DUAL EXAM: CONTROLLED EVALUATION OF MULTIMODAL AGENTS AND WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

World Models (WMs) can support Multimodal Agents (MAs) through simulated experience and action-conditioned predictions, while MA interactions can provide feedback for refining WMs. This potential reciprocal relationship motivates assessing both components. A pilot study of one MA–WM pairing across ten scenarios and 150 cases illustrates a diagnostic limitation: coupled task outcomes alone cannot identify the source of failure. We introduce Dual Exam, a unified framework that evaluates the two components separately under controlled conditions, assessing MAs under a fixed environment response mechanism and WMs under fixed action interventions. Across four representative domains including physical, digital, social, and scientific, we evaluate ten MA configurations on 1,088 cases spanning 15 reporting categories and 16 WMs across eight scenarios, with 15 evaluated on the full set of 400 cases. The highest MA Index and WM overall scores are 31.3 and 21.7 on their respective 0–100 scales, revealing substantial capability gaps under the evaluated task requirements. These findings position Dual Exam as a diagnostic framework for assessing capabilities relevant to general-purpose multimodal co-evolution.

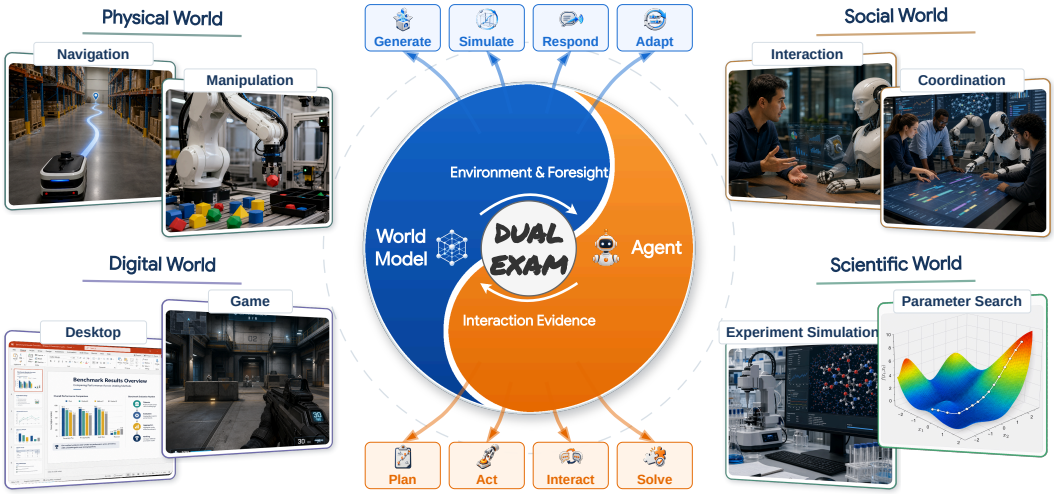


Figure 1: Overview of the Dual Exam between world models and agents.

1 INTRODUCTION

World Models (WMs) and Multimodal Agents (MAs) can improve each other in complementary ways. WMs support MAs through simulated experience and action-conditioned predictions (Ha & Schmidhuber, 2018; Hafner et al., 2025; Gu et al., 2025), while MA interaction trajectories support WMs by providing feedback for refining WMs (Zhang et al., 2026b; Yu et al., 2026). Recent work has begun to couple the two in specific settings (Fang et al., 2025; Guo et al., 2025; Liu et al., 2026). For example, a web agent can use a world model to anticipate the effects of candidate clicks before acting, while its interaction trajectories provide data for refining the model (Fang et al., 2025). Yet it

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.

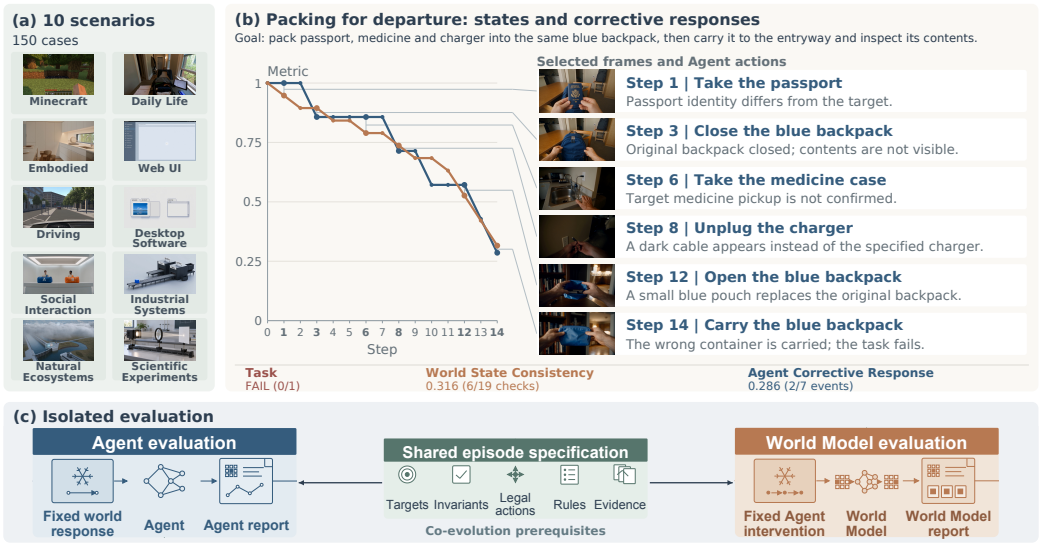


Figure 2: **From a failed packing rollout to controlled evaluation.** (a) Coverage of 150 cases across ten scenarios. (b) Six selected frames and diagnostic curves covering all 14 steps of a GPT-5.6 Sol–Seedance 2.5 rollout. Blue text gives MA commands; images show WM responses. Gray lines mark same-step correspondence, not causal attribution. World State Consistency (orange) and Agent Corrective Response (blue) use cumulative deductions from an assigned unit baseline with fixed full-trajectory denominators. Later packing and carrying use a different blue pouch, violating the same-backpack requirement. (c) Dual Exam holds the environment response mechanism fixed to evaluate MAs and fixes action interventions to evaluate WMs under a shared episode specification.

remains unclear whether current MAs and WMs already possess the basic capabilities needed for such reciprocal improvement to extend beyond these specialized settings. We study this question across four representative multimodal domains, including physical, digital, social, and scientific, to examine the prerequisites for broader MA–WM co-evolution (Chu et al., 2026).

However, reciprocal improvement requires both MAs and WMs to remain reliable over repeated interaction. MAs still struggle with visual grounding, action sequencing, and sustained task completion (Xie et al., 2024; Chen et al., 2025; Sun et al., 2026). WMs also face challenges in memory consistency and action control (Ye et al., 2026). Visual fidelity or state similarity alone does not guarantee downstream utility (Shang et al., 2026a; Huang et al., 2026). These weaknesses become more apparent when MAs and WMs are coupled. The output of one repeatedly becomes the input to the other. WM-based simulation can improve planning (Gu et al., 2025), but the planning benefits depend on behavioral coverage, environment complexity, and whether the MA can use the predicted outcomes (Li et al., 2025b; Qian et al., 2026a). Model bias may also accumulate over repeated rollouts (Janner et al., 2019). End-to-end performance can show that the coupled system fails, but not why. Failures may come from MA decisions, WM responses, or their interaction. A component-level evaluation is therefore needed to characterize the capabilities that end-to-end outcomes alone cannot reveal.

Our preliminary study tests one MA–WM pairing on 150 cases across ten scenarios, with 0% task success under the stated protocol (Section 2.1). The trajectories reveal state discrepancies and limited responses, but do not identify which component caused each failure. This motivates **Dual Exam**: controlled, separate assessments of MAs and WMs rather than causal attribution from coupled outcomes.

We instantiate **Dual Exam** across physical, digital, social, and scientific domains. Both tracks check execution evidence against explicit goals and state constraints. The MA track covers ten configurations, 1,088 items, and 15 reporting categories. The WM track covers 16 configurations: 15 on all 400 cases and Seedance 2.5 on 300 cases in six scenarios (Section 3). The highest Agent Index and WM Overall are 31.3 and 21.7 on their respective 0–100 scales. Separate evaluations diagnose

Table 1: **Coupled MA–WM evaluation across ten scenarios.** GPT-5.6 Sol interacts with Seedance 2.5 on 15 cases per scenario (150 total), with one attempt of at most 20 steps and no episode-level retries. Initialization and technical failures are excluded. Scenario rates are case-averaged; Overall is their unweighted mean. Section 2.1 defines the metrics.

Metric (%)	Minecraft	Daily Life	Embodied Tasks	Web UI	Driving	Desktop Software	Industrial Systems	Natural Ecosystems	Scientific Experiments	Social Interaction	**Overall**
World State Consistency	12.7	37.9	16.5	18.7	14.7	22.9	34.4	39.5	29.5	38.9	26.6
Agent Corrective Response	0.0	13.3	16.7	6.7	0.0	0.0	20.0	10.0	0.0	25.0	9.2
Task success	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

task-specific limitations; whether addressing them enables sustained co-evolution requires learning experiments (Section 4).

In summary, our contributions are:

- We introduce **Dual Exam**, a unified diagnostic framework for MAs and WMs across four representative multimodal domains. It evaluates MAs with controlled environment responses and WMs with fixed action interventions, allowing the two components to be examined separately.
- We evaluate ten MA configurations on 1,088 cases spanning 15 reporting categories and 16 WM configurations across eight scenarios: 15 cover all 400 cases and one covers 300 cases in six scenarios. The results reveal substantial and uneven capability gaps, including incomplete MA solutions despite locally correct decisions and incorrect WM responses despite apparent visual continuity.
- We provide a 150-case pilot across ten scenarios, with trajectory audits and a packing example that motivate separate assessments by illustrating the limits of diagnosis from coupled task outcomes.

2 DUAL EXAM

We first present a preliminary study of coupled MA–WM interaction to motivate the design of Dual Exam. The framework evaluates Agents under fixed, validated environment responses and World Models under fixed action interventions. Both tracks assess task-relevant behavior against specified goals and constraints, while retaining track-specific cases, interaction protocols, and verification procedures. The two tracks need not use one-to-one matched episodes. Evaluation relies on recorded execution evidence rather than self-reported completion. Across physical, digital, social, and scientific domains, Dual Exam comprises 1,088 MA cases and 400 WM cases.

2.1 PRELIMINARY STUDY: MOTIVATION FOR DUAL EXAM

As a pilot study, we evaluate GPT-5.6 Sol with Seedance 2.5 on 150 cases across ten scenarios, with 15 cases per scenario. Each case has one attempt of at most 20 interaction steps, with no episode-level retries. Initialization and technical failures are excluded from the 150 cases and the task-success denominator. Table 1 reports a retrospective trajectory audit; Appendix C details the protocol.

World State Consistency measures preservation of task-relevant identity and state and realization of requested action effects, averaging supported checks over applicable checks, then over steps and cases. *Agent Corrective Response* measures eligible visible discrepancies receiving targeted verification, repositioning, or correction within two actions; it measures response behavior, not successful repair. *Task Success* requires satisfying the full goal and all constraints. Overall is the unweighted mean of the ten scenario rates: 26.6%, 9.2%, and 0%, respectively, for this model pairing and protocol.

The packing rollout illustrates the gap between issuing task commands and meeting task requirements (Figure 2(b)). Three items must be packed into the same blue backpack and carried to the entryway for inspection. Commands continue despite unconfirmed acquisition, and later packing and carrying use a smaller blue pouch, violating the same-backpack requirement. Six selected frames accompany curves over all 14 steps, ending at 0.316 (6/19 state checks) and 0.286 (2/7 response events). Both responses are navigation retries; none of the four identity discrepancies receives a targeted response.

Table 2: **Comparison of benchmark scope.** Representative Agent and World Model benchmarks are compared along interaction, diagnostic, and evaluation-scope dimensions.

Benchmark	Interaction		Diagnosis		Scope		
	Multimodal Input	Interactive Rollout	Environment-grounded	Component Isolation	Persistent State	All Four Domains	# Eval. Units
Agent Benchmarks							
GAIA (Mialon et al., 2024)	✓	✗	✗	✗	✗	✗	466
AgentBench (Liu et al., 2024)	✗	✓	✓	✗	✗	✗	1,141
WebArena (Zhou et al., 2024)	✗	✓	✓	✗	✗	✗	812
VisualWebArena (Koh et al., 2024)	✓	✓	✓	✗	✗	✗	910
OSWorld 2.0 (Yuan et al., 2026)	✓	✓	✓	✗	✗	✗	108
Agents' Last Exam (Sun et al., 2026)	✓	✓	✓	✗	✗	✗	1,500+
Dual Exam (Agent)	✓	✓	✓	✓	✓	✓	**1,088**
World Model Benchmarks							
VBench (Huang et al., 2024)	✗	✗	✗	✗	✗	✗	946
WorldScore (Duan et al., 2025)	✓	✓	✗	✗	✗	✗	3,000
WorldModelBench (Li et al., 2025a)	✓	✗	✗	✗	✗	✗	350
MIND (Ye et al., 2026)	✓	✓	✗	✗	✓	✗	250
WorldArena (Shang et al., 2026a)	✓	✓	✓	✗	✗	✗	500
Omni-WorldBench (Wu et al., 2026)	✓	✓	✗	✗	✓	✗	1,068
WorldMark (Xu et al., 2026)	✓	✓	✗	✗	✓	✗	500
WBench (Ying et al., 2026)	✓	✓	✗	✗	✓	✗	289
WorldExam (Yang et al., 2026)	✓	✓	✗	✗	✓	✗	1,474
PlayWorld (Ding et al., 2026)	✓	✓	✗	✗	✓	✗	171
Dual Exam (World Model)	✓	✓	✓	✓	✓	✓	**400**

Notation. ✓ / ✗ indicate explicit / no explicit coverage, not model capability; coverage need not span all tasks. **Multimodal Input**: visual inputs paired with language or action/control inputs, initially or during interaction. **Interactive Rollout**: actions or controls affect subsequent observations or states, using predefined or adaptive sequences. **Environment-grounded**: evaluation against execution outcomes or independently verifiable environment states. **Component Isolation**: separate component evaluation with fixed environment responses or action interventions; not causal attribution of individual coupled failures. **Persistent State**: dedicated evaluation of state preservation or evolution across revisits, occlusion, or interaction turns, beyond generic temporal video quality. **All Four Domains**: coverage of physical, digital, social, and scientific domains. **# Eval. Units**: benchmark-specific counts of questions, tasks, prompts, videos, or scenarios, not directly comparable.

Unlike the table's audit, the single-case state score excludes unknown checks. Its curves accumulate deductions from an assigned unit baseline using fixed full-trajectory denominators. Their decline follows from this definition, not measured monotonic capability degradation; they are retrospective diagnostics rather than isolated model scores. Appendix C gives the scoring details and limitations.

The pilot motivates a diagnostic question: how does each component perform when the other side is controlled? Dual Exam holds a validated environment response mechanism fixed to test MAs and fixes action interventions to test WMs against specified goals and constraints (Figure 2(c)). These tests measure role-specific performance rather than retrospectively assigning causes to the pilot failures. Table 2 compares this scope with existing benchmarks.

2.2 Dual Examination Framework

Agent decisions and World Model responses jointly determine coupled outcomes. Dual Exam evaluates them separately: the Agent track fixes environment responses, and the World Model track fixes action interventions. Specifically, the Agent track evaluates a candidate A_θ against a fixed, validated environment response process:

$$a_t \sim A_\theta(h_t), \qquad o^*_{t+1} = \mathcal{T}^*(s_t, a_t),$$

where h_t is the interaction history and s_t is the controlled environment state at step t; $\mathcal{T}^*$ returns the reference observation after action a_t. The World Model track instead fixes a validated intervention sequence $a^*_{1:T}$ and evaluates the resulting observations:

$$\hat{o}_{t+1} = M_\phi(\hat{s}_t, a^*_t).$$

Here, $\hat{s}_t$ denotes the generated world state as represented by the preceding observation and permitted history, not an observed internal model state; $\hat{o}_{t+1}$ is the next generated observation. "Validated" means consistent with the task specification and verifier, not universally perfect. Both tracks assess task-relevant behavior against specified goals and constraints, while retaining track-specific cases, interaction protocols, and verification procedures. Table 2 compares this scope with representative Agent and World Model benchmarks.

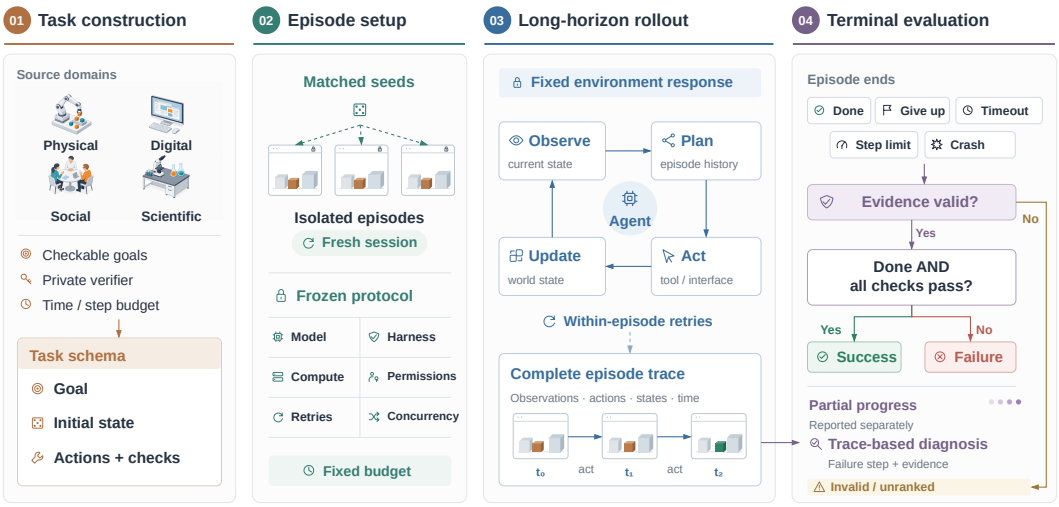


Figure 3: **Agent task construction and evaluation.** Tasks specify goals, initial states, allowed actions, rules, and required checks. Candidate Agents interact with controlled environments under predefined settings and budgets, with observations, actions, states, and artifacts recorded throughout. Success requires task completion and satisfaction of all required checks.

Diagnostic scope. The two tracks identify component-level capability gaps under controlled conditions. They do not by themselves establish the cause of a particular coupled failure or whether joint training will yield sustained reciprocal improvement. These questions require matched interaction experiments and evaluation across successive training rounds.

2.3 Agent Examination: Control the World, Test the Agent

The Agent Examination tests whether a candidate MA can complete tasks in a controlled, specification-consistent environment through observation interpretation, state reasoning, planning, action selection, progress verification, and error recovery. Figure 3 summarizes benchmark construction and evaluation.

Benchmark construction. The collection comprises 1,088 cases in 15 reporting categories across physical, digital, social, and scientific domains (Chen et al., 2026; 2025; Sun et al., 2026; Yuan et al., 2026), combining generated episodes, benchmark adaptations, and human-authored stress tests. Each case defines a goal, initial state, allowed actions, rules, invariants, execution budget, and evidence checks. Existing task-specific verifiers are retained and supplemented with action traces, hidden-state checks, or follow-up tests as needed.

Evaluation. Before each episode, we fix model and execution settings, permissions, budgets, retry policies, and environment initialization. The Agent interacts until completion or termination, with execution traces recorded. Strict success requires task completion and all required checks; graded task scores are retained where applicable. The Agent Index weights the 15 category scores by item count: physical 524 (48.16%), digital 255 (23.44%), social 200 (18.38%), and scientific 109 (10.02%). Scientific results include 40 added binary-scored and six graded items. Appendix A provides evaluation details. Fixed environment responses allow Agent capabilities to be assessed without variation from a candidate World Model.

2.4 World Model Examination: Control the Actions, Test the World Model

The World Model Examination fixes action selection and tests whether a candidate WM correctly realizes specification-valid interventions while preserving prior state, identities, and shared-state consistency and obeying active rules. Figure 4 summarizes case construction and evaluation.

Benchmark construction. Behavior traces from four domains are expanded into intervention sequences and case-specific rubrics covering action effects, state persistence, cross-agent consistency, and rule or event updates. Automated screening and expert review check executability, rubric consistency, and evidential support before case specifications and verifier bindings are frozen. The

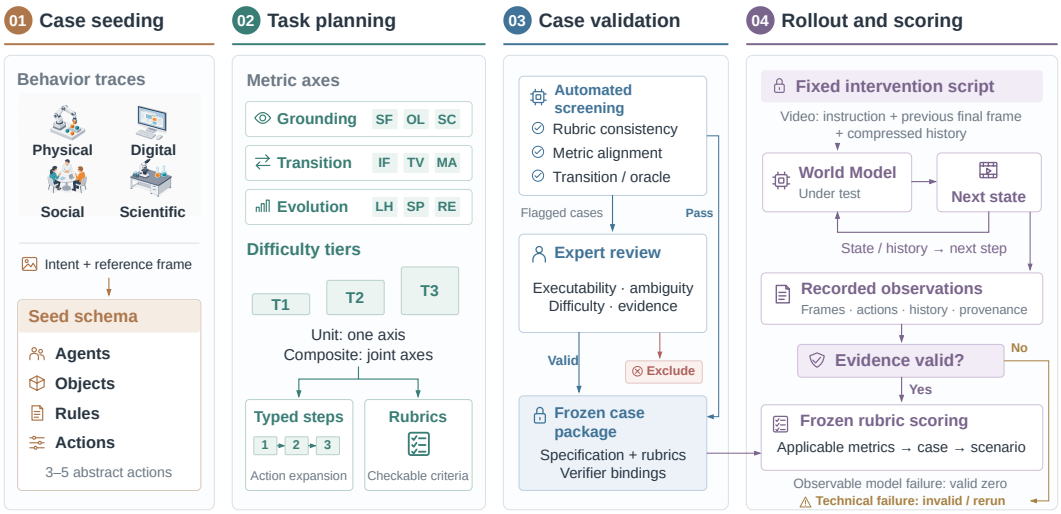


Figure 4: **World Model task construction and evaluation.** Cases and rubrics are screened before fixed-intervention evaluation. Metric labels: SF = Setting Fidelity; OL = Observational Legibility; SC = Shared-State Coherence; IF = Interaction Fidelity; TV = Transition Validity; MA = Multi-Agent State Coupling; LH = Long-Horizon Consistency; SP = State Persistence; RE = Rule Evolution. MA here denotes a metric, not the Multimodal Agent abbreviation used in the text.

benchmark contains 400 cases, with 50 in each of eight scenarios: Embodiment, Driving, and Daily Life (physical); Web and Game (digital); Human Interaction (social); and Natural Science and Expert Science (scientific).

Evaluation. Each rollout uses fixed interventions: the model receives the current intervention, preceding observation, and permitted history to generate the next observation. Case-specific binary checks evaluate requested effects, state/identity preservation, shared-state consistency, and rule adherence. Technical evaluation failures are invalid; observable model errors remain valid. Scores are normalized to $[0, 1]$ and aggregated across metrics, cases, and scenarios; WM Overall is the unweighted mean of eight scenario scores, scaled to 0–100. Fifteen configurations cover all 400 cases. Seedance 2.5 covers six scenarios (300 cases); its six-scenario mean is shown separately and excluded from full-coverage Overall ranking. Missing scenarios are not scored as zero. Fixed interventions isolate World Model evaluation from upstream Agent action-selection variability.

3 Experiments

Tables 3 and 4 report ten Agent configurations on 1,088 items and 16 WM configurations on eight scenarios (15 full-coverage; one partial). Agent Index and WM Overall use different tasks and metrics and are not directly comparable. Scores describe the evaluated model–harness configurations; protocol and evidence limits are detailed in Appendices A and B.

3.1 Agent Results

Reliable environments do not guarantee effective execution. Holding the environment response mechanism fixed removes variation from a candidate WM. Yet the highest scores are only 9.5 on Manipulation, 11.2 on Action Sequencing, and 5.6 on Desktop Application Interaction. Web Interaction reaches 29.6. These results identify Agent-side execution limitations under the evaluated settings; improving WM fidelity alone would not resolve them. Scores across task families use different verifiers and should not be treated as equally difficult measures.

Task completion requires globally compatible decisions. Navigation traces contain correct feasibility decisions but routes with nonexistent edges; Action Sequencing traces contain correct clips in the wrong order. Locally correct elements therefore need not form a valid solution. Selected scientific and resource-governance cases illustrate revision after explicit feedback, suggesting consistency

Table 3: **Agent leaderboard by overall Index.** Ten configurations, 15 categories, and 1,088 items; scores are on a 0–100 scale. N is the item count. Index weights categories by N (524 Physical, 255 Digital, 200 Social, 109 Scientific). Columns follow Index; bold and underline mark the highest and second-highest distinct scores, including ties.

Category	N	GPT 6 Astra	GPT 5.6 Sol	Gemini 3.6 Flash	Qwen 3.8 Max	Kimi K3	DeepSeek V4.1 Flash	GLM 5.3 Flash	Seed 2.1 Pro	MiniMax M3	MiMo V2.5
		OpenAI (b)	OpenAI (a)	Google	Alibaba	Moonshot AI	DeepSeek	Zhipu	ByteDance	MiniMax	Xiaomi
Physical World											
PL	30	**63.3**	16.7	10.0	20.0	16.7	23.2	10.0	3.3	3.3	6.7
MAN	42	7.1	**9.5**	4.8	7.1	4.8	4.8	7.1	0.0	0.0	7.1
NAV	282	**49.7**	36.2	37.6	36.2	33.3	23.1	23.1	8.9	22.0	15.6
AS	170	**11.2**	6.5	0.0	6.5	4.7	2.4	4.1	3.5	0.6	0.0
Digital World											
WI	74	**29.6**	24.9	7.8	20.1	14.3	14.1	20.7	18.5	22.0	11.8
DAI	95	**5.6**	2.1	2.1	3.2	1.1	4.5	1.6	1.1	2.1	1.6
MAI	35	**42.9**	8.6	22.9	20.0	28.6	8.6	8.6	2.9	8.6	14.3
GEI	51	**47.1**	13.3	25.4	27.1	15.9	20.3	20.0	15.1	10.4	18.0
Social World											
CSI	87	31.5	15.7	48.0	18.1	32.5	10.1	14.2	**50.6**	13.5	13.5
WSC-RG	56	8.3	10.9	8.9	11.7	9.3	**43.9**	9.5	11.0	10.2	12.9
TEH-RG	57	9.6	13.1	10.4	11.2	10.4	**21.1**	11.5	12.8	12.0	14.9
Scientific World											
MC	15	**60.8**	45.2	38.5	37.1	49.5	45.5	40.0	32.7	10.3	31.9
HM	25	44.3	**45.4**	10.2	22.8	39.1	37.8	41.8	22.8	8.6	18.6
LS	32	48.9	46.6	28.2	30.1	37.8	**53.7**	36.5	29.5	19.2	25.9
PS	37	**52.2**	35.0	25.3	19.7	20.9	39.2	39.1	14.5	7.9	15.2
Index ↑	1,088	**31.3**	20.7	20.1	20.0	19.8	18.3	16.1	12.7	11.7	11.4
Index comparison		35 20 0									

Categories are reporting groups. Index uses the original displayed category means plus added item-level scores before final rounding; small differences do not imply statistical significance. Scientific results include all 109 items, with binary and graded scoring as specified in Appendix A. The bars use a shared zero-based 0–35 axis, with a reference at 20; scores remain on a 0–100 scale.

Abbreviations: PL = Planning; MAN = Manipulation; NAV = Navigation; AS = Action Sequencing; WI = Web Interaction; DAI = Desktop Application Interaction; MAI = Mobile Application Interaction; GEI = Game Environment Interaction; CSI = Coordination & Strategic Interaction; WSC-RG = Water, Soil & Climate Resource Governance; TEH-RG = Terrestrial Ecosystem & Health Resource Governance; MC = Mathematics & Computation; HM = Health & Medicine; LS = Life Sciences; PS = Physical Sciences.

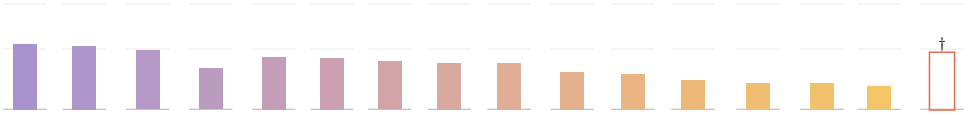
checking and feedback use as targets for improvement. These examples do not establish the dominant causes of aggregate failures.

Average strength can obscure task-critical weaknesses. GPT-6 Astra leads Index at 31.3, followed by GPT-5.6 Sol (20.7) and Gemini 3.6 Flash (20.1). Four configurations lead at least one category: Astra leads nine, DeepSeek V4.1 Flash three, Sol two, and Seed 2.1 Pro one. The case-weighted Index allows strengths to offset weaknesses, whereas a task requiring several skills jointly may not permit such compensation. These descriptive rankings concern the evaluated model–harness configurations; small differences derived from rounded baseline means do not establish significant superiority.

3.2 World Model Results

Action responsiveness must preserve accumulated progress. The fixed-intervention track tests whether valid action inputs produce valid state evolution. Its illustrative failures show that responding to the current instruction is only part of this requirement. Qwen-Image-3.0 restores the apparatus view following an instruction in the pendulum case (Appendix Figure 19), but introduces additional cups in the cup-stacking case, violating the initial quantity constraint (Appendix Figure 21). MiniMax H3 preserves the cup count yet introduces an unrequested red tray, whereas GPT-Image-2 preserves the original tray colors (Appendix Figure 22). These cases illustrate distinct failures in realizing requested changes and preserving unaffected state; they do not establish the causes of aggregate

Table 4: **World Model scenario scores.** Sixteen configurations; eight scenarios (50 cases each). The 15 full-coverage configurations are ordered by Overall within image/video groups. Seedance† covers six scenarios. Scores: 0–100; cross-modality comparisons are descriptive.

	Image outputs: full coverage															Video Partial†
Scenario	GPT Image 2 OpenAI (2026)	Qwen Image 3.0 Alibaba (2026)	Nano Banana 2 Google (2026)	Seedream 5.0 Pro ByteDance (2026b)	MiniMax H3 MiniMax (2026)	Kling 3.0 Kuaishou (2026)	Cosmos3 Super NVIDIA (2026)	LTX 2.5 Lightricks (2026)	Alaya World Team et al. (2026)	Zing 0.5 Seedleap.ai (2026)	YUME 1.5 Mao et al. (2025)	Matrix Game 3.5 Qian et al. (2026b)	Echo WM Zhang et al. (2026a)	LingBot World v2 Gao et al. (2026)	SANA WM Zhu et al. (2026)	Seedance 2.5† ByteDance (2026a)
Physical World																
Emb.	20.9	20.5	25.8	13.7	**53.8**	38.7	44.7	41.9	21.1	12.8	27.9	14.3	14.4	25.9	21.6	51.2
Drive	15.0	**19.8**	17.3	12.0	12.7	11.3	7.3	10.5	3.9	8.0	6.0	6.2	6.2	3.0	1.1	12.3
Daily	45.2	**46.1**	32.6	33.9	23.3	28.7	31.5	25.1	40.7	22.5	20.9	20.4	19.4	16.5	19.5	–
Digital World																
Web	**16.5**	10.7	11.8	9.5	3.7	2.7	1.8	5.0	5.8	3.4	5.9	4.0	0.7	1.5	2.4	6.0
Game	**13.3**	11.5	**13.3**	12.9	8.5	6.0	5.9	7.8	9.3	7.2	6.0	9.4	1.6	0.4	4.7	9.6
Social World																
Human	**22.9**	21.6	20.0	4.7	10.7	18.2	12.6	11.5	18.9	19.3	10.1	7.7	12.1	6.5	2.6	–
Scientific World																
Nat. Sci.	**21.5**	**21.5**	19.5	14.3	14.6	14.0	11.5	10.9	9.0	10.5	10.2	3.6	8.2	8.6	7.9	15.1
Exp. Sci.	**18.6**	17.4	15.5	7.6	11.2	17.3	11.9	10.7	14.3	15.7	5.5	12.9	8.1	6.7	2.3	17.9
Overall ↑	**21.7**	21.1	19.5	13.6	17.3	17.1	15.9	15.4	15.4	12.4	11.6	9.8	8.8	8.6	7.8	18.7†
Overall comparison																

Overall averages eight scenarios (400 cases). Bold/underline mark the highest/second-highest distinct scores, including ties. †Seedance covers six scenarios (300 cases); its six-scenario mean is excluded from full-coverage ranking and shown with an outlined bar. Daily and Human are unavailable due to face-generation restrictions; missing values are not zero. Bars use a zero-based 0–35 axis, with a reference at 20.

Emb. = Embodiment; Drive = Driving; Daily = Daily Life; Web/Game/Human = Web/Game/Human Interaction; Nat./Exp. Sci. = Natural/Expert Science.

ranking differences. In coupled interaction, such state errors can become misleading observations for the Agent.

Digital scenarios remain challenging across configurations. All evaluated configurations score at most 16.5 on Web and 13.3 on Game (Table 4). Even models with stronger results elsewhere remain weak in these scenarios: MiniMax H3 scores 53.8 on Embodiment but only 3.7 on Web and 8.5 on Game, while Qwen-Image-3.0 scores 46.1 on Daily Life but 10.7 on Web and 11.5 on Game. These results identify digital scenarios as a shared weakness under the evaluated protocol, although differences in task difficulty and scoring prevent attributing the gap solely to the domain.

Scenario averages do not establish suitability for a partner Agent. GPT-Image-2 leads WM Overall with 21.7 versus 17.3 for MiniMax H3, but their ordering reverses on Embodiment, at 20.9 versus 53.8. WM Overall weights the eight scenarios equally, whereas a particular Agent may encounter a different distribution of states and actions. A higher overall score therefore need not imply more reliable responses along that Agent's trajectories. Scenario profiles narrow this uncertainty, but selecting an effective Agent–World Model pairing requires evaluation on the interactions the pair actually produces.

Joint perspective. The two tracks reveal complementary obstacles to reliable interaction: Agents may fail to complete tasks despite locally correct decisions, while World Models may produce visually plausible outputs that violate task-relevant state constraints. These gaps can compound in coupled rollouts: incorrect world states may mislead subsequent decisions, and inadequate Agent verification may allow inconsistencies to persist. Reliable interaction therefore requires both effective task execution and faithful state evolution. Improving either component alone may leave failures on the other side unresolved, motivating evaluation of both capabilities when assessing the potential reliability of an Agent–World Model system. The separate evaluations characterize these capabilities, but do not establish how component scores predict coupled performance.

4 DISCUSSION

4.1 RELATED WORK

Existing benchmarks typically assess one component. Agent benchmarks measure tool use, collaboration, workplace execution, transferable skills, and long-horizon computer operation (Liu et al., 2024; Yao et al., 2024; Barres et al., 2025; Xu et al., 2024; Li et al., 2026; Sun et al., 2026), whereas World Model benchmarks evaluate prediction, embodied functionality, latent dynamics, memory, action control, and behavioral consistency (Shang et al., 2026b; Chen et al., 2025; Vafa et al., 2024; Ye et al., 2026; Huang et al., 2026). These benchmarks assess competence, not reciprocal improvement.

Prior work supports both directions. Dyna and learned World Models use simulated trajectories to improve planning and policies (Sutton, 1990; Ha & Schmidhuber, 2018; Hafner et al., 2025; Gu et al., 2025; Wang et al., 2026). Conversely, Agent-directed exploration and policy rollouts improve held-out World Model learning or adapt simulators (Kim et al., 2020; Chebotar et al., 2019), while recent methods refine World Models from interaction feedback (Zhang et al., 2026b; Yu et al., 2026). These gains are demonstrated within bounded task distributions. Narrow-domain systems begin to close the loop. WebEvolver (Fang et al., 2025) couples Agent experience with an evolving World Model, GenEnv co-evolves LLM Agents and environment simulators, and COMAP alternates World Model and policy updates (Fang et al., 2025; Guo et al., 2025; Liu et al., 2026).

4.2 TOWARD AGENT–WORLD MODEL CO-EVOLUTION

World Models can improve Agents through simulated experience, as DreamerV3 demonstrates across diverse control tasks (Hafner et al., 2025). Such benefits depend on controlling model bias: MBPO uses short simulated rollouts to support policy learning (Janner et al., 2019). The relevant test is whether this experience improves Agent behavior beyond the model used for training.

Agents can, in turn, improve World Models by collecting informative interaction data. Directed exploration improves predictive model learning (Kim et al., 2020), while SimOpt uses real policy rollouts to refine simulator parameter distributions and improve policy transfer (Chebotar et al., 2019). These updates improve the model or simulator; the Agent's task score alone does not measure the collected data's training value.

Connecting these directions offers a route to co-evolution, but separate gains do not establish a stable loop. WebEvolver links them by training a World Model on Agent trajectories and using it for policy training and look-ahead planning (Fang et al., 2025). We hypothesize that reciprocal improvement across broader multimodal tasks may become sustainable once both components reach sufficient task-specific competence, but neither this transition nor the required capability levels has been established.

Dual Exam diagnoses component capabilities rather than co-evolution itself. Our observed execution and state-consistency failures identify possible obstacles to useful interaction, not evidence of limited learnability. In the packing case, commands continue after the container changes (Figure 2); this observation motivates diagnosis but does not test whether either component can learn from the interaction. Future work should compare joint updates with Agent-only, World-Model-only, and no-update controls, using fixed held-out tests across rounds and seeds to track both components and closed-loop task success. This would test whether both components improve over successive rounds and whether these gains translate into more reliable joint task performance.

5 CONCLUSION

Dual Exam evaluates Multimodal Agents and World Models separately under controlled conditions, covering ten Agent configurations on 1,088 items across 15 categories and 16 World Model configurations, of which 15 cover all 400 cases. GPT-6 Astra leads Agent Index at 31.3, while GPT-Image-2 leads WM Overall at 21.7. Performance remains uneven: locally plausible Agent decisions need not produce valid solutions, and visually plausible World Model outputs can violate requested effects or previously established state. These complementary challenges motivate evaluating task execution and state-consistent generation separately, while coupled success and sustained co-evolution remain to be tested through joint interaction and learning experiments.

AI USE STATEMENT

We used generative AI tools, including OpenAI Codex, to assist with literature retrieval, drafting and revising portions of the manuscript and appendix, improving readability and organization, preparing LaTeX tables and figures, and developing data-processing and consistency-checking scripts. AI assistance also supported organizing evaluation records, recalculating aggregate scores under author-specified rules, qualitative analysis of recorded trajectories, and interpretation of results. AI-generated illustrative material is identified in the relevant captions and distinguished from recorded observations. The use of models as evaluation subjects is described separately in the experimental protocol. The authors are responsible for verifying AI-assisted text, code, calculations, citations, and interpretations against the underlying evidence, and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 2 describes task construction, evaluation protocols, and aggregation; Section 3 reports the main results. Appendix A provides task counts, available execution settings, scoring conventions, raw optimization objectives, and weighting sensitivity. Appendix B describes World Model coverage, scenario sensitivity, and case-evidence limits. Appendix C specifies the pilot protocol and packing-case calculations. These appendices also identify missing execution details, incomplete historical records, and differences between model–harness configurations that limit exact reproduction and comparison. Reported aggregates use the available baseline means and retained task outcomes; they do not imply access to complete item-level records for every configuration.

ETHICS STATEMENT

Dual Exam evaluates AI systems, not the abilities of people depicted in example scenes. Human-interaction and resource-governance cases are simplified task settings and do not establish findings about real populations or institutions. Benchmark scores and selected cases do not establish safety for autonomous deployment, including in physical, medical, or resource-allocation settings. Recorded observations, artifact replays, and AI-generated illustrations are distinguished in the captions to avoid misleading evidential claims. Evaluation traces may contain identifiers or interface content; any redistribution requires privacy review and compliance with the applicable third-party licenses and permissions.

REFERENCES

Alibaba. Alibaba unveils Qwen3.8-Max: Its largest and most capable flagship model to date, August . URL https://www.alibabagroup.com/en-US/document-2021044032125272064.

Alibaba. Qwen-Image-3.0: Rich content, authentic details, deep knowledge. https://www.alibabacloud.com/blog/qwen-image-3-0-rich-content-authentic-details-deep-knowledge_603385, July 2026. Official release announcement, republished on the Alibaba Cloud blog. Accessed September 12, 2026.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment. *CoRR*, abs/2506.07982, 2025. doi: 10.48550/arXiv.2506.07982. URL https://arxiv.org/abs/2506.07982.

ByteDance. Seed2.1 officially released: Advancing AI productivity, June . URL https://seed.bytedance.com/en/blog/seed2-1-officially-released-advancing-ai-productivity.

ByteDance. One-take creation, flexible referencing: Introducing Seedance 2.5. https://seed.bytedance.com/en/blog/one-take-creation-flexible-referencing-introducing-seedance-2-5, July 2026a. Official release announcement. Accessed September 12, 2026.

ByteDance. Beyond generation, it understands design: Introducing Seedream 5.0 Pro. https://seed.bytedance.com/en/blog/beyond-generation-it-understands-design-introducing-seedream-5-0-pro, July 2026b. Official release announcement. Accessed September 12, 2026.

Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan D. Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8973–8979. IEEE, 2019. doi: 10.1109/ICRA.2019.8793789. URL https://doi.org/10.1109/ICRA.2019.8793789.

Delong Chen, Willy Chung, Yejin Bang, Ziwei Ji, and Pascale Fung. Worldprediction: A benchmark for high-level world modeling and long-horizon procedural planning. *arXiv preprint arXiv:2506.04363*, 2025.

Tianxing Chen, Yue Chen, Zixuan Li, Junyuan Tang, Kailun Su, Haoran Lu, Weijie Wan, Baijun Chen, Songling Liu, Haowen Yan, Honghao Su, Zhiyang Dou, Kaixuan Wang, Dandan Zhang, Yunze Liu, Yan Qin, Qiwei Liang, Qiwei Wu, Zijian Lin, Wenwei Lin, Yuran Wang, Minghua He, Tianshu Wu, Ruihai Wu, Jingquan Zhou, Kai-Chong Lei, Haibao Yu, Yuanfeng Ji, Weiyang Jin, Guanyu Lin, Xiaofan Li, Qi Xiong, Renjing Xu, Zhongyu Li, Wenhao Chai, Enze Xie, Ziwei Wang, Yao Mu, Hao Dong, Wojciech Matusik, Mingyu Ding, Wenbo Ding, Ping Luo, and Masayoshi Tomizuka. Robodojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies, 2026. URL https://arxiv.org/abs/2607.04434.

Meng Chu, Xuan Billy Zhang, Kevin Qinghong Lin, Lingdong Kong, Jize Zhang, Teng Tu, Weijian Ma, Ziqi Huang, Senqiao Yang, Wei Huang, et al. Agentic world modeling: Foundations, capabilities, laws, and beyond. *CoRR*, abs/2604.22748, 2026. doi: 10.48550/arXiv.2604.22748. URL https://arxiv.org/abs/2604.22748.

DeepSeek. Deepseek-v4.1-flash: Pushing the limits of kv cache compression. URL https://arxiv.org/abs/2609.19969.

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*, pp. 27713–27724, 2025.

Tianqing Fang, Hongming Zhang, Zhisong Zhang, Kaixin Ma, Wenhao Yu, Haitao Mi, and Dong Yu. WebEvolver: Enhancing web agent self-improvement with coevolving world model. *CoRR*, abs/2504.21024, 2025. doi: 10.48550/arXiv.2504.21024. URL https://arxiv.org/abs/2504.21024.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, et al. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

Google. Introducing gemini 3.6 flash, 3.5 flash-lite, and 3.5 flash cyber, July . URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-6-flash-3-5-flash-lite-3-5-flash-cyber/.

Google. Build with Nano Banana 2, Our Best Image Generation and Editing Model. https://blog.google/innovation-and-ai/technology/developers-tools/build-with-nano-banana-2/, 2026. Accessed 2026-08-15.

Yu Gu, Kai Zhang, Yuting Ning, Boyuan Zheng, Boyu Gou, Tianci Xue, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun, and Yu Su. Is your LLM secretly a world model of the internet? model-based planning for web agents. *CoRR*, abs/2411.06559, 2025. doi: 10.48550/arXiv.2411.06559. URL https://arxiv.org/abs/2411.06559.

Jiacheng Guo, Ling Yang, Peter Chen, Qixin Xiao, Yinjie Wang, Xinzhe Juan, Jiahao Qiu, Ke Shen, and Mengdi Wang. GenEnv: Difficulty-aligned co-evolution between LLM agents and environment simulators. *CoRR*, abs/2512.19682, 2025. doi: 10.48550/arXiv.2512.19682. URL https://arxiv.org/abs/2512.19682.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pp. 1–7, 2025.

Youling Huang, Guanqiao Chen, Junchi Yao, Lu Wang, Fangkai Yang, Chao Du, Chenzhuo Zhao, Pu Zhao, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Beyond state consistency: Behavior consistency in text-based world models. *arXiv preprint arXiv:2604.13824*, 2026.

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*, pp. 21807–21818, 2024.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12498–12509, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5faf461eff3099671ad63c6f3f094f7f-Abstract.html.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active world model learning with progress curiosity. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5306–5315. PMLR, 2020. URL https://proceedings.mlr.press/v119/kim20e.html.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 881–905, 2024.

Kuaishou. Kling AI. https://kling.ai/, 2026. Official product website. Version evaluated: [FILL IN, e.g., Kling 2.x via API]. Accessed September 12, 2026.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025a.

Xiangyi Li, Yimin Liu, Wenbo Chen, Bingran You, Zonglin Di, Yifeng He, Shenghan Zheng, Kyoung Whan Choe, Jiankai Sun, Shuyi Wang, et al. Skillsbench: Benchmarking how well agent skills work across diverse tasks. *CoRR*, abs/2602.12670, 2026. doi: 10.48550/arXiv.2602.12670. URL https://arxiv.org/abs/2602.12670.

Yixia Li, Hongru Wang, Jiahao Qiu, Zhenfei Yin, Dongdong Zhang, Cheng Qian, Zeping Li, Pony Ma, Guanhua Chen, and Heng Ji. From word to world: Can large language models be implicit text-based world models? *CoRR*, abs/2512.18832, 2025b. doi: 10.48550/arXiv.2512.18832. URL https://arxiv.org/abs/2512.18832.

Lightricks. LTX-2.5 Supported Models. https://docs.ltx.io/models, 2026. Accessed 2026-08-15.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, 2024.

Youwei Liu, Jian Wang, Hanlin Wang, and Wenjie Li. COMAP: Co-evolving world models and agent policies for LLM agents. *CoRR*, abs/2606.02372, 2026. doi: 10.48550/arXiv.2606.02372. URL https://arxiv.org/abs/2606.02372.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A Text-Controlled Interactive World Generation Model. *arXiv preprint arXiv:2512.22096*, 2025. URL https://arxiv.org/abs/2512.22096.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: A benchmark for general ai assistants. In *International Conference on Learning Representations*, 2024.

MiniMax. MiniMax M3: Frontier coding, 1m context, native multimodality—all in one model, June . URL https://www.minimax.io/blog/minimax-m3.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. https://www.minimax.io/blog/minimax-h3, July 2026. Official release announcement. Accessed September 12, 2026.

Moonshot AI. Kimi K3: Open frontier intelligence, July . URL https://www.kimi.com/en/blog/kimi-k3.

NVIDIA. Cosmos 3: Omnimodal World Models for Physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf.

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

OpenAI. GPT-6 Astra: A new generation of intelligence, July b. URL https://openai.com/index/gpt-6-astra/.

OpenAI. GPT Image 2 Model. https://developers.openai.com/api/docs/models/gpt-image-2, 2026. Accessed 2026-08-15.

Cheng Qian, Emre Can Acikgoz, Bingxuan Li, Xiusi Chen, Yuji Zhang, Bingxiang He, Qinyu Luo, Dilek Hakkani-Tur, Gokhan Tur, Yunzhu Li, and Heng Ji. Current agents fail to leverage world model as tool for foresight. *CoRR*, abs/2601.03905, 2026a. doi: 10.48550/arXiv.2601.03905. URL https://arxiv.org/abs/2601.03905.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, et al. Matrix-game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026b.

Seedleap.ai. Zing-0.5: An efficient real-time interactive world model. https://huggingface.co/seedleap/zing-0.5, 2026. Model card. Accessed 2026-09-11.

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

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

Yiyou Sun, Xinyang Han, Weichen Zhang, Yuanbo Pang, Tianyu Wang, Yuhan Cao, Yixiao Huang, Chris Duroiu, Haoyun Zhang, Jeffrey Lin, et al. Agents' last exam. *CoRR*, abs/2606.05405, 2026. doi: 10.48550/arXiv.2606.05405. URL https://arxiv.org/abs/2606.05405.

Richard S. Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pp. 216–224. Morgan Kaufmann, 1990. doi: 10.1016/B978-1-55860-141-3.50030-4. URL https://doi.org/10.1016/B978-1-55860-141-3.50030-4.

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Mingliang Zhai, et al. Alayaworld: Interactive long-horizon world modeling–full technical report. *arXiv preprint arXiv:2607.18367*, 2026.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. *arXiv preprint arXiv:2406.03689*, 2024.

Zhaoyang Wang, Canwen Xu, Boyi Liu, Yite Wang, Siwei Han, Zhewei Yao, Huaxiu Yao, and Yuxiong He. Agent world model: Infinity synthetic environments for agentic reinforcement learning. *CoRR*, abs/2602.10090, 2026. doi: 10.48550/arXiv.2602.10090. URL https://arxiv.org/abs/2602.10090.

Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, Jing Tang, Lei Sun, Jiahong Wu, Xiangxiang Chu, Zeming Liu, and Kaiqi Huang. Omni-worldbench: Towards a comprehensive interaction-centric evaluation for world models. *arXiv preprint arXiv:2603.22212*, 2026.

Xiaomi. MiMo-V2-5: A leap in agency and multimodality, July . URL https://mimo.xiaomi.com/mimo-v2-5/.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*, 2024.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, et al. Theagentcompany: Benchmarking llm agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*, 2024.

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

Yuxue Yang, Shuyao Shang, Jiahe Wang, Zitong Zhou, Liang Tan, Junhan Zeng, Ruizhi Li, Junyan Li, Yu Liu, Xiao Yang, Yong Li, Jun Zhu, Hongsheng Li, Tieniu Tan, Lue Fan, and Zhaoxiang Zhang. Worldexam: Benchmarking world models from apparent appearance to inherent reactivity. *arXiv preprint arXiv:2608.02603*, 2026.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

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

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

Xiao Yu, Baolin Peng, Ruize Xu, Yelong Shen, Pengcheng He, Suman Nath, Nikhil Singh, Jiangfeng Gao, and Zhou Yu. Reinforcement world model learning for LLM-based agents. *CoRR*, abs/2602.05842, 2026. doi: 10.48550/arXiv.2602.05842. URL https://arxiv.org/abs/2602.05842.

Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, et al. Osworld2.0: Benchmarking computer use agents on long-horizon real-world tasks. *arXiv preprint arXiv:2606.29537*, 2026.

Songchun Zhang, Yaowei Li, Junhao Zhuang, Weiyang Jin, Haoyu Wang, Xin Lu, Yilang Sun, Shiyi Zhang, Haoran Li, Xiaoxiao Ma, et al. Echowm: Open and enterable omnimodal world models. *arXiv preprint arXiv:2608.23189*, 2026a.

Xuan Zhang, Wenxuan Zhang, See-Kiong Ng, and Yang Deng. Self-evolving world models for LLM agent planning. *CoRR*, abs/2606.30639, 2026b. doi: 10.48550/arXiv.2606.30639. URL https://arxiv.org/abs/2606.30639.

Zhipu. GLM-5.3-Flash: Frontier intelligence, flash cost, July . URL https://z.ai/blog/glm-5.3-flash.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

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. *arXiv preprint arXiv:2605.15178*, 2026.

A Details of Agent Evaluation

The current Agent leaderboard covers 1,088 items and 10 configurations. The following visual cases distinguish reference demonstrations, historical runs, submitted-artifact replays, and illustrative reconstructions. They explain task requirements and observed outcomes; they are not additional leaderboard trials or estimates of failure frequency.

Reference conditions. Fixed environments provide task-specific references, not a guarantee of error-free execution. We do not report an aggregate validation-error rate across environments. Interpretation therefore depends on the retained verifiers and execution records, including the Scientific limitations below.

A.1 Task Classification and Counting Units

The four domains organize task settings, not independent cognitive abilities. Physical categories describe action demands; Digital categories describe interfaces; Social categories describe interaction and resource settings; Scientific categories describe subject matter. Each item has one primary assignment for aggregation, even when its requirements overlap other categories.

Domain	Task category	Items
Physical	Planning	30
	Manipulation	42
	Navigation	282
	Action Sequencing	170
Digital	Web Interaction	74
	Desktop Application Interaction	95
	Mobile Application Interaction	35
	Game Environment Interaction	51
Social	Coordination & Strategic Interaction	87
	Water, Soil & Climate Resource Governance	56
	Terrestrial Ecosystem & Health Resource Governance	57
Scientific	Mathematics & Computation	15
	Health & Medicine	25
	Life Sciences	32
	Physical Sciences	37
Total	15 reporting categories	1,088

Table 5: Current agent inventory with descriptive category names. Counts follow the current leaderboard configuration and do not imply equal difficulty or statistical independence. Godot and GameWorld are included within Game Environment Interaction and are not counted again.

Physical and Digital assignments. Planning requires organizing dependent household actions to satisfy goals. Action Sequencing instead requires assigning shuffled clips to two tasks and reconstructing both orders. Navigation checks agent-specific traversability and routes; Manipulation checks object configurations. Digital categories identify Web, Desktop, Mobile, and Game environments, which share perception, planning, and recovery requirements. Browser-based games retain one primary assignment and are not counted twice.

Social assignments. The 200 cases instantiate 40 templates. Coordination & Strategic Interaction has 87 cases from 17 templates, covering bargaining, congestion, public goods, negotiation, and persuasion. Water, Soil & Climate Resource Governance has 56 cases from 11 templates; Terrestrial Ecosystem & Health Resource Governance has 57 cases from 12 templates. The latter includes agricultural, forest, and antibiotic resources. Assignments follow the recorded primary resource setting, despite ecological overlap; shared templates also limit statistical independence.

Scientific assignments. Scientific categories follow the task objective: mathematical or computational output, health, biological investigation, or physical science. Source assignments are retained for traceability, including the quantum Ising-chain task in Mathematics & Computation. Physical

Sciences includes physics, chemistry, materials, and Earth/space applications, distinct from the action-oriented Physical domain. Its original 13 ALE tasks span molecular calculations, climate simulation, lake calibration, orbital transfer, Hubble-image processing, and chemical kinetics. Reassignment would require recomputing category scores.

Counting units. The Scientific inventory combines 99 full workflows and ten tool-decision checkpoints (Table 9). Checkpoints inherit their parent repair task's subject and are separate evaluation items, not independent full-workflow trials. Six EdgeBench tasks are included; 84 other historical ALE tasks remain excluded.

A.2 Sensitivity to Domain Weighting

To test sensitivity to task composition, we first compute a case-weighted score within each domain, then average the four domain scores equally. The primary Index instead weights every item equally. Both use the same rounded baseline means and added-task outcomes. Extra decimal places explain ordering, not improved source precision; reweighting is not an uncertainty estimate.

Configuration	Index	Rank	Equal-domain	Rank
GPT 6 Astra	31.26	1	32.48	1
GPT 5.6 Sol	20.73	2	22.74	3
Gemini 3.6 Flash	20.15	3	20.84	5
Qwen 3.8 Max	20.01	4	19.67	6
Kimi K3	19.77	5	21.53	4
DeepSeek V4.1 Flash	18.35	6	23.16	2
GLM 5.3 Flash	16.08	7	19.47	7
Seed 2.1 Pro	12.72	8	16.83	8
MiniMax M3	11.74	9	11.63	10
MiMo V2.5	11.42	10	13.53	9

Table 6: Sensitivity of Agent rankings to domain weights. Within each domain, category scores retain their case-count weights. Values combine the rounded baseline means and added task-level scores; rank changes do not establish statistical significance.

GPT 6 Astra ranks first under both mixtures, but several other positions change substantially. DeepSeek V4.1 Flash moves from sixth to second and Kimi K3 from fifth to fourth. The case-weighted Index should therefore be interpreted as performance on the configured inventory rather than a distribution-independent ordering of general Agent ability. Confidence intervals and paired model comparisons require the underlying evaluation records and a resampling unit that accounts for shared templates and the repair/checkpoint parent links. In particular, the 200 Social cases instantiate 40 templates and should not automatically be treated as independent observations.

Dataset attribution. The scientific baseline uses Agents' Last Exam (ALE) results. The 30 repair tasks were supplied as a local SWE-bench collection (Jimenez et al., 2023); the ten derived tool-decision checkpoints are local evaluation items, not an official split. The optimization tasks use EdgeBench[1]. The selected sequencing example AS-002 uses clips identified as COIN in the supplied archive. Task-type labels elsewhere describe the evaluation format rather than a newly collected source dataset.

A.3 Physical Tasks: Goal and Constraint Satisfaction

Planning: satisfy the complete goal. Thirty tasks combine household actions with final-state constraints. Figure 5 shows four reference demonstrations: 105, 108, 65, and 73 steps satisfy six, five, four, and one goal groups, respectively. The last group contains 23 conditions. These demonstrations establish intended behavior, not model success rates.

[1]D. Zhu et al., *EdgeBench: Unveiling Scaling Laws of Learning from Real-World Environments*, 2026. https://arxiv.org/abs/2607.05155.



Figure 5: **Planning reference demonstrations.** Four household workflows pair a scene view with selected actions and final goal-group indicators from the same replay. Scene and event timestamps can differ. The counts describe satisfied goal groups, not model success rates; the fourth workflow has one group containing 23 conditions.

Manipulation: progress is not completion. Among 42 tasks, selected seed-0 episodes show successful ball pouring and bowl stacking (score 1.0), but failed cloth folding (0.2) and key insertion (0.15). Figure 6 contrasts these outcomes. The historical InternVLA_A15 / arx_x5 controller label does not independently identify its upstream planner.

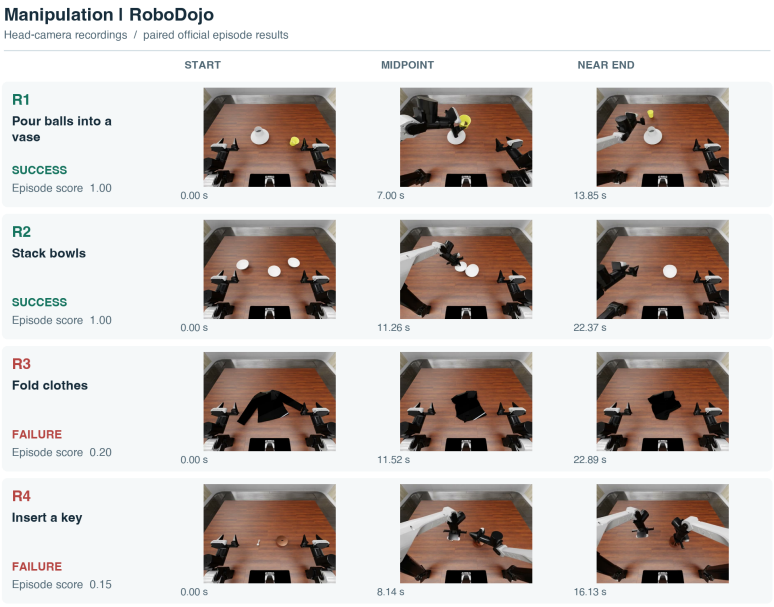


Figure 6: Representative manipulation outcomes from the source material. Ball pouring and bowl stacking succeed; cloth folding and key insertion receive partial credit without success. Start, middle, and late frames illustrate behavior, while the recorded evaluator supplies the outcome. Historical controller differences prevent treating this panel as a matched model comparison.

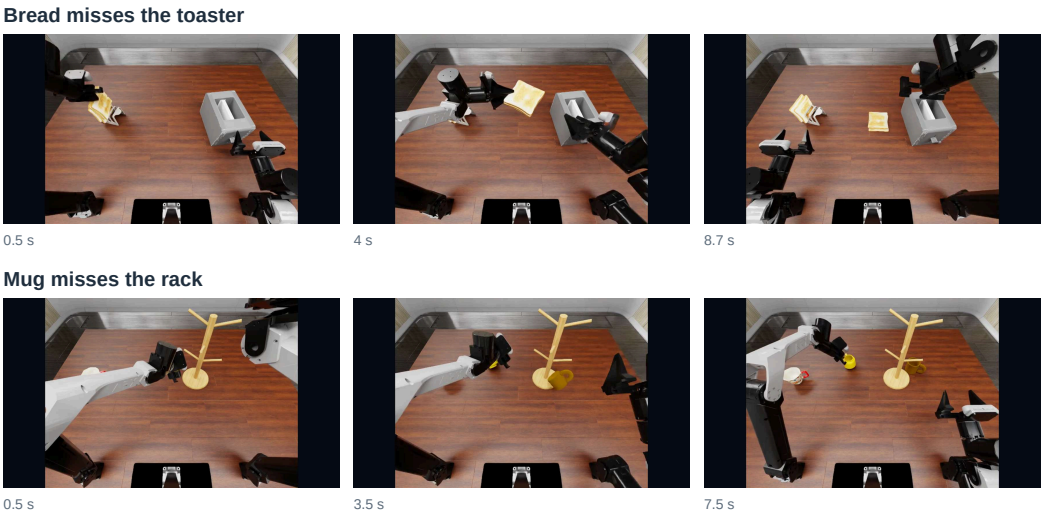


Figure 7: **Two manipulation failures in original robot footage.** Bread lands beside the toaster (top); a mug misses the rack (bottom). Times refer to the supplied clips. The archive labels both InternVLA_A15 episodes as failures; selected frames do not establish the precise controller fault or a leaderboard failure rate.

Navigation: provide a valid route. Each of 282 questions requires correct feasibility decisions for five agent types and valid traversable routes. In Figure 8, saved Gemini 3.6 Flash predictions

can classify feasibility correctly yet fail through nonexistent edges (6–7 in HM3D00002; 45–40 in MP3D00003). The references are valid examples, not unique routes; scene images provide context. Original API responses are unavailable, so the examples rely on saved predictions checked against task maps.

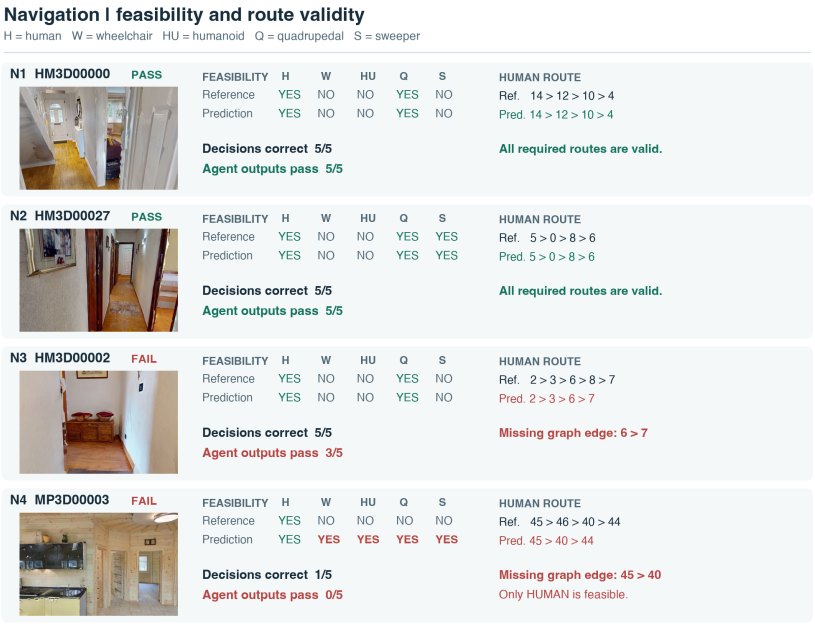


Figure 8: Navigation examples distinguish feasibility from valid route construction. A route can fail through a nonexistent edge even when all feasibility judgments are correct. The displayed reference is one valid route, not the only permitted answer; scene images are context rather than exact endpoint views.

Sequencing: membership and order must both hold. The 170 problems require partitioning shuffled clips into two tasks and reconstructing both reference orders. In the saved Qwen3.8-max-preview case AS-002, membership is correct but both orders are wrong (Figure 9). A plausible collection of clips is not an exact reconstruction.

Task family	Locally plausible evidence	Requirement still to verify
Planning	A salient household subgoal is reached	All goal groups and restoration constraints hold together
Manipulation	Objects move toward the target	Final arrangement satisfies the official objective
Navigation	Destination is judged reachable	Every edge and agent-specific route is valid
Sequencing	Relevant clips are recognized	Both assignments and complete orders match

Table 7: Physical-task lessons grounded in the recorded task rules and selected examples. The rows describe diagnostic distinctions, not measured failure frequencies.

A.4 Digital Tasks: Verified Interface Outcomes

Web: verify the requested evidence. The shopping tasks require 30 qualifying products and three priced purchase pages per product. Their checks separate 30 attribute credits from 90 page credits. A complete-looking list or an accepted evidence snippet does not establish that every requirement is met.

Desktop and mobile: verify the final state. The inventory retains 95 Desktop and 35 Mobile tasks. A completion message is insufficient: the brightness example claims success but the verifier

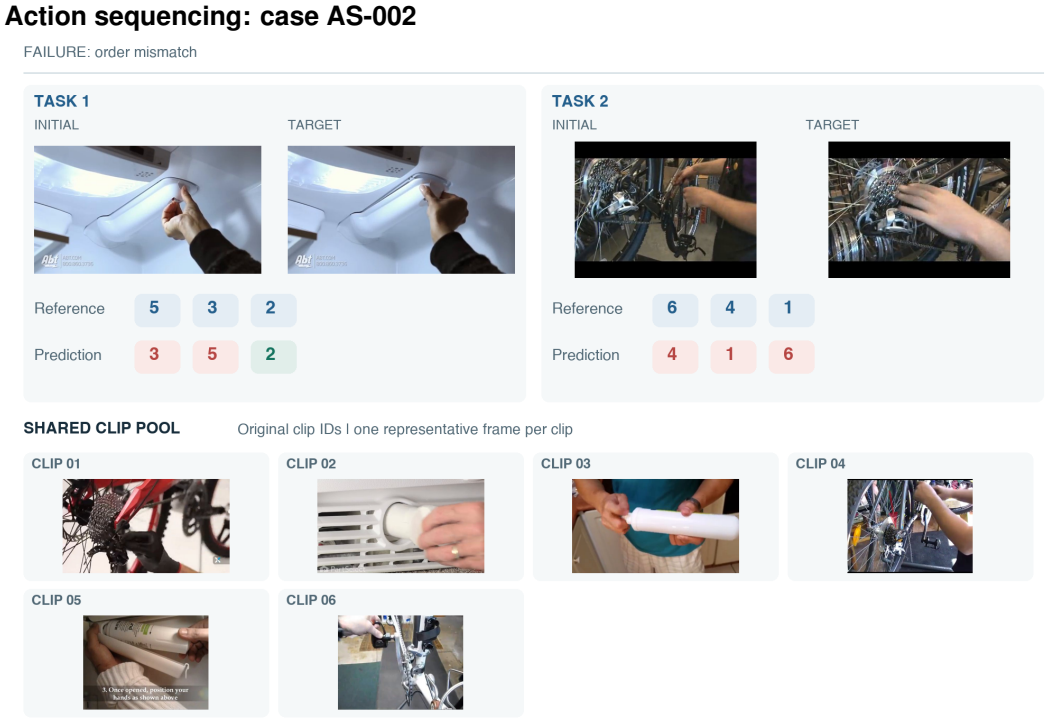


Figure 9: **Correct clip membership, incorrect sequence.** In the historical Qwen3.8-max-preview prediction for case AS-002, the predicted orders $[3, 5, 2]$ and $[4, 1, 6]$ differ from references $[5, 3, 2]$ and $[6, 4, 1]$. The images show task endpoints and one representative frame per clip; they do not measure failure frequency in the main leaderboard.

records 183/255 (Figure 10). A separate Gemini archive successfully looks up repository statistics and submits a benchmark email (Figure 11).

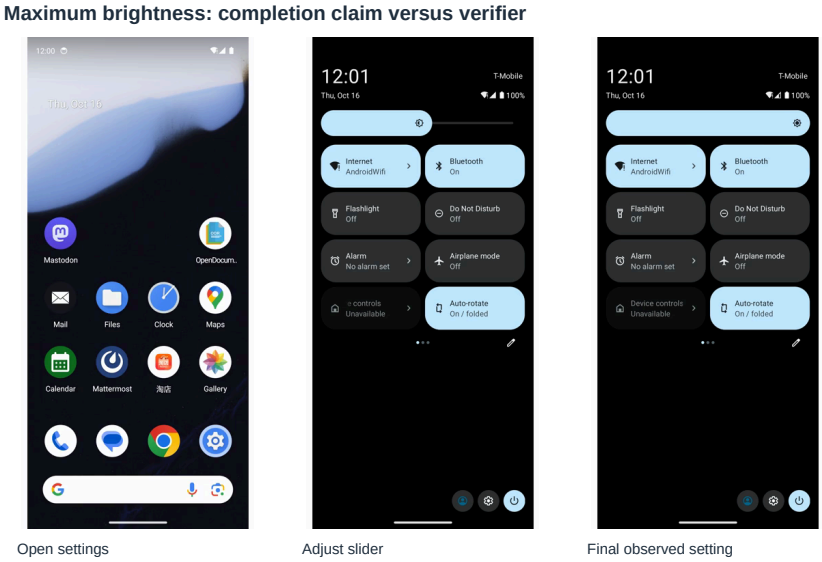


Figure 10: **Claimed maximum brightness, failed verification.** Cropped UI observations from the supplied MobileWorld replay show opening settings, adjusting the slider, and the final screen. The Agent answers that brightness is maximal; the result file records 183/255 and score 0. The supplied evidence does not establish a model identity.

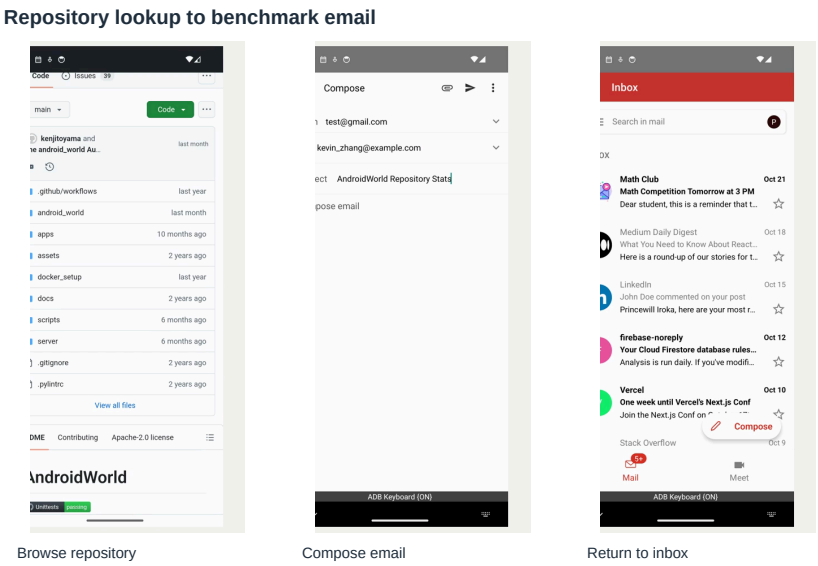
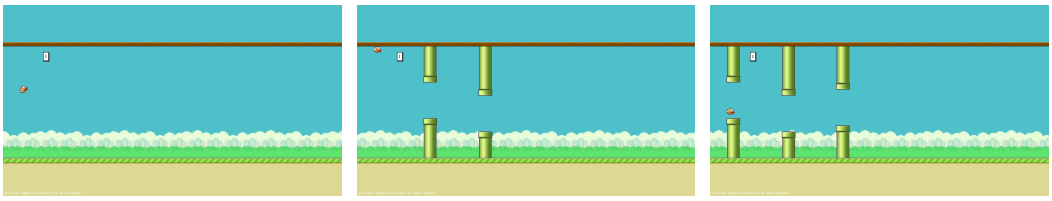


Figure 11: **Repository lookup and benchmark email.** Cropped observations from CheckGithubInfoTask show browsing, composing, and returning to the inbox. The archive labels Gemini 3.6 Flash, contains 40 observations, and records score 1.0. Reported values (854 stars, 15 contributors) belong to that historical run, not current GitHub statistics. This is a saved-observation replay, not continuous screen capture.

Games: inspect the transition as well as the action. The MiniMax M3 example clears one pipe after nine valid actions (Figure 12). A separate GPT-5.6 Sol archive resets after collision and records zero pipes across 100 actions (Figure 13). These historical cases illustrate feedback and termination; they are not a matched model comparison.



Step 1: flap | Step 6: wait | Step 9: flap

Figure 12: **Selected observations from the Flappy Bird case.** Frames from the nine-action MiniMax M3 trace show the changing obstacle configuration and the corresponding control choices. These are observations before the labeled actions; completion and score 1 come from the recorded evaluator, not from the last screenshot alone.

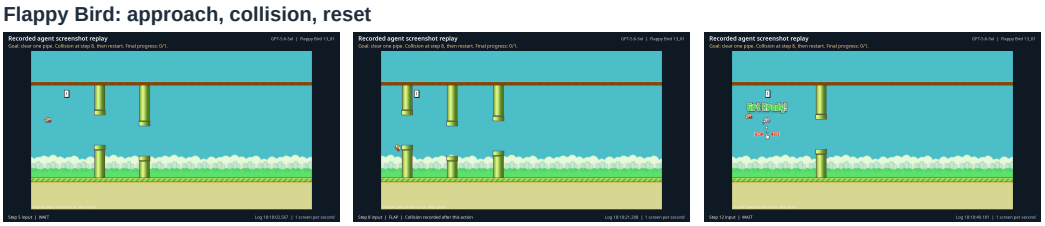


Figure 13: **Approach, collision, and reset in a failed Flappy Bird run.** The supplied GPT-5.6 Sol (high) archive replays saved observations. A collision is logged after the middle action; the later screen returns to Get Ready. The manifest reports 100 actions and no pipe cleared. This is distinct from the successful MiniMax example.

Reconstruction: an endpoint is not a trajectory. The supplied GPT-5.6 Sol Go artifact contains five opening placements, 163 passes, and a setup edit, rather than a reconstructed 168-move game (Figure 14). The source manifest records score 0.

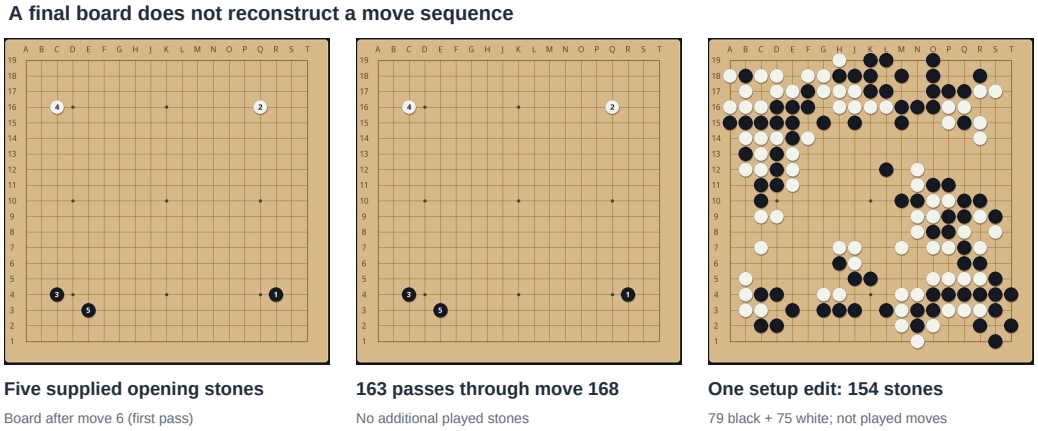


Figure 14: **Submitted SGF omits the move history.** The opening board persists through 163 passes; a separate setup node then clears all 361 points and places 79 black plus 75 white stones. Cropped boards come from deterministic playback of the submitted artifact, not an execution recording or AI reconstruction. This historical example is not an additional leaderboard item.

A.5 Resource-Governance Examples: Groundwater and Forests

Inventory and mechanisms. Social comprises 200 cases: Coordination & Strategic Interaction (87), Water, Soil & Climate Resource Governance (56), and Terrestrial Ecosystem & Health Resource Governance (57), from 40 templates (17/11/12). Among the 113 resource-labelled cases, 87 use capacity allocation and 26 use resource extraction; scene labels are not independent governance abilities. Figure 15 visualizes one documented allocation, while Table 8 records two extraction traces.

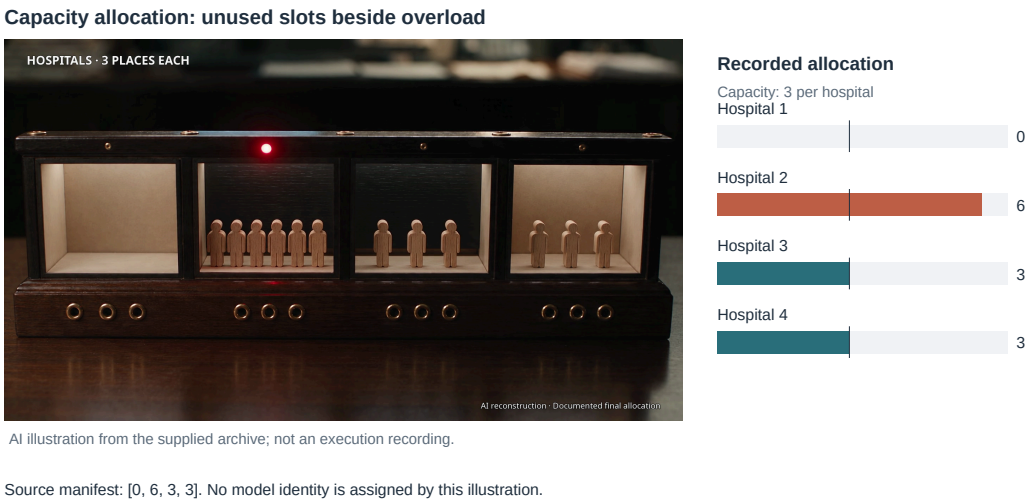


Figure 15: **Unused capacity beside overload.** The source manifest for HOS-F5 records allocation $[0, 6, 3, 3]$ with capacity three per hospital. The left panel is an AI illustration from the supplied package, not an execution recording; the right panel plots the manifest values. No model identity is supplied, and the picture cannot establish the actual action sequence.

Two resource-extraction traces. The historical groundwater and forest records retain identifier `anthropic/claude-fable-5`; no equivalence to a leaderboard configuration is inferred. Each has twelve operators, six rounds, initial stock 100, and sustainable extraction 50 per round. A fixed industrial request of 46 leaves four sustainable units for all participants. First-round participant requests total 129 (groundwater) and 157 (forest), exceeding supply; these are requests, not realized yields. Both logs end round one with zero stock and record no later participant extraction, `survived=false`, `constraint_ok=false`, and raw score 0.05. These simplified games are not validated ecological simulators.

Source identifiers: `social.commons.aquifer.v0`, strict100 suffix `018d88e51fdb6e3d`; and `social.commons.forest.v0`, archive `case1`. Each contains 72 prompt–response entries.

Recorded quantity	Groundwater	Forest
Initial stock	100	100
Sustainable total per round	50	50
Fixed industrial request per round	46	46
Participant requests in round 1	129	157
Stock after round 1	0	0
Participant extraction in each of rounds 2–6	0	0
Final stock	0	0
Raw recorded score	0.05	0.05

Table 8: Two source-linked resource-extraction failure traces. Groundwater quantities are in megaliters and forest quantities in hectares within the respective game. First-round participant requests are sums of the twelve JSON responses, not realized yields; stock and later extraction follow the recorded state logs. Scores are on the original 0–1 scale and are not category averages.

A.6 Scientific Workflows and Verification

Task types and classification. Table 9 maps the 109 Scientific items to subjects and task types. Repair tasks are classified by scientific objective, and each tool-decision checkpoint inherits its parent assignment. The inventory includes 99 full workflows and ten related decision checkpoints; the latter are not independent workflow trials. Dataset attribution is given in Section A.2.

Scientific category	Baseline	Repair	Decision	Optimization	Total
Mathematics & Computation	9	2	1	3	15
Health & Medicine	21	2	2	0	25
Life Sciences	20	9	3	0	32
Physical Sciences	13	17	4	3	37
Total	63	30	10	6	109

Table 9: Scientific item counts by subject category and task type. All 109 items enter the current leaderboard. Tool-decision items are decision checkpoints derived from repair workflows and inherit the parent subject assignment; parent linkage is not an independence claim.

Scoring and execution settings. Repair and tool-decision items receive 100 for the recorded strict pass and 0 otherwise; optimization uses the graded mappings below. Table 10 retains the archived outcomes. Six configurations use Claude Code 2.1.240 for repair and an OpenAI-compatible decision interface (11 tools, 300 seconds, high reasoning, one attempt, no retries). The four-configuration extension uses native Codex for Astra and API execution for GLM, DeepSeek, and MiMo. Context, tools, and budgets differ, so scores compare model–harness configurations.

Interpretation limits. Six decision prompts (v01, v03, v05, v06, v08, v10) omit a required `task_id`; their strict scores remain included and cannot isolate tool-selection ability. The repair collection lacks upstream instance IDs, and one title differs from its instruction (both map to Physical Sciences). Retained outcomes include 30 Seed repair runs without task-tool use, 26 Kimi runs with interrupted tools, 29 extension timeouts, and one Astra tool-cap termination. These records do not by themselves distinguish model errors from harness effects. Excluded-attempt totals for infrastructure replacements, extension repair caps, and exact provider snapshots are unavailable; historical decision-token limits also differ from Astra's native interface.

Configuration	Repair passes / 30	Decision passes / 10
GPT-6 Astra	14	2
GPT-5.6 Sol	10	1
Gemini 3.6 Flash	5	2
Qwen 3.8 Max	7	2
Kimi K3	0	2
DeepSeek V4.1 Flash	8	2
GLM 5.3 Flash	8	1
Seed 2.1 Pro	0	2
MiniMax M3	3	2
MiMo V2.5	4	2

Table 10: Strict outcomes on the 40 added code-repair and tool-decision items. All 10 configurations have 30 code-repair and ten tool-decision records (400 task–configuration observations). Harness and specification limitations described above apply.

Tool choice: selecting is not executing. TC-02 asks for one `read_source` call. GPT-5.6 Sol matches the path, bounds, and call constraints for score 1.0 (Figure 16); the selected call is not executed.

Numerical science: check the submitted artifacts. The ten-qubit Ising workflow produces four files and score 1.0 (Figure 17). Its residual is 8.093×10^{-13}; a second calculation differs in observables by less than 5.71×10^{-4} against tolerance 10^{-2}. Shared helper code makes this a consistency check, not independent replication.

TASK INPUT / STARTING CONTEXT

Original checkpoint: a scoped search has localized MACS3 consumption-cursor and score-attribution logic to PeakDetect.py lines 410-505, but omitted surrounding state transitions. Choose exactly one next call; do not search again or edit from the truncated snippet.

TC-02 | gpt-5.6-sol | one-round chat tool calling | 11 supplied tool definitions

01 What is known before the call

Known: repository file and line range.

Missing: surrounding transitions needed to inspect the logic.

The previous search is described in the request; its execution trace is not supplied.

Original system/user input. Do not invent unobserved evidence.

02 The actual submitted call

The response carries the supplied path and line bounds into a single source read.

```
read_source({
  "path":
    "source/MACS3/Signal/PeakDetect.py",
  "start_line": 410,
  "end_line": 505
})
```

03 Compare the submitted arguments with the strict contract

Field	Required	Recorded
Tool	read_source	read_source
Path	Exact PeakDetect.py path	Exact path matches
start_line	390-430	410
end_line	490-530	505
Call count	Exactly 1	1

Why this action follows: the location is already known, while source context is missing. Re-searching or editing would violate the supplied checkpoint constraints.

RECORDED GRADE / FULL_PASS = TRUE / SCORE = 1.0

Tool name, path, line bounds and call count satisfy the contract; no forbidden tools are used. The selected call is the answer. The harness does not execute it, so there is no next-round tool return.

Figure 16: **TC-02: verifying a proposed next action.** The source context, GPT-5.6 Sol call, and strict field checks are displayed together. The score of 1.0 applies to this decision checkpoint; the selected tool is not executed and no completed repair is claimed.



Figure 17: **NW-01: numerical workflow and evaluated artifacts.** The historical 41-step trace shows method selection, dependency recovery, numerical checks, and the four retained output files. The archived evaluation assigns 1.0. Shared helper code limits independence of the second calculation; the panel's source identifiers distinguish the execution framework, API alias, and archive model group.

Code repair: validate the specified behavior. The DFT patch propagates an ignored parameter and rejects invalid values (Figure 18). Three tests pass; energy changes by -0.34018629 Hartree. This verifies software behavior, not physical accuracy.

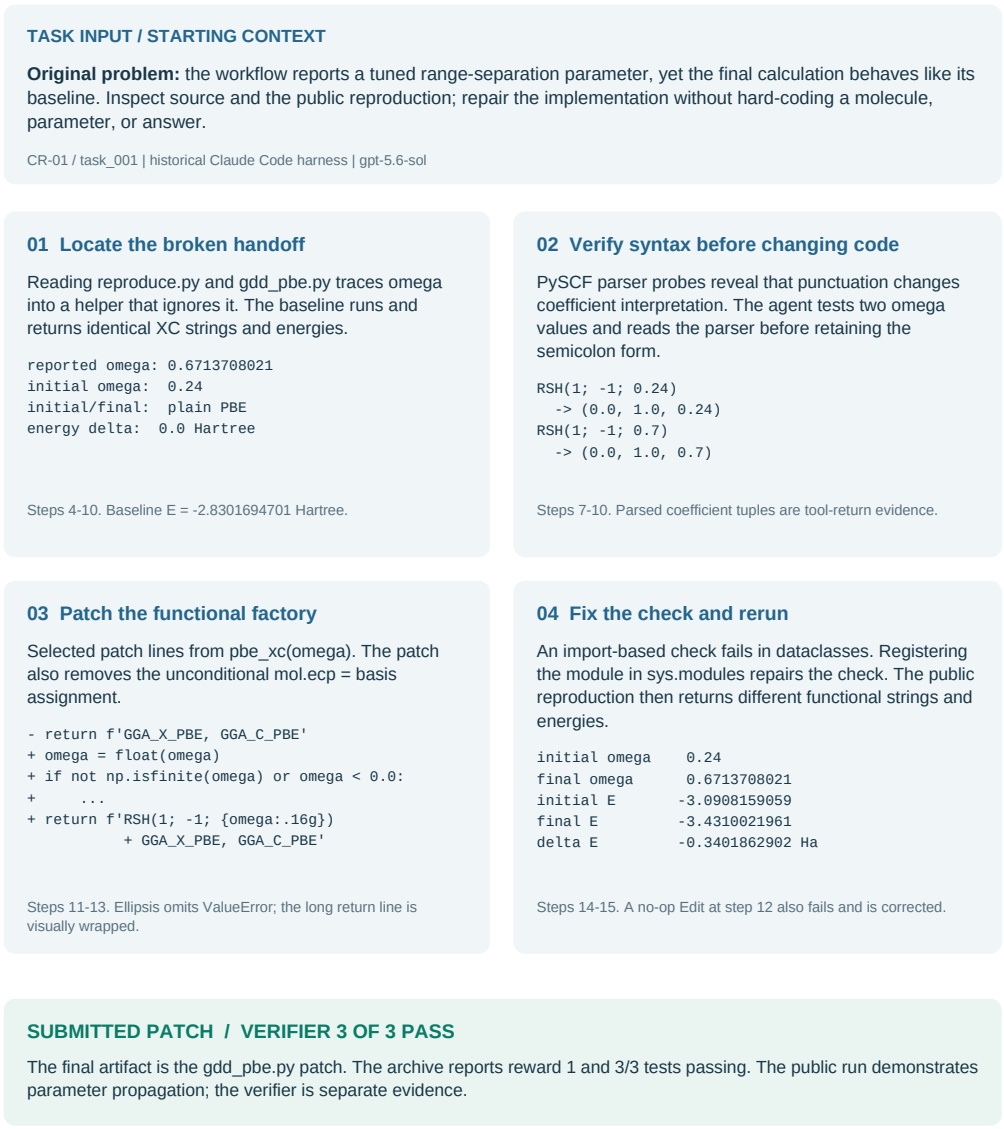


Figure 18: **CR-01: making a parameter affect the calculation.** The historical GPT-5.6 Sol / Claude Code trace connects source inspection and parser checks to the patch and changed output. The external verifier reports three passing tests. Omitted code and wrapped lines are marked; passing these tests establishes the specified software behavior, not improved physical accuracy.

A.7 Edge Task Aggregation and Raw Scores

We include six tasks from EdgeBench. Wireless Electricity Layout, BipedalWalker Locomotion RL, and Graph Node Classification are assigned to Mathematics & Computation; Borden Source Inversion, Molecular Self Assembly, and DABIC Gravity Inversion are assigned to Physical Sciences. These primary reporting assignments do not establish disciplinary exclusivity.

Dual Exam score convention. Wireless, Molecular, Graph, and BipedalWalker retain their archived mappings to 0–100. Borden and DABIC have no separate rescaling in the public task configurations. We clip their raw scores below 0 to 0 and above 100 to 100, leaving values within that range unchanged. This is a **Dual Exam aggregation convention, not an official EdgeBench normalization rule**. It bounds each item's contribution without equating task difficulty. Table 11 retains the raw values; Astra's Borden score of 100.493 contributes 100. The reason for the raw value exceeding 100 remains unresolved.

Budget and evidence. We retain the best valid submission within the selected run's 43,200-second allowance, including submissions before early termination. Confirmed absence of a valid submission receives zero; missing records do not automatically receive zero. Six configurations use frozen run IDs; the four additions use the supplied summary, with all four Wireless endpoints checked against best-submission judge reports. DeepSeek and MiMo Wireless reached context/resume limits but passed 50/50 tests; their mapped scores are 19.9172 and 0, respectively. MiniMax Wireless has conflicting validity fields, but either interpretation maps to zero. The transferred archive is truncated: the remaining added-task endpoints and evaluator-version equivalence are not all independently verified. Astra uses native Codex; these are model–harness results.

Configuration	Wireless	Borden	Molecular	Bipedal	Graph	DABIC
GPT-6 Astra	1.30310e+15	100.493	3.20779e+08	28.34	0.85042	15.877115
GPT-5.6 Sol	3.43635e+15	34.69	2.36472e+08	22.76	0.704273	16.797531
Gemini 3.6 Flash	3.03795e+16	7.224	1.73814e+08	17.79	0.513631	17.460083
Qwen 3.8 Max	3.45418e+15	7.298	2.56779e+08	20.35	0.414559	15.042482
Kimi K3	3.08015e+15	75.276	1.53309e+08	22.33	0.669031	9.45664
DeepSeek V4.1 Flash	2.46882e+15	50.646	2.85375e+08	30	0.610576	17.480643
GLM 5.3 Flash	1.51011e+16	41.463	2.13459e+08	29.41	0.500476	15.929143
Seed 2.1 Pro	6.63101e+17	7.432	6.92801e+07	2.5	0.284502	8.837336
MiniMax M3	1.00000e+17	7.048	1.86909e+08	13.6	0.369085	–
MiMo V2.5	6.41351e+16	38.967	1.48430e+08	19.83	0.596243	8.5838

Table 11: Raw Edge objectives before Dual Exam aggregation. Wireless is minimized; the other five objectives are maximized. Large objectives are abbreviated in scientific notation; exact source values are retained in the audit files. A dash denotes no recorded valid score (MiniMax DABIC), assigned zero in our aggregation. This table reports objective values, not pass rates; early-terminated valid results are retained under the stated budget rule.

A.8 Evidence Summary

Across these examples, local progress can fail a final constraint, a completion claim can disagree with the verifier, and a plausible endpoint can omit the required process. The figures support these concrete distinctions. They do not estimate failure frequencies, isolate model-only ability, or establish which intervention improves leaderboard performance.

B Details of World Model Evaluation

The current evaluation covers eight scenarios with 50 cases each (400 cases for full coverage). Fifteen configurations cover all eight scenarios; Seedance 2.5 covers six (300 cases), with Daily Life and Human Interaction unavailable. Its six-scenario mean is reported separately from the full-coverage ranking.

Six case panels examine requested action effects, preserved state, and rule adherence under fixed interventions. They illustrate errors that aggregate scenario scores cannot distinguish; their source and observation limits are specified below.

Reporting scope. Scores are descriptive under the specified case rubrics. We do not report independent evaluator agreement, adjudication or invalid-run rates, or per-configuration execution settings sufficient for exact replication. Confidence intervals, repeated-run variability, and no-op, shuffled-action, or copied-frame controls are also not reported. The following weighting analyses and selected cases assess sensitivity and illustrate behavior; they do not establish scoring reliability or significant model differences.

B.1 Sensitivity to the Scenario Mixture

The full-coverage ordering in Table 4 depends on the scenario mixture. MiniMax H3 outperforms GPT-Image-2 on Embodiment (53.8 versus 20.9), but has a lower Overall (17.3 versus 21.7). A partner Agent concentrated on embodied tasks may therefore require a different weighting from the benchmark's eight-scenario mean.

If Embodiment receives more than about 22.9% of the weight and the other seven scenarios share the remainder equally, MiniMax H3 overtakes GPT-Image-2. This calculation uses rounded scenario scores and unchanged within-scenario mixtures; it is a weighting sensitivity result, not a threshold for partner suitability. Digital scores remain low (Web at most 16.5; Game at most 13.3), subject to task and rubric differences.

B.2 Case Evidence and Observation Limits

Each panel shows six checkpoints from the complete instructed history. Columns align instructed steps; rows track four image and eleven video configurations. The reference is an initial scene, not a ground-truth trajectory. Historical labels are preserved: the first four cases use Seedream 5.0 Pro, the interface cases use Seedream 5.0; sources also include physical Yume `runs_hard_v2`, pendulum/mentorship `yume-5b`, and September 2026 Kling reruns. These batches need not match the main leaderboard settings.

Selected checkpoints can reveal visible discrepancies but cannot establish continuous-time behavior, failure prevalence, or a controlled model comparison. Table 12 separates each case's requested effect, invariant, and evidence requirement. Similarity alone is insufficient: a copied frame may omit an action, while an altered frame may realize it and corrupt unrelated state.

Case	Requested effect	State to preserve	Evidence limit
Pendulum	Evolve under edited gravity	Bob, rod, support	Period needs video
Mentorship	Named person's pose/gaze	Identity and other attributes	Timing needs video
Cup manipulation	Assigned arm actions	Cup count and roles	Concurrency needs video
Dual trays	Transfer and rule changes	Jar, trays, prior state	Final image is not completion
Minesweeper	Edit specified cells	Other cells and flags	Layout alone is insufficient
Shopping	Query and filter operations	Earlier selections	Imagery is not UI state

Table 12: A common reading of the World Model cases: desired change, invariant, and observation needed for verification. These are case-specific checks rather than independent quantitative metrics.

B.3 Rule Changes and Delayed Effects

The pendulum case changes Earth gravity to lunar gravity at step 4 without requesting immediate motion. Steps 5–6 require evolution under the new rule, and step 7 restores the view. The copper bob, rod, and support must persist. Alaya's final checkpoint no longer shows the original apparatus, whereas Qwen-Image-3.0 restores its view. Neither static observation verifies the oscillation period.

Immediate stability after a rule edit can be valid; compliance must also be checked during later motion. Yume's step-4 checkpoint is carried forward and provides no new evidence of a response.

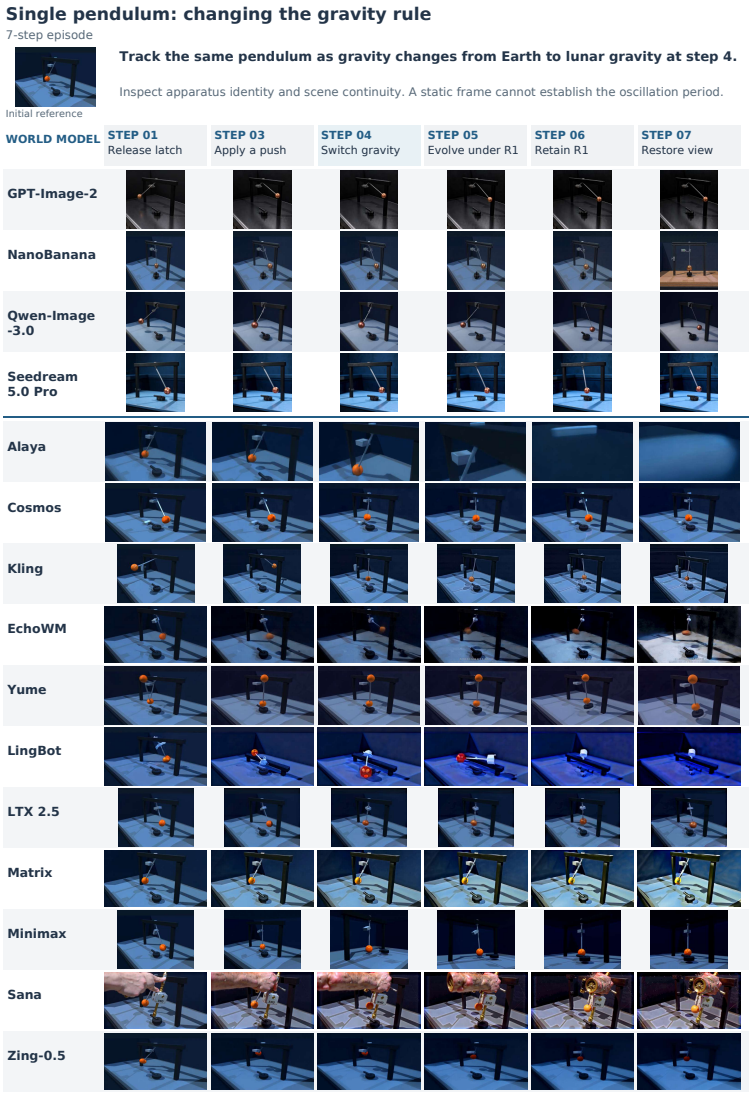


Figure 19: **Single pendulum: rule editing and persistent apparatus.** Steps 1, 3, 4, 5, 6, and 7 cover release, a push, lunar-gravity editing, subsequent evolution, and view restoration. Static checkpoints reveal identity changes; oscillation period requires video. Yume's step-4 checkpoint is carried forward.

B.4 PARTICIPANT-SPECIFIC STATE

In the painting-studio case, A is the standing mentor and B the seated student. A establishes gaze toward B and raises a hand; B later nods. At step 9, A must lower the hand while retaining the gaze. The new instruction changes one attribute without resetting the others.

Changes in framing or appearance make participant-specific checks ambiguous in several rows. The displayed states support checks of visible identity and pose, but response timing requires video. A socially plausible image can still assign the action to the wrong person; uncertain gaze or identity cannot be treated as verified compliance.



Figure 20: **Painting-studio mentorship: entity-specific updates.** Steps 1, 3, 4, 5, 8, and 9 show baseline, restored view, mentor gaze/hand updates, student response, and lowering of the mentor's hand while gaze persists. The panel supports state checks, not within-video timing claims.

B.5 Coordination and State Across Revisits

Cup manipulation. Two arms manipulate three cups through concurrent motion, nesting, hand-over, and role exchange. The step-7 role change must persist into step 8, when the right arm waits. EchoWM and Yume introduce human hands, and Qwen-Image-3.0 adds cups; MiniMax H3 retains three cups at step 10. The first two observations illustrate entity and inventory changes, while retained cup count verifies only one task constraint. A correct final arrangement would still not verify arm assignment or within-turn concurrency, which require temporal evidence.

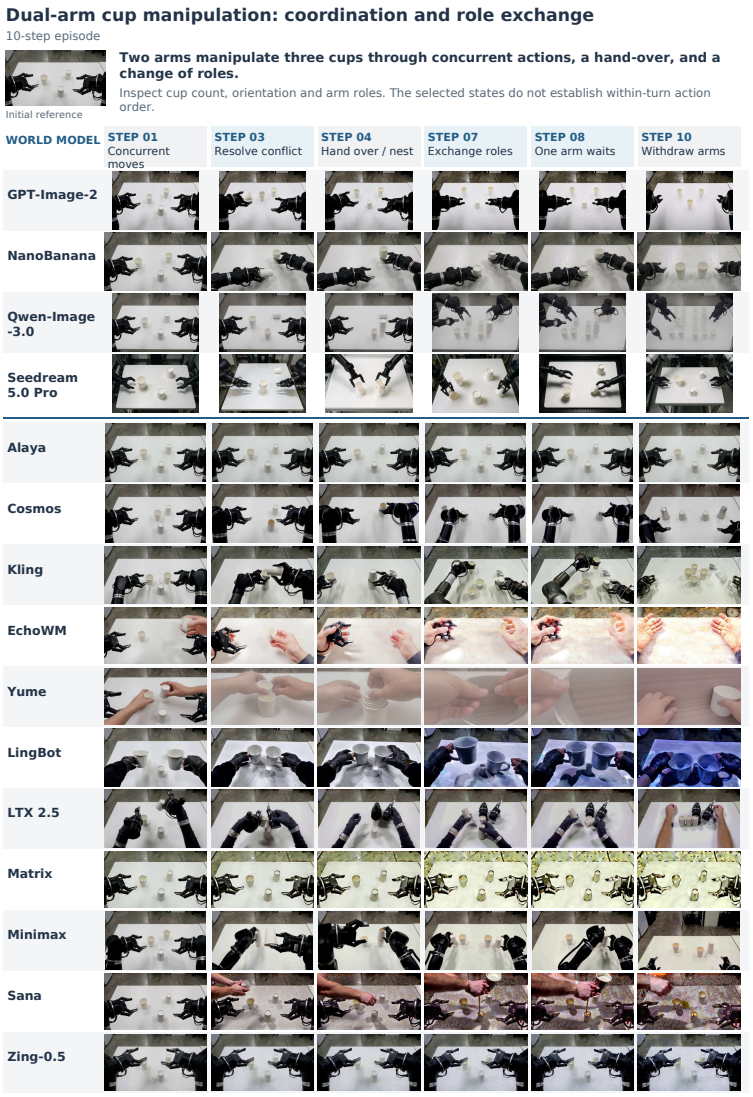


Figure 21: **Dual-arm cup manipulation: coordination with persistent roles.** Steps 1, 3, 4, 7, 8, and 10 sample concurrent actions, contested nesting, hand-over, role exchange, one-arm action, and withdrawal. Images can expose object changes; order and sustained nonmovement require video.

Dual trays. The 20-step task requires moving a glass jar to the blue tray and aligning the trays while retaining state through rule, role, and viewpoint changes. Steps 7 and 10 bracket a visit to another bench; steps 5 and 18 edit rules without requesting immediate motion. At step 10, MiniMax H3 introduces a red tray, whereas GPT-Image-2 retains the original white and blue trays. Yume's final checkpoint shows a wooden bench. Tray appearance alone does not verify jar placement, alignment, or rule compliance.

Revisits test whether a workspace agrees with its earlier state plus authorized changes. Adjacent-frame similarity cannot establish this persistence. The observed drift does not, however, identify memory as its sole cause; generation and instruction interpretation may also contribute.

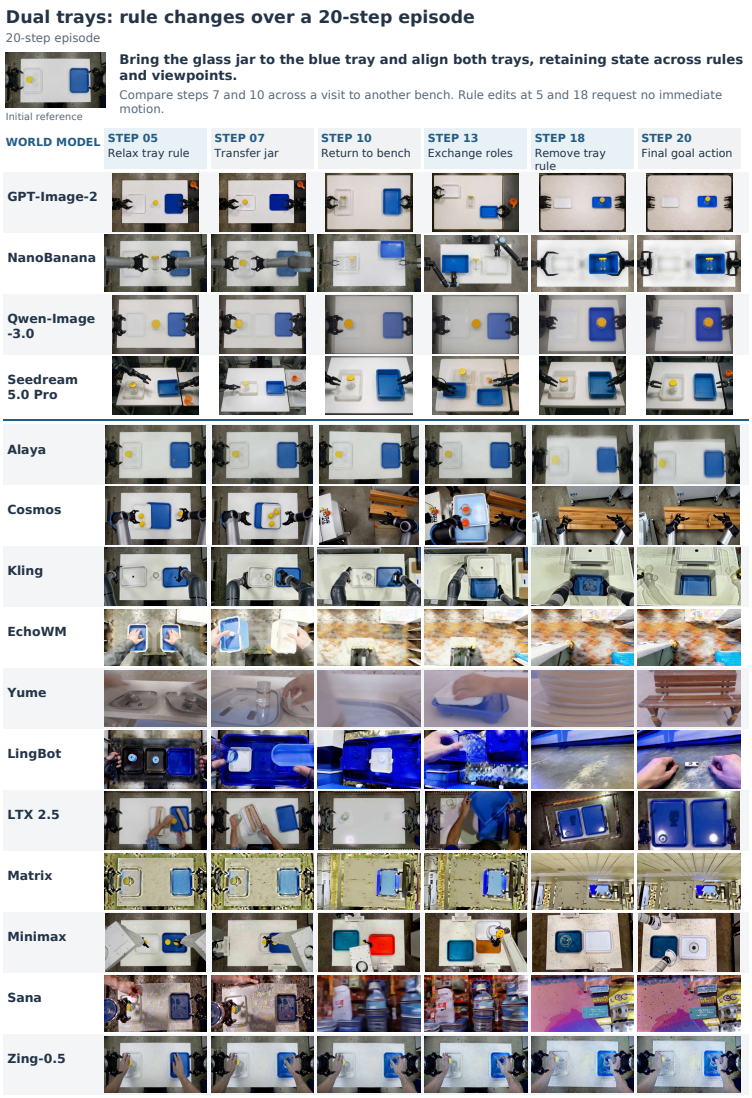


Figure 22: **Dual trays: memory across interventions and revisits.** Steps 5, 7, 10, 13, 18, and 20 cover rule relaxation, jar transfer, return from another bench, role exchange, rule removal, and final action. Rule edits at 5 and 18 do not request immediate object motion.

B.6 Exact Updates in Digital Interfaces

Minesweeper. Actions flag g1 and i1, reveal e5, then flag g2, h2, and i2. The omitted second step flags h1, which must also persist in later checkpoints. Checks therefore use the complete seven-step history rather than only the six displayed actions. LingBot replaces the grid with a different graphic. Less conspicuous errors can preserve the layout while omitting a requested flag or erasing a previous one. Validity requires both the target edit and unchanged non-target cells.

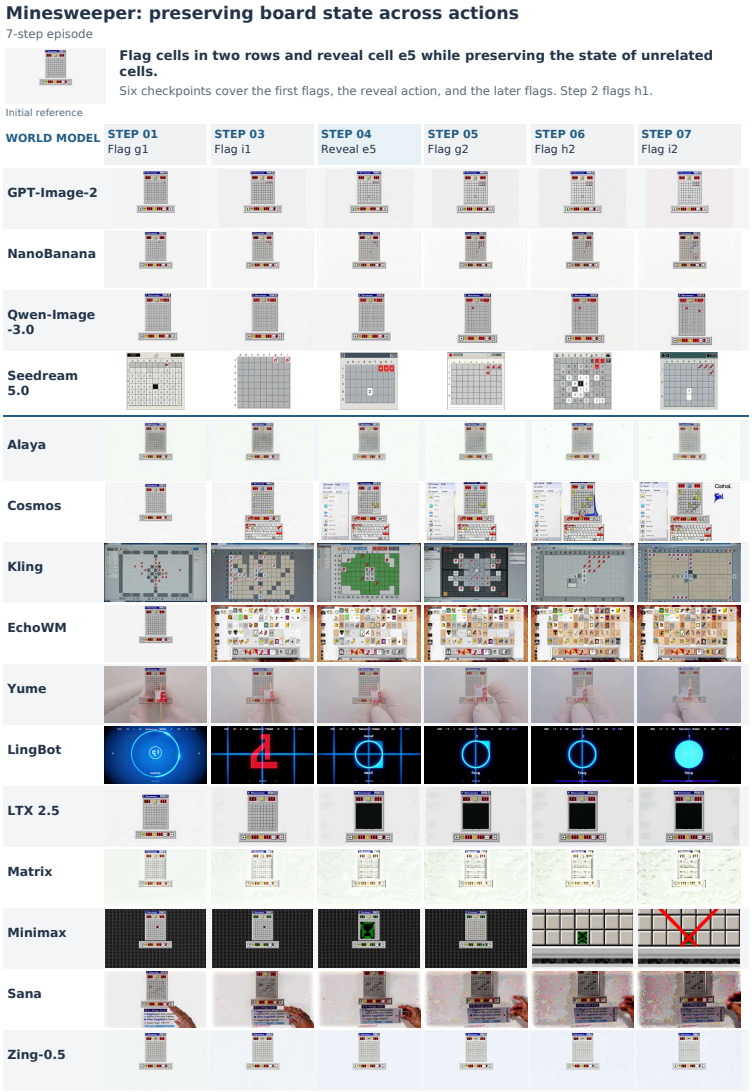


Figure 23: **Minesweeper: exact cell edits and persistent board state.** Steps 1, 3, 4, 5, 6, and 7 show first-row flags, revealing e5, and later second-row flags. Omitted step 2 flags h1 and remains in the history. Geometry preservation alone does not establish cell-level correctness.

Shopping. The six instructions enter and select the query Google Pixel 7, enable Buy on Google, choose 128 GB and Snow, and open the sorting menu. Each action must preserve previous selections and use a visible, actionable control. MiniMax ultimately replaces the page with a dialog. Plausible product imagery cannot verify query contents, filters, or menu state, although these small regions determine the meaning of subsequent actions.

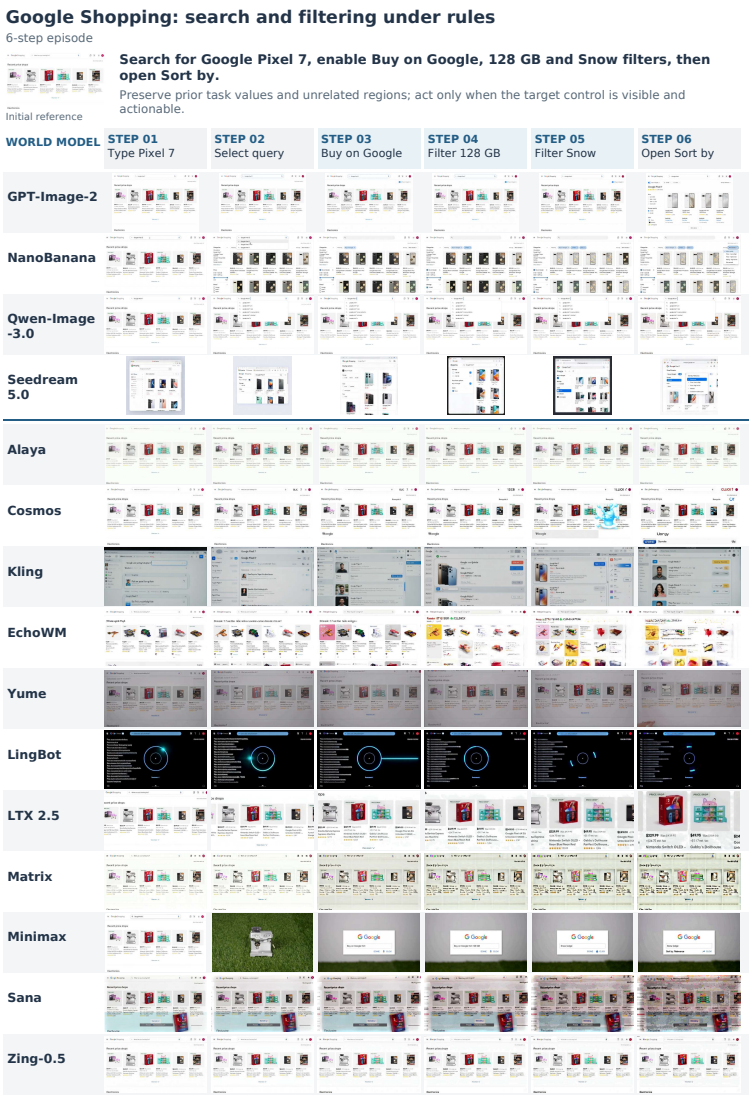


Figure 24: **Google Shopping: persistent query and filter state.** Steps 1–6 cover query entry and selection, the Buy on Google, 128 GB, and Snow filters, and opening the sorting menu. Every operation must preserve previous selections and unaffected interface state.

B.7 Implications for Evaluation

A valid WM response realizes the requested change while preserving unaffected entities, attributes, and earlier commitments. The cases distinguish three requirements: immediate action effects, persistence across the instructed history, and relevance to states encountered by a partner Agent. The controlled benchmark measures the first two; partner-specific utility requires matched interaction tests. Historical and modality differences remain material: static frames cannot establish timing, and selected examples cannot estimate error frequencies.

C Preliminary Study Protocol and Packing-Case Diagnostics

Study configuration. The preliminary study uses the same model pairing throughout: GPT-5.6 Sol selects actions, and Seedance 2.5 generates the resulting observations. The 150 cases are evenly distributed across Minecraft, Daily Life, Embodied Tasks, Web UI, Driving, Desktop Software, Industrial Systems, Natural Ecosystems, Scientific Experiments, and Social Interaction, with 15 cases per scenario. Overall is the unweighted arithmetic mean of the ten displayed scenario rates, yielding 26.6% World State Consistency, 9.2% Agent Corrective Response, and 0% Task Success after rounding.

Excluded runs. Initialization failures and technical failures are excluded from the reported 150 cases and from the task-success denominator; they are not counted as task failures. The reported 0% success rate therefore corresponds to zero successes among the 150 included cases, rather than among all attempted runs.

Interaction budget. Each case is run once, with a maximum of 20 interaction steps and no episode-level retries. Repeating or revising an action within the ongoing trajectory is distinct from restarting a case and consumes the same step budget. The navigation retries discussed below are such within-trajectory actions, not additional evaluation attempts.

Response versus repair. Agent Corrective Response records a targeted action following a visible discrepancy, not whether that action restores the required state. Successful repair would require separate evidence from subsequent observations that the same discrepancy has been resolved; task success additionally requires the complete goal and all constraints. Thus the 25.0% response rate for Social Interaction is a case-averaged behavioral rate, not a repair-success rate or the fraction of successful tasks. It cannot be converted into event counts without the underlying case-level denominators. Differences in discrepancy types and response opportunities also prevent interpreting a higher response rate as stronger recovery ability. The current aggregate results do not establish which responses produce repairs, and response behavior alone is insufficient as an optimization objective: ineffective responses may still receive credit. Evaluating repair outcomes and checking credited responses against the original discrepancies are needed to assess this risk.

Evidence and scope. Figure 2(b) uses one recorded 14-step GPT-5.6 Sol–Seedance 2.5 rollout, not an aggregate over the preliminary collection. The six displayed frames are from steps 1, 3, 6, 8, 12, and 14; the curves use all 14 recorded steps. The task specifies a burgundy passport with a gold band, a red medicine case with a white cross, and a white charger with an orange cable, all packed into the same blue backpack with a yellow star. A single automated retrospective review inspected eight frames per four-second segment at 2 fps, together with the initialization image and selected full-resolution tail frames. The labels have not been independently adjudicated, and the rubric was specified after inspecting the case. The offline failure assessment rests on the use of a different blue pouch for later packing and carrying, rather than on an assumption that occluded items are missing; it does not certify completion of the original runner's technical validation.

Checks and response windows. Each step is assessed for original-backpack identity, persistence of visibly acquired items, and task-conditioned action results. Of 42 candidate checks, 6 pass, 13 fail, 22 are unknown, and 1 is inapplicable. Unknown and inapplicable checks are excluded, giving an assessable coverage of $19/41$ and a World State Consistency score of $6/19$. For Agent Corrective Response, a qualifying response is targeted verification, reselection, replanning, or a retry within the next two actions after a visible discrepancy available to the MA. A persistent identity discrepancy is counted once; an event with only one subsequent action is excluded. Two of seven eligible events receive a response, both navigation retries. A retry does not establish successful repair.

Curve construction. Let F_t count failed state checks through step t, and let M_t count eligible discrepancies whose complete two-action windows have ended without a corrective response. The plotted retained scores for World State Consistency (S_t) and Agent Corrective Response (R_t) are

$$S_t = 1 - \frac{F_t}{19}, \qquad R_t = 1 - \frac{M_t}{7}.$$

Both denominators are fixed using the complete trajectory. Failures contribute one deduction each at steps 1, 2, 4, 6, 8, 9, and 11, and two each at steps 12–14. Missed-response deductions occur at steps 3, 8, 10, 13, and 14, when the corresponding windows close. Thus $S_{14} = 6/19$ and $R_{14} = 2/7$. Unknown checks incur no deductions and are not labeled successful. Each score remains unchanged unless its corresponding failure or missed response is newly counted. Straight segments connect integer-step values only as visual guides.

Interpretation and sensitivity. The initial value of 1 is assigned, not measured: initialization already deviates from the specification, and the wrong-colored passport is visible at step 1. Monotonicity follows from cumulative deductions and does not demonstrate monotonic degradation of either model's capability. Navigation checks use arrival inside the requested room by segment end. Treating doorway-only outcomes at steps 4 and 11 as unknown and excluding their response events changes the final rates to $6/17 = 0.353$ and $1/5 = 0.200$. The absence of responses to the four identity discrepancies and the final wrong-container failure are unchanged.