocky game world, Mario jumping over pipes and blocks	✗	✓	✗	✗	✗	✗	✓	2
2	foggy British street at night, handheld mockumentary view	✗	✓	✗	✗	✗	✓	✓	3
3	rainy anime city street, young woman walking amid neon	✓	✓	✓	✗	✗	✓	✓	5
4	realistic soccer pitch, player tracked in third person	✗	✓	✗	✗	✗	✗	✓	2
5	crowded anime subway interior, camera swaying with train	✗	✓	✗	✗	✗	✓	✓	3
6	living room with clothing rack, heated iron on table	✗	✓	✗	✓	✗	✓	✗	3
7	busy urban street, dark rain cloud above pedestrians	✗	✓	✓	✗	✗	✓	✓	4
8	sterile chemistry lab, glassware and pressure vial under lights	✗	✓	✗	✗	✗	✓	✓	3
9	Olympic red track, athlete approaching hurdles in lane	✗	✓	✗	✗	✗	✗	✓	2

Table 9: Per-case Track 2 dimension activation (continued).

ID	Scene description	F	C	S	D	W	R	H	#Dims
10	Antarctic ice sheet, emperor penguin walking on snow	✗	✗	✓	✗	✓	✗	✓	3
11	Tibetan monastery courtyard at golden hour, maroon-robed monk	✗	✓	✗	✗	✓	✓	✓	4
12	autumn park with fallen leaves, toddler in red running	✗	✓	✗	✗	✓	✗	✓	3
13	African savanna at sunset, elephant on dusty trail	✓	✗	✓	✗	✓	✓	✗	4
14	white office desk, small plush keychain with green accents	✗	✓	✗	✓	✗	✗	✗	2
15	Mediterranean stone plaza, young man walking in sunlight	✗	✓	✗	✗	✗	✓	✓	3
16	mountain lake shore at dawn, hiker with green backpack	✓	✓	✓	✗	✗	✓	✓	5
17	muddy rural road after rain, dark green off-road jeep	✓	✗	✓	✗	✗	✓	✗	3
18	African savanna trail, chestnut horse with safari rider	✗	✓	✓	✗	✓	✗	✓	4
19	misty Finnish lake at dawn, man rowing wooden boat	✓	✓	✗	✗	✗	✓	✓	4
20	wooden beach boardwalk at golden hour, young man walking	✓	✗	✗	✗	✓	✓	✓	4
21	city park walkway, woman in red jacket walking in sun	✗	✓	✗	✗	✗	✓	✓	3
22	seaside stone promenade, young skateboarder with board in hand	✓	✓	✗	✗	✗	✓	✓	4
23	tree-lined campus pathway, college student jogging in sun	✗	✓	✗	✗	✗	✗	✓	2
24	desert highway shoulder at sunset, backpacker walking	✗	✗	✓	✗	✗	✗	✓	2
25	wide snowy field in winter, person on fat-tire bicycle	✗	✗	✓	✗	✗	✗	✓	2
26	hot desert midday, monitor lizard crawling on golden sand	✗	✗	✓	✗	✗	✗	✗	1
27	outdoor basketball court, orange ball bouncing on concrete	✗	✓	✗	✗	✗	✓	✗	2
28	subway platform, FPP view of vending machine by tiled wall	✗	✓	✗	✗	✗	✓	✓	3
29	cozy antique bookstore, brass globe on wooden stand	✗	✓	✗	✗	✗	✓	✓	3
30	bustling Asian night market, red lanterns and moving crowd	✓	✓	✗	✗	✗	✓	✓	4
31	science classroom interior, human skeleton model near blackboard	✗	✓	✗	✗	✗	✓	✓	3
32	coastal city aerial view, FPP seabird flight with beak	✓	✓	✗	✗	✗	✗	✗	2
33	abandoned warehouse, racing quadcopter flying between pillars	✗	✓	✗	✗	✗	✓	✗	2
34	countryside farmland, colorful hot air balloon rising	✗	✗	✗	✗	✗	✗	✓	1
35	calm ocean at sunset, white seagull diving above water	✓	✗	✗	✗	✗	✓	✗	2
36	supermarket aisle, tall shelves with stacked red cans	✗	✓	✗	✗	✗	✓	✓	3
37	cherry blossom garden in spring, viewer with camera viewfinder	✗	✓	✗	✗	✓	✗	✓	3
38	war-torn city ruins, FPP view holding military grenade	✓	✓	✗	✓	✓	✗	✓	5
39	misty lakeside at dawn, viewer holding bamboo fishing rod	✓	✓	✗	✗	✗	✓	✓	4
40	sunlit artist studio, viewer holding paintbrush near canvas	✗	✓	✗	✗	✗	✓	✓	3
41	snowy tundra under aurora borealis, green purple lights	✗	✓	✓	✗	✗	✓	✗	3
42	floating sky island anime, fairy girl with dragonfly wings	✓	✗	✗	✗	✓	✓	✓	4
43	lavender field at golden hour, woman in white, oil-painting style	✗	✗	✓	✗	✓	✓	✓	4
44	magical toy workshop, wooden puppet walking among stations	✗	✓	✗	✗	✗	✓	✓	3
45	underwater ancient ruins (CG), mermaid among marble columns	✓	✓	✓	✗	✗	✓	✓	5
46	moonlit rooftops, ink-wash ninja running across tiles	✗	✓	✗	✗	✗	✗	✓	2
47	Middle Eastern bazaar (flat), fruit merchant pushing cart	✗	✓	✗	✗	✗	✗	✓	2
48	surreal melting landscape, brass robot among dripping clocks	✗	✓	✓	✓	✗	✗	✓	4
49	spooky haunted garden, glowing ghost dog among jack-o-lanterns	✗	✓	✗	✗	✗	✓	✗	2
50	volcanic crater edge (CG), dragon hatchling flying along rim	✓	✗	✗	✗	✓	✓	✓	4
Total	**(50 cases)**	15	38	14	4	11	32	39	153

C.6.2 VISUAL PLAUSIBILITY

Definition. Visual plausibility measures whether a generated video is visually realistic and physically reasonable. It focuses on the overall appearance quality, temporal consistency, object structure, motion

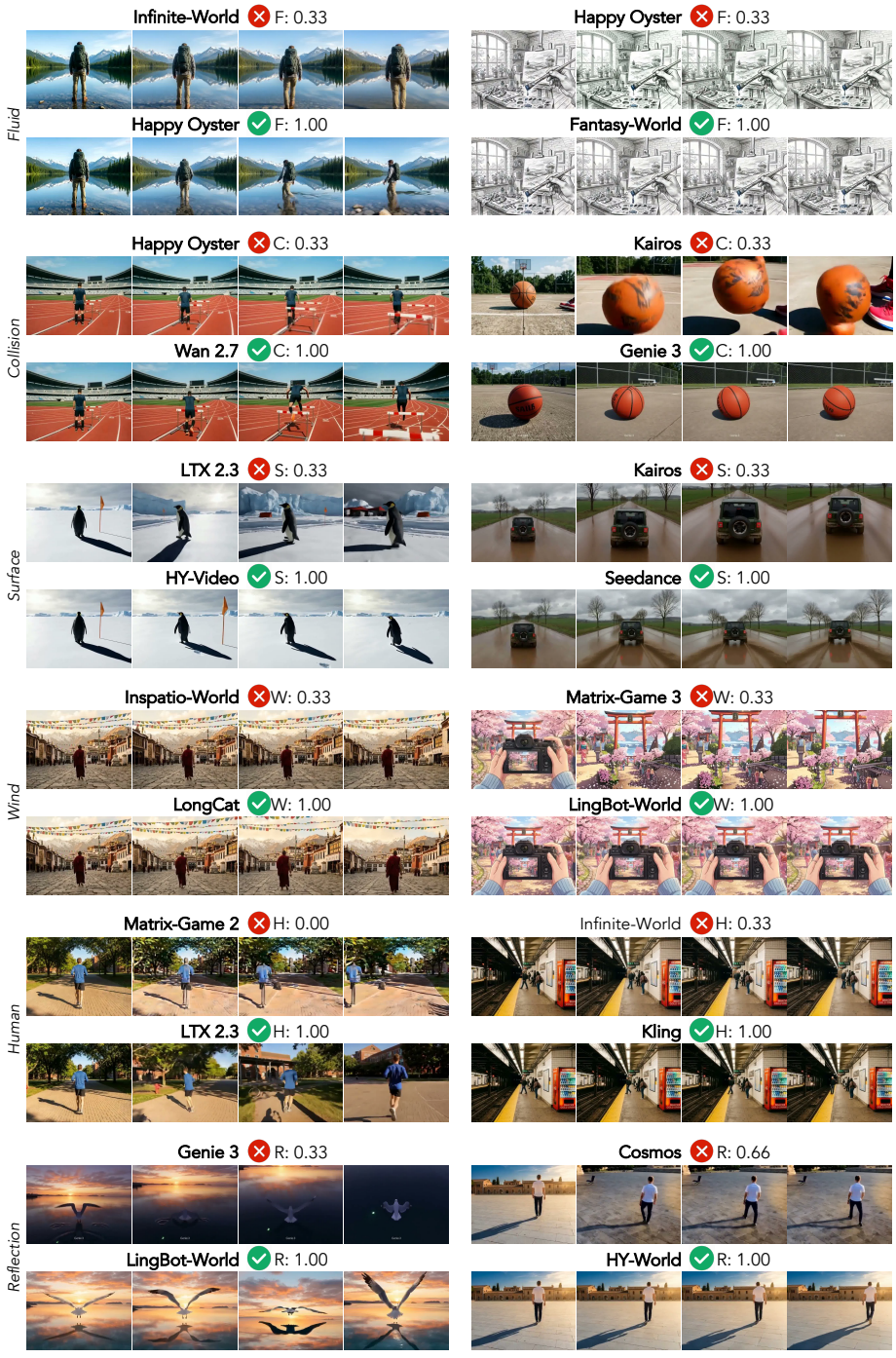


Figure 21: Qualitative comparisons on physics compliance.

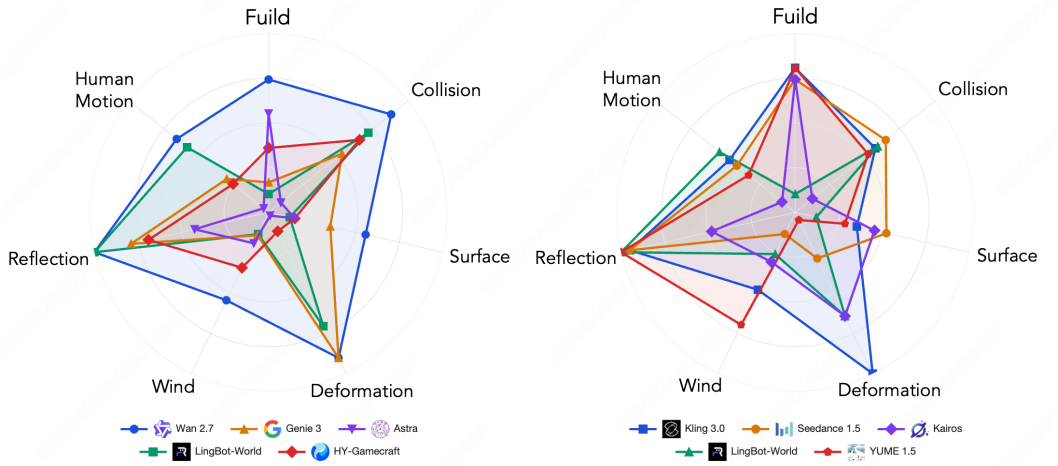


Figure 22: Causal Fidelity decomposed along the seven Track 2 sub-dimensions. *Left:* a top-to-bottom tier across paradigms. *Right:* models at a similar overall tier that trade off strengths across sub-dimensions. Each axis is rescaled to its own visible range to make the per-sub-dimension differences easier to read.

coherence, and whether the depicted dynamics conform to common-sense physical principles. We evaluate visual plausibility using a fine-tuned Qwen3-VL-30B-A3B model.

Prompt. All videos are evaluated with a fixed prompt. The input video is sampled at 2 FPS, and the maximum number of pixels per frame is set to 602,112. The prompt used for both training and inference is shown below.

Visual Plausibility — Fine-tuned Qwen3-VL-30B-A3B

```
Suppose you are an expert in judging and evaluating the quality of AI-generated videos,
please watch the above provided video and give scores for the video's truthfulness and
rationality, i.e., whether the video's overall appearance and motion are consistent with
our common-sense and physical principles.

Your rating should be chosen from the following five categories:  Perfect, Good, Fair,
Poor, and Bad.  Now please rate this video:
```

Training Data and Annotation. We train the visual plausibility scorer on an internal set of approximately 7K videos from open-source and closed-source generators, covering photorealistic and stylized content. Eight expert annotators follow shared rating guidelines. Each video receives three ratings on a 1–5 scale, and their mean is used as the training target.

Score Prediction and Optimization. We conduct full-parameter fine-tuning upon the pre-trained Qwen3-VL-30B-A3B checkpoint to regress the human-annotated visual plausibility score. Given a video and the above prompt, we extract the next-token prediction distribution at the final prompt token. Let the five rating tokens be $\{\texttt{Perfect}, \texttt{Good}, \texttt{Fair}, \texttt{Poor}, \texttt{Bad}\}$, corresponding to scores $\{5, 4, 3, 2, 1\}$, respectively. We take the model probabilities assigned to these five tokens and renormalize them over the rating set:

$$\tilde{p}_c = \frac{p_c}{\sum_{c' \in \mathcal{C}} p_{c'}}, \quad \mathcal{C} = \{\texttt{Perfect}, \texttt{Good}, \texttt{Fair}, \texttt{Poor}, \texttt{Bad}\}.$$

The predicted visual plausibility score is then computed as the expectation over the five ordered categories:

$$\hat{s} = 5\tilde{p}_{\texttt{Perfect}} + 4\tilde{p}_{\texttt{Good}} + 3\tilde{p}_{\texttt{Fair}} + 2\tilde{p}_{\texttt{Poor}} + 1\tilde{p}_{\texttt{Bad}}.$$

Given the ground-truth human score s, the model is optimized with mean squared error:

$$\mathcal{L}_{\text{VP}} = (\hat{s} - s)^2.$$

This formulation preserves the ordinal structure of the five rating categories while allowing the model to produce a continuous visual plausibility score aligned with averaged human judgments.

For reporting, the implementation uses $S_{\mathrm{VP}} = 100\hat{s}/5$. Thus the five-category expectation has a theoretical reporting range of $[20, 100]$, with no subtraction of the minimum category score.

Held-Out Validation and Overlap. The held-out set contains 623 videos from ten text-to-video generators and is disjoint from training at both the video and prompt levels. Pearson correlation with human ratings is 0.92 across generator means and 0.47–0.71 within individual generators. These results support aggregate model comparisons, while indicating limited precision for ranking individual videos. No videos evaluated in WBench are included in scorer training, although training generators and evaluated models share some model families and backbones. Thus, video-level separation does not imply generalization to unseen model families.

Coverage and Limitations. Both splits include photorealistic and stylized content. This coverage alone does not establish comparable scoring accuracy across styles. Separate validation by rendering style and model-family overlap, as well as validation on image-conditioned and interactive outputs, is needed to assess these forms of generalization.

Availability. The fine-tuned checkpoint and the scoring protocol above are public. Training and held-out annotations remain internal, so the release supports scorer inference but not independent reproduction of training or held-out validation. Table 22 lists the released resources.

D Additional Experimental Results

D.1 Full Split Results on Text-Driven Models

Table 10 reports the complete per-sub-metric scores for all nine text-driven models evaluated on the **full test set** (289 cases). Unlike Table 2, which is restricted to the navigation subset shared across paradigms, this table includes all interaction types (navigation, event editing, subject action, and perspective switching) and uses the full-split consistency metrics (gated spatial consistency, geometric consistency, and photometric consistency).

Several patterns emerge. Kling 3.0 and Wan 2.7 achieve the strongest interaction adherence, particularly in event editing and subject action, while perspective switching remains universally difficult (average 30.7). LongCat-Video and HunyuanVideo lead consistency metrics but exhibit limited dynamic degree, confirming a trade-off between motion magnitude and temporal stability. Wan 2.7 leads causal fidelity by a wide margin (83.3 *vs.* the 74.7 average), while visual plausibility scores cluster within a narrow 4.6-point range across all models. No single text-driven model dominates across all evaluation axes.

Table 10: Full per-sub-metric results for text-driven models on the **full test set** (289 cases). All scores $\in [0, 100]$, higher is better. **Bold** = best, underline = second best per row.

	Metrics	Seedance 1.5	Wan 2.7	Kling 3.0	YUME 1.5	HY-Video 1.5	LTX 2.3	LongCat-Video	Kairos	Cosmos 2.5	*Average*
Video Quality	Aesthetic	59.7	59.6	61.3	59.3	61.9	56.9	**64.7**	58.4	60.1	60.2
	Imaging	69.8	68.1	67.7	65.7	67.4	62.3	**69.8**	63.6	67.2	66.8
	Flickering	93.4	93.0	94.5	94.8	95.5	94.1	94.9	**96.3**	96.0	94.7
	Dynamic	98.3	**99.3**	89.9	86.1	68.8	94.4	59.7	63.5	42.4	78.0
	Smoothness	97.6	96.6	97.9	97.7	**98.8**	96.8	97.7	97.9	98.3	97.7
	HPSv3-Norm	72.9	69.4	68.8	62.0	67.5	57.7	**76.3**	58.8	65.9	66.6
	Average	**82.0**	81.0	80.0	77.6	76.6	77.1	77.2	73.1	71.7	77.4
Setting	Scene	71.6	88.3	**89.0**	53.1	77.6	81.3	53.1	52.2	72.4	71.0
	Subject	94.3	**94.6**	92.9	91.7	93.6	89.2	91.5	88.5	94.3	92.3
	Average	82.9	**91.5**	91.0	72.4	85.6	85.2	72.3	70.3	83.3	81.6
Interaction	Navigation	66.5	64.7	69.6	**72.0**	71.8	66.0	62.7	64.0	62.9	66.7
	Event Editing	80.4	**84.0**	81.4	57.8	63.8	53.0	50.4	46.8	48.2	62.9
	Subject Action	80.0	83.4	**85.6**	47.0	55.6	51.8	48.4	41.4	41.6	59.4
	Persp. Switching	45.0	**55.0**	**55.0**	16.7	27.6	25.0	18.3	13.3	20.0	30.7
	Average	68.0	71.8	**72.9**	48.4	54.7	49.0	44.9	41.4	43.2	54.9
Consistency	Background	89.6	89.5	92.7	92.0	92.4	89.3	**94.7**	91.8	92.4	91.6
	Spatial	72.7	71.0	75.3	71.5	79.2	70.2	**83.3**	76.8	78.1	75.3
	Gated Spatial	72.4	71.0	**75.1**	71.4	**75.1**	70.2	66.2	62.0	74.3	70.9
	Perspective	62.7	62.2	76.8	48.0	**86.6**	69.8	81.5	76.3	84.3	72.0
	Segment	92.4	65.6	92.7	**99.3**	**99.3**	77.8	98.6	94.1	93.1	90.3
	Geometric	83.6	82.6	89.4	91.1	94.4	81.1	**94.7**	91.5	94.2	89.2
	Photometric	76.7	75.5	80.4	**84.1**	81.4	79.4	81.5	82.1	82.1	80.4
	Subject Cross	89.3	88.7	88.5	89.4	91.5	86.7	**92.4**	90.7	91.8	89.9
	Average	79.9	75.8	83.9	80.9	**87.5**	78.1	86.6	83.2	86.3	82.4
Physical	Causal Fidelity	76.0	**83.3**	78.0	72.7	75.0	74.0	76.0	62.7	74.7	74.7
	Visual Plausibility	60.5	59.8	60.4	58.1	59.3	56.3	**60.8**	58.2	59.3	59.2
	Average	68.2	**71.6**	69.2	65.4	67.1	65.1	68.4	60.5	67.0	67.0

D.2 Human-Preference Annotation Platform and Protocol

Platform and task design. We use a web-based platform for blind pairwise comparisons of model outputs on the same benchmark case (Fig. 23). Annotators view synchronized videos and select *A better*, *B better*, *Tie*, or *Discard*. Video order is randomized, and dimension-specific instructions guide annotators to assess the target capability independently of unrelated quality signals. Model coverage depends on the evaluation dimension and supported control interface.

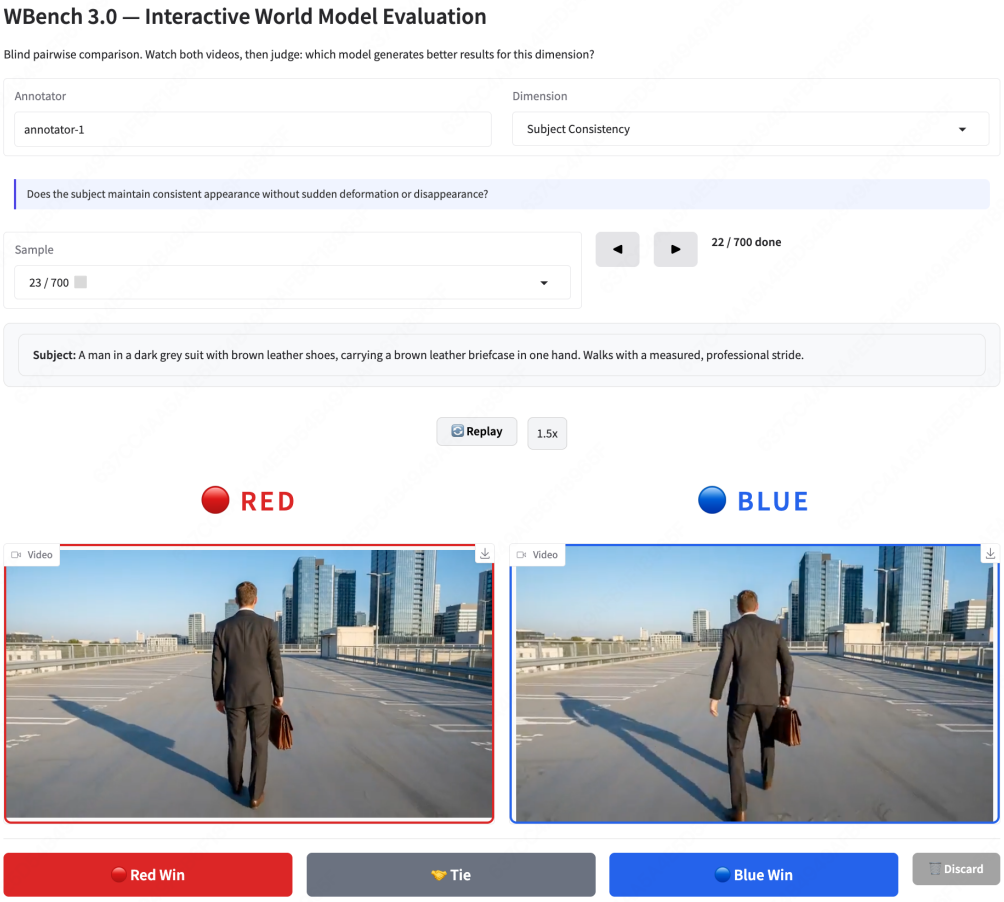


Figure 23: Human-preference annotation platform. Each task is a blind pairwise comparison between two models, with the annotator choosing among *A better*, *B better*, *Tie*, or *Discard*.

Annotators and quality control. The study involves a pool of 400 annotators who receive task-specific training and complete practice comparisons before annotation. Each comparison is independently assessed by three annotators, with the final preference determined by majority vote. Gold-standard checks and expert review of sampled annotations are used to identify unreliable judgments and request re-annotation where needed.

Study scope. The study comprises 13,515 pairwise comparison tasks across 10 evaluation dimensions. Model-level human–metric alignment is reported in Fig. 5, and case-level agreement is examined in Appendix D.6.

D.3 Long-Horizon Navigation Pilot

To examine longer interaction horizons, we conduct a supplementary navigation experiment on long-horizon variants of two source scenarios, a third-person campus scene with lateral translation and a first-person cathedral scene with yaw rotation. These sequences are evaluated separately from the benchmark test set. Each scenario uses ten sequence lengths from 2 to 20 turns, with k leftward

actions followed by k rightward actions to return to the starting position or orientation. We hold the initial frame and world setting constant across sequence lengths within each scenario. Two camera-controlled models, LingBot-World-2-Fast and HY-World 1.5, generate videos for all 20 settings, and we evaluate them on the available metrics.

Tables 11 and 12 report scores at all ten horizons. Slopes are fitted by ordinary least squares against the actual turn counts $2, 4, \ldots, 20$. Both tables report 19 sub-metrics on a 0–100 scale. Dashes indicate inapplicable or unavailable scores. The three semantic interaction metrics are omitted because they are inapplicable to these navigation-only sequences. Causal Fidelity results are incomplete in the archived evaluations, so we report only available scores. Slopes are reported only for complete ten-horizon series. Three observations emerge:

Model ranking can invert with horizon. Averaged over the two scenarios, HY-World 1.5 leads on NavScore at t2 (91.95 vs. 87.15), but its fitted decline is steeper (-1.47 vs. -0.52 points per turn). At t20, LingBot-World-2-Fast leads (70.7 vs. 62.4), and it has higher mean navigation scores at every tested horizon beyond t10. The reversal shows that the relative navigation performance of these two models depends on the evaluated horizon.

Degradation varies across metrics and models. From t2 to t20, geometric consistency averaged over the two scenarios drops by 22.6 points for LingBot-World-2-Fast and 27.0 for HY-World 1.5, while gated spatial consistency drops by 14.35 and 33.8 points, respectively. Motion smoothness changes by at most 2.5 points in absolute value across the four model–scenario combinations. Aesthetic quality changes by less than one point in each scenario for LingBot-World-2-Fast, but drops by 17.7 and 14.3 points for HY-World 1.5. These results show that long-horizon navigation and consistency degradation can coexist with model-dependent losses in visual quality.

Existing weaknesses can widen. In the campus scenario, HY-World 1.5 has lower perspective consistency at t2 than LingBot-World-2-Fast (38.9 vs. 88.1), and the gap widens at t20 (22.3 vs. 80.9). This case illustrates how longer sequences can expose persistent differences in a specific capability. The pilot is limited to these two roundtrip navigation scenarios and does not establish long-horizon robustness across other settings or interaction types.

From an operational standpoint, extending an existing case to more turns requires only appending additional interaction entries to its JSON specification, and the evaluation pipeline processes them without modification. The primary practical constraints are the sequential generation cost per turn and the human annotation cost for dimension-specific labels.

D.4 Reproducibility Analysis

We assess reproducibility through repeated generation and evaluation, distinguishing changes in generated videos from changes in scores when the same videos are re-evaluated. For every case and model, we generate two samples and evaluate each sample twice through the same pipeline. The final score for each case and metric is the average of the four scores. We examine variability in detail for two representative models, LongCat-Video and HY-World 1.5. Comparisons within a generation run isolate evaluation variability, while comparisons across generation runs additionally reflect sampling variability.

Table 13 reports the LongCat-Video results. For each metric, we compute the range across the four aggregate scores, and the mean of these ranges over 22 metrics is 0.57 points on a 0–100 scale. Non-VLM metrics and the fine-tuned Visual Plausibility scorer return identical scores within each generation run in these repetitions. For the six API-based VLM metrics, the larger of the two within-run absolute differences ranges from 0.48 to 1.10 points. HY-World 1.5 has a mean per-metric range of 0.87 points. These results indicate limited aggregate-score variation in the tested repetitions. They do not establish identical case-level scores or stable rankings for closely matched models, so small score differences should be interpreted cautiously.

D.5 Case Construction and Quality Control

The 289 cases in WBench are constructed through a multi-stage pipeline designed to balance diversity, controllability, and visual quality.

Table 11: Long-horizon navigation on the campus scenario, third-person lateral translation. Δ is t20 minus t2. Slope is measured in score points per turn. L = LingBot-World-2-Fast and H = HY-World 1.5.

Case 26: Campus

Metric	Model	t2	t4	t6	t8	t10	t12	t14	t16	t18	t20	Δ	Slope
Aesthetic Quality	L	61.0	61.0	62.3	62.3	62.5	59.3	62.9	61.3	59.6	61.4	+0.4	−0.04
	H	57.3	53.5	57.3	57.9	53.3	48.8	47.7	43.8	44.2	39.6	−17.7	−0.99
Imaging Quality	L	68.7	65.3	66.6	66.8	66.0	61.6	66.9	65.1	64.5	64.6	−4.1	−0.16
	H	66.4	67.7	64.5	63.7	59.3	55.3	56.1	50.4	54.7	49.9	−16.5	−1.02
Temporal Flickering	L	88.4	88.8	89.4	88.0	90.0	89.7	90.1	89.5	89.3	90.8	+2.4	+0.10
	H	94.2	93.6	94.6	93.7	93.6	93.4	92.2	92.3	90.3	89.9	−4.3	−0.24
Dynamic Degree	L	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	+0.0	+0.00
	H	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	+0.0	+0.00
Motion Smoothness	L	93.7	94.8	95.4	94.9	95.6	95.2	95.5	95.5	95.6	96.2	+2.5	+0.09
	H	98.3	98.1	98.3	97.8	97.6	97.9	97.5	97.0	97.0	95.9	−2.4	−0.11
HPSv3-Norm	L	62.0	61.8	65.9	62.9	67.8	64.6	66.5	61.7	63.9	64.9	+2.9	+0.08
	H	72.4	61.0	65.4	59.5	53.1	39.7	38.0	29.7	32.9	24.0	−48.4	−2.69
Scene Adherence	L	100.0	100.0	100.0	100.0	90.0	100.0	100.0	90.0	100.0	100.0	+0.0	−0.12
	H	50.0	100.0	50.0	50.0	100.0	50.0	50.0	100.0	100.0	50.0	+0.0	+0.61
Subject Adherence	L	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	50.0	−50.0	−1.36
	H	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	+0.0	+0.00
Navigation Score	L	81.3	67.2	69.9	65.9	52.0	75.2	70.6	64.9	65.6	54.1	−27.2	−0.74
	H	85.8	84.1	82.5	82.9	79.1	60.6	66.1	69.7	71.8	50.2	−35.6	−1.63
Background Cons.	L	86.6	84.3	86.2	83.0	86.1	83.7	85.1	83.9	83.1	83.8	−2.8	−0.12
	H	91.2	88.9	87.5	84.5	83.5	81.0	79.5	78.2	78.4	76.5	−14.7	−0.82
Spatial Consistency	L	75.0	71.1	69.5	69.4	61.9	67.0	65.6	63.5	61.5	65.7	−9.3	−0.57
	H	80.2	78.6	74.7	78.8	69.2	62.4	56.4	58.5	52.1	53.2	−27.0	−1.77
Gated Spatial Cons.	L	75.0	71.1	69.5	69.4	61.9	67.0	65.6	63.5	61.5	65.7	−9.3	−0.57
	H	80.2	78.6	74.7	78.8	69.2	62.4	56.4	58.5	52.1	53.1	−27.1	−1.77
Perspective Cons.	L	88.1	87.5	82.7	81.4	81.3	78.6	81.5	84.9	80.9	80.9	−7.2	−0.31
	H	38.9	55.4	47.7	54.4	28.4	24.0	11.3	28.2	30.8	22.3	−16.6	−1.68
Segment Continuity	L	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	+0.0	+0.00
	H	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	+0.0	+0.00
Subject Cons.	L	82.4	82.5	82.5	79.3	85.0	79.1	82.1	81.7	82.1	80.0	−2.4	−0.08
	H	85.3	86.4	86.7	84.3	79.8	79.8	78.9	82.5	80.6	78.2	−7.1	−0.43
Geometric Cons.	L	95.7	91.3	81.9	75.6	75.5	65.4	75.8	78.5	79.5	67.1	−28.6	−1.11
	H	98.4	97.1	96.3	95.3	68.4	72.0	61.1	27.7	43.0	53.4	−45.0	−3.71
Photometric Cons.	L	78.4	75.8	74.5	71.9	73.9	71.9	72.9	71.8	74.9	71.8	−6.6	−0.24
	H	83.9	81.9	82.2	79.0	77.7	76.4	74.4	74.3	69.4	71.3	−12.6	−0.77
Causal Fidelity	L	33.3	50.0	33.3	50.0	33.3	41.7	–	–	–	–	–	–
	H	100.0	58.3	58.3	58.3	75.0	75.0	58.3	–	–	–	–	–
Visual Plausibility	L	58.0	57.2	56.8	54.4	57.8	56.9	58.3	60.1	56.0	58.0	+0.0	+0.06
	H	59.5	53.7	56.7	54.8	54.0	52.5	54.1	51.6	50.1	51.3	−8.2	−0.39

Case design and sampling. We define candidate cases using the taxonomy of world settings and interaction types in Section 3. Each candidate specifies a scene, a perspective, an optional subject, and an ordered interaction sequence. Manual screening checks whether the specification is coherent and whether its interactions are meaningful within the declared setting.

Keyframe generation and selection. Initial frames are generated or collected as described in Section 3. Manual inspection checks image validity, agreement with the visible scene and subject specification, and suitability as the starting state for the interaction sequence. Offscreen elements are

Table 12: Long-horizon navigation on the cathedral scenario, first-person yaw rotation. Δ is t20 minus t2. Slope is measured in score points per turn. L = LingBot-World-2-Fast and H = HY-World 1.5.

Case 73: Cathedral

Metric	Model	t2	t4	t6	t8	t10	t12	t14	t16	t18	t20	Δ	Slope
Aesthetic Quality	L	66.4	64.9	62.7	65.1	63.4	64.8	64.0	63.7	64.3	65.5	−0.9	−0.03
	H	62.1	52.9	47.5	46.8	51.7	48.7	51.2	50.9	47.9	47.8	−14.3	−0.41
Imaging Quality	L	72.6	72.8	71.3	70.3	70.8	71.0	68.5	68.1	68.1	68.6	−4.0	−0.27
	H	71.1	64.1	64.8	54.7	63.4	63.2	64.0	64.5	58.9	61.4	−9.7	−0.30
Temporal Flickering	L	81.7	82.4	83.9	85.6	85.5	85.2	86.8	87.8	88.4	88.3	+6.6	+0.38
	H	83.1	85.7	83.6	90.7	87.8	86.6	87.5	86.2	87.7	87.1	+4.0	+0.16
Dynamic Degree	L	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	+0.0	+0.00
	H	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	+0.0	+0.00
Motion Smoothness	L	95.5	95.7	95.8	96.9	96.3	96.6	97.1	97.3	97.4	97.5	+2.0	+0.12
	H	97.2	97.8	97.0	98.3	98.0	97.4	97.9	97.6	98.0	97.6	+0.4	+0.02
HPSv3-Norm	L	86.1	84.3	80.3	81.7	78.6	80.5	79.8	73.6	78.2	77.9	−8.2	−0.47
	H	71.8	50.4	32.7	21.0	47.3	36.9	47.8	42.6	28.5	30.7	−41.1	−1.22
Scene Adherence	L	100.0	50.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	+0.0	+1.06
	H	50.0	100.0	100.0	100.0	100.0	50.0	50.0	50.0	100.0	50.0	+0.0	−1.36
Subject Adherence	L	–	–	–	–	–	–	–	–	–	–	–	–
	H	–	–	–	–	–	–	–	–	–	–	–	–
Navigation Score	L	93.0	95.5	93.8	90.6	89.6	93.4	91.6	90.7	89.4	87.3	−5.7	−0.31
	H	98.1	97.5	92.1	84.8	95.2	86.7	83.8	82.3	74.6	74.6	−23.5	−1.31
Background Cons.	L	96.3	94.3	91.2	91.4	89.8	91.0	89.2	89.6	89.4	89.5	−6.8	−0.33
	H	94.1	89.6	83.0	83.8	85.8	84.9	87.4	87.7	83.6	83.7	−10.4	−0.31
Spatial Consistency	L	96.3	90.8	80.8	77.1	75.2	70.7	76.1	63.5	72.0	70.4	−25.9	−1.39
	H	97.7	96.9	95.8	95.4	79.7	74.5	77.8	77.4	57.1	57.2	−40.5	−2.40
Gated Spatial Cons.	L	89.7	90.8	80.8	77.1	75.2	70.7	76.1	63.5	72.0	70.3	−19.4	−1.21
	H	97.7	96.9	95.8	95.3	79.7	74.5	77.8	77.4	57.1	57.2	−40.5	−2.40
Perspective Cons.	L	–	–	–	–	–	–	–	–	–	–	–	–
	H	–	–	–	–	–	–	–	–	–	–	–	–
Segment Continuity	L	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	+0.0	+0.00
	H	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	+0.0	+0.00
Subject Cons.	L	–	–	–	–	–	–	–	–	–	–	–	–
	H	–	–	–	–	–	–	–	–	–	–	–	–
Geometric Cons.	L	97.4	96.3	94.8	94.4	91.3	88.0	88.8	86.2	80.1	80.8	−16.6	−0.99
	H	95.5	94.7	93.3	89.1	87.8	71.8	91.3	87.6	62.8	86.5	−9.0	−1.04
Photometric Cons.	L	85.2	80.9	77.6	77.7	73.9	70.4	75.6	75.6	73.5	74.8	−10.4	−0.50
	H	79.1	80.3	74.5	85.7	76.9	72.7	78.5	75.8	74.9	75.0	−4.1	−0.28
Causal Fidelity	L	–	–	–	–	–	–	–	–	–	–	–	–
	H	–	–	–	–	–	–	–	–	–	–	–	–
Visual Plausibility	L	65.1	64.9	62.1	61.3	62.5	61.9	58.4	59.9	62.9	60.9	−4.2	−0.22
	H	55.7	52.4	53.2	51.2	49.7	49.2	49.0	49.6	51.9	50.0	−5.7	−0.24

checked against their intended role in subsequent interactions rather than required to appear in the initial frame.

Candidate review and final selection. Twenty contributors produce 1,000 candidate cases in the first stage. Each case is assigned to three reviewers, and reviewers do not review cases they authored. Reviewers assess image–specification agreement, clarity of the requested interactions, and coherence across turns. A candidate enters the pool if more than half of its reviewers approve it, which requires at least two approvals out of three. This screening yields a pool of 598 cases, corresponding to an

Table 13: Reproducibility of WBench evaluation scores on LongCat-Video. Two generation runs (R1, R2), each evaluated twice (E1, E2). Scores are aggregated over cases. Max $|\Delta|$ is the largest minus the smallest of the four scores for each metric, and Mean averages this range over the 22 metrics.

Metric	**Run 1**		**Run 2**		**Max $\|\Delta\|$**
	Eval 1	**Eval 2**	**Eval 1**	**Eval 2**	
Aesthetic Quality	66.42	66.42	66.59	66.59	0.17
Imaging Quality	69.89	69.89	69.32	69.32	0.57
Temporal Flickering	94.94	94.94	94.67	94.67	0.27
Dynamic Degree	46.07	46.07	45.74	45.74	0.33
Motion Smoothness	97.76	97.76	98.04	98.04	0.28
HPSv3-Norm	77.50	77.50	77.71	77.71	0.21
Scene Adherence	53.30	53.93	52.64	52.52	1.41
Subject Adherence	91.91	91.16	91.46	91.47	0.75
Navigation Score	62.99	62.99	62.42	62.42	0.57
Event Editing	50.38	49.90	50.62	50.69	0.79
Subject Action	48.32	47.87	48.46	48.96	1.09
Perspective Switching	18.21	17.67	18.55	18.77	1.10
Background Consistency	95.20	95.20	95.01	95.01	0.19
Spatial Consistency	83.51	83.51	83.10	83.10	0.41
Gated Spatial Consistency	65.70	65.70	66.71	66.71	1.01
Perspective Consistency	81.75	81.75	81.26	81.26	0.49
Segment Continuity	99.50	99.50	99.30	99.30	0.20
Subject Consistency	93.64	93.64	93.16	93.16	0.48
Geometric Consistency	95.19	95.19	95.61	95.61	0.42
Photometric Consistency	82.14	82.14	82.26	82.26	0.12
Causal Fidelity	75.28	76.38	76.11	76.25	1.10
Visual Plausibility	61.52	61.52	62.09	62.09	0.57
Mean					**0.57**

acceptance rate of 59.8%. The counts of candidates receiving zero, one, two, or three approvals are reported in Table 14.

From the pool of 598 cases, final selection balances coverage across world settings and interaction types, maintains diversity among the selected cases, and considers overall case quality, yielding the released set of 289 cases. The remaining 309 candidates are not selected under these combined criteria and should not all be interpreted as defective cases. These procedures document quality control during dataset construction. Table 14 summarizes the review and selection counts.

Table 14: Construction review records.

Item	**Value**
Case contributors	20
First-stage candidates reviewed	1,000
Reviewers per candidate	3
Minimum approvals required	2
Candidates with 0 / 1 / 2 / 3 approvals	111 / 291 / 181 / 417
Candidates admitted to the pool	598
First-stage acceptance rate	59.8%
Final selected cases	289
Pool candidates not selected	309

D.6 Case-Level Human-Metric Agreement

To complement the main-text human-preference study in Fig. 5, we conducted an additional case-level human annotation study covering 15 sub-metrics spanning setting adherence, interaction, consistency, and physical evaluation. The six video-quality sub-metrics and Background Consistency are omitted here, as related metrics have been studied in VBench (Huang et al., 2024) and other prior work.

Protocol and reporting. The analysis includes Seedance, Kling, Happy Oyster, HY-World, Matrix-Game 3.0, LingBot-World, YUME, and LongCat-Video. Each comparison concerns a pair of videos generated for the same benchmark case. Human–metric agreement is the proportion of distinct matched video pairs for which human judgments and automatic scores favor the same video. For each sub-metric, we retained 1,000 valid video pairs after excluding human ties, unavailable metric scores, and equal automatic scores. We estimate 95% confidence intervals by bootstrap resampling the valid video pairs. These intervals do not account for dependence among pairs from the same case.

Coverage and interpretation. Table 15 indicates reasonable agreement between automatic scores and human judgments on the evaluated cases, complementing the model-ranking analysis. These results provide additional support for using the metrics to compare model performance across the evaluated aspects of interactive world modeling.

Table 15: Case-level human-metric agreement. Agreement is computed over distinct matched video pairs. Agreement is reported as a percentage with its 95% confidence interval.

Metric	Human–metric agreement [95% CI]
Scene Adherence	72.0% [69.6–74.2%]
Subject Adherence	78.9% [75.6–82.0%]
Navigation Score	83.4% [81.0–85.7%]
Event Editing	70.5% [66.7–74.0%]
Subject Action	81.4% [78.1–84.2%]
Perspective Switching	80.5% [78.1–82.7%]
Spatial Consistency	70.0% [67.6–72.3%]
Gated Spatial Consistency	88.0% [86.2–89.6%]
Perspective Consistency	79.3% [76.9–81.7%]
Segment Continuity	94.5% [91.9–97.0%]
Subject Consistency	74.3% [71.7–76.8%]
Geometric Consistency	73.2% [71.3–74.9%]
Photometric Consistency	71.3% [69.3–73.2%]
Causal Fidelity	67.3% [64.6–69.9%]
Visual Plausibility	71.7% [69.1–74.2%]

D.7 Confidence Intervals for Model Scores

Tables 16 to 21 report scores and marginal 95% confidence intervals for 20 models and 22 sub-metrics. Non-navigation interaction metrics are evaluated only for the nine text-driven models.

The bootstrap procedure uses 10,000 case-level resamples, drawn separately within navigation and non-navigation cases and shared across models and metrics. Scores, including composite scores, are recomputed for each resample, and percentile intervals use the 2.5th and 97.5th percentiles. The intervals describe uncertainty from case sampling and do not constitute pairwise significance tests.

Table 16: Confidence intervals for 20 models. Navigation-split summary. Scores and marginal 95% intervals are estimated with a case-level bootstrap. See the protocol and source limitations in the accompanying text.

Model	Overall Navi	Video Quality	Setting	Navigation Consistency	Physical
Seedance 1.5	76.2 [75.4, 76.9]	82.1 [81.6, 82.6]	82.9 [80.5, 85.3]	66.5 [64.8, 68.1]	81.3 [80.3, 82.3]	68.3 [65.8, 70.9]
Wan 2.7	78.2 [77.3, 79.1]	81.5 [80.9, 82.1]	91.4 [88.8, 93.4]	64.7 [62.7, 66.8]	81.6 [80.5, 82.7]	71.8 [69.8, 73.7]
Kling 3.0	79.0 [78.1, 79.9]	81.4 [80.7, 82.1]	91.0 [88.6, 93.0]	69.6 [67.6, 71.5]	83.7 [82.6, 84.7]	69.3 [66.6, 71.8]
YUME 1.5	73.5 [72.5, 74.5]	77.6 [76.7, 78.4]	72.4 [70.0, 74.7]	72.0 [70.4, 73.6]	80.1 [79.0, 81.2]	65.2 [62.2, 68.2]
HY-Video 1.5	77.9 [77.0, 78.8]	77.6 [76.4, 78.8]	85.5 [82.6, 88.1]	71.8 [69.7, 73.9]	87.4 [86.7, 88.0]	67.3 [64.9, 69.7]
LTX 2.3	74.1 [73.2, 75.0]	77.1 [76.4, 77.8]	85.2 [82.5, 87.8]	66.0 [64.3, 67.8]	77.2 [75.7, 78.6]	64.8 [62.6, 67.1]
LongCat-Video	73.3 [72.3, 74.3]	75.4 [74.0, 76.8]	72.3 [69.9, 74.6]	62.7 [60.7, 64.6]	87.1 [86.2, 87.9]	68.9 [66.4, 71.3]
Kairos 3.0	70.3 [69.3, 71.3]	74.0 [72.9, 75.1]	70.3 [67.7, 72.8]	64.0 [62.2, 65.9]	82.6 [81.5, 83.7]	60.4 [57.7, 63.3]
Cosmos 2.5	74.6 [73.7, 75.4]	72.9 [71.6, 74.1]	83.3 [80.7, 85.8]	62.9 [61.0, 64.7]	86.5 [85.6, 87.3]	67.4 [65.3, 69.6]
LingBot-World	78.6 [77.7, 79.5]	78.9 [77.7, 80.1]	72.6 [70.7, 74.3]	80.2 [78.1, 82.2]	89.9 [89.0, 90.7]	71.2 [68.4, 73.9]
HY-World 1.5	78.2 [77.2, 79.1]	78.1 [77.1, 79.1]	72.2 [69.7, 74.5]	87.5 [85.4, 89.5]	86.9 [85.8, 87.9]	66.3 [63.5, 69.1]
Fantasy-World	73.8 [72.9, 74.7]	72.4 [71.4, 73.4]	71.2 [68.7, 73.5]	72.0 [70.4, 73.6]	86.4 [85.4, 87.4]	66.8 [64.1, 69.4]
InSpatio-World	73.9 [72.9, 74.9]	71.5 [70.5, 72.5]	71.4 [69.0, 73.7]	73.1 [71.2, 75.1]	88.4 [87.1, 89.6]	65.2 [62.2, 68.2]
Astra	63.7 [62.3, 65.1]	67.1 [65.8, 68.4]	59.7 [55.9, 63.3]	66.7 [65.1, 68.2]	73.3 [71.9, 74.8]	51.5 [47.7, 55.6]
Happy Oyster	76.8 [75.9, 77.8]	77.3 [76.4, 78.2]	74.2 [71.4, 76.9]	84.7 [82.9, 86.5]	84.3 [83.0, 85.6]	63.5 [60.7, 66.3]
Matrix-Game 3.0	71.3 [70.0, 72.5]	75.5 [74.9, 76.2]	63.7 [59.9, 67.4]	83.6 [82.2, 85.1]	74.5 [73.4, 75.6]	59.4 [56.6, 62.2]
Genie 3	74.0 [72.8, 75.1]	75.2 [74.2, 76.2]	72.5 [68.6, 76.0]	74.1 [72.0, 76.1]	82.6 [81.3, 83.8]	65.7 [63.5, 67.9]
Matrix-Game 2.0	68.7 [67.5, 69.8]	73.8 [72.6, 75.1]	67.2 [64.4, 69.9]	80.3 [78.4, 82.1]	65.1 [63.9, 66.4]	57.1 [53.9, 60.4]
HY-GameCraft	68.2 [67.1, 69.3]	73.0 [72.1, 73.8]	66.5 [63.1, 69.5]	66.3 [64.6, 67.9]	72.6 [71.9, 73.3]	62.4 [59.7, 65.1]
Infinite-World	72.8 [71.7, 73.9]	77.0 [76.0, 78.0]	69.2 [65.7, 72.3]	75.6 [73.9, 77.2]	79.9 [78.8, 81.1]	62.1 [59.2, 65.1]

Table 17: Video quality. Confidence intervals for 20 models on the navigation split. Values are score and 95% interval.

Model	Aesthetic	Imaging	Temporal Flickering	Dynamic	Motion Smoothness	HPSv3-Norm
Seedance 1.5	61.0 [59.8, 62.2]	69.3 [68.6, 70.0]	92.4 [92.0, 92.8]	99.4 [96.5, 99.9]	97.5 [97.3, 97.6]	73.0 [71.2, 74.8]
Wan 2.7	61.4 [60.0, 62.9]	68.0 [67.1, 68.9]	92.2 [91.8, 92.6]	100.0 [97.6, 100.0]	96.3 [96.1, 96.6]	71.1 [69.0, 73.3]
Kling 3.0	63.0 [61.6, 64.3]	68.1 [67.2, 69.0]	93.2 [92.7, 93.6]	97.5 [93.6, 99.0]	97.6 [97.5, 97.8]	69.1 [66.9, 71.2]

Table 17: Continued from the preceding page.

Model	Aesthetic	Imaging	Temporal Flickering	Dynamic	Motion Smoothness	HPSv3-Norm
YUME 1.5	58.7 [57.3, 60.2]	63.3 [62.1, 64.5]	93.0 [92.6, 93.4]	96.8 [92.8, 98.6]	97.0 [96.8, 97.2]	57.0 [54.3, 59.8]
HY-Video 1.5	63.4 [61.9, 64.9]	67.4 [66.5, 68.2]	94.2 [93.7, 94.7]	73.9 [66.5, 80.1]	98.7 [98.6, 98.8]	68.0 [65.9, 70.0]
LTX 2.3	57.9 [56.6, 59.1]	61.0 [59.9, 62.1]	93.2 [92.9, 93.5]	98.1 [94.5, 99.3]	96.4 [96.2, 96.6]	56.1 [53.9, 58.3]
LongCat-Video	66.5 [65.1, 67.9]	69.6 [68.8, 70.3]	94.8 [94.3, 95.2]	45.9 [38.3, 53.7]	97.9 [97.7, 98.1]	77.6 [74.8, 80.4]
Kairos 3.0	59.9 [58.1, 61.6]	62.7 [61.5, 63.9]	95.4 [94.9, 95.9]	70.1 [62.5, 76.7]	97.5 [97.2, 97.8]	58.5 [55.5, 61.4]
Cosmos 2.5	61.8 [60.3, 63.2]	66.9 [66.1, 67.7]	94.8 [94.4, 95.2]	49.0 [41.3, 56.8]	98.2 [98.0, 98.3]	66.5 [64.2, 68.7]
LingBot-World	66.9 [65.6, 68.2]	67.9 [67.0, 68.8]	94.1 [93.6, 94.6]	66.2 [58.5, 73.2]	96.9 [96.6, 97.2]	81.4 [79.1, 83.7]
HY-World 1.5	60.1 [58.4, 61.7]	65.4 [64.4, 66.5]	93.5 [93.0, 94.1]	91.1 [85.6, 94.6]	98.1 [98.0, 98.2]	60.5 [57.4, 63.5]
Fantasy-World	63.0 [61.4, 64.5]	62.8 [61.7, 63.9]	95.8 [95.3, 96.3]	49.0 [41.3, 56.8]	97.9 [97.6, 98.1]	65.8 [62.8, 68.7]
InSpatio-World	64.4 [62.9, 65.9]	67.6 [66.8, 68.3]	96.0 [95.6, 96.5]	26.1 [19.9, 33.5]	98.8 [98.7, 98.9]	76.1 [73.9, 78.2]
Astra	48.6 [47.0, 50.2]	52.5 [50.9, 54.2]	96.0 [95.8, 96.3]	79.6 [72.6, 85.2]	97.7 [97.5, 97.8]	28.0 [24.3, 31.7]
Happy Oyster	56.6 [55.2, 58.1]	63.9 [63.0, 64.8]	94.0 [93.4, 94.6]	94.2 [89.4, 96.9]	97.0 [96.6, 97.3]	58.3 [55.4, 61.0]
Matrix-Game 3.0	46.4 [45.4, 47.3]	70.0 [69.4, 70.6]	86.3 [85.4, 87.2]	97.5 [93.6, 99.0]	95.4 [95.0, 95.7]	57.1 [54.5, 59.7]
Genie 3	51.6 [50.2, 53.0]	59.3 [57.6, 60.9]	95.0 [94.6, 95.3]	92.4 [87.1, 95.6]	97.8 [97.6, 98.0]	55.2 [52.4, 57.9]
Matrix-Game 2.0	54.0 [52.4, 55.7]	60.3 [58.8, 61.8]	94.6 [94.2, 95.0]	94.9 [90.3, 97.4]	98.2 [98.0, 98.3]	41.0 [37.2, 44.6]
HY-GameCraft	52.6 [51.5, 53.7]	58.7 [57.5, 59.8]	93.7 [93.2, 94.1]	96.8 [92.8, 98.6]	97.6 [97.4, 97.8]	38.3 [35.6, 40.8]
Infinite-World	58.7 [57.1, 60.3]	66.1 [65.2, 66.9]	94.1 [93.7, 94.5]	82.8 [76.1, 87.9]	98.0 [97.9, 98.1]	62.3 [59.3, 65.2]

Table 18: Setting, navigation, and semantic interaction for 20 models. Values are scores and intervals.

Model	Scene Adherence	Subject Adherence	Navigation	Event Editing	Subject Action	Perspective Switching
Seedance 1.5	71.6 [67.8, 75.6]	94.2 [91.2, 96.6]	66.5 [64.8, 68.1]	80.4 [75.3, 85.2]	80.0 [75.9, 83.9]	45.0 [31.6, 58.7]
Wan 2.7	88.3 [85.9, 90.5]	94.6 [90.8, 97.6]	64.7 [62.7, 66.8]	84.0 [79.6, 87.8]	83.4 [79.1, 87.3]	55.0 [43.4, 65.6]
Kling 3.0	89.0 [86.2, 91.7]	92.9 [89.0, 96.1]	69.6 [67.6, 71.5]	81.4 [77.5, 85.1]	85.6 [81.7, 89.1]	55.0 [41.0, 69.2]
YUME 1.5	53.1 [50.5, 55.9]	91.7 [87.7, 95.2]	72.0 [70.4, 73.6]	57.8 [53.0, 62.8]	47.0 [42.2, 52.1]	16.7 [12.4, 21.8]
HY-Video 1.5	77.5 [73.7, 81.4]	93.6 [89.9, 96.6]	71.8 [69.7, 73.9]	63.8 [58.3, 69.3]	55.6 [49.9, 61.2]	27.6 [16.6, 40.2]
LTX 2.3	81.3 [77.9, 84.5]	89.2 [84.8, 93.1]	66.0 [64.3, 67.8]	53.0 [48.4, 57.6]	51.8 [46.2, 57.7]	25.0 [16.4, 35.0]
LongCat-Video	53.1 [50.8, 55.7]	91.5 [87.3, 95.2]	62.7 [60.7, 64.6]	50.4 [44.5, 56.7]	48.4 [42.9, 54.1]	18.3 [11.1, 26.1]
Kairos 3.0	52.2 [49.6, 55.0]	88.5 [84.1, 92.5]	64.0 [62.2, 65.9]	46.8 [40.7, 53.2]	41.4 [36.6, 46.5]	13.3 [6.7, 20.6]

Table 18: Continued from the preceding page.

Model	Scene Adherence	Subject Adherence	Navigation	Event Editing	Subject Action	Perspective Switching
Cosmos 2.5	72.4 [68.6, 76.2]	94.2 [90.5, 97.4]	62.9 [61.0, 64.7]	48.2 [42.4, 54.0]	41.6 [35.8, 47.5]	20.0 [10.7, 30.7]
LingBot-World	51.6 [49.8, 53.6]	93.6 [90.0, 96.5]	80.2 [78.1, 82.2]	–	–	–
HY-World 1.5	53.5 [50.8, 56.4]	90.8 [86.4, 94.7]	87.5 [85.4, 89.5]	–	–	–
Fantasy-World	52.4 [50.3, 54.7]	90.1 [85.6, 94.1]	72.0 [70.4, 73.6]	–	–	–
InSpatio-World	51.7 [49.9, 53.8]	91.1 [86.7, 95.1]	73.1 [71.2, 75.1]	–	–	–
Astra	43.4 [41.3, 45.5]	75.9 [69.0, 82.2]	66.7 [65.1, 68.2]	–	–	–
Happy Oyster	57.4 [54.3, 60.7]	91.1 [86.4, 95.1]	84.7 [82.9, 86.5]	–	–	–
Matrix-Game 3.0	48.9 [46.4, 51.6]	78.4 [71.6, 84.9]	83.6 [82.2, 85.1]	–	–	–
Genie 3	61.1 [57.2, 65.0]	83.8 [77.7, 89.2]	74.1 [72.0, 76.1]	–	–	–
Matrix-Game 2.0	49.4 [47.1, 51.9]	84.9 [79.9, 89.3]	80.3 [78.4, 82.1]	–	–	–
HY-GameCraft	50.6 [48.1, 53.3]	82.5 [76.3, 87.9]	66.3 [64.6, 67.9]	–	–	–
Infinite-World	54.0 [51.2, 57.0]	84.5 [78.4, 90.1]	75.6 [73.9, 77.2]	–	–	–

Table 19: Consistency, part 1. Confidence intervals for 20 models on the navigation split. Values are score and 95% interval.

Model	Background	Spatial	Gated Spatial	Segment
Seedance 1.5	89.6 [89.1, 90.0]	72.7 [70.6, 74.8]	72.4 [70.4, 74.4]	96.2 [91.9, 98.2]
Wan 2.7	89.4 [88.9, 90.0]	71.0 [69.2, 72.8]	71.0 [69.2, 72.8]	92.4 [87.1, 95.6]
Kling 3.0	92.3 [91.8, 92.8]	75.2 [73.6, 76.9]	75.1 [73.5, 76.7]	93.0 [87.9, 96.0]
YUME 1.5	90.3 [89.7, 90.9]	71.5 [69.3, 73.7]	71.4 [69.3, 73.6]	99.4 [96.5, 99.9]
HY-Video 1.5	92.1 [91.7, 92.6]	79.2 [77.3, 81.0]	75.1 [73.2, 76.9]	99.4 [96.5, 99.9]
LTX 2.3	88.3 [87.8, 88.9]	70.2 [68.5, 71.9]	70.2 [68.5, 71.9]	75.8 [68.5, 81.8]
LongCat-Video	95.1 [94.5, 95.7]	83.3 [81.0, 85.6]	66.2 [61.3, 70.7]	99.4 [96.5, 99.9]
Kairos 3.0	91.1 [90.2, 92.0]	76.8 [73.8, 79.9]	62.0 [56.4, 67.1]	94.3 [89.5, 97.0]
Cosmos 2.5	92.3 [91.7, 92.8]	78.1 [75.9, 80.4]	74.3 [71.8, 76.6]	94.3 [89.5, 97.0]
LingBot-World	96.9 [96.6, 97.2]	92.7 [91.0, 94.2]	67.1 [60.7, 73.0]	99.4 [96.5, 99.9]
HY-World 1.5	92.7 [92.1, 93.3]	90.6 [88.6, 92.4]	84.9 [81.4, 88.2]	100.0 [97.6, 100.0]
Fantasy-World	94.2 [93.6, 94.8]	80.6 [77.7, 83.4]	64.2 [60.4, 67.9]	100.0 [97.6, 100.0]
InSpatio-World	95.0 [94.5, 95.5]	93.8 [91.5, 95.7]	66.5 [58.5, 74.0]	100.0 [97.6, 100.0]

Table 19: Continued from the preceding page.

Model	Background	Spatial	Gated Spatial	Segment
Astra	85.3 [84.5, 86.1]	64.7 [62.2, 67.4]	63.3 [61.2, 65.5]	86.6 [80.4, 91.1]
Happy Oyster	91.4 [90.8, 92.0]	77.7 [75.2, 80.2]	75.8 [73.1, 78.4]	96.2 [91.9, 98.2]
Matrix-Game 3.0	85.7 [85.1, 86.3]	81.0 [78.6, 83.4]	80.4 [77.8, 82.8]	89.8 [84.1, 93.6]
Genie 3	90.7 [90.1, 91.2]	79.9 [77.6, 82.1]	78.4 [76.1, 80.6]	93.6 [88.7, 96.5]
Matrix-Game 2.0	86.9 [86.1, 87.6]	64.5 [62.8, 66.3]	64.5 [62.8, 66.3]	21.0 [15.4, 28.0]
HY-GameCraft	86.5 [85.9, 87.1]	60.5 [59.4, 61.7]	60.5 [59.4, 61.7]	99.4 [96.5, 99.9]
Infinite-World	88.8 [87.9, 89.6]	74.9 [72.3, 77.4]	74.4 [71.9, 76.8]	100.0 [97.6, 100.0]

Table 20: Consistency, part 2. Confidence intervals for 20 models on the navigation split. Values are score and 95% interval.

Model	Perspective	Subject	Geometric	Photometric
Seedance 1.5	70.5 [65.9, 74.8]	90.1 [89.2, 90.9]	82.4 [80.5, 84.1]	76.8 [76.1, 77.6]
Wan 2.7	78.2 [73.7, 82.3]	90.7 [89.9, 91.6]	83.7 [81.9, 85.3]	76.4 [75.6, 77.1]
Kling 3.0	76.8 [73.0, 80.2]	88.5 [87.1, 89.8]	88.9 [87.5, 90.3]	79.9 [79.1, 80.6]
YUME 1.5	48.0 [42.6, 53.3]	88.8 [87.8, 89.7]	88.0 [86.1, 89.7]	83.3 [82.9, 83.8]
HY-Video 1.5	86.6 [83.1, 89.4]	91.6 [90.7, 92.6]	94.6 [93.9, 95.3]	80.3 [79.5, 81.0]
LTX 2.3	69.8 [64.0, 75.0]	87.2 [85.8, 88.3]	76.9 [74.2, 79.5]	79.2 [78.4, 80.0]
LongCat-Video	81.5 [77.4, 85.3]	93.4 [92.5, 94.2]	95.4 [94.4, 96.3]	82.2 [81.4, 83.0]
Kairos 3.0	76.3 [71.7, 80.5]	90.8 [89.7, 91.9]	89.0 [87.0, 90.9]	80.8 [79.5, 82.2]
Cosmos 2.5	84.3 [82.1, 86.3]	92.3 [91.5, 93.1]	94.6 [93.8, 95.3]	81.6 [80.8, 82.4]
LingBot-World	90.9 [88.7, 92.8]	93.5 [92.7, 94.2]	95.4 [94.1, 96.4]	83.3 [82.4, 84.2]
HY-World 1.5	62.5 [56.4, 68.6]	89.1 [88.2, 90.1]	92.0 [89.8, 93.9]	83.1 [82.3, 84.0]
Fantasy-World	79.8 [74.8, 84.7]	92.5 [91.5, 93.5]	95.3 [94.4, 96.1]	84.8 [83.8, 85.9]
InSpatio-World	72.5 [66.3, 78.3]	94.4 [93.3, 95.4]	97.3 [96.6, 97.8]	87.4 [86.5, 88.2]
Astra	30.0 [25.6, 34.6]	83.5 [82.1, 85.0]	85.6 [83.2, 87.8]	87.5 [87.0, 88.1]
Happy Oyster	75.0 [68.8, 80.4]	91.5 [90.7, 92.4]	87.2 [85.2, 88.9]	79.8 [78.8, 80.7]
Matrix-Game 3.0	13.3 [10.3, 16.7]	83.0 [81.6, 84.4]	87.6 [85.3, 89.7]	75.3 [74.2, 76.3]
Genie 3	54.5 [47.0, 61.9]	90.4 [89.2, 91.6]	88.6 [87.0, 90.0]	84.5 [83.6, 85.3]
Matrix-Game 2.0	29.2 [25.1, 33.7]	87.2 [86.4, 88.0]	86.1 [83.4, 88.5]	81.3 [80.5, 82.1]

Table 20: Continued from the preceding page.

Model	Perspective	Subject	Geometric	Photometric
HY-GameCraft	17.9 [14.5, 21.3]	82.6 [80.9, 84.2]	88.3 [86.5, 90.0]	85.0 [84.1, 85.8]
Infinite-World	33.8 [27.5, 40.0]	88.4 [87.3, 89.4]	94.3 [93.3, 95.0]	85.1 [84.7, 85.6]

Table 21: Physical scores and full interaction. Confidence intervals for 20 models on the navigation split. Values are score and 95% interval.

Model	Causal Fidelity	Visual Plausibility	Interaction Full
Seedance 1.5	76.0 [70.9, 80.9]	60.7 [60.2, 61.3]	68.0 [64.1, 71.8]
Wan 2.7	83.3 [79.4, 87.2]	60.3 [59.8, 60.8]	71.8 [68.5, 74.9]
Kling 3.0	78.0 [72.8, 82.9]	60.7 [60.1, 61.3]	72.9 [69.0, 76.8]
YUME 1.5	72.7 [66.8, 78.5]	57.7 [57.0, 58.4]	48.4 [46.3, 50.7]
HY-Video 1.5	75.0 [70.4, 79.8]	59.7 [59.1, 60.3]	54.7 [51.2, 58.5]
LTX 2.3	74.0 [69.6, 78.4]	55.7 [55.1, 56.4]	49.0 [46.1, 52.0]
LongCat-Video	76.0 [71.2, 80.7]	61.8 [60.9, 62.6]	44.9 [41.9, 48.0]
Kairos 3.0	62.7 [57.3, 68.1]	58.0 [57.1, 58.9]	41.4 [38.7, 44.2]
Cosmos 2.5	74.7 [70.5, 78.9]	60.1 [59.4, 60.8]	43.2 [39.9, 46.7]
LingBot-World	77.7 [72.2, 83.0]	64.8 [64.2, 65.4]	–
HY-World 1.5	74.0 [68.6, 79.2]	58.6 [57.9, 59.3]	–
Fantasy-World	74.0 [68.7, 79.2]	59.7 [58.8, 60.5]	–
InSpatio-World	67.3 [61.5, 73.2]	63.1 [62.4, 63.8]	–
Astra	48.3 [41.1, 55.9]	54.6 [53.8, 55.4]	–
Happy Oyster	69.3 [63.9, 74.7]	57.6 [56.9, 58.3]	–
Matrix-Game 3.0	64.7 [59.0, 70.3]	54.0 [53.4, 54.6]	–
Genie 3	71.7 [67.4, 76.1]	59.7 [59.1, 60.2]	–
Matrix-Game 2.0	59.3 [53.0, 65.6]	55.0 [54.4, 55.6]	–
HY-GameCraft	68.3 [63.2, 73.5]	56.5 [55.8, 57.3]	–
Infinite-World	67.0 [61.3, 72.7]	57.2 [56.4, 58.0]	–

D.8 Data and Code Availability

Table 22 summarizes the released resources and the annotations that remain internal. The scorer's training and held-out annotations and the human-preference annotations are not released. Their protocols and reported statistics are described in Appendix C.6.2, Section 5.4, and Table 15.

E Broader Impact

WBench is intended to support more informative assessment of interactive world models. By exposing failures in controllability, consistency, and physics compliance alongside visual quality, it can help researchers identify limitations before considering downstream applications. Potential beneficiaries include research on simulation, robotics, gaming, and education, where visually convincing outputs alone are insufficient to establish dependable behavior. The benchmark provides evidence about the evaluated scenarios, not a substitute for application-specific testing. In particular, strong video-generation scores do not demonstrate that a model can safely control a physical system or accurately predict real-world consequences.

Progress informed by these evaluations may also improve the production of realistic synthetic media, with potential misuse in deception or misrepresentation. Generated outputs should therefore be

Table 22: Open-source resource checklist. Video releases and aggregate leaderboard scores are listed separately. Third-party checkpoints retain their original licenses, as detailed in the license notice.

Resource	Contents	Availability
Benchmark dataset	289 cases with images and metadata	Public
Evaluation code	22 metrics and evaluation prompts	Public
Visual plausibility scorer	Fine-tuned checkpoint	Public
Expert model checkpoints	Evaluation model weights	Public
Model-generated videos	Released model outputs	Public
Leaderboard	Aggregate model scores	Public
Visual plausibility annotations	Training and held-out video ratings	Internal
Human-preference annotations	Pairwise human judgments	Internal

identified as synthetic when shared, and reuse should respect source-specific rights and the terms of the underlying models. Benchmark optimization also risks emphasizing measured capabilities while overlooking unmeasured failures or underrepresented settings. We encourage users to report dimension-level results, metric applicability, and uncertainty rather than treating an aggregate ranking as a universal measure of capability.