LONGWM-BENCH: BENCHMARKING PERSISTENT AND NAVIGABLE WORLDS OVER LONG-HORIZON INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Short-term visual quality and action following offer limited evidence of whether an interactive world model remains reliable during extended use. We introduce LongWM-Bench, a benchmark of 300 tasks with target rollout durations of 60–180 seconds. Two task suites evaluate sustained control and revisit memory through combined inputs, repeated changes in control, and six families of trajectories with delayed and repeated revisits. We combine automatic visual and motion measurements with checklists tailored to each scene to assess visual quality, temporal consistency, revisit retention, physical and causal validity, and camera responses to actions. Evaluating 12 interactive world models reveals distinct capability profiles: models with similar appearance consistency differ substantially in revisit retention, while accurate camera responses can coexist with weak physical and causal validity. Coherent scene evolution remains a common weakness under our evaluation. These findings highlight the need to evaluate whether generated worlds remain controllable, retain previously observed content, and evolve coherently throughout interaction. Evaluation code and example data are included in the supplementary material.

1 INTRODUCTION

Interactive world models are moving beyond video generation toward environments that users and agents can continuously explore and interact with (Valevski et al., 2025; Zhang et al., 2025; Mao et al., 2025). Systems such as HappyOyster already support real-time exploration and interaction within generated worlds (HappyOyster, 2026). Yet sustaining interaction does not by itself ensure that the generated world remains reliable over time (Wang et al., 2026; Zhang et al., 2026a). A sequence of movements and turns may appear correct locally, while scene layouts gradually drift, object states change unexpectedly, or traces of earlier actions disappear upon return. When generated observations guide subsequent actions, evaluation must consider whether the world remains coherent throughout interaction, beyond the quality of individual frames and immediate responses to actions.

Recent benchmarks have expanded interactive world-model evaluation beyond visual quality to multi-turn control, scene consistency, physics, and memory (Xu et al., 2026b; Fang et al., 2026). WBench evaluates a broad range of interactive-generation capabilities (Ying et al., 2026), while WorldRoamBench examines action, vision, physics, and memory over 10–60-second interactions (Xu et al., 2026a). These efforts provide an important foundation, but evaluations at such horizons cannot determine whether reliability persists as interaction extends from tens of seconds to several minutes. For sustained control, the question is whether correct responses can be maintained under long sequences of combined and changing actions. For revisit memory, the question is whether previously established appearance, spatial structure, and state information survive prolonged absence and repeated returns. More generally, long-horizon evaluation should test whether later observations and interactions remain grounded in the world established earlier, rather than merely whether generation can continue for longer.

We introduce **LongWM-Bench**, a benchmark for long-horizon navigability comprising 300 tasks with 60–180-second rollouts. The benchmark targets two complementary challenges of prolonged interaction. **Complex control** tests whether models remain responsive under combined actions,

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.

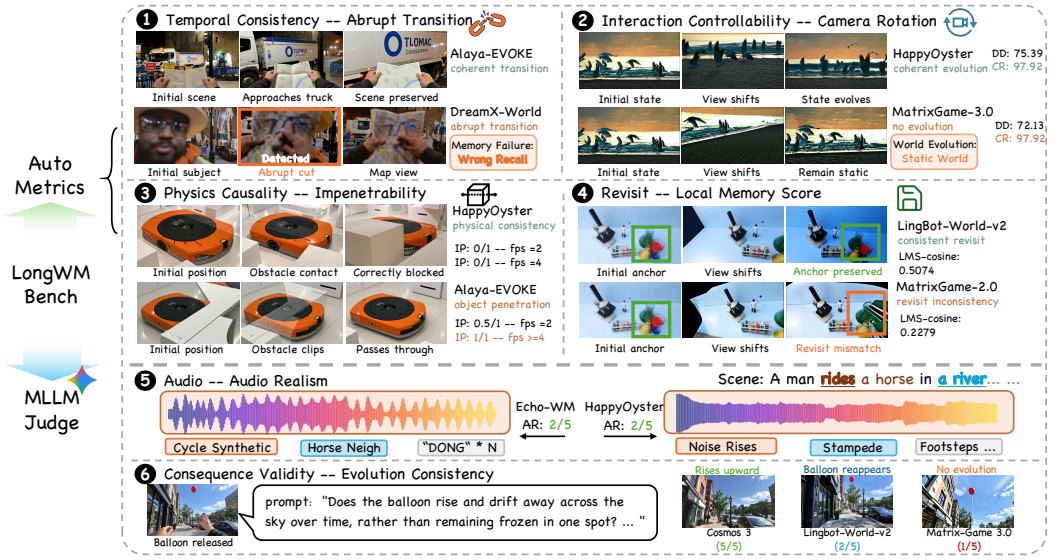

Figure 1: Overview of LongWM-Bench with representative evaluation cases. The benchmark combines automatic metrics and MLLM-based judgments to assess long-horizon reliability. Examples highlight temporal discontinuities, control failures, physical penetration, revisit mismatch, and long-form audio degradation, with successful and failure cases shown side by side.

direction reversals, and repeated changes in control. **Revisit memory** tests whether previously observed scenes remain visually consistent after extended periods out of view and repeated returns. Together, the two suites assess sustained controllability and visual persistence during prolonged interaction.

Figure 1 provides an overview of our evaluation framework. LongWM-Bench evaluates interactive world models along five complementary capability dimensions: **Action Controllability**, **Perceptual Quality**, **Physical & Causal Validity**, **Temporal Consistency**, and **Revisit Memory**. We combine visual and motion measurements, revisit-based comparisons, and event-level checks to assess these capabilities throughout each rollout. Importantly, consistency does not require the world to remain static: action-induced changes should persist unless later events justify otherwise, while ongoing processes should continue to evolve coherently. This enables us to diagnose a central failure mode of current interactive world models: a model may respond correctly to individual actions while gradually drifting away from the world established earlier. Our evaluation therefore examines not only whether local responses are correct, but also whether later states remain consistent with prior observations, actions, and their consequences.

Our contributions are threefold:

- ***LongWM-Bench.*** We introduce a benchmark for long-horizon navigability with 300 tasks and 60–180-second rollouts. It targets two complementary challenges—sustained control and world persistence—through complex action sequences, delayed revisits, and repeated revisits.

- ***Diagnostic evaluation.*** We develop a multidimensional evaluation framework that combines automatic visual and motion measurements, revisit-based comparisons, and MLLM-based event-level judgments. It evaluates five complementary capabilities—action controllability, perceptual quality, physical and causal validity, temporal consistency, and revisit memory— using observable evidence accumulated throughout each rollout.

- ***Long-horizon analysis.*** We evaluate 12 interactive world models and analyze their capabilities during extended interaction. The results reveal distinct weaknesses in camera control, revisit retention, physical and causal validity, and ongoing scene evolution, showing why these capabilities should be evaluated together.

Table 1: Comparison of LongWM-Bench with representative world model benchmarks. Revisit denotes the number of revisit categories and unique trajectories; Turns is the average number of interaction turns per instance, and Duration is measured in seconds. Capabilities include A (action control), S (single revisit), M (multiple revisits), P (physics), and E (ongoing evolution). Audio indicates audio support, and a lock marks unavailable information for non-open-source benchmarks.

Benchmark	Revisit	Instances	Domains	Turns	Duration	Audio	Capabilities
WorldScore (Duan et al., 2025)	–	3,000	15	1.3	2–10	✗	A E
WorldMark(Xu et al., 2026b)	2 / 6	500	3	1.7	20/40/60	✗	A S
iWorld-Bench (Fang et al., 2026)	1 / 8	4,900	2	1.0	–	✗	A S
WBench (Ying et al., 2026)	2 / 23	289	6	3.7	10–45	✗	A S P
WorldOlympiad (Zhao et al., 2026)	–	1,000	3	3.2	–	✗	A P E
WorldRoamBench (Xu et al., 2026a)	🔒	🔒	🔒	🔒	10–60	✗	A S P
LongWM-Bench	**6 / 150**	**300**	**9**	**40.8**	**60–180**	✓	A S M P E

2 Related Work

Interactive World Models Interactive world models extend video generation into environments that respond to actions. Genie learns latent controls from videos without action labels (Bruce et al., 2024), while GameNGen demonstrates diffusion-based real-time game simulation (Valevski et al., 2025). Matrix-Game 2.0 and Hunyuan-GameCraft advance continuous interaction through keyboard and mouse control, history conditioning, and efficient inference (He et al., 2025; Li et al., 2025). Genie 3 and LingBot-World support minute-scale interaction across diverse scenes (Team et al., 2026b), while SANA-WM improves efficiency with hybrid linear attention and dual-branch camera control (Zhu et al., 2026). HappyOyster broadens user intervention through its Wandering and Directing modes, supporting exploration and real-time editing of characters, cameras, and scenes (HappyOyster, 2026). Echo-WM unifies discrete commands and continuous poses as relative 6-DoF trajectories, enabling first- and third-person navigation with joint audio–video generation (Zhang et al., 2026b). These advances improve efficiency, control, and interaction length, but do not by themselves establish whether the generated world remains consistent.

World Model Evaluation Table 1 compares LongWM-Bench with representative world-model evaluation benchmarks. Early video-generation benchmarks such as VBench and EvalCrafter focus on visual quality, motion, temporal consistency, and text alignment (Huang et al., 2024; Liu et al., 2024; Wu et al., 2025; Wen et al., 2026), while VBench-2.0 further introduces physical and commonsense evaluation (Zheng et al., 2025). Recent benchmarks increasingly target interactive settings: WorldScore evaluates controllability, quality, and dynamics under specified camera trajectories; WorldMark and iWorld-Bench examine action responses, view consistency, control, and memory (Xu et al., 2026b; Fang et al., 2026); and WBench and WorldRoamBench extend evaluation to multi-turn interaction, physics, and memory (Ying et al., 2026). PlayWorld further introduces an agent that adapts actions to visual feedback (Ding et al., 2026). Together, these benchmarks provide broad diagnostics, but long-horizon reliability remains less characterized. Existing evaluations typically focus on interaction horizons of around 60 seconds or shorter, leaving open whether controllability, world consistency, and memory persist over longer interactions and repeated revisits. LongWM-Bench complements these efforts by evaluating sustained controllability and world persistence over 60–180-second rollouts.

3 LongWM-Bench

3.1 Dataset Construction

Figure 2 summarizes our three-stage data construction pipeline, which combines multi-model generation, automated verification, and human review to construct diverse interaction trajectories and reliable evaluation checklists. All prompts are provided in Appendix I.

Stage 1: Initial-frame and caption construction. We curate initial frames from real images, game screenshots, and GPT-Image-2 generations across nine scene domains (OpenAI, 2026b). Gemini 3.8 Flash first produces *environment*, *subject*, and *perspective* descriptions (Google, 2026b). The de-

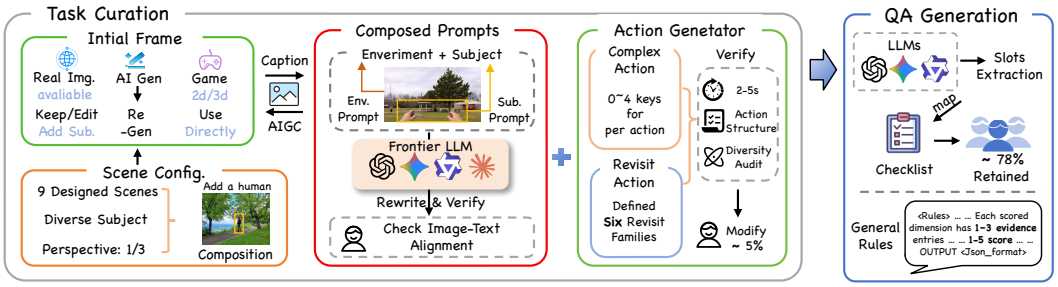

Figure 2: Data curation pipeline of LongWM-Bench. The pipeline comprises three stages: task curation with prompt rewriting and verification, interaction-sequence generation and validation, and checklist-based QA generation with human review.

scriptions are then independently rewritten and verified using GPT-5.6 Sol, Qwen3.8-Max, Gemini 3.8 Flash, and Claude Sonnet 5 (OpenAI, 2026a; Alibaba Cloud, 2026; Google, 2026b; Anthropic, 2026). Human reviewers check image–text alignment and correct 10% of the samples. The resulting image–text pairs provide the scene specification used as input for interaction generation.

Stage 2: Interaction construction. Using the same four-model pool, we generate complex-control sequences and six predefined revisit families to improve trajectory diversity. Generated trajectories are automatically checked for action structure, duration, revisit validity, and diversity, followed by human revision of invalid, repetitive, or degenerate sequences; approximately 5% are modified.

Stage 3: Checklist construction. GPT-5.6 Sol, Qwen3.8-Max, and Gemini 3.8 Flash extract capability-specific evaluation slots from each scene and trajectory and convert them into candidate checklist items (OpenAI, 2026a; Alibaba Cloud, 2026; Google, 2026b). Human reviewers remove redundant, speculative, ambiguous, or visually unverifiable items, while ensuring that the retained checks are grounded in observable evidence and relevant to the target capability; approximately 78% of the candidates are retained. The final per-sample checklists are supplemented with shared evidence requirements and 1–5 scoring rules for consistent evaluation across models.

3.2 Data Statistics

LongWM-Bench contains 150 complex-control cases and 150 revisit cases, with requested durations of 60–180 seconds. Figure 3 summarizes the distribution of scenes, viewpoints, subjects, and control programs. First- and third-person perspectives account for 60% and 40% of the cases, respectively. Across all annotated action segments, including idle intervals, 38.2% contain simultaneous key inputs. The six revisit families range from in-place turns and short out-and-back movements to radial excursions, closed loops, and interleaved waypoint visits. These patterns distinguish repeated observation of one anchor from returns after longer or multi-location excursions; the detailed trajectory definitions and examples are given in Appendix B.

3.3 Evaluation Metrics

Building on the capability taxonomy in Figure 3, we evaluate interactive world models along five complementary dimensions: **Perceptual Quality**, **Temporal Consistency**, **Revisit Memory**, **Physical & Causal Validity**, and **Action Controllability**. Together, these dimensions cover both the local quality of generated observations and the long-horizon reliability of the evolving world under sustained interaction. The taxonomy defines the capabilities expected of a persistent and navigable world, while the evaluation metrics operationalize these capabilities through observable evidence in generated rollouts. Our evaluation combines perceptual, motion, visual-feature, geometric, and vision–language measurements, and reports each metric separately to preserve diagnostic interpretability. Detailed definitions and evaluator configurations are provided in Appendix D.1.

Perceptual Quality. This dimension characterizes whether generated observations remain visually plausible while exhibiting meaningful visual dynamics. Following the capability taxonomy, we assess both visual quality and motion magnitude. **Visual Quality (VQ)** uses MUSIQ (Ke et al., 2021) to score individual frames and averages the frame-level predictions, providing a reference-free mea-

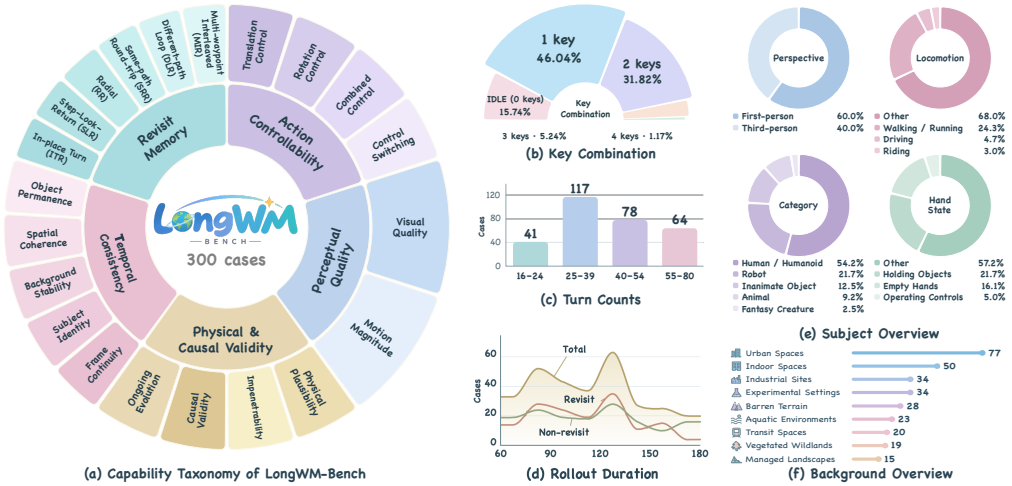

Figure 3: Statistics of LongWM-Bench, including capability coverage, interaction patterns, rollout duration, background domains, and subject distributions.

sure of visual fidelity throughout the rollout. **Dynamic Degree (DD)** estimates motion magnitude from dense optical flow, with normalization for temporal sampling and image scale and trimming to reduce the influence of extreme flow estimates. Reporting VQ and DD together distinguishes visually convincing but nearly static generations from outputs with sustained visual dynamics, while DD itself reflects motion magnitude rather than physical correctness.

Temporal Consistency. This dimension evaluates whether the visible world remains coherent throughout continuous interaction, including observation continuity and the temporal stability of both subjects and backgrounds. **Jump-Based Continuity (JP)** uses TransNetV2 (Soucek & Lokoc, 2024) to detect abrupt cut-like transitions, with the strongest predicted cut probability used to measure temporal continuity. **Subject Consistency (SC)** measures the stability of the foreground subject by combining adjacent-frame DINOv2 similarity with CLIP similarity to the initial observation, capturing both short-term appearance consistency and accumulated subject drift. **Background Consistency (BC)** separately measures the stability of the background using adjacent-frame and initial-observation comparisons over background regions. Subject masks are used where available to separate foreground subjects from background content.

Revisit Memory. This dimension evaluates whether previously observed locations remain consistent after leaving the field of view and being revisited later. We evaluate 150 predefined revisit trajectories with designated observation anchors. **Global Memory Score (GMS)** uses DINOv2 features (Oquab et al., 2024) to compare each revisit with the anchor's initial observation, measuring retention of the originally observed scene. **Local Memory Score (LMS)** instead compares consecutive revisits, excluding the initial-to-first-revisit pair, and therefore applies to trajectories with at least three observations. Both scores weight comparisons by the elapsed time since the initial observation, emphasizing consistency over longer intervals. GMS captures long-term consistency with the original view, whereas LMS captures stability across successive returns.

Physical & Causal Validity. This dimension evaluates whether the generated world follows scene-appropriate physical constraints, causal relationships, and ongoing state evolution. **Physical Causality (PC)** assesses physical plausibility through properties and responses such as gravity, support, buoyancy, collision, deformation, material behavior, thermodynamics, and non-penetration of solid boundaries. **Content Causality (CC)** evaluates whether a scene-supported cause produces a visually consistent effect on another object or region, such as illumination inducing corresponding shadows or reflections. **Evolution Consistency (EC)** measures whether spontaneous processes and state changes, such as drifting, flowing, or melting, evolve coherently over time rather than remaining unnaturally static. PC, CC, and EC are evaluated by Gemini 3.1 Pro Preview (Google, 2026a) using case-specific checklists, with each item scored on a 1–5 scale based on visual evidence.

Table 2: Results on LongWM-Bench across five evaluation dimensions. Closed-source models are listed first, followed by open-weight models; all models are compared jointly for each metric (higher is better). The highest score in each metric is **bold** with a category-matched pale background; the second highest is underlined. Params. denotes the parameter count; 🔒 indicates closed source. All scores are normalized, higher is better.

Model	Params.	Perceptual Quality		Temporal Consistency			Revisit Memory		Physical & Causal Validity			Action Controllability	
		VQ	DD	JP	SC	BC	GMS	LMS	PC	CC	EC	TR	CR
Closed-source models													
HappyOyster	🔒	63.91	69.77	59.54	76.75	84.70	62.78	62.50	50.72	47.84	21.77	79.64	93.33
Genie 3†	🔒	64.60	**87.81**	79.95	76.70	82.57	N/A	N/A	**67.44**	55.28	21.55	N/A	N/A
Open-weight models													
DreamX-World	5B	62.89	58.28	85.39	77.59	84.02	68.92	72.25	46.11	56.25	**35.70**	**94.26**	90.32
Lyra 2.0	14B	66.90	55.83	79.71	79.39	84.04	**79.81**	**86.78**	36.62	36.38	12.86	69.03	**95.88**
Echo-WM	19B	65.99	62.61	79.79	78.86	84.31	56.80	62.61	36.59	37.50	17.96	84.49	92.89
Alaya-EVOKE	14B	**68.61**	51.42	86.25	78.00	84.44	69.46	77.33	63.11	**60.79**	29.67	59.50	42.67
LingBot-World-v2	14B	60.43	55.01	73.38	76.32	81.54	46.41	52.72	44.13	32.59	14.15	85.38	88.21
Matrix-Game 3.0	5B	64.95	70.99	57.18	78.09	80.48	39.41	51.02	8.15	4.54	0.93	81.76	95.55
Cosmos 3	16B	66.06	25.30	**88.87**	**82.19**	**86.91**	79.27	82.57	36.30	33.33	10.56	42.49	23.64
Yume-1.5	5B	68.06	49.36	85.08	79.02	83.35	47.89	58.36	24.32	24.63	12.28	58.03	24.44
ABot-World	0.5B	50.34	69.55	57.78	76.61	77.64	34.57	41.33	17.44	11.16	4.09	60.36	81.73
Matrix-Game 2.0	1.8B	53.85	52.72	80.33	79.43	79.00	24.78	34.65	2.08	0.89	0.00	79.94	90.23

†Genie 3 is evaluated through its official hosted interface with a maximum rollout duration of 60 seconds; unavailable long-horizon metrics are denoted by "N/A".

Action Controllability. This dimension evaluates whether camera motion reliably follows the prescribed control inputs throughout sustained interaction, including translation, rotation, combined controls, and repeated control switching. Using camera trajectories recovered by ViPE (Huang et al., 2025), **Translation Response (TR)** measures whether the net camera displacement of each action segment agrees with the commanded translation direction in the action segment's initial camera coordinate system. **Camera Rotation Response (CR)** measures whether the estimated yaw and pitch changes follow the commanded rotation directions with sufficient angular magnitude. Both metrics report the proportion of correctly executed action segments, with combined translation–rotation commands contributing to both TR and CR. These metrics characterize directional control accuracy rather than requiring an exact displacement or rotation magnitude.

Together, these metrics provide complementary views of long-horizon reliability. Perceptual and Temporal metrics characterize the quality and continuity of the generated observations, while revisit, physical–causal, and controllability metrics test whether the world preserves previously established states, evolves plausibly, and remains responsive to sustained interaction. We therefore report these metrics separately throughout our experiments to facilitate fine-grained diagnostic analysis across complementary long-horizon capabilities.

4 Experiments

4.1 Experimental Setup

We evaluate 12 interactive world models: ABot-World (Jiang et al., 2026), Alaya-EVOKE (Yin et al., 2026), DreamX-World (Team et al., 2026a), Genie 3 (Parker-Holder & Fruchter, 2025), HappyOyster (HappyOyster, 2026), Echo-WM (Zhang et al., 2026b), LingBot-World-v2 (RobbyAnt, 2026), Lyra 2.0 (Shen et al., 2026), Matrix-Game 2.0 (He et al., 2025), Matrix-Game 3.0 (Wang et al., 2026), Cosmos 3 (NVIDIA, 2026), and Yume-1.5 (Mao et al., 2026). For each case, the initial input, task requirements, and interaction budget are held fixed across models, with controls mapped to each model's native interface. Complex control and revisit memory use prescribed action sequences. We evaluate the generated videos using the automatic measurements and VLM-based checks defined in Section 3.3. Model details and inference settings are in Appendix D.

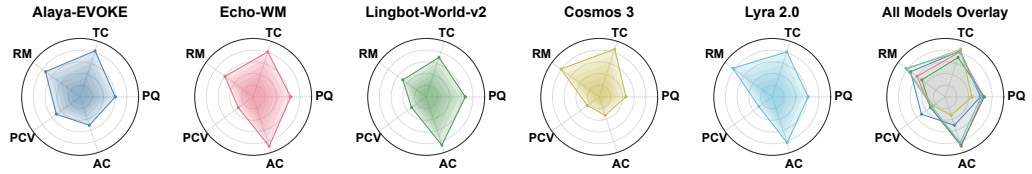

Figure 4: Multi-metric radar comparison on five evaluation dimensions. Each axis reports the average of the corresponding dimension's submetric scores.

Table 3: Agreement between automatic evaluators and human annotations on LongWM-Bench. A–H measures the agreement between each automatic evaluator and the consensus of two human annotators, while H–H measures inter-annotator agreement. All values are mean L_1 distances on the 1–5 scale (lower is better).

Agreement ($L_1 \downarrow$)	VQ	DD	JP	SC	BC	GMS	LMS	PC	CC	EC	TR	CR	Mean
A–H	0.50	0.47	1.25	1.16	1.23	0.82	1.10	1.06	1.07	1.16	1.10	1.26	1.02
H–H	0.24	0.33	0.72	0.45	0.38	0.59	0.71	0.34	0.36	0.50	0.54	0.55	0.48

4.2 Main Results

Table 2 summarizes performance on LongWM-Bench. We report visual quality, world consistency, physical and causal validity, control accuracy, and revisit similarity separately rather than combining them into an overall score. Metrics are computed wherever their required evidence is available, including control and revisit measurements. Scores are aggregated from case-level results, with per-suite breakdowns and evaluation coverage provided in the appendix.

Uneven capabilities, with world evolution as a shared weakness. Table 2 and Figure 4 show distinct capability profiles rather than a single model leading throughout. Visual quality is relatively concentrated: 10 of the 12 models score between 60 and 70 on VQ. SC and BC also occupy narrow ranges of 76.32–82.19 and 77.64–86.91, respectively. The evolution checks reveal a different picture: EC never exceeds 35.70, and eight models score below 20. This weakness is not confined to models with poor physical or causal scores. Alaya-EVOKE achieves PC/CC of 63.11/60.79, but its EC remains 29.67; DreamX-World obtains the highest EC, yet still scores only 35.70. The results identify coherent state evolution as an unresolved requirement beyond maintaining appearance or satisfying individual physical and causal checks.

Similar appearance consistency can mask large differences in revisit retention. Although SC and BC vary relatively little across models, GMS spans 24.78–79.81 and LMS spans 34.65–86.78. Lyra 2.0 and Matrix-Game 2.0 illustrate this separation particularly clearly: their SC scores are nearly identical at 79.39 and 79.43, whereas their GMS scores are 79.81 and 24.78. Thus, similar temporal-consistency scores can mask very different abilities to recover previously observed scenes after navigation. Moreover, LMS exceeds GMS for 10 of the 11 models with available revisit results. This pattern is consistent with closer agreement between successive returns than with the initial observation. It highlights why recent visual continuity and preservation of the original scene should be evaluated separately, rather than treated as a single notion of consistency.

Strong control does not ensure a reliable world. Control and world retention reveal another separation. Cosmos 3 leads JP, SC, and BC and achieves a GMS of 79.27, but its TR/CR scores are only 42.49/23.64. Matrix-Game 3.0 shows the opposite profile: TR/CR reaches 81.76/95.55, while GMS is 39.41 and PC/CC/EC is 8.15/4.54/0.93. Preserving a scene is therefore insufficient if the requested movement is not reliably executed; responding to movement is likewise insufficient if the resulting world lacks memory, physical validity, or coherent evolution. These are not necessarily competing objectives: DreamX-World combines strong TR/CR of 94.26/90.32 with substantially better physical and causal results than Matrix-Game 3.0. Rather, the differing profiles show that control, retention, and world dynamics remain separate requirements that must be verified together.

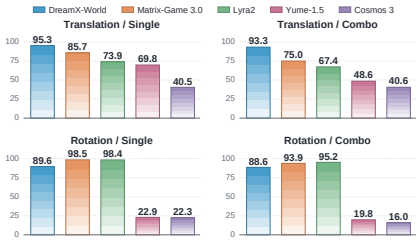

Figure 5: **Action-type analysis.** Control accuracy (%) for translation (top) and rotation (bottom), under single-key (left) and composite-key (right) inputs.

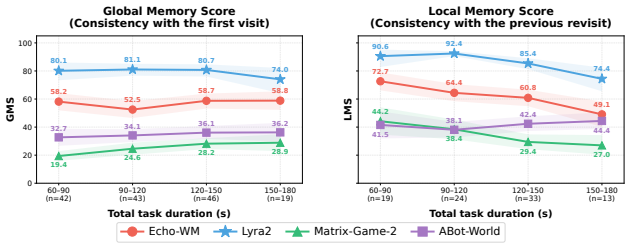

Figure 6: **Revisit consistency.** Mean GMS and LMS (0–100) across task durations. Shading shows 95% bootstrap confidence intervals (10,000 case-level resamples); n denotes the number of shared evaluation cases per group.

4.3 Further Analysis

Human Agreement. We assess agreement with human judgments on 250 generated videos, each independently rated by two annotators. Table 3 reports mean absolute differences on the 1–5 scale. The mean difference is 0.48 between annotators and 1.02 between the automatic scores and the mean human ratings. Automatic agreement varies across metrics, with larger discrepancies for several consistency and control measures. These results support using the metrics as diagnostic signals, while indicating that automated scores do not fully reproduce human judgments.

Variation Across LLM Evaluators. We further examine whether this agreement is sensitive to the choice of LLM evaluator by comparing several candidate judges against the same human consensus. On the signal-level metrics, the mean L_1 distances are 0.988 for Gemini-3.1-Pro-Preview, 0.808 for Gemini-2.5-Pro, 0.690 for Gemini-3.7-Flash, and 0.673 for GPT-5.6-Sol. On the checklist-based metrics, the corresponding distances are 1.097, 0.815, 0.817, and 0.574, respectively. Despite evaluator-specific variation, the candidate judges show broadly similar trends relative to the same human reference, indicating that the agreement analysis is not specific to a single LLM evaluator.

Action-Type Analysis. Figure 5 shows that aggregate control scores can obscure differences between translation and rotation. Lyra 2.0 performs better on rotation, whereas Yume-1.5 performs better on translation. Combined inputs reduce translation accuracy for most of the selected models, while DreamX-World maintains strong performance across all four settings. These results distinguish accuracy on individual control types from the ability to execute combined commands.

Long-Horizon Performance. Figure 6 compares revisit consistency across task durations from 60 to 180 seconds. Echo-WM exhibits a pronounced decline in LMS, from 72.7 to 49.1, while its GMS remains comparatively stable. Lyra 2.0 maintains the highest scores among the four models shown, but its LMS decreases from 92.4 in the 90–120 s group to 74.4 in the longest-duration group. Matrix-Game 2.0 shows contrasting trends: GMS increases from 19.4 to 28.9, whereas LMS decreases from 44.2 to 27.0. ABot-World remains relatively stable at lower absolute scores. These results illustrate that consistency with the initial observation and consistency across successive revisits can vary differently with task duration. We interpret these trends descriptively, since duration groups contain different scenes and trajectory structures.

Metric Complementarity and Ranking Robustness. Pairwise Spearman correlations indicate limited metric redundancy (mean $|\rho| = 0.22$; Figure 10, Appendix C.1). Correlations are stronger within dimensions than across them (0.38 vs. 0.20), with the highest values for GMS–LMS (0.76), PC–CC (0.67), and TR–CR (0.60). DD correlates negatively with several consistency and memory metrics, suggesting a potential motion–consistency trade-off. Rankings remain stable across 20 translation/rotation threshold configurations ($\rho \geq 0.964$), with a maximum shift of four positions.

Audio Evaluation Results. We evaluate Echo-WM and HappyOyster, the two audio-generating world models, on audio quality (AQ), audio–visual semantic alignment (IB), predicted synchronization offset (SF), acoustic realism (AR), and action–sound correspondence (AAC). Table 4 summarizes the results, with evaluation details provided in Appendix E.

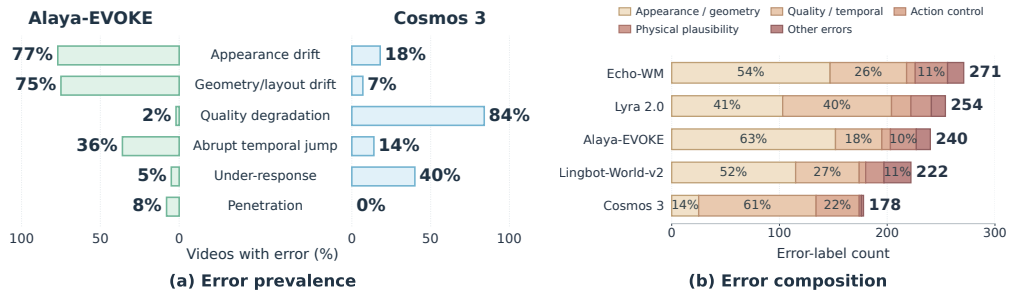

Figure 7: Error analysis on five models. (a) Video-level prevalence of the six most frequent pooled labels for two contrasting models. (b) Error-label counts and category composition across all five models; internal percentages denote shares of each model's total labels. See more in Figure 11.

Echo-WM achieves higher audio quality, audio–visual semantic alignment, and action–sound correspondence, whereas HappyOyster obtains higher acoustic realism and a lower predicted synchronization offset. Despite these differences, long-form audio remains challenging for both models. Qualitative inspection reveals noise, temporal drift, and repeated synthetic sound patterns in later portions of some rollouts. Footsteps often coarsely follow walking, but such local correspondence does not imply precise synchronization and does not consistently extend to other sound events. These observations highlight the importance of evaluating audio throughout the full long-horizon rollout.

Table 4: Audio evaluation results for Echo-WM and HappyOyster. Higher is better except for SF.

Model	AQ↑	IB↑	SF↓	AR↑	AAC↑
Echo-WM	6.573	0.1545	1.039	44.09	35.22
HappyOyster	5.641	0.1354	0.912	48.09	28.91

4.4 Discussion of Error Patterns

We use Gemini 3.1 Pro Preview to annotate 500 videos generated by five models on the same 100 tasks using 25 predefined error labels. Each video is represented by 20 uniformly sampled frames, each tagged with its source timestamp, together with the camera-control instructions. The judge outputs video-level failure labels only and does not produce temporal localizations. The complete prompt is provided in Appendix I.4.

Aggregate counts obscure distinct failure profiles. Figure 7 contrasts Alaya-EVOKE's appearance and geometry drift (77% and 75%) with Cosmos 3's quality degradation and under-response (84% and 40%). Despite fewer labels per video (1.78 versus 2.40), Cosmos 3 exhibits more quality degradation. Lower aggregate counts therefore do not imply uniformly stronger reliability; evaluation must distinguish scene consistency, visual quality, and control responsiveness.

Appearance and geometry failures frequently coincide. Appearance and geometry drift co-occur in 203 videos (Jaccard overlap: 0.60; Figure 11e), motivating joint evaluation of persistent appearance and spatial layout. However, models exhibit overlapping weaknesses: Lyra 2.0 combines substantial drift with frequent quality degradation. Improving one capability alone is therefore insufficient to ensure reliable rollouts, as other bottlenecks may persist.

5 Conclusion

We introduced LongWM-Bench to evaluate whether interactive world models remain controllable and preserve previously observed scenes during extended interaction. The benchmark comprises 300 tasks with target durations of 60–180 seconds, covering sustained control and delayed and repeated revisits. Our evaluation reveals that similar appearance consistency can mask substantial differences in revisit retention, while accurate camera responses can coexist with weak physical and causal validity. Coherent scene evolution remains a common weakness across the evaluated models. These findings highlight the need to assess control, persistence, and evolution together when evaluating interactive worlds. LongWM-Bench provides a diagnostic framework for this purpose and supports progress toward generated environments that remain reliable throughout continued interaction.

LIMITATIONS

LongWM-Bench evaluates observable behavior rather than access to a model's internal world state. Its visual similarity scores can be affected by viewpoint changes and navigation errors, while sparse temporal sampling may miss brief events. Differences in model interfaces also limit the equivalence of control inputs across systems. Automated judgments remain subject to evaluator bias, and agreement with human ratings varies across metrics. Audio evaluation currently covers only two models. Finally, the 300 cases span a range of scenes and interaction patterns but do not exhaust the diversity of possible environments, actions, or durations.

GENERATIVE AI DISCLOSURE

In this work, we used generative AI tools in benchmark construction, evaluation, and manuscript preparation. During benchmark construction, GPT-Image-2 was used for prompt-based initial-frame generation, while generative models were also used for prompt rewriting and verification, interaction sequence generation, and candidate evaluation-item generation. All AI-generated or revised benchmark content was subsequently reviewed, filtered, or corrected by the authors or human annotators before inclusion in the final benchmark.

Generative AI also forms part of the evaluation pipeline. Vision-language models were used as automated evaluators for several physical, causal, and evolution-related metrics under predefined case-specific checklists and scoring rules. Their agreement with human judgments was separately evaluated to assess the reliability of the automated evaluation.

Generative AI tools were not used to design or determine the research methodology or experimental protocol, nor to formulate the conclusions of this work. Additionally, generative AI tools were used to assist with language editing, code development and debugging, figure and data-visualization preparation, and the formatting of tables, captions, and other manuscript elements.

All AI-assisted outputs were reviewed and revised as necessary by the authors. The authors take responsibility for the final content of this work, including the benchmark, experiments, analyses, figures, and manuscript.

ETHICS STATEMENT

This work focuses on benchmarking interactive world models and does not involve human-subject experiments or the collection of sensitive personal information. Benchmark data are constructed from curated real images, game screenshots, and AI-generated content, with automated processing followed by human review to reduce inappropriate, ambiguous, or unsupported content. Our evaluation is intended to provide a more transparent understanding of the capabilities and limitations of current interactive world models, rather than to promote deployment in safety-critical settings. While advances in interactive world modeling may enable beneficial applications in simulation, embodied intelligence, and content creation, such technologies may also be misused to generate deceptive or harmful synthetic environments. We therefore encourage responsible use of the benchmark and careful consideration of downstream risks when deploying interactive generative systems.

REPRODUCIBILITY.

We provide detailed specifications to support reproducibility of the benchmark and evaluation. The benchmark construction procedure, including initial-frame preparation, interaction generation, and checklist construction, is described in Section 3.1. Model-specific inference protocols and configurations are provided in Appendix D, while the definitions, implementation details, temporal sampling, and preprocessing settings of the video evaluation metrics are documented in Appendix D.1. The audio evaluation protocol is detailed in Appendix E, and the complete prompts used for benchmark construction and model-based evaluation are provided in Appendix I.

REFERENCES

Alibaba Cloud. Alibaba unveils Qwen3.8-Max: Its largest and most capable flagship model to date. Alibaba Cloud, August 2026. URL https://www.alibabacloud.com/en/press-room/alibaba-unveils-qwen3-8-max. Accessed September 26, 2026.

Anthropic. Introducing Claude Sonnet 5. Anthropic, June 2026. URL https://www.anthropic.com/news/claude-sonnet-5. Accessed September 26, 2026.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. URL https://arxiv.org/abs/2402.15391.

Zhe Cao, Tao Wang, Jiaming Wang, Yanghai Wang, Yuanxing Zhang, Jialu Chen, Miao Deng, Jiahao Wang, Yubin Guo, Chenxi Liao, Yize Zhang, Zhaoxiang Zhang, and Jiaheng Liu. T2AV-Compass: Towards unified evaluation for text-to-audio-video generation. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2512.21094.

Kaixin Ding, Xi Chen, Minghong Cai, Zhiyuan Xu, Yiyang Wang, Yuxiang Lu, Junyi Li, Shuyang Chen, Yuan Gao, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Playworld: Benchmarking world models with agent players over long-horizon objectives, 2026. URL https://arxiv.org/abs/2608.13552.

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 (ICCV)*, pp. 27713–27724, October 2025.

Jianjie Fang, Yingshan Lei, Qin Wan, Ziyou Wang, Yuchao Huang, Yongyan Xu, Baining Zhao, Weichen Zhang, Chen Gao, Xinlei Chen, and Yong Li. iworld-bench: A benchmark for interactive world models with a unified action generation framework, 2026. URL https://arxiv.org/abs/2605.03941.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pp. 363–370. Springer, 2003.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023. doi: 10.1109/CVPR52729.2023.01457. URL https://openaccess.thecvf.com/content/CVPR2023/html/Girdhar_ImageBind_One_Embedding_Space_To_Bind_Them_All_CVPR_2023_paper.html.

Google. Gemini 3.1 Pro Preview. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview, February 2026a. Accessed: 2026-09-26.

Google. Introducing Gemini 3.8 Flash and Gemini 3.8 Flash Cyber. Google, September 2026b. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/. Accessed September 26, 2026.

HappyOyster. Happy oyster: Real-time world model for interactive creation. https://www.happyoyster.com/, 2026. Accessed: 2026-09-22.

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.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception, 2025. URL https://arxiv.org/abs/2508.10934.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.17982.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10448489. URL https://arxiv.org/abs/2401.16423.

Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, Wenlin Liu, Yun Wang, Xue Zheng, Rui Sun, et al. Abot-world-0: Infinite interactive world rollout on a single desktop gpu. *arXiv preprint arXiv:2607.19191*, 2026.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer, 2021. URL https://arxiv.org/abs/2108.05997.

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, 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 (CVPR)*, pp. 22139–22149, June 2024.

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, 2025. URL https://arxiv.org/abs/2507.17744.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, and Kaipeng Zhang. Yume1. 5: A text-controlled interactive world generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7752–7761, 2026.

NVIDIA. Cosmos 3: Omnimodal world models for physical ai, 2026. URL https://arxiv.org/abs/2606.02800.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. OpenAI, July 2026a. URL https://openai.com/index/gpt-5-6/. Accessed September 26, 2026.

OpenAI. GPT-Image-2 model. OpenAI API Documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-image-2. Accessed September 26, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

J. Parker-Holder and S. Fruchter. Genie 3: A new frontier for world models. Google DeepMind, August 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

RobbyAnt. Lingbot-world-v2. https://huggingface.co/collections/robbyant/lingbot-world-v2, 2026. Hugging Face Collection.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026.

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.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026a.

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*, 2026b.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound, 2025. URL https://arxiv.org/abs/2502.05139.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 73754–73776, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b71ecea210f7159f31e46631fe5c838f-Paper-Conference.pdf.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory, 2026. URL https://arxiv.org/abs/2604.08995.

Yuqing Wen, Yukai Huang, Qianqian Xie, Jiangtao Wu, Yibin Lin, Yikai Gu, Jialu Chen, Yuanxing Zhang, and Jiaheng Liu. Ave-compass: Towards holistic evaluation for audio-video editing abilities. *arXiv preprint arXiv:2607.24821*, 2026.

Jiangtao Wu, Shihao Li, Zhaozhou Bian, Jialu Chen, Runzhe Wen, An Ping, Yiwen He, Jiakai Wang, Yuanxing Zhang, and Jiaheng Liu. Vidic: Video difference captioning. *arXiv preprint arXiv:2512.03405*, 2025.

Ting-Bing Xu, Jiacheng Sui, Zhe Gao, Kewei Shi, Wenjin Yang, Zhicheng Liu, Zhaoxu Sun, Mingchao Sun, Hongyu Pan, Fan Jiang, Mu Xu, Qi Fan, Yang Gao, Yong Li, and Baoquan Chen. Worldroambench: An open-world benchmark for long-horizon stability of interactive world models, 2026a. URL https://arxiv.org/abs/2606.31672.

Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Yongtao Ge, and Kaipeng Zhang. Worldmark: A unified benchmark suite for interactive video world models, 2026b. URL https://arxiv.org/abs/2604.21686.

Yuanyang Yin, Gongxuan Wang, Yifan Zhan, Chuanhao Li, Kaipeng Zhang, and Feng Zhao. Alaya-evoke: From linear-scaling supervision to endless world. *arXiv preprint arXiv:2608.13546*, 2026.

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

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Hanyang Wang, Chensheng Dai, Min Chen, Yifan Li, Yuxin Li, Yingjie Chen, Hao Liu, Chen Li, Jing Lyu, and Yueqi Duan. Mbench: A comprehensive benchmark on memory capability for video world models, 2026a. URL https://arxiv.org/abs/2606.00793.

Songchun Zhang, Yaowei Li, Junhao Zhuang, Weiyang Jin, Haoyu Wang, Xin Lu, Yilang Sun, Shiyi Zhang, Haoran Li, Xiaoxiao Ma, Yuming Li, Yijun Liu, Yaofeng Su, Yanwen Ma, Haoyu Wu, Zihan Su, Yue Ma, Lvmin Zhang, Haoyang Huang, Zeyue Xue, Anyi Rao, and Nan Duan. Echowm: Open and enterable omnimodal world models, 2026b. URL https://arxiv.org/abs/2608.23189.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, and Yahui Zhou. Matrix-game: Interactive world foundation model, 2025. URL https://arxiv.org/abs/2506.18701.

Yuke Zhao, Wangbo Zhao, Weijie Wang, Zeyu Zhang, Dakai An, Akide Liu, Yinghao Yu, Jiasheng Tang, Fan Wang, Wei Wang, and Bohan Zhuang. Worldolympiad: Can your world model survive a triathlon?, 2026. URL https://arxiv.org/abs/2606.11129.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. URL https://arxiv.org/abs/2503.21755.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer, 2026. URL https://arxiv.org/abs/2605.15178.

APPENDIX

A HUMAN ANNOTATION PROTOCOL

We recruited 10 human annotators to annotate a total of 300 items, with 30 items assigned to each annotator. The annotation task was completed over one day, and each annotator received US$50 for their work. As shown in Figure 8, the annotation interface presents one reference image at a time, together with fields for an environment prompt, a subject prompt, and a time-indexed action sequence. The environment prompt describes the scene layout and background, while the subject prompt specifies the controlled subject's appearance, position, and relationship to the scene. The action sequence specifies the start and end times of each segment and its associated keyboard inputs: `W`, `A`, `S`, and `D` encode movement, while `L`, `I`, `J`, and `K` encode rightward, upward, leftward, and downward view changes, respectively. Simultaneous inputs are joined by `+`, and `IDLE` denotes a segment without input. Annotators save the current annotation before advancing to the next assigned item.

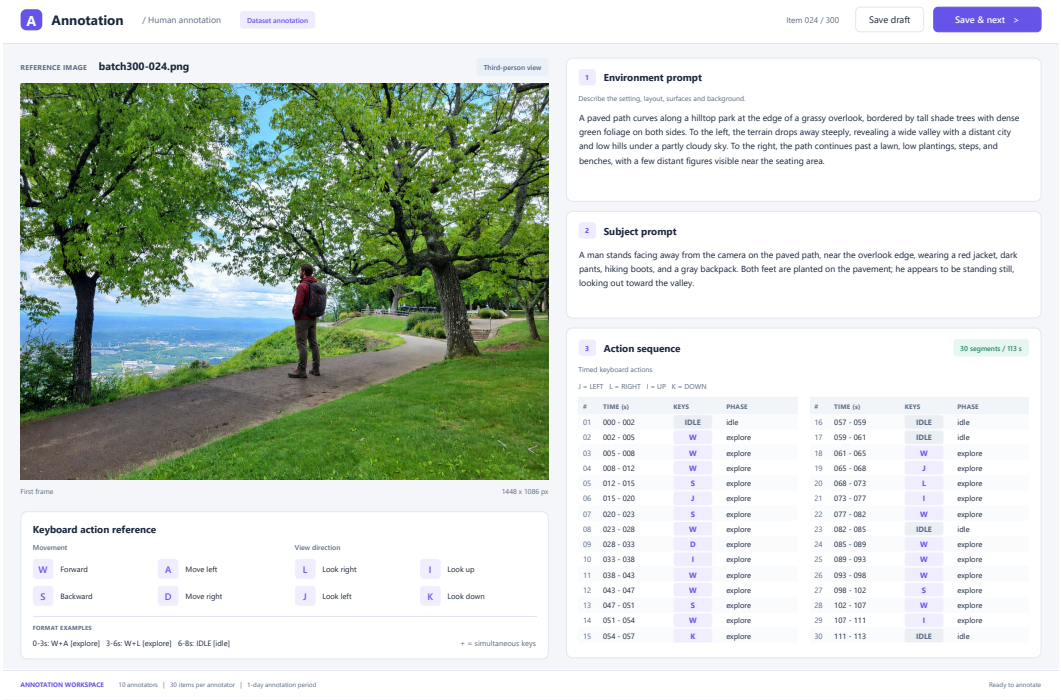

Figure 8: Human annotation interface for the 300-item annotation task. The interface displays the current reference image, environment and subject prompts, and a time-indexed action sequence, together with the WASD movement and LIJK view-control mappings.

B REVISIT PATH TYPES

We categorize revisit paths into the following six types, covering both simple viewpoint changes and complex multi-waypoint navigation.

- **In-place Turn Revisit (ITR):** The camera rotates in place to revisit surrounding views, including full 360° panning. Typical patterns: $L \rightarrow R$, $L \rightarrow R \rightarrow L$, or a full 360° yaw rotation.
- **Step–Look–Return (SLR):** The agent steps forward, looks around, and then steps back to the original position. Typical pattern: $W \rightarrow$ look left/right $\rightarrow S$.

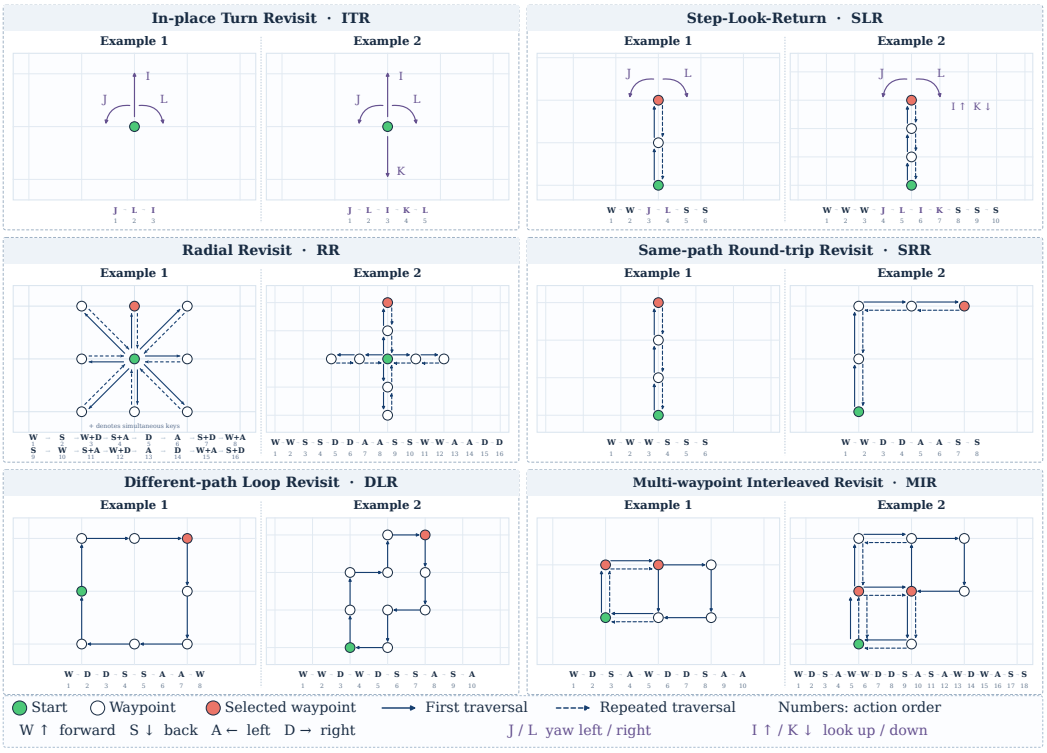

Figure 9: Six categories of revisit trajectories, with two examples per category. WASD controls translation, while IJKL controls camera rotation. Each action sequence corresponds to the illustrated trajectory. Green, white, and red circles denote starting positions, intermediate waypoints, and selected waypoints, respectively. Solid and dashed arrows indicate first and repeated traversals of a path segment; curved arrows indicate camera rotation.

- **Radial Revisit (RR):** The agent departs from a central point in multiple directions and returns to the same center. Typical patterns: $W \to S$, $A \to D$, $W + A \to S + D$, all starting from the same origin.
- **Same-path Round-trip Revisit (SRR):** The agent follows a path and returns along the exact same route. Typical patterns: $W \to S$, $A \to D$, $L \to R$, or longer symmetric sequences.
- **Different-path Loop Revisit (DLR):** The agent returns to the starting area via a different route, forming a closed loop. Typical pattern: $W \to A \to S \to D$, or other non-reversing loops.
- **Multi-waypoint Interleaved Revisit (MIR):** The agent visits multiple waypoints in sequence and later revisits them in an interleaved order. Typical patterns: $A \to B \to C \to B \to A$, or $A \to B \to A \to B$.

C Further Analysis.

C.1 Metric Redundancy and Score Robustness.

Pairwise Spearman correlations indicate limited redundancy among the metrics (Figure 10 in Appendix C.1), with a mean absolute correlation of only 0.22 across valid off-diagonal pairs. The strongest correlation is between GMS and LMS ($\rho = 0.76$), as expected from their shared DINOv2-based similarity formulation, followed by PC–CC (0.67) and TR–CR (0.60), both within semantically related dimensions. Overall, within-dimension correlations are higher than cross-dimension correlations (0.38 vs. 0.20), suggesting that the five dimensions are internally coherent yet largely



Figure 10: **Metric correlation.**

complementary. Dynamic Degree is a notable exception, showing negative correlations with several consistency and memory metrics (e.g., SC -0.48, JP -0.34, GMS/LMS $-0.30/-0.34$), highlighting a motion–consistency trade-off and motivating joint reporting of these metrics. The evaluation is also robust to metric parameterization: sweeping the translation tolerance ($40°$–$80°$) and rotation threshold ($2°$–$8°$) over 20 settings preserves model rankings with $\rho \geq 0.964$ and shifts no model by more than four positions.

C.2 Additional Error Analysis

Figure 11 extends the main-text analysis with the remaining model profiles, deduplicated category prevalence, temporal annotations, metric associations, and label co-occurrence. Echo-WM exhibits substantial appearance and geometry/layout drift (74% and 73%), whereas Lingbot-World-v2 has the highest observed prevalence of abrupt temporal jumps (42%). Lyra 2.0 exhibits a mixed profile, combining quality degradation (56%) with appearance and geometry/layout drift (55% and 48%). These differences further support treating model failures as multidimensional profiles rather than a single ordering of reliability.

The timing records in panel 11(c) were obtained through a separate coarse temporal-localization annotation step and were not produced by the video-level failure-type prompt.The supplementary analyses provide further diagnostic context. Quality degradation is localized late in 86.1% of the supplied timing records and 86.6% of records matching the video-level labels. In contrast, the largest temporal share for under-response shifts from early to middle after matching, preventing a firm interpretation as an initialization failure. The supplied metric differences associate under-response with lower TR and CR, and geometry/layout drift with lower GMS and LMS. These patterns, together

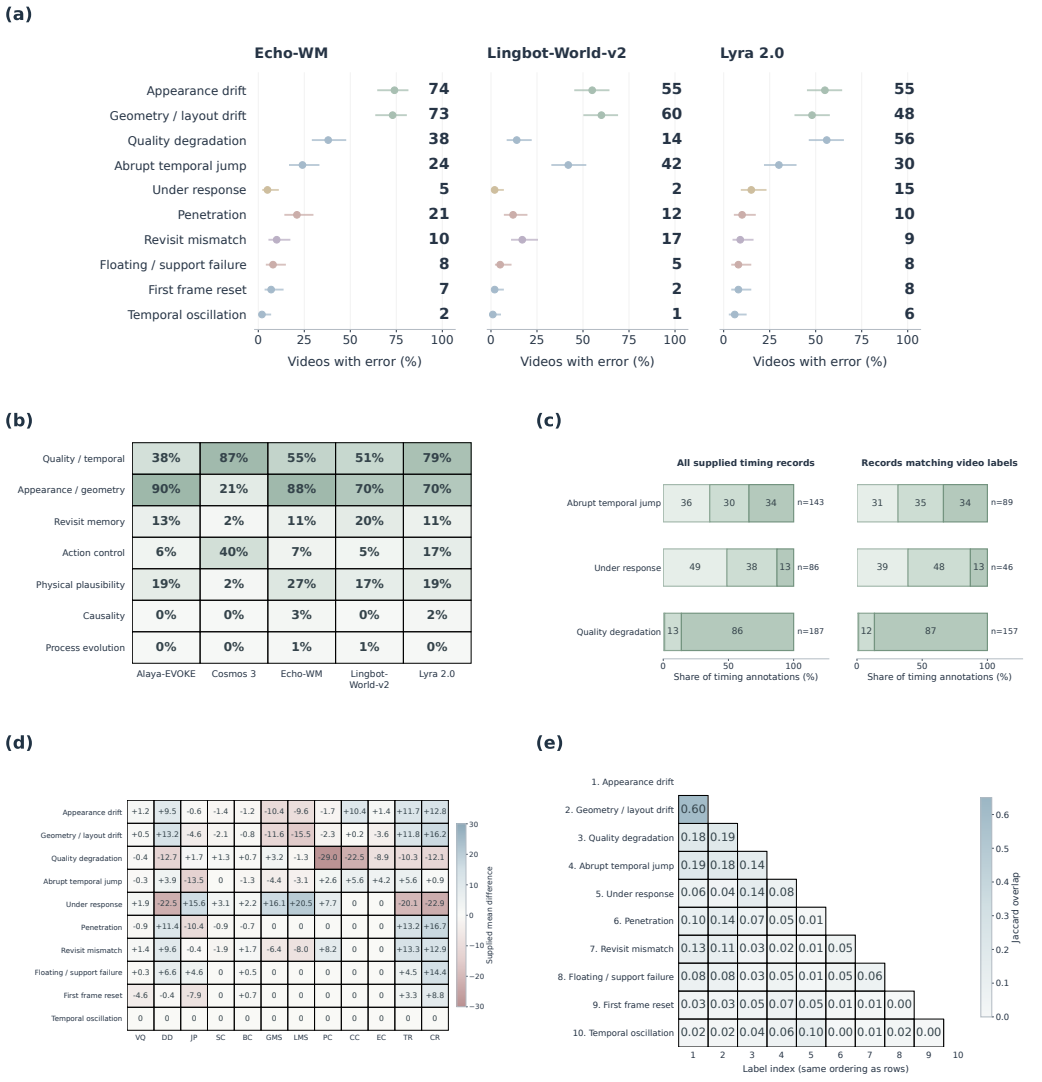

Figure 11: Supplementary error analysis. (a) Remaining model profiles with marginal 95% Wilson intervals. (b) Video-level category prevalence, deduplicated within categories. (c) Early/middle/late distributions from a separate auxiliary temporal-localization annotation, shown for all timing records (left) and records matching the video-level labels (right); n denotes the annotation count. (d) Exploratory mean metric differences between videos with and without each label; these auxiliary results are reported descriptively without significance claims. (e) Pairwise Jaccard overlap of the ten most frequent labels.

with the co-occurrence analysis, provide descriptive support for evaluating control responsiveness and scene consistency as complementary dimensions.

These findings require cautious interpretation. Only 772 of 1,114 timing records match the video-level labels after one spelling correction, and the matching subset is a sensitivity check rather than verified ground truth. Per-video metric values and coverage are unavailable, so the metric associations support neither significance nor causal claims. Moreover, sparse causality and process-evolution labels cannot establish that these capabilities are solved without accounting for observation opportunities and evaluator sensitivity.

Table 5: Inference configurations of the evaluated interactive world models. "Conditioning" denotes the input reference image (Img.) and text, while "Control" denotes the native interface used for benchmark actions. "–" indicates unavailable or unspecified information. Latency is the average inference time per case; resolution and FPS refer to generated videos used for evaluation.

Model	Conditioning	Control	Latency (s)	Resolution	FPS
Genie 3 (Parker-Holder & Fruchter, 2025)	Img. + Text	Keyboard	🔒	$\sim 1280 \times 704$	16
HappyOyster (HappyOyster, 2026)	Img. + Text	Action	🔒	$\sim 832 \times 480$	24
LingBot-World-v2 (RobbyAnt, 2026)	Img. + Text	Camera pose	587	$\sim 832 \times 480$	16
Matrix-Game 2.0 (He et al., 2025)	Img.	Keyboard + Mouse	160	640×352	12
Matrix-Game 3.0 (Wang et al., 2026)	Img. + Text	Keyboard + Mouse	550	1280×704	16
Yume-1.5 (Mao et al., 2026)	Img. + Text	Control tokens	1802	1280×704	16
DreamX-World (Team et al., 2026a)	Img. + Text	Camera pose	769	1280×704	16
ABot-World (Jiang et al., 2026)	Img. + Text	Keyboard	191	832×480	12
Cosmos 3 (NVIDIA, 2026)	Img. + Text	Relative pose	1480	480p short side	16
Echo-WM (Zhang et al., 2026b)	Img. + Text	Camera trajectory	1134	1280×704	24
Alaya-EVOKE (Yin et al., 2026)	Img. + Text	Camera pose	962	640×384	24
Lyra 2.0 (Shen et al., 2026)	Img. + Text	Camera pose	6446	832×480	16

D Inference Protocols and Model Configurations

The evaluated models differ in their conditioning inputs and control interfaces. We therefore translate each benchmark case into the native inference format of each model, preserving the reference image, textual context, prescribed controls, and requested duration whenever supported. Table 5 summarizes the configurations used in our evaluation.

For models with native keyboard, mouse, or action interfaces, the benchmark controls are mapped directly to the corresponding movement and camera commands. For trajectory-conditioned models, the prescribed controls are converted into their native camera representations, including camera poses, Plücker embeddings, PRoPE, or UCPE where applicable. Translation and rotation are applied jointly when simultaneous controls are supported, while intervals without prescribed commands are represented by neutral or zero-motion inputs.

Several model-specific adaptations are necessary because the available interfaces are not identical. Genie 3 is evaluated through the hosted Project Genie interface, where controls are issued through browser automation. The deployed interface does not expose direct control over the underlying checkpoint, random seed, sampling steps, or guidance scale. Matrix-Game 2.0 is evaluated with image conditioning and keyboard/mouse actions without additional textual conditioning, whereas Matrix-Game 3.0 additionally accepts a text prompt. DreamX-World represents the prescribed actions through camera motion and therefore does not provide an independently commanded character state. Lyra 2.0 similarly uses camera trajectories; for third-person cases, it constructs a look-at trajectory around a proxy subject position. Cosmos 3 is evaluated through its forward-dynamics action interface: yaw is supported, whereas pitch has no corresponding action in this configuration, and turning in place is approximated with a small forward motion.

Natural-language interventions also differ across interfaces. LingBot-World-v2 and Echo-WM can incorporate event descriptions as global text conditioning, but the descriptions are not activated at their prescribed execution times. We therefore denote these interfaces as *Global*. For models whose evaluated interfaces do not support timed natural-language interventions, such cases are treated as unsupported rather than approximated using unrelated movement or camera controls.

D.1 Evaluation Metrics

We evaluate generated videos along five complementary dimensions: perceptual quality, temporal consistency, revisit memory, physical and causal validity, and action controllability. Table 6 summarizes the evaluation methods and input configurations.

Table 6: Evaluation methods and input configurations for video metrics. Listed FPS values denote target sampling rates; native FPS retains all decoded frames. Spatial preprocessing preserves aspect ratio unless a fixed size is specified. A dash indicates an unspecified setting.

Dimension	Metric	Method / Backbone	Temporal Sampling	Spatial Preprocessing
Perceptual Quality	Visual Quality (VQ)	MUSIQ Ke et al. (2021)	2 FPS	Long side $\leq$ 512 px
	Dynamic Degree (DD)	Farnebäck flow Farnebäck (2003)	12 FPS	Long side $\leq$ 512 px
Temporal Consistency	Jump-Based Continuity (JP)	TransNetV2 Soucek & Lokoc (2024)	Native FPS	48×27 px ($W \times H$)
	Subject Consistency (SC)	DINOv2-B/14 + CLIP-B/16	3 FPS	DINO-S256; CLIP: -
	Background Consistency (BC)	CLIP-B/32	2 FPS	-
Revisit Memory	Global Memory Score (GMS)	DINOv2-B/14	Observation midpoints[a]	DINO-S256
	Local Memory Score (LMS)	DINOv2-B/14	Observation midpoints[a]	DINO-S256
Physical & Causal Validity	Physical Causality (PC)	Gemini 3.1 Pro Preview	8 global + interval samples[b]	Long side $\leq$ 640 px
	Content Causality (CC)	Gemini 3.1 Pro Preview	8 global + interval samples[b]	Long side $\leq$ 640 px
	Evolution (EC)	Gemini 3.1 Pro Preview	8 global + interval samples[b]	Long side $\leq$ 640 px
Action Controllability	Translation Response (TR)	ViPE poses Huang et al. (2025)	Poses within action segments	-
	Camera Rotation Response (CR)	ViPE poses Huang et al. (2025)	Poses within action segments	-

Notes. DINO-S256 denotes resizing to a 256-pixel short side without center cropping, followed by normalization and padding to multiples of 14. [a] The nearest valid frame to each observation-window midpoint is selected. [b] Eight uniformly spaced video-wide timestamps are supplemented with each temporal interval's start, midpoint, and end.

D.1.1 Perceptual Quality

Visual Quality (VQ). VQ measures frame-level perceptual quality using MUSIQ Ke et al. (2021). The video score is the mean prediction over sampled frames, capturing image quality without directly assessing temporal coherence or instruction adherence.

Dynamic Degree (DD). DD measures motion intensity using Farnebäck optical flow Farnebäck (2003). For each adjacent sampled-frame pair, we remove the largest 1% of flow magnitudes and average the remainder. After normalization by the nominal sampling interval and the image short-side length relative to 256 pixels, the mean over frame pairs gives $d(v)$. We define

$$\mathrm{DD}(v) = 100\sqrt{\min\left(1, \frac{d(v)}{d_{\mathrm{ref}}}\right)}, \qquad d_{\mathrm{ref}} = 102.889, \tag{1}$$

where d_{ref} is the fixed 95th-percentile reference from the calibration pool. DD quantifies the magnitude of camera and scene motion. Physical correctness is evaluated separately by PC.

D.1.2 Temporal Consistency

Jump-Based Continuity (JP). JP measures abrupt discontinuities using frame-wise cut probabilities from TransNetV2 Soucek & Lokoc (2024). Predictions from overlapping temporal windows are averaged. For the resulting probabilities p_t,

$$\mathrm{JP}(v) = 100\left(1 - \max_t p_t\right). \tag{2}$$

Higher scores indicate weaker evidence of a discontinuity. The score is determined by the strongest predicted transition.

Subject Consistency (SC). SC combines adjacent-frame appearance stability with retention relative to the initial frame. Valid subject masks are used to gray out the background and crop around the subject; otherwise, the full frame is retained. For N sampled frames, let d_i and c_i denote DINOv2 ViT-B/14 features Oquab et al. (2024) and CLIP ViT-B/16 features Radford et al. (2021). Using $s(x, y) = \max\{0, \cos(x, y)\}$,

$$\mathrm{SC}(v) = \frac{100}{2(N-1)} \sum_{i=1}^{N-1} \left[s(d_{i-1}, d_i) + s(c_0, c_i)\right]. \tag{3}$$

SC is evaluated on subject-mask-eligible cases.

Background Consistency (BC). BC applies the same adjacent-frame and initial-frame comparison structure to CLIP ViT-B/32 background features b_i:

$$\mathrm{BC}(v) = \frac{100}{2(N-1)} \sum_{i=1}^{N-1} \left[s(b_{i-1}, b_i) + s(b_0, b_i) \right]. \tag{4}$$

Subject regions are grayed out where masks are available; otherwise, the full frame is used. SC and BC assess subject and background appearance consistency across frames.

D.1.3 Revisit Memory

Revisit memory evaluates appearance retention at predefined observation anchors. We extract L2-normalized, full-frame DINOv2 ViT-B/14 features from the nearest valid frames to observation-window midpoints. For anchor a, let $z_{a,0}$ denote the initial observation and $z_{a,1}, \ldots, z_{a,K_a}$ its revisits. Comparisons remain within the same anchor.

Global Memory Score (GMS). GMS compares each revisit with the initial observation:

$$G_a = \frac{\sum_{k=1}^{K_a} w_{a,k}\, \bar{s}(z_{a,0}, z_{a,k})}{\sum_{k=1}^{K_a} w_{a,k}}, \qquad w_{a,k} = t_{a,k} - t_{a,0}, \tag{5}$$

where $t_{a,k}$ is the scheduled observation time and $\bar{s}$ is cosine similarity clipped to $[0, 1]$. Elapsed-time weighting gives later revisits greater influence.

Local Memory Score (LMS). LMS compares successive revisits:

$$L_a = \frac{\sum_{k=2}^{K_a} w_{a,k}\, \bar{s}(z_{a,k-1}, z_{a,k})}{\sum_{k=2}^{K_a} w_{a,k}}. \tag{6}$$

The initial-to-first-revisit pair is excluded to avoid duplicating its contribution in GMS. Both metrics use time since the initial observation as the weight. Anchor scores are averaged equally within each video; LMS is unavailable when there are insufficient revisits. Model-level aggregation uses complete cases with all prescribed comparison pairs available. These metrics assess visual retention at scheduled checkpoints.

D.1.4 Physical and Causal Validity

PC, CC, and EC use scene-specific checklists shared across models for each case. Gemini 3.1 Pro Preview evaluates each checklist item on a 1–5 scale, from complete failure to correct and natural behavior, and provides supporting visual evidence. Persistent constraints use video-wide observations; temporal checks additionally request interval-specific samples, as summarized in Table 6. Checks requiring change are judged by comparing states over time, rather than by initial object presence.

Physical Causality (PC). PC evaluates whether scene contents obey applicable physical constraints, including gravity, buoyancy, collision response, deformation, material behavior, thermodynamics, and non-penetration of solid boundaries. Checks distinguish persistent properties, such as support, from event-specific responses, such as deformation under impact.

Content Causality (CC). CC evaluates whether a scene-supported cause is accompanied by a visually consistent effect on another object or region, such as illumination producing corresponding shadows or reflections. The causal source is a scene object or natural phenomenon, rather than the commanded camera or agent action.

Evolution Consistency (EC). EC evaluates spontaneous motion and state changes independent of the commanded action sequence, such as drifting, flowing, or melting. It requires visible progression over the specified interval; maintaining a static property does not constitute evolution.

D.1.5 Action Controllability

We evaluate camera-action following using trajectories recovered by ViPE Huang et al. (2025). Each non-idle action segment is assessed using its first and last available camera-to-world poses, (R_s, t_s) and (R_e, t_e).

Translation Response (TR). TR compares the commanded direction with displacement in the segment's initial camera coordinates:

$$\Delta t_{\text{cam}} = R_s^\top (t_e - t_s). \tag{7}$$

Forward, backward, left, and right correspond to $+z$, $-z$, $-x$, and $+x$; combined commands sum their direction vectors. A segment is correct when the recovered and expected directions differ by at most 60°. Zero or undefined displacements do not count as correct.

Camera Rotation Response (CR). CR evaluates yaw and pitch extracted from $R_s^\top R_e$. A commanded rotation is correct when its sign matches the instruction and its absolute magnitude is at least 4°. Commands involving both axes must satisfy both conditions.

For correct counts C_T, C_R and eligible segment counts N_T, N_R,

$$\text{TR}(v) = 100\frac{C_T}{N_T}, \qquad \text{CR}(v) = 100\frac{C_R}{N_R}. \tag{8}$$

Combined translation–rotation segments contribute to both metrics. Segments with insufficient pose coverage remain in the denominator, while segments entirely within an identified corrupt prefix are excluded. Model scores average eligible videos equally. TR and CR quantify directional agreement between the prescribed controls and the recovered camera motion.

Aggregation Protocol. The main table reports video-equal macro averages, whereas Figure 5 pools valid action segments across videos within each action type (micro averaging). Consequently, the two results may differ because videos contain different numbers of action segments. For example, DreamX-World obtains a pooled rotation accuracy of 89.04% $(6497/7297)$, compared with the video-equal score of 90.32% in Table 2.

E Audio Evaluation

We evaluate audio generated by Echo-WM and HappyOyster using five complementary measures: Audio Quality (AQ), ImageBind (Girdhar et al., 2023) similarity (IB), Synchformer (Iashin et al., 2024) offset (SF), Audio Realism (AR), and Action–Audio Correspondence (AAC). These measures characterize the sound itself and its relationship with visual content. Gemini 3.1 Pro Preview serves as the judge for AR and AAC.

Audio Quality (AQ). Audiobox Aesthetics (Tjandra et al., 2025) evaluates the complete audio track and predicts Production Quality (PQ) and Content Usefulness (CU). We define a composite score as their unweighted mean:

$$\text{AQ}_v = \frac{\text{PQ}_v + \text{CU}_v}{2}. \tag{9}$$

ImageBind Similarity (IB). IB measures audio–visual semantic similarity. Within annotated non-idle action intervals, we exclude 0.25 seconds at each boundary and sample six-second windows where sufficient duration remains. Each window pairs its audio with the midpoint video frame. For audio and visual embeddings $\mathbf{a}_i$ and $\mathbf{v}_i$, the window score is

$$c_i = \frac{\mathbf{a}_i^\top \mathbf{v}_i}{\|\mathbf{a}_i\|_2 \|\mathbf{v}_i\|_2}. \tag{10}$$

The video score averages valid windows using newly covered duration as the weight, reducing the influence of an overlapping final window.

Synchformer Offset (SF). SF estimates audio–visual temporal offset in five-second windows using the same action-interval boundary rule. For predicted offset Δ_i and newly covered duration w_i, the video-level mean absolute offset is

$$\text{SF}_v = \frac{\sum_{i \in \mathcal{W}_v} w_i |\Delta_i|}{\sum_{i \in \mathcal{W}_v} w_i}, \tag{11}$$

where $\mathcal{W}_v$ contains valid windows. SF is treated as exploratory because its predictions did not reliably recover known synthetic shifts during validation.

Audio Realism (AR). We divide each video into consecutive segments of up to 20 seconds. Gemini 3.1 Pro Preview evaluates each segment using audio alone, assessing acoustic texture, continuity, and signal integrity on a 1–5 scale. AR measures acoustic realism, while AAC evaluates correspondence with visible actions.

Action–Audio Correspondence (AAC). Within the same segments, Gemini 3.1 Pro Preview independently describes audible events in four-second audio windows and visible actions in matching windows of a silent video sampled at 2 FPS. The same model then compares these descriptions and assigns a 1–5 score when sufficient evidence is available. AAC measures whether identifiable actions have appropriate sound responses and whether their changes correspond within the same temporal window.

For $m \in \{\mathrm{AR}, \mathrm{AAC}\}$, segment scores are aggregated by duration:

$$m_v = \frac{\sum_{j \in \mathcal{S}_v^m} \ell_j s_j^m}{\sum_{j \in \mathcal{S}_v^m} \ell_j}, \tag{12}$$

where $\mathcal{S}_v^m$ contains scored segments, s_j^m is the segment rating, and ℓ_j is its duration. For all window- and segment-based audio metrics, aggregation uses only valid scores. Videos with no valid observations receive no score, and temporal coverage is reported alongside the averages.

F Human Agreement

F.1 Experimental Setup

We construct the human-agreement subset by sampling 50 cases from five representative world models: Alaya-EVOKE, Echo-WM, LingBot-World-v2, Cosmos 3, and Lyra 2.0. Each case is evaluated for all five models, resulting in 250 generated videos (50 cases $\times$ 5 models).

Each video is independently evaluated by two human annotators. The annotators are blind to the identity of the generating model and to the corresponding automatic evaluation scores, and the samples are presented in randomized order. Both annotators evaluate the full set of 250 videos, such that every applicable evaluation dimension receives two independent human annotations.

F.2 Human Annotation Protocol

All dimensions use a unified five-point ordinal scale from 1 to 5, while the interpretation of the scale follows the semantics of each metric. For quality-, consistency-, physics-, causality-, evolution-, and controllability-oriented dimensions, higher scores indicate stronger satisfaction of the corresponding property. Specifically, 1 denotes a clear violation of the target property, 2 indicates that the property is mostly violated, 3 indicates partial satisfaction, 4 indicates that the property is largely satisfied, and 5 indicates that it is fully satisfied.

DD is treated differently because it measures *motion magnitude* rather than motion quality. For DD, a low score indicates little visible motion and a high score indicates substantial visible motion, regardless of whether that motion is perceptually desirable.

Each valid annotation must be accompanied by a brief rationale. If the required evidence is absent or the corresponding property cannot be reliably judged, the annotation is marked as `null`. Such samples are excluded from the corresponding agreement computation.

The human annotation criteria are designed to match the intended semantics of the automatic metrics as closely as possible. We use separate protocols for signal-level and checklist-based metrics.

F.3 Metric-Specific Annotation Rubrics

Signal-level metrics. For VQ, DD, JP, SC, BC, GMS, LMS, TR, and CR, annotators assign one dimension-level score based on temporal samples distributed uniformly over the full video duration. We use 16 uniformly sampled timestamps for each video. For the controllability metrics TR and CR, the corresponding camera-action sequence is additionally provided.

Table 7: Human annotation rubric for the nine signal-level metrics. The annotation criterion follows the intended semantics of each automatic metric.

Metric	Human annotation criterion	Automatic metric
VQ	Visual clarity, texture quality, blur, and compression artifacts	MUSIQ
DD	Amount of visible motion; measures motion magnitude rather than quality	Dense optical flow: $100\sqrt{\min(1, \overline{u_{99}}/102.889)}$
JP	Absence of abrupt cuts or discontinuous temporal jumps; no cut is better	TransNetV2: $100(1 - \max_t p_t)$
SC	Stability of subject identity, appearance, and geometry	DINOv2 subject-feature similarity
BC	Stability of background appearance and structure	DINOv2 background-feature similarity
GMS	Consistency between a revisited anchor and its first observation	DINOv2 cosine similarity between revisit and first observation
LMS	Consistency across repeated revisits of the same anchor	DINOv2 cosine similarity between successive revisits
TR	Whether camera translation follows the commanded W/S/A/D direction	VIPE translation; angular error $\leq 60^\circ$
CR	Whether camera yaw/pitch rotation follows the commanded direction	VIPE relative rotation; correct sign and magnitude $\geq 4^\circ$

Table 8: Human annotation rubric for checklist-based metrics. Only items enabled by the case-specific applicability gate are scored; the dimension score is the mean over valid checklist items.

Dimension	Checklist family	Human annotation criterion
PC	`gravity`	Objects respect gravitational support and do not unnaturally float, sink, or tilt.
PC	`material`	Material appearance and rigidity remain plausible without unintended liquefaction, stretching, or jitter.
PC	`collision`	Contacting objects respect physical boundaries without penetration or interpenetration.
PC	`buoyancy`	Floating objects exhibit plausible buoyancy rather than unexplained sinking or upward motion.
PC	`thermodynamics`	Phase and thermal state remain compatible with the environment, without unsupported melting, freezing, or boiling.
PC	`deformation`	Deformable materials respond plausibly to force without remaining unrealistically rigid or undergoing implausible breakage.
CC	`cross_object_cause`	Effects remain causally coupled to their causes, e.g., reflections follow viewpoint changes and attached objects remain connected.
EC	`motion_change`	Ongoing dynamic processes continue to evolve rather than freezing into static appearance.
EC	`state_change`	Slowly evolving states progress smoothly without unexplained reversal, reset, or discontinuity.

Table 7 summarizes the annotation target and the corresponding automatic metric. Importantly, annotators follow the semantics of each metric rather than a generic notion of video quality. For DD, an otherwise high-quality but nearly static video should still receive a low score because DD measures the amount of motion. For JP, the scoring direction is opposite to that of a shot-boundary detector: the absence of abrupt cuts or temporal jumps corresponds to a higher JP score.

Applicability and exclusion rules. Annotations are marked as `null` whenever the required evidence is absent or the corresponding property cannot be reliably judged. A `null` sample is excluded from both sides of the corresponding agreement computation.

Revisit-based metrics are defined only when the required anchor is actually revisited; LMS additionally requires at least three observations of the anchor. TR and CR are defined only for trajectories containing translation and rotation commands, respectively. For trajectories with multiple action

segments, both the human annotation and the automatic metric are computed over the complete relevant trajectory rather than an isolated action window.

F.4 Agreement Computation

Let $S_h^{(d)}(i,m) \in \{1,\dots,5\}$ denote the score assigned by human annotator h to dimension d for sample (i,m), and let $\mathcal{S}^{(d)}$ denote the set of samples for which valid scores are available for dimension d. Following T2AV-Compass (Cao et al., 2026), we quantify agreement using mean L_1 distance on the 1–5 scale, where lower values indicate stronger agreement.

Human–human agreement (H–H). Inter-annotator agreement is computed as

$$L_1^{\text{H-H}}(d) = \frac{1}{|\mathcal{S}^{(d)}|} \sum_{(i,m)\in\mathcal{S}^{(d)}} \left| S_{h_1}^{(d)}(i,m) - S_{h_2}^{(d)}(i,m) \right|. \tag{13}$$

Automatic–human agreement (A–H). We first define the human consensus as

$$C^{(d)}(i,m) = \frac{1}{2}\left(S_{h_1}^{(d)}(i,m) + S_{h_2}^{(d)}(i,m) \right). \tag{14}$$

Agreement between the original automatic metric and the human consensus is then computed as

$$L_1^{\text{A-H}}(d) = \frac{1}{|\mathcal{S}^{(d)}|} \sum_{(i,m)\in\mathcal{S}^{(d)}} \left| C^{(d)}(i,m) - \hat{S}_{\text{auto}}^{(d)}(i,m) \right|. \tag{15}$$

For PC, CC, and EC, the automatic evaluator and human annotators use the same checklist structure and the same 1–5 item-level scale, so no additional transformation is required. For the nine signal-level metrics, whose normalized automatic outputs lie in $[0,100]$, we apply the fixed linear transformation

$$\hat{S}_{\text{auto}} = 1 + \frac{4 s_{\text{auto}}}{100}, \qquad s_{\text{auto}} \in [0,100], \tag{16}$$

which maps the automatic scores onto the same 1–5 scale while preserving their ordering.

We report L_1 distance rather than exact-match accuracy because the ratings are ordinal and L_1 retains the magnitude of disagreement. The overall mean reported in Table 3 is the macro-average over the twelve evaluation metrics.

G Variation Across LLM Evaluators

G.1 Experimental Setup

We further evaluate the sensitivity of the agreement analysis to the choice of LLM evaluator. This experiment uses the same 250 generated videos and the same human consensus defined in Section F.

We consider four candidate LLM evaluators: Gemini-3.1-Pro-Preview, Gemini-2.5-Pro, Gemini-3.7-Flash, and GPT-5.6-Sol. Each candidate evaluator independently scores the same evaluation dimensions using the human annotation criteria described in Section F.2 and Section F.3. Its predictions are then compared against the same human consensus using the mean L_1 distance.

For the nine signal-level dimensions (VQ, DD, JP, SC, BC, GMS, LMS, TR, and CR), each candidate LLM judge produces a 1–5 score according to the corresponding metric-specific rubric. For PC, CC, and EC, each judge answers the same applicability-gated checklist used in the human annotation protocol, and the dimension score is obtained by averaging the valid item scores.

We report the mean L_1 distance separately for the nine signal-level dimensions and the three checklist-based dimensions. Each group-level value is computed as a macro-average over its constituent dimensions. All candidate evaluators use the same samples, human reference, annotation rubric, score range, applicability rules, and aggregation procedure, so that the comparison isolates variation associated with the choice of LLM evaluator.

H QUALITATIVE CASE STUDIES

We examine three cases from the two task suites: ongoing scene evolution, world persistence across repeated revisits, and sustained subject control under combined actions.

Case 1: Ongoing Scene Evolution. Figure 12 presents `case 1`, an 80-second street scene with a freely floating red balloon. Cosmos 3 shows coherent upward motion, while the other examples exhibit color changes, reappearance after disappearance, scene drift, or little visible progression.

Case 2: World Persistence across Repeated Revisits. Figure 13 presents `case 2`, which repeatedly leaves and returns to a nighttime roadwork anchor. LingBot-World-v2 preserves the anchor across the displayed revisits, while other examples exhibit identity changes, viewpoint mismatch, anchor loss, or scene drift.

Case 3: Sustained Control under Combined Actions. Figure 14 presents `case 3`, which requires a maintenance robot to execute combined actions and direction changes. HappyOyster shows coherent subject motion, while other examples exhibit reversed motion, identity changes, scene drift, inactivity, or subject loss.



Figure 12: Qualitative comparison of ongoing scene evolution. Cosmos 3 shows coherent balloon ascent, while other examples exhibit color changes, reappearance after disappearance, scene drift, or little visible progression.

Case 2

Env Prompt: A nighttime roadwork site occupies the street behind linked blue plastic barriers and traffic cones. A large white tipper truck stands left of center, with a white excavator partly visible to the right. Road-closure and CCTV warning signs face the foreground, while workers in orange high-visibility clothing appear to be attending to the area near the truck. A black lamppost partially occludes the excavator. Tall brick buildings with windows form the background under work lights.

Perspective Prompt: First-person viewpoint from the controlled observer, with hands visibly holding the object identified in the subject prompt.

Sub Prompt: Two bare hands extending from dark ribbed sleeves hold the left and right edges of a folded city route map across the lower foreground, leaving the truck, excavator, and warning signs visible.

Action list

W+I (2s) → S+K (2s) → IDLE (2s) → A+I (2s) → A+K (2s) → IDLE (4s) → IDLE (2s) → D+I (2s) → D+K (2s) → IDLE (2s) → W+I (2s) → W+K (2s) → W+K (2s) → S+I (2s) → IDLE (2s) → K (2s) → I (2s) → K (2s) → I (2s) → J+I (2s) → ...

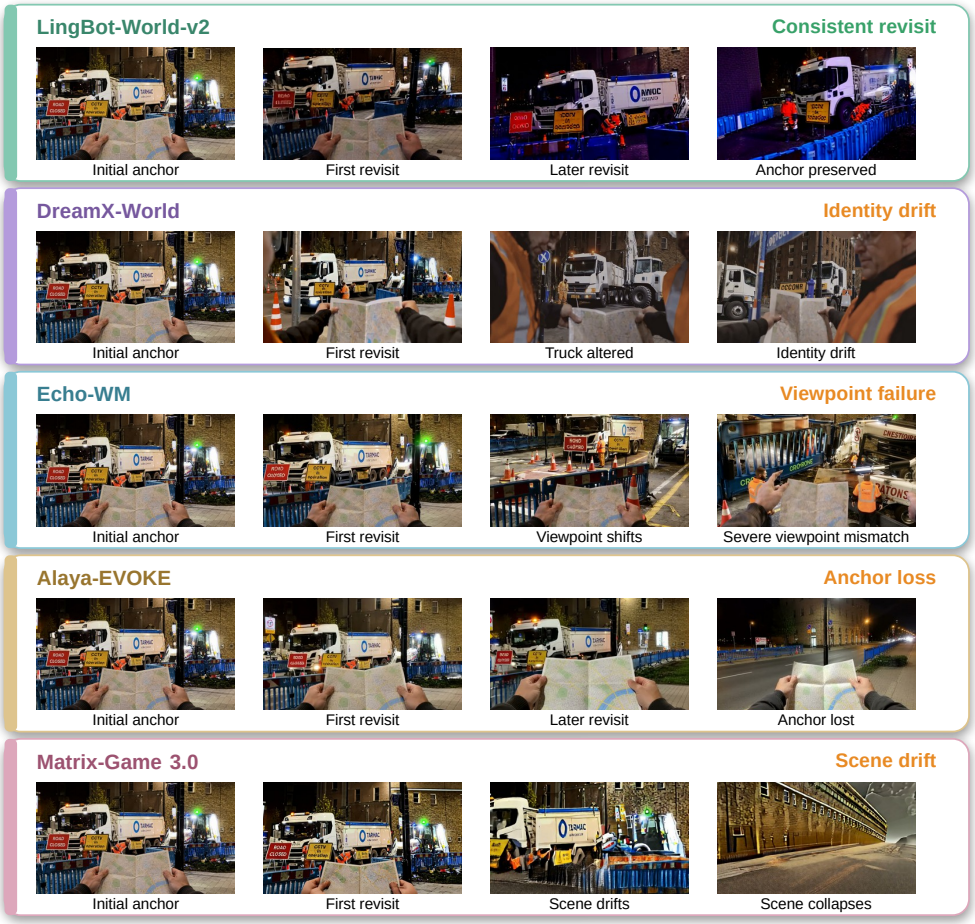

Figure 13: Qualitative comparison on world persistence across repeated revisits. LingBot-World-v2 preserves the initial anchor most consistently, whereas other models exhibit identity drift, viewpoint failure, anchor loss, or scene drift.

Case 3

Env Prompt: Inside a large shipyard maintenance hall, a rust-colored steel hull rests on supports and is surrounded by multi-level scaffolding. Red welding curtains form mid-ground partitions; a worker appears to be welding in front of one, with visible sparks. Gas cylinders, hoses, tool carts, pallets, and steel frames occupy the foreground and sides. Yellow lines define a clear passage across the concrete floor. Overhead crane rails and a suspended hook span the hall, while a wide open doorway reveals cranes and a docked ship outside.

Perspective Prompt: Third-person camera view with one visibly identified controlled subject; use the subject prompt for its appearance, location, scale, and support relationship.

Sub Prompt: A medium-sized yellow-paneled bipedal maintenance robot occupies the marked passage on the right side of the hall, facing away from the camera in a rear-oblique orientation. It is mid-step, with one foot contacting the concrete and the other lifted.

Action list

K (3s) → W+A+J+K (3s) → S+A+J (3s) → S (3s) → W+D+L (3s) → W+D+L+K (3s) → IDLE (2s) → W+J (2s) → D+L+K (3s) → W (5s) → S+D+L (2s) → W+D+L+I (3s) → S+D+L+K (2s) → S+I (4s) → W (5s) → D+L+I (4s) → ...

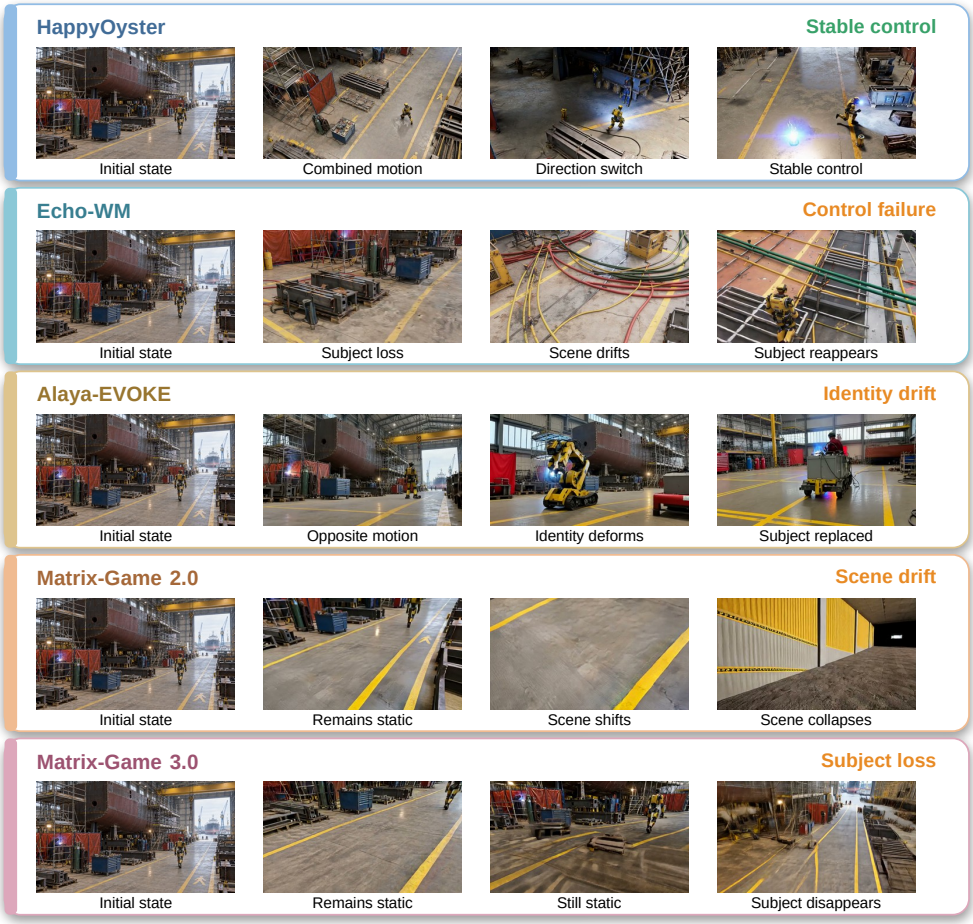

Figure 14: Qualitative comparison on sustained control under complex action sequences. HappyOyster maintains stable control throughout the trajectory, while other models exhibit control failure, identity drift, scene drift, or subject loss.

I PROMPTS

I.1 PROMPT REWRITING AND VERIFICATION

To improve the descriptive completeness of the initial-frame prompts, we use a multi-model rewriting and verification procedure involving GPT-5.6 Sol (OpenAI, 2026a), Gemini 3.8 Flash (Google, 2026b), Claude Sonnet 5 (Anthropic, 2026), and Qwen3.8-Max (Alibaba Cloud, 2026). For each case, the initial image and the original perspective, environment, and subject descriptions are provided as input. Two models independently rewrite the environment and subject fields, while a third model verifies the candidates against the image and source descriptions. Model assignments follow a balanced fixed rotation across data splits, source origins, viewpoints, and subject presentations. Candidate selection prioritizes source-fact preservation, visual grounding, spatial relations, and subject consistency. The perspective description remains unchanged, and all candidate outputs, verification decisions, and model identifiers are retained for auditing.

We apply this procedure uniformly to 300 cases, with the four models receiving approximately equal numbers of candidate-generation and verification assignments. After rewriting, the median combined length of the environment and subject descriptions increases from approximately 47 to 108 words, with the environment description increasing from 25 to 75 words and the subject description from 22 to 35 words. The rewritten prompts provide richer descriptions of stable landmarks, spatial layout, foreground–background relations, occlusions, support and contact relations, and the observable states of dynamic objects. This improves the completeness and visual verifiability of the initial scene descriptions for long-horizon navigation, state persistence, revisit consistency, and interactive behavior evaluation.

Image-grounded rewriting prompt

You are an image-grounded prompt writer for a long-horizon interactive world-model benchmark.
Given the initial image and the source perspective, environment, and subject descriptions, rewrite only the environment and subject fields. Keep the perspective description unchanged.
Preserve important source facts and add only details supported by the image. Describe stable landmarks, spatial layout, foreground/background relations, paths, openings, occlusions, support, contact, and visible object states.
For potentially dynamic elements, describe their visible current state. Describe an apparent activity only when supported by visual evidence such as pose, motion blur, displacement, flowing, or waving. Do not infer hidden intentions, goals, or future actions. A car on a road is not necessarily driving, and a person near a table is not necessarily eating.
For first-person views, describe only visible hands, arms, held objects, or interfaces. For third-person views, describe only the visible controlled subject. Do not invent unseen body parts, objects, identities, text, exact counts, hidden spaces, or unsupported functionality. Avoid poetic or cinematic filler.
Return JSON only in the following format:

```
{
  "rewrite_environment_prompt_en": "...",
  "rewrite_subject_prompt_en": "..."
}
```

Use soft length targets of approximately 60–85 words for the environment and 20–40 words for the subject.

Independent verification prompt

You are an independent visual-grounding verifier, not a stylistic editor.
Given the initial image, the source descriptions, and two rewrite candidates, evaluate each candidate independently. Check whether each candidate preserves important source facts, adds only image-supported details, maintains spatial relations and viewpoint constraints, and describes current activities only when visually supported.
Mark a candidate as FAIL if it contains a material hallucination, an incorrect spatial or contact relation, an unsupported activity, a viewpoint or visible-body inconsistency, or an important omitted source fact. Do not reward verbosity and do not rewrite or paraphrase the candidates.
If exactly one candidate passes, select it. If both pass, select the one with fewer unsupported additions; if they are equivalent, select Candidate A. If both fail, return `NONE` and request fallback generation.

Return JSON only in the following format:

```
{
  "candidate_a": {
    "status": "PASS",
    "issues": []
  },
  "candidate_b": {
    "status": "FAIL",
    "issues": [
      "unsupported activity"
    ]
  },
  "selected": "A",
  "fallback_required": false
}
```

I.2 Physical and Causal Validity Evaluation Prompts

We use a shared two-stage prompting pipeline for **Physical Causality (PC)**, **Content Causality (CC)**, and **Evolution Consistency (EC)**. The first prompt extracts a scene-specific checklist from the initial-frame descriptions and the timed action sequence. The second prompt evaluates the generated video against the applicable checklist items. The complete prompts used in both stages are provided below to specify the evaluation criteria, scoring rules, and required output formats.

PC/CC/EC checklist extraction prompt

You are a checklist extraction expert for world-model evaluation.
You are given a case description consisting of environment, subject, and perspective prompts, which describe the static first frame, together with a timestamped action sequence describing camera or observer movement. Extract checklist items for three dimensions: Physical Causality (PC), Content Causality (CC), and Evolution Consistency (EC). All items will be evaluated on a 1–5 scale.
Core principle. This is an image-to-video setting. The generated video should animate the static scene while preserving physically correct behavior. The checklist should capture the physical laws and natural dynamics that the scene should exhibit. Infer these requirements from physical common sense, not only from explicit motion verbs in the prompt. For example, a floating balloon should drift or rise, and visible water should exhibit physically plausible water behavior.
Step 1: determine applicability.

- **PC** is applicable when the scene contains objects or phenomena governed by physical constraints, including gravity and support, buoyancy, collision, deformation, material behavior, thermodynamics, fluids, soft bodies, smoke, or fire. A scene containing ordinary physical objects is usually applicable because gravity, support, and material behavior can be checked.
- **CC** is applicable when the scene supports a cross-object causal relation in which a scene object or natural event causes an effect on another object or region. Examples include light producing shadows or reflections, wind moving an object, fire igniting material, or water making a surface wet. Scene-supported physical inference is allowed, but do not invent a causal source absent from the prompt or image.
- **EC** is applicable when the scene contains an object or phenomenon that should move or change state spontaneously, such as a drifting balloon, flowing or rippling water, moving clouds, walking people, wind-driven flags or plants, dispersing smoke or fog, melting ice, spreading fire, or growth. Decide based on whether the object would naturally change in the physical world, not only on whether the prompt contains a motion verb. Treat an explicitly static state such as calm, static, still, moored, or seated as non-evolving when appropriate. Exclude motion directly driven by the action sequence, because commanded camera or subject motion is not spontaneous evolution.

Step 2: extract items only for dimensions with `applicable=true`.
1. Physical Causality (PC). Each item must express one physical law as a positive assertion of how an object should behave; do not write conditional "if–then" statements. Allowed subtypes are `gravity`, `buoyancy`, `collision`, `deformation`, `material`, `thermodynamics`, and `impenetrability`. Extract one item for each applicable subtype, normally three to six items.

Persistent laws such as gravity, buoyancy, material behavior, and thermodynamics use `temporal=null`. Event-specific collision, deformation, and impenetrability checks use a time interval derived from the action sequence.
Use `deformation` only for event-driven deformation. Static draping, bending, rigid shape preservation, and similar persistent properties belong to `material`.
For `impenetrability`, extract at most one item when the camera, controlled subject, vehicle, or a rigid attached object moves near a solid boundary. Check that it does not pass through an intact wall, floor, terrain, closed door, fence, furniture item, or other solid object. Ordinary occlusion, valid contact, stopping, detouring, and passage through a genuine opening are not penetration.
Examples include: "The controlled subject should remain outside the intact wall while approaching it," and "The first-person viewpoint should pass through the open doorway rather than through the adjacent wall."
2. Content Causality (CC). Each item must express one cross-object causal chain in the form "because A occurs, B should appear or change." The cause A must be a scene object or natural event, not an agent action. Extract approximately three items when supported. Every CC item must include a temporal interval. A continuously active relation may use the full video interval.
3. Evolution Consistency (EC). Each item must express one spontaneous process that changes over time, either `motion_change` or `state_change`, and must include a temporal interval. Write a positive assertion specifying how the element should gradually or continuously change between `since` and `until`. Extract one or two items per applicable subtype, normally two to four items.
A persistent state is not an evolution: continuously glowing, remaining inflated, remaining upright, or remaining still should not be extracted as EC.
Hallucination policy. Do not introduce objects or events absent from the scene, such as an unsupported fire, jump, object twitch, or traffic flow. Do not contradict an explicitly static description, such as requiring ripples on water described as calm. Standard physical consequences of visible scene elements are allowed, including balloon drift, water motion, cloud drift, cloth draping, and light-induced shadows.
Return strict JSON without Markdown:

```
{
  "case_id": "...",
  "applicability": {
    "PC": {"applicable": true, "reason": "...", "cues": []},
    "CC": {"applicable": false, "reason": "...", "cues": []},
    "EC": {"applicable": true, "reason": "...", "cues": []}
  },
  "items": [
    {
      "id": "PC-01",
      "dimension": "PC",
      "metric_key": "gravity",
      "check_type": "physics",
      "question": "...",
      "verify": "...",
      "expected": "...",
      "source": "prompt|inferred",
      "temporal": null,
      "required": true
    }
  ]
}
```

The allowed `metric_key` values are:

```
PC: gravity | buoyancy | collision | deformation |
    material | thermodynamics | impenetrability
CC: cross_object_causal
EC: motion_change | state_change
```

The `temporal` field must be either `null` or an object of the form `{"since":"...","until":"..."}`. All EC items, all CC items, and event-specific PC collision, deformation, or impenetrability items require a temporal object. Persistent PC gravity, buoyancy, material, and thermodynamics items use `temporal=null`. Do not output items for a dimension marked `applicable=false`.

PC/CC/EC video judging prompt

You are a world-model video-evaluation judge. You are given sampled video frames, each labelled with a timestamp, and a set of checklist items. Evaluate every item and return an integer score from 1 to 5 together with one sentence of visual evidence.
Use the following scale:

```
5 = completely correct and natural
4 = mostly correct, with only a very minor defect
3 = a clear but non-severe problem
2 = a severe problem
1 = completely incorrect or corrupted
```

Temporal evaluation is essential.
For `temporal=null`, evaluate representative frames across the full video and determine whether the persistent property, such as gravity, support, or material behavior, holds throughout.
For `temporal={since,until}`, compare the beginning and ending states of the specified interval. If an item requires an object to change over time, such as drifting, melting, moving, or otherwise evolving, verify that the ending state actually differs in the required way. The object's correct presence in the initial frame is not sufficient.
If the required change is absent at the end of the interval, assign a score of 1 or 2. For example, if an item requires a balloon to rise and drift away from 0 to 80 seconds, but the balloon remains at its initial location in the final interval frame, assign a score of 1 or 2.
Use only visible evidence from the supplied frames. Do not infer unsupported events. If the critical evidence is unclear, score conservatively. Ignore ordinary changes caused by viewpoint, illumination, or occlusion, and evaluate only the physical, causal, or evolutionary phenomenon named by the item.
Return strict JSON without Markdown:

```
{
  "verdicts": [
    {"id": "EC-01", "score": 4, "evidence": "..."}
  ]
}
```

Every `score` must be an integer from 1 to 5.

I.3 Audio Evaluation Prompts

We use four prompts to support the audio evaluation pipeline, covering **Audio Realism (AR)** and **Action–Audio Correspondence (AAC)**.

AR evaluates the perceptual realism and recording fidelity of the generated audio from the audio track alone, focusing on generation artifacts, temporal continuity, and signal integrity. The AR judge directly inspects the complete audio recording and assigns an integer score from 1 to 5.

AAC evaluates whether visible actions are accompanied by plausible audible responses and whether their coarse temporal changes are consistent. To reduce cross-modal bias, AAC is implemented in three stages. An **audio-only event observer** independently identifies acoustic events within labelled temporal windows, while a **silent-video action observer** independently describes visible actions, motion changes, and potentially sound-producing events from the corresponding video windows. An **evidence-linked AAC judge** then compares the two observation streams and scores their correspondence based only on supported cross-modal evidence.

This separation prevents the audio observer from inferring actions from visual context and prevents the visual observer from imagining sounds. The final AAC judge therefore operates on independently extracted evidence rather than on the raw multimodal input. The complete prompts used for each stage are provided below.

Audio Realism (AR) Judge

You are an audio-only judge of recording realism for AI-generated world-model audio. Evaluate audible generation artifacts and recording fidelity from the supplied audio alone. Judge the received sound, not an imagined video or action instruction.

Inspect the entire supplied recording for the following three types of evidence before assigning one Audio Realism (AR) score:
1. Texture and transients. Identify unnatural metallic or robotic coloration in voices and environmental sounds; phasey, hollow, or warbling textures; smeared impacts; or severe muffling or bandwidth degradation audible within this recording.
2. Continuity. Identify artificial clicks, bursts, dropouts, splice or loop seams, or a noise floor that pumps or changes abruptly. Regular footsteps, flowing water, and repeated actions are not defects merely because they repeat. If alleging a loop, identify the audible seam or unnatural repeated texture.
3. Signal integrity. Identify clipping, harsh distortion, intrusive electronic interference, or collapse into unstable or mechanical noise. Distinguish natural wind, water, and friction from electronic artifacts using their audible texture, not a guessed scene.
World-model adaptation. Common content includes locomotion, breathing, clothing or prop friction, water, and environmental ambience. It may also contain speech or music. Do not assume any of these must occur. Do not score whether a footstep matches a visible action or material, whether an environment is appropriate, or whether a voice has a visible speaker; those require other metrics.
Do not excuse a distorted voice, stutter, or metallic texture merely by calling it an intentional vocoder, glitch effect, or sound design. Intention is not available from the input. Conversely, an electronic instrument or musical passage is not automatically defective. Describe the actual evidence and state uncertainty where legitimate musical timbre and a generation defect cannot be distinguished.
The common evaluation format is 24 kHz mono. Do not penalize mono presentation or assume lost frequencies above its available bandwidth. Do not require studio silence; natural recording noise and quiet scenes can be realistic.
For a long recording, inspect the beginning, middle, and end, as well as any audible transitions. A clean majority does not erase a clearly defective passage. Report approximate regions or times only when supported, and do not claim coverage of intervals that were not reliably inspected.
Scoring: integer 1–5.
1: Pervasive severe corruption or artificial texture makes most meaningful sound difficult to recognize.
2: Pronounced, sustained, or recurring mechanical coloration, distortion, muffling, or unstable noise substantially damages recording realism.
3: Sound remains intelligible, but identifiable generation-like texture, phase effects, noise, or continuity defects are clearly noticeable in relevant passages.
4: Convincing recording quality overall; remaining imperfections are small and isolated, such as a slightly softened impact or faint brief artifact.
5: Convincingly recorded textures, resolved transients, and stable signal continuity throughout the supplied recording, with no identifiable generation artifact. Support this with concrete audible observations; clear or intelligible sound alone is insufficient.
Choose the anchor supported by severity, recurrence, and temporal extent. Do not force a distribution, a target mean, or a low score. Do not replace uncertainty with invented defects. No N/A. A score of 5 must not coexist with acknowledged clear unresolved defects in the reason.
Return JSON only. Keep each observation brief and evidence-based; do not provide hidden reasoning.

```
{
  "texture_observation": "...",
  "continuity_observation": "...",
  "signal_observation": "...",
  "defect_regions": [
    {
      "region": "approximate portion of the recording",
      "evidence": "audible property, not presumed intent"
    }
  ],
  "uncertainties": [],
  "reason": "brief justification of the selected anchor",
  "AR": 1
}
```

AAC Audio-only Event Observer

You are an audio-only event observer. Listen to the labelled, successive audio clips and process **each clip independently**. No video or scene description is available.
Report acoustic events in that clip and split distinct sounds into distinct events. An impact, an animal-like vocalization, and steady ambience must be separate entries. A source label attached to one event cannot be copied onto another event merely because they co-occur.

For each event, report:

- its acoustic description;
- its event type: impact, vocalization, ambience, speech, music, or other;
- a specific source only if it is recognizable from that event alone.

In particular, rhythmic thuds do not prove a horse, a gallop, human walking, or any contact material. A broadband rush does not prove wind or water. When identity is uncertain, write `unresolved` and describe the sound itself instead.
Contact material is `unresolved` unless it is independently audible from the impact texture; a guess is not enough. A short clip caption is not exhaustive, and omitted events are not verified absent. Do not infer an exact onset inside a labelled clip.
Return one JSON object only. Include each provided window ID exactly once, in the given order. Use an empty list for an apparently quiet clip. Do not copy a source guess from one segment into another without hearing it there.

```
{
  "segments": [
    {
      "id": "W0",
      "events": [
        {
          "type": "impact|vocalization|ambience|speech|music|other",
          "heard": "audible texture only",
          "source": "specific source or unresolved",
          "source_confidence": "high|medium|low",
          "contact_material": "identifiable or unresolved"
        }
      ],
      "limitations": []
    }
  ]
}
```

AAC Silent-video Action Observer

You are a silent-video action observer. Inspect only this {`duration_seconds`}-second video sampled at {`video_fps`} fps. The labelled windows are {`window_boundaries`}.
For **each window**, report specific visible actors, actual actions, motion, pauses, changes, and directly visible contact or other potentially sound-producing events.
Camera bobbing alone does not prove footsteps. Subject motion together with scene displacement may support locomotion. Do not imagine sounds.
Report only visible evidence. If a subject leaves the frame or its mouth is occluded, record the visibility limitation rather than inferring an associated sound or its absence.
Return JSON only. Return exactly {`number_of_windows`} segments in order, one per supplied window. Make no audio claims.

```
{
  "segments": [
    {
      "id": "W0",
      "time_s": "...",
      "visible_actors": "...",
      "visible_actions": "...",
      "changes": "...",
      "direct_sound_action_evidence": "... or unresolved",
      "limitations": []
    }
  ],
  "scene": "brief setting"
}
```

For the reported experiment, `video_fps = 2`. The placeholders are filled separately for each segment of up to 20 seconds, and `window_boundaries` enumerates consecutive 4-second windows `W0`, `W1`, and so on.

Action–Audio Correspondence (AAC) Judge

You are an evidence-based judge of **Action–Audio Correspondence (AAC)**. Assign an integer score from 1 to 5 using only the supplied audio-only and silent-video observations.
The input consists of independent audio-only event observations and silent-video action observations for the same labelled time windows. These observations are evidence, not ground truth. Do not fill gaps with imagined sounds or actions.
Assess audible responses to visible actions and their coarse temporal relation across the full excerpt.
Scope. First identify concrete visible actions for which an audible response can reasonably be evaluated, such as foot or object contact, splashing, mechanical operation, or directly visible vocalization.
Confirmed locomotion may also qualify when subject motion and scene displacement jointly establish travel, even if feet are occluded; camera bobbing alone is insufficient.
Test whether a plausible rhythmic or event sound appears, changes, pauses, or stops in the corresponding windows. A sound can continue while its source is offscreen. Do not require every visible movement to make sound.
A scene label or mere coexistence of a source and sound does not establish action–audio correspondence. A flying bird together with bird calls is not evidence that a visible vocalization caused a call. Generic rushing together with visible waves is not evidence of a response change.
Do not judge general acoustic fidelity, synthetic artifacts, or scene ambience without an identifiable action; those are outside AAC.
Evidence rules. Claim a missing sound only if the action is visible, the sound would reasonably be audible, absence was explicitly checked, and masking is considered.
Claim a mismatch only from supported observations, not from a sparse caption's omission.
An animal may vocalize while its mouth is hidden or too small to resolve. Lack of visible mouth movement is **not** evidence that an audible call is detached.
A transition between water, stone, grass, or other materials does not require an audible timbre change unless contact and the specific sound relation are independently verified; material matching is outside AAC.
Mechanical repetition, muffling, and other acoustic defects belong to AR and must not lower AAC by themselves.
Do not infer exact contact timing inside a window, precise source location, or frame-level synchronization from 2 fps video and window-level audio. State uncertainty when an audible event's source cannot be identified. Do not use the prompt's intended action as evidence that the video performed it.
Rating.
5: Multiple supported action/sound changes align, including starts, stops, pauses, or intensity changes, with no meaningful contradiction.
4: At least one specific supported action/sound change aligns at the available time resolution, with one minor lapse or limited further variation.
3: At least one specific supported action/sound response exists, but timing evidence is weak or mixed. Sustained, visually verified locomotion plus plausible rhythmic contact sounds without observable starts or stops is at most 3, and source uncertainty must be stated.
2: Major supported missing, displaced, or detached responses dominate, though some relationship remains.
1: Clear, pervasive detachment across the assessable actions and time windows.
Do not force a target distribution or penalize uncertainty as failure.
To assign a score of 3–5, cite an identifiable visible sound-producing action and a plausible corresponding audible event. Sustained confirmed locomotion with plausible rhythmic contacts is allowed but capped at 3 when footfalls are unseen and timing changes are absent.
To assign a score of 4–5, additionally cite a supported temporal change in both modalities, with window IDs.
If the evidence consists only of birds and calls coexisting, or generic waves and rushing coexisting, return `status="needs_review"` with no score, regardless of apparent scene plausibility.
If no action–audio relation is actually assessable, return `status="needs_review"` and omit the score.
Return JSON only. Omit `score` when `status="needs_review"`. Every cited ID must exist in the supplied observations.

```
{
  "metric": "AAC",
  "status": "scored|needs_review",
```

```
  "score": 1,
  "evidence": [
    {
      "window": "W0",
      "audio_ids": ["W0:0"],
      "video_ids": ["W0"],
      "relation": "supported action/sound relation"
    }
  ],
  "reason": "brief evidence-based anchor justification",
  "limitations": []
}
```

I.4 Failure-Type Diagnostic Prompt

To support qualitative diagnosis beyond aggregate evaluation scores, we use a vision–language judge to identify recurrent observable failure modes in long-horizon interactive rollouts. The judge performs multi-label classification over a predefined failure taxonomy covering temporal instability, appearance and geometry drift, revisit inconsistency, control response errors, physical violations, causal-state errors, and failures of ongoing scene evolution. Multiple failure types may be assigned to the same video when they correspond to distinct observable phenomena.

The prompt requires the judge to rely only on directly observable video evidence, avoid inferring hidden model mechanisms, and abstain from assigning a label when the evidence is weak or ambiguous. It also specifies boundaries between closely related failure modes, such as global temporal jumps versus entity-level impossible motion, revisit-content mismatch versus loop-closure failure, and global first-frame resets versus local state reversion. The output is restricted to a JSON list containing only labels from the predefined taxonomy; an empty list indicates that no clearly supported failure is observed.

Long-Horizon Failure-Type Judge

You are an evidence-based evaluator of failure modes in **long-horizon interactive world models**.
Your task is to identify which observable failure types occur in the generated interactive video.
A video may contain **ZERO, ONE, or MULTIPLE** failure types. Select all clearly supported failures, but use only the predefined labels below. Do not invent, rename, merge, or rewrite labels.
Allowed failure labels.

```
[
  "Quality_Degradation",
  "Abrupt_Temporal_Jump",
  "First_Frame_Reset",
  "Temporal_Oscillation",
  "Repeated_Segment_Loop",
  "Appearance_Drift",
  "Geometry_Layout_Drift",
  "Revisit_Mismatch",
  "Repeated_Revisit_Drift",
  "Loop_Closure_Failure",
  "Under_Response",
  "Over_Response",
  "Delayed_Response",
  "Residual_Action",
  "Control_Interference",
  "Subject_Camera_Confusion",
  "Penetration",
  "Floating_Support_Failure",
  "Impossible_Motion",
  "Missing_Consequence",
  "Spurious_Consequence",
  "State_Reversion",
  "Process_Freeze",
```

```
  "Process_Restart",
  "Temporal_Reversal"
]
```

Failure definitions.
1. Quality_Degradation. The overall visual quality becomes noticeably worse as the rollout continues. Typical evidence includes increasing blur or noise, texture corruption, structural artifacts, loss of fine details, or general visual collapse. Use this label only for degradation of visual quality, not for identity changes, scene-layout changes, or world-state errors.
2. Abrupt_Temporal_Jump. The overall scene, viewpoint, or world state changes abruptly between adjacent moments without a continuous transition explained by the intended interaction or normal scene evolution. Typical evidence includes sudden scene replacement, an abrupt viewpoint jump, a large unexplained visual discontinuity, or globally incompatible neighboring moments. Use this label for a *global* temporal discontinuity.
3. First_Frame_Reset. The generated world abruptly returns to the initial scene, initial viewpoint, or a clearly near-initial global state, effectively discarding accumulated interaction history. Use this label only when the reset is specifically toward the initial state.
4. Temporal_Oscillation. The generated world repeatedly alternates between two or more distinct previously observed states instead of evolving continuously, e.g., $A \to B \to A \to B$ or $A \to B \to C \to B \to C$. Use this label when distinct states repeatedly switch back and forth.
5. Repeated_Segment_Loop. The same or nearly identical frame sequence is repeatedly regenerated, forming a short temporal cycle. Typical evidence includes the same few frames or motion pattern repeatedly restarting. Unlike `Temporal_Oscillation`, this failure repeats nearly the same frames or short sequence rather than alternating among distinct world states.
6. Appearance_Drift. A persistent subject, object, or scene element changes its stable visual identity or attributes without justification. Examples include changes in identity, clothing, color, shape, texture, or material. Use this label when the entity should remain the same but its appearance changes.
7. Geometry_Layout_Drift. The spatial structure of the world changes inconsistently over time. Examples include roads changing shape, walls or buildings moving, room layouts changing, landmarks shifting, or previously stable spatial relationships becoming inconsistent. Use this for geometry and layout changes rather than appearance changes.
8. Revisit_Mismatch. After successfully returning to approximately the same previously observed location or viewpoint, the world content is inconsistent with the earlier observation. Objects may disappear or appear, landmarks may change, or geometry and spatial relations may differ. Use this only when the revisit pose is approximately recovered but the world content is inconsistent.
9. Repeated_Revisit_Drift. Across multiple revisits to the same place or viewpoint, consistency progressively deteriorates. Use this only when multiple revisits are available and the mismatch accumulates across them.
10. Loop_Closure_Failure. After a control sequence that should return the camera or subject to an earlier pose, the generated result fails to recover the expected position and/or orientation. The final viewpoint may be spatially offset, have an incorrect heading, or show expected landmarks from the wrong pose. Unlike `Revisit_Mismatch`, the failure is in recovering the pose itself rather than preserving world content at a recovered pose.
11. Under_Response. The observed translation or rotation is substantially weaker than requested by the control input. Examples include insufficient forward, backward, or lateral movement, insufficient camera rotation, or only partial execution of the requested motion.
12. Over_Response. The observed translation or rotation is substantially stronger than requested. Examples include excessive translation or rotation, or a short control producing an disproportionately large response.
13. Delayed_Response. The intended movement or rotation begins noticeably later than the corresponding control input. Use this when the control is issued but the expected response appears only after a clear delay.
14. Residual_Action. The effect of a previous control continues after that control has ended. Examples include continued camera rotation or subject movement after the corresponding input stops, or a previous action persisting into the next control segment.
15. Control_Interference. Two or more simultaneous or sequential controls incorrectly interfere with one another. Examples include one component of a combined action being suppressed, previous and current controls becoming mixed, or translation and rotation becoming incorrectly entangled. Use this only when the failure is specifically attributable to interaction between multiple controls.
16. Subject_Camera_Confusion. A subject-directed action incorrectly affects the camera, or a camera-directed action incorrectly affects the subject. Examples include the camera moving when only the controlled subject should move, the subject moving when only the camera should rotate, or the two becoming incorrectly coupled.

17. Penetration. The camera, subject, or object passes through a solid surface or obstacle. Examples include walking through walls, the camera passing through geometry, objects intersecting solid obstacles, or a subject sinking through the floor.
18. Floating_Support_Failure. A subject or object loses physically plausible support. Examples include a subject floating above the ground, an unsupported object remaining suspended, or feet, wheels, or objects visibly detaching from their supporting surface.
19. Impossible_Motion. A specific subject or object exhibits physically or temporally impossible motion while the overall scene does not necessarily reset. Examples include entity teleportation, discontinuous displacement, implausible acceleration, an impossible trajectory, or a sudden object-position jump. Unlike `Abrupt_Temporal_Jump`, this label concerns a specific moving entity rather than a global scene or viewpoint discontinuity.
20. Missing_Consequence. An observable action or physical interaction occurs, but its expected visible consequence does not. Examples include an object being pushed without responding, visible contact without an expected effect, or an interaction that should alter the world but produces no visible change.
21. Spurious_Consequence. A world-state change occurs without a corresponding visible action or event. Examples include an object moving by itself, a door opening without interaction, or another unexplained state change.
22. State_Reversion. A previously established local or semantic world-state change is undone without a later event that justifies the reversal. Examples include a moved object returning to its previous location, an opened door becoming closed again, or an earlier action consequence disappearing. Unlike `First_Frame_Reset`, this affects a specific state or consequence rather than globally resetting the world toward its initial state.
23. Process_Freeze. An ongoing dynamic process unexpectedly stops. Examples include walking characters freezing, traffic stopping, moving background elements becoming static, or another evolving process halting without explanation.
24. Process_Restart. An ongoing dynamic process unexpectedly returns to an earlier stage and starts evolving again from that stage. Examples include a motion sequence restarting or a dynamic process repeatedly resetting. Use this for dynamic processes, not for restoration of a static world state.
25. Temporal_Reversal. An ongoing process or state evolution moves backward in an implausible manner. Examples include visible rewinding, a process evolving backward, objects traversing previously observed intermediate states in reverse, or a completed transition reversing without a plausible cause.
General decision rules.
Judge only failures that are directly observable in the video. Do not infer hidden model mechanisms, internal causes, or unobserved events.
Do not treat intended camera motion, normal viewpoint changes, temporary occlusion, ordinary motion blur, legitimate physical interaction, or justified scene evolution as failures.
Select **all** clearly supported failure labels. Multiple labels may be assigned to the same video when they describe genuinely distinct observable failures.
Do not assign multiple labels to the same phenomenon merely because several labels could loosely describe it. Prefer the most specific applicable label.
If the evidence for a failure is weak, ambiguous, or depends on an unsupported inference, do not select that label.
If no clear failure is visible, return an empty list.
Important label boundaries.
Use `Quality_Degradation` when global visual quality becomes worse, and `Appearance_Drift` when a persistent entity changes identity or visual attributes.
Use `Geometry_Layout_Drift` when stable world geometry or spatial layout changes.
Use `Abrupt_Temporal_Jump` for a sudden global scene or viewpoint discontinuity, and `Impossible_Motion` for impossible displacement of a specific entity.
Use `First_Frame_Reset` when the whole generated world abruptly returns toward its initial state.
Use `Temporal_Oscillation` when two or more distinct states repeatedly alternate, and `Repeated_Segment_Loop` when essentially the same short frame sequence repeats.
Use `Revisit_Mismatch` when a previously observed pose is approximately recovered but its world content is inconsistent. Use `Repeated_Revisit_Drift` when such inconsistencies progressively increase over multiple revisits. Use `Loop_Closure_Failure` when the expected return pose itself is not correctly recovered.
Use `State_Reversion` when a specific established state or action consequence is unexpectedly undone.
Use `Process_Freeze` when an ongoing process stops, `Process_Restart` when it returns to an earlier stage and begins again, and `Temporal_Reversal` when its evolution visibly proceeds backward.
Output format.
Return exactly one JSON list. Each item must be exactly one label from the predefined label set.
Example with multiple failures:

```
[
  "First_Frame_Reset",
  "Temporal_Oscillation",
  "Geometry_Layout_Drift",
  "State_Reversion"
]
```

Example with one failure:

```
[
  "Quality_Degradation"
]
```

Example with no detected failure:

```
[]
```

Strict output constraints.
Output valid JSON only. Output exactly one JSON list. Do not output Markdown, explanations, timestamps, confidence scores, severity scores, category names, or any additional text. Do not output duplicate labels. Do not invent or rename labels. Every output label must exactly match one item in the predefined label set.