WORLDROAMBENCH: EVALUATING LONG-HORIZON STABILITY OF INTERACTIVE WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite rapid progress in interactive world models (IWMs), their *long-horizon stability* under sustained action inputs remains insufficiently evaluated. We introduce WorldRoamBench, an open-world benchmark addressing this gap across four complementary dimensions: (i) **Action**: pose-based discrete-action accuracy reducing sensitivity to cross-model semantic scale disparity and exposing frame-level errors hidden by trajectory scores; (ii) **Vision**: sliding-window drift metric capturing non-monotonic mid-sequence collapse missed by start-vs-end comparisons; (iii) **Physics**: evaluation of physical plausibility across mechanics, optics, and 3D consistency, gated by camera-motion and subject-tracking checks; (iv) **Memory**: a trajectory-aware protocol reducing confounding from action-following errors, evaluating scene memory via transition-localized 3D point-cloud reconstruction and subject memory via tracking-plus-VLM reasoning. The benchmark comprises 1000+ test cases across Nature, Urban, and Indoor scenes in first/third-person views with WASD 10–60 s continuous interaction. Evaluating 12 open/closed-source models reveals none reliably satisfies all dimensions; even the best long-horizon first-person overall score is only 67.17/100 (Genie 3). Advances on WorldRoamBench are steps toward IWMs that are stable, physically grounded, memory-faithful, and deployable in real-world applications.

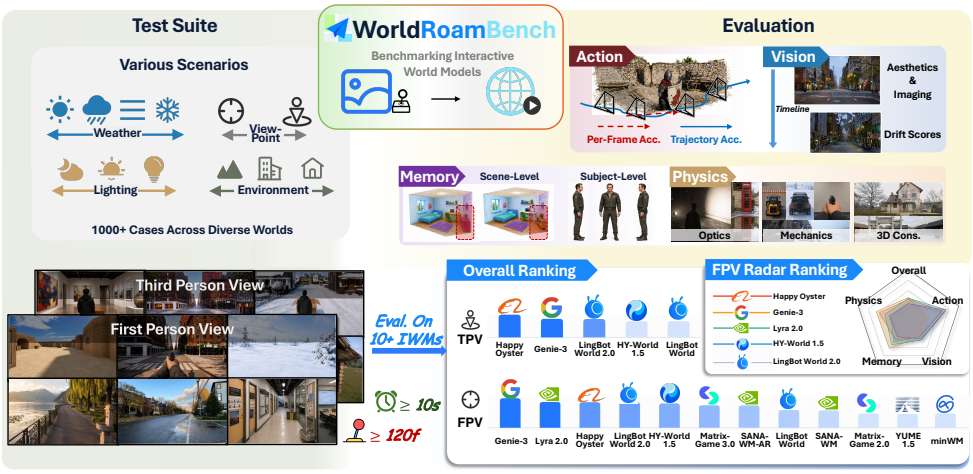


Figure 1: **WorldRoamBench overview.** Evaluation of open- and closed-source IWMs across action following, visual quality, interaction physics, and memory in diverse scenes and views.

1 INTRODUCTION

Imagine exploring an infinite, AI-generated world by pressing WASD keys on your keyboard: moving forward through forests, turning to gaze at distant mountains, and returning to find the same landscape as you left it. This vision of interactive world models (Bruce et al., 2024; He et al.,

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

Table 1: **Interactive world model benchmarks.** 3D Scene Matching compares separately reconstructed observation/revisit scenes after alignment. Turn/chunk durations vary by model. †Counts include general video generators. ‡TPV cases, no dedicated TPV leaderboard. *Cosmos example.

Benchmark	Eval. Metrics							Eval. Models		Leaderboards		Categories	
	Action Eval. Granularity	Discrete-action Acc.	Visual Drift	Interaction Physics	Revisit Memory	3D Scene Matching	Eval. Duration	Open-Source	Closed-Source	First-Person	Third-Person	Game World	Real World
WildBench	Segment	×	×	×	×	×	∼5–10s	5	0	✓	×	✓	×
iWorld-Bench†‡	Traj.	×	×	×	✓	×	∼5–10s	10+	0	✓	×	✓	✓
WorldOlympiad‡	Multi-chunk	×	×	✓	×	×	∼3.3 s/chunk*	8	0	✓	×	✓	✓
MIND	Traj.	×	×	×	✓	×	∼5–10s	2	0	✓	✓	✓	×
WBench†‡	Multi-turn	×	×	✓	✓	×	∼5 s/turn	10+	2	✓	×	✓	✓
WorldMark	Frame/segment	×	×	×	✓	×	20/40/60s	10	0	✓	✓	×	✓
WorldRoamBench (Ours)	**Traj. + per-frame**	✓	✓	✓	✓	✓	**10–60s**	**10**	**2**	✓	✓	✓	✓

2025; Sun et al., 2025) is rapidly approaching reality: models generate photorealistic environments at interactive frame rates, accepting discrete keyboard inputs to control camera movement.

Yet as these models proliferate, a central question is: *How well do they respond to user inputs over extended interactions?* Existing benchmarks (Table 1) have advanced interactive evaluation. Many protocols use short clips (Fang et al., 2026), action windows (Ye et al., 2026), action segments (Li et al., 2026), interaction turns (Ying et al., 2026), or generation chunks (Zhao et al., 2026b), often spanning only a few seconds, while WorldMark explicitly evaluates 20–60 s rollouts (Xu et al., 2026). However, evaluation over sustained interactions must track how local action errors, transient visual degradation, and physical inconsistencies emerge or accumulate, while reducing confounding between imperfect control and memory failure during revisits.

We argue the challenge for interactive world models is *long-horizon stability*: faithful, physically plausible, visually coherent, and memory-consistent generation as users continuously roam the world. Current models are most fragile in four aspects: 1) **Long-horizon action following.** Trajectory-based benchmarks (Ye et al., 2026; Fang et al., 2026; Ying et al., 2026) compare camera trajectories via RPE and ATE[1], but cross-model semantic scale disparity (different motion magnitudes and generation difficulty for identical actions) can complicate fair comparison, and trajectory alignment can mask per-frame failures (e.g., ignoring a keystroke then over-compensating). Figure 2(a) shows discrete-action errors hidden by trajectory scores, especially at action-switch boundaries. 2) **Long-horizon visual stability.** Existing benchmarks measure average visual quality (Imaging and Aesthetic) but rarely assess *temporal degradation*. Start-to-end comparisons can also miss mid-rollout degradation followed by partial recovery. Figure 2(b) shows visual collapse despite correct action following. 3) **Long-horizon interaction physics.** Unlike recent benchmarks (Ying et al., 2026; Zhao et al., 2026b), we explicitly gate physics evaluation on camera-motion response or subject-tracking continuity, covering mechanics, optics, and rigid-body 3D coherence under sustained action inputs. Figure 2(c) shows a model fol-



Figure 2: **Long-horizon failures.** (a) Discrete-action errors hidden by high trajectory scores; (b) visual degradation; (c) physical-constraint violations; (d) inconsistent scene revisits.

[1]RPE: Relative Pose Error; ATE: Absolute Trajectory Error.

lowing the intended direction yet passing through obstacles. 4) **Long-horizon memory.** MIND (Ye et al., 2026) and iWorld-Bench (Fang et al., 2026) use symmetric revisit tests comparing frames t and $T-t$, but this assumes perfectly executed actions; cumulative drift spatially offsets the "return" frame. Figure 2(d) shows inconsistent scene revisits even under symmetric return commands.

To fill these gaps, we introduce WorldRoamBench—**R**esponsive **O**pen-world **A**ction **M**odel **Bench**mark—which focuses on *long-horizon interactive stability* (Figure 1). WorldRoamBench evaluates models as they *roam* through generated open-domain worlds for ∼10–60s of continuous WASD/IJKL interaction, covering all four dimensions: action following, visual quality, interaction physics, and memory. Contributions are as follows:

Action. We propose a keystroke-level per-frame action metric that bypasses cross-model semantic scale disparity and reveals per-step failures hidden by trajectory.
Vision. We introduce a sliding-window temporal drift metric over imaging and aesthetic scores that captures non-monotonic mid-sequence collapse missed by per-frame or start-vs-end measures.
Physics. We design a validity-gated physics evaluation over *mechanics*, *optics*, and *3D consistency*, scoring only rollouts with usable motion evidence: pose-verified response in FPV and continuous subject tracking in TPV.
Memory. We present a trajectory-aware dual-track protocol, designed to *reduce* the sensitivity of memory scoring to action-following error, that evaluates *scene memory* via transition-localized 3D point-cloud reconstruction and *subject memory* via tracking-plus-Qwen3-VL-PLUS reasoning.
Leaderboard. We build the most comprehensive leaderboard to date, evaluating 10+ state-of-the-art open- and closed-source IWMs on 1000+ test cases, with dedicated first- and third-person leaderboards and short-, mid-, and long-horizon rankings for both views.

Throughout our evaluation, we identify three recurring weaknesses in current IWMs: **(i) limited long-horizon stability.** From short to long rollouts, visual quality declines, action accuracy drops for most models, and memory deteriorates most sharply. Sensitivity to action switching also varies substantially across models; **(ii) weak adherence to physical constraints.** Models can follow the commanded direction yet generate physically implausible motion, such as passing through solid obstacles. Even leading models achieve only moderate mechanics scores, while open-source models lag behind; and **(iii) unreliable long-term memory.** Most models struggle to retain previously observed scene content, often regenerating inconsistent geometry upon revisit. Reliable long-term memory is therefore still rare among IWMs.

2 WorldRoamBench Benchmark Design

2.1 Overview

WorldRoamBench evaluates interactive world models through four questions (pipeline in Figure 4, Appendix A): (1) **Action Following**: does the model respond correctly to each keystroke? (2) **Visual Quality**: does the generated world remain perceptually faithful and temporally stable? (3) **Interaction Physics**: does the world obey basic physical laws during interaction? (4) **Memory**: does the model preserve what it previously generated? Together, these dimensions provide a comprehensive assessment of interactive world models, with each capturing capabilities not fully reflected by the others. The test suite comprises 1,006 cases: 648 action, 211 memory, and 147 physics. Appendix B details image and action construction, horizon/switch design, and per-dimension statistics.

2.2 Evaluation Dimension 1: Action Following

We evaluate action following using estimated camera trajectories: *discrete action accuracy* measures frame-level command agreement, and *TrajScore* measures camera-path geometry (Figure 3, left).

2.2.1 Per-frame Discrete Action Accuracy

Let $\mathcal{A} = (a_1, \ldots, a_T)$ be the command sequence, with a_t controlling the transition from I_{t-1} to I_t. Commands comprise translations (WASD), camera rotations (IJKL), valid combinations, or `no_op`; W denotes `forward` and L denotes `look_right`. The full vocabulary is given in Appendix G.3.
Pose-based action decoding. From the aligned video, we estimate camera-to-world poses $\hat{\mathcal{T}} =$

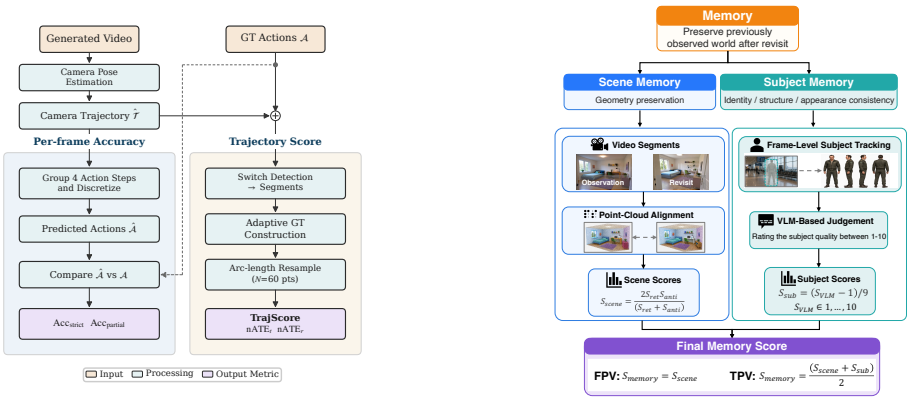


Figure 3: **Evaluation pipelines.** Left: Action following evaluation. Right: Memory evaluation.

$(\hat{P}_0, \ldots, \hat{P}_T)$ using MegaSaM (Li et al., 2025b) with a DA-v3 (Lin et al., 2025) depth prior (highest mean action-decoding ceiling among 6 pose pipelines, Appendix F). Inspired by WorldCompass (Wang et al., 2026b), we decode relative camera motion in non-overlapping K-frame windows and score each frame using its window's action label (window-size ablation and decoding rules in Appendices G.1 and G.3). To accommodate motion-scale differences across models and scenes, we calibrate multiple fixed translation-threshold candidates via cross-model statistics. All models and scenes share the same four candidates; Appendix G details calibration and selection.

Strict and partial accuracy. Strict accuracy requires exact command agreement, including every compound-action component; partial accuracy requires at least one shared atomic component. With $\mathcal{P}(a)$ denoting the atomic components of action a:

$$\text{Acc}_{\text{strict}} = \frac{1}{T}\sum_{t=1}^{T}[\![\hat{a}_t = a_t]\!], \qquad \text{Acc}_{\text{partial}} = \frac{1}{T}\sum_{t=1}^{T}[\![\mathcal{P}(\hat{a}_t) \cap \mathcal{P}(a_t) \neq \emptyset]\!]. \tag{1}$$

where $[\![\cdot]\!]$ is the indicator function. Predicting `forward+look_right` for `look_right` yields a partial match, but extra translation prevents strict matching; the partial–strict gap measures how often predictions are only partially correct, and separate move/look accuracies assess translation and rotation agreement.

2.2.2 Trajectory-level Metrics: TrajScore

Discrete labels miss geometric deviations (a path labeled `forward` may still drift laterally). Inspired by WBench (Ying et al., 2026), TrajScore evaluates camera-path geometry via adaptive references and arc-length alignment, complementing discrete metrics.

Adaptive reference construction. For each constant-action segment, we construct an *adaptive reference trajectory*: subject to minimum-motion safeguards, it retains the model's motion magnitude while imposing the commanded direction and shape. This separates *how far the model moves* from *whether it follows the commanded geometry*, without fixing movement distances or turning angles across models. For example, a `forward` command defines a straight path matching the predicted displacement magnitude. Appendix H.1 details rotation and compound-action constructions.

Arc-length alignment. To separate geometry from traversal speed, we pair predicted and reference samples within each constant-action segment at equal fractions of their respective arc lengths; ATE_t and ATE_r are the mean Euclidean position and angular orientation errors over these pairs (Appendix H details construction, safeguards, and sampling).

Normalized geometric error. We normalize translation and rotation errors by predicted path length L_{path} and accumulated rotation Φ_{total}, respectively, with denominator floors $\ell_{\min}$ and $\phi_{\min}$ stabilizing low-motion cases:

$$\text{nATE}_t = \text{clamp}\Big(\frac{\text{ATE}_t}{\max(L_{\text{path}}, \ell_{\min})}, 0, 1\Big), \qquad \text{nATE}_r = \text{clamp}\Big(\frac{\text{ATE}_r}{\max(\Phi_{\text{total}}, \phi_{\min})}, 0, 1\Big). \tag{2}$$

Higher values of $\text{TrajScore} = 1 - (\text{nATE}_t + \text{nATE}_r)/2$ indicate closer geometric alignment.

2.2.3 ACTION SCORE

We aggregate the complementary scores by view:

$$S_{\text{action}}^{\text{FPV}} = \frac{\text{Acc}_{\text{strict}} + \text{Acc}_{\text{partial}} + \text{TrajScore}}{3}, \qquad S_{\text{action}}^{\text{TPV}} = \frac{\text{Acc}_{\text{partial}} + \text{TrajScore}}{2}. \tag{3}$$

In TPV, subject-centered camera orbits couple rotation and translation, potentially yielding compound predictions that fail strict matching; partial matching still credits the correctly decoded look action. We also report normalized translation and rotation errors as diagnostics.

2.3 EVALUATION DIMENSION 2: VISUAL QUALITY

We evaluate visual quality along two axes: *absolute quality* and *temporal drift*.

Aesthetic and Imaging Quality. Aesthetic quality uses a CLIP-based predictor trained on LAION-Aesthetics (Schuhmann et al., 2022), while imaging quality uses SPAQ-pretrained MUSIQ (Ke et al., 2021). Their outputs, divided by 10 and 100, respectively, yield normalized frame scores s_t^{aes} and s_t^{img} (Appendix I). We average all frame scores:

$$S_{\text{aes}} = \frac{1}{T}\sum_{t=1}^{T} s_t^{\text{aes}}, \qquad S_{\text{img}} = \frac{1}{T}\sum_{t=1}^{T} s_t^{\text{img}}. \tag{4}$$

Temporal Drift Metrics. Autoregressive rollouts can degrade progressively: crisp early frames give way to artifacts, color shifts, or detail loss. Degradation is not always monotonic: we frequently observe *mid-rollout collapse followed by partial recovery*, so contrasting a clip's start and end can underestimate visual instability. Inspired by Helios (Yuan et al., 2026), we quantify drift over the rollout using a sliding-window formulation adapted to capture the largest quality excursion.

We average aesthetic or imaging scores within $W = 3\,\text{s}$ windows with stride $R = 1\,\text{s}$, obtaining $\bar{s}_i$ for window i. With $\bar{s}_{\text{best}} = \max_i \bar{s}_i$ and $\bar{s}_{\text{worst}} = \min_i \bar{s}_i$, drift is the relative best-to-worst drop:

$$D = \frac{\bar{s}_{\text{best}} - \bar{s}_{\text{worst}}}{|\bar{s}_{\text{best}}|}. \tag{5}$$

Normalizing by the best-window score expresses drift relative to each rollout's quality level, facilitating comparison across models. Appendix I details window construction and parameter selection.

Fixed-duration windows in *seconds* preserve a common temporal scale across duration buckets, avoiding the additional dilution of brief collapses caused by windows that grow with rollout duration. *Overlapping windows* ($R < W$) reduce dilution when collapses straddle boundaries between disjoint windows. Because D is a range statistic, more windows in longer rollouts increase its expected spread; compare drift within, not across, duration buckets. This formulation yields lower-is-better *Drift*$_{\textit{Aesthetic}}$ ($D_{\text{aes}}\downarrow$) and *Drift*$_{\textit{Imaging}}$ ($D_{\text{img}}\downarrow$) from aesthetic and imaging scores.

2.3.1 VISUAL SCORE

Absolute quality reflects average perceptual quality, while drift captures within-rollout quality variation. For non-negative quality scores with $\bar{s}_{\text{best}} > 0$, drift satisfies $D \in [0, 1]$. We convert drift to stability scores $1 - D \in [0, 1]$ and average them with the two absolute-quality scores:

$$S_{\text{visual}} = \frac{S_{\text{aes}} + (1 - D_{\text{aes}}) + S_{\text{img}} + (1 - D_{\text{img}})}{4}. \tag{6}$$

2.4 EVALUATION DIMENSION 3: INTERACTION PHYSICS

WorldRoamBench evaluates the physical plausibility of user-driven interactions across three domains: *mechanics*, *optics*, and *3D consistency*. All VLM-based judgments use Qwen3-VL-PLUS. We exclude rollouts without usable motion evidence. FPV requires pose-verified camera translation; TPV requires continuous subject tracking, with a Qwen3-VL-PLUS direction check when masks are unavailable (Appendix J.1).

2.4.1 PHYSICS PROTOCOLS

Mechanics tests collision, clipping, deformation, and gravity, including falls, sinking, and ground contact on slopes and steps. FPV uses reconstructed camera trajectories to assess environmental effects on commanded motion. TPV uses an agent that locates critical events, inspects dense cropped samples, and re-checks the collected evidence before judging. Scores are binary unless specified otherwise (Appendix J.2.1).

Optics tests reflection and occlusion. Reflection combines a coarse plausibility check with per-frame scoring over valid frames, normalized to $[0, 1]$. Occlusion is assessed through shadow presence, direction, and shape relative to scene lighting.

3D consistency tests rigid spatial structure under controlled camera motion by comparing sampled frames with source-engine reference views within landmark-defined regions (Appendix J.2).

2.4.2 PHYSICS SCORE

Each valid case receives a score $s_c \in [0, 1]$. We average scores within each domain and weight available domains equally:

$$S_d = \frac{1}{|\mathcal{C}_d|} \sum_{c \in \mathcal{C}_d} s_c, \qquad S_{\text{physics}} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} S_d, \tag{7}$$

where $\mathcal{C}_d$ contains valid cases in domain d and $\mathcal{D} \subseteq \{\text{mech}, \text{opt}, \text{3d}\}$ contains domains with valid cases. Scores depend on model-specific valid subsets; TPV Control Rate reports complementary coverage but does not equalize these subsets.

2.5 EVALUATION DIMENSION 4: MEMORY

WorldRoamBench evaluates memory as *revisit stability under imperfect control*: a model should preserve the previously observed world when the user returns to a region, even if the generated trajectory does not exactly match the ground-truth path. As summarized in Figure 3 (right), we separate memory into two complementary tracks: *scene memory*, whether observed scene geometry is retained upon revisit, and *subject memory*, whether the third-person protagonist preserves identity, structure, and appearance.

Scene Memory. Symmetric frame-pair comparisons can be confounded when action execution changes the revisit viewpoint. We evaluate geometric scene memory by splitting the rollout at the estimated observation–revisit transition, reconstructing both segments into point clouds, and aligning the clouds before scoring. The score measures agreement of reconstructed geometry and remains sensitive to visible coverage and reconstruction quality.

A reconstructed point cloud is a *discrete* sample whose self-agreement is bounded by its sampling density, so we score the cross-segment discrepancy *relative to* this sampling floor—a signal-to-noise ratio rather than an absolute distance. Let $\mathcal{P}_{\text{obs}}$ and $\mathcal{P}_{\text{rev}}$ be the aligned observation and revisit clouds, $\sigma(\cdot)$ the sampling floor (the mean nearest-neighbor distance between two disjoint subsets of a cloud), and $\delta(\mathcal{X} \to \mathcal{Y})$ the directed nearest-neighbor discrepancy from $\mathcal{X}$ to $\mathcal{Y}$ (both defined in Appendix K.1). Retention and anti-hallucination are the clamped ratios

$$S_{\text{ret}} = \min\left(1,\ \frac{\sigma(\mathcal{P}_{\text{obs}})}{\delta(\mathcal{P}_{\text{obs}} \to \mathcal{P}_{\text{rev}})}\right), \quad S_{\text{anti}} = \min\left(1,\ \frac{\sigma(\mathcal{P}_{\text{rev}})}{\delta(\mathcal{P}_{\text{rev}} \to \mathcal{P}_{\text{obs}})}\right). \tag{8}$$

Analogous to *recall* and *precision*, S_{ret} and S_{anti} assess retention of observed structure and avoidance of unsupported structure, respectively. Their per-case harmonic mean $S_{\text{scene}} = 2S_{\text{ret}}S_{\text{anti}}/(S_{\text{ret}}+S_{\text{anti}})$ is averaged across cases. A clamped score of 1 indicates discrepancy at or below the estimated sampling floor, not identical point clouds. Appendices K.1 and M give full definitions and scoring details and explain why frame-pair alternatives remain unreliable.

Subject Memory. Subject memory captures a complementary TPV failure mode: the protagonist may follow the scene trajectory yet drift in identity, structure, or appearance. As a small articulated region unsuited to point clouds, it uses a tracking-and-VLM protocol: a controllability gate discards cases with insufficient subject-tracking continuity, then Qwen3-VL-PLUS holistically rates the tracked subject to yield a subject score $S_{\text{sub}} \in [0, 1]$. Appendix K.2 details the gate, tracking, scoring rubric, failure taxonomy, normalization, and prompt.

2.5.1 MEMORY SCORE

The memory score depends on the view: FPV has no visible protagonist, whereas TPV additionally requires a stable protagonist,

$$S^{\text{FPV}}_{\text{memory}} = S_{\text{scene}}, \qquad S^{\text{TPV}}_{\text{memory}} = \tfrac{1}{2}\left(S_{\text{scene}} + S_{\text{sub}}\right). \tag{9}$$

Both TPV terms are averaged only over cases passing the controllability gate, and we report the overall TPV control rate separately (Table 5).

2.6 SCORING AND LEADERBOARD

The overall WorldRoamBench score is a weighted composite:

$$S_{\text{ROAM}} = w_1 \cdot S_{\text{action}} + w_2 \cdot S_{\text{visual}} + w_3 \cdot S_{\text{physics}} + w_4 \cdot S_{\text{memory}}. \tag{10}$$

Results are reported with equal default weights $w_1{=}w_2{=}w_3{=}w_4{=}0.25$, together with per-dimension rankings. For TPV, the leaderboard overall score accounts for whether a model can maintain a continuously trackable third-person subject: $S^{\text{TPV}}_{\text{ROAM}} = R_{\text{ctrl}} \cdot S_{\text{ROAM}}$, where the Control Rate, R_{ctrl}, is the fraction of evaluated TPV cases passing the controllability gate (Appendix J.1).

3 EXPERIMENTS

3.1 EVALUATED MODELS

We evaluate representative frontier interactive world models under a unified image-and-action interface: 2 closed-source products, Genie 3 and Happy Oyster, and 10 publicly released or open-source models covering the dominant design choices in video world models. These models span four control-interface families—native keyboard input, continuous camera-pose conditioning, hybrid action-plus-pose conditioning, and caption- or instruction-driven control—so to ensure comparability every WorldRoamBench test case is converted into the closest supported control format for each model while preserving the same first-frame image, action order, and evaluation pipeline. Appendix E provides the model summary (Table 6), inference speed and throughput, per-model interface families, evaluation adapters, and a detailed efficiency analysis.

3.2 MAIN RESULTS

Overall ranking. Table 2 reports first-person and third-person results. Genie 3 leads first-person view (70.15), narrowly ahead of Lyra 2.0 (69.61) and Happy Oyster (68.53), while Happy Oyster leads third-person view (70.66) after control-rate weighting. The rankings show that **no model dominates all dimensions**: closed-source models are generally stronger in physics and memory, whereas frontier open-source models are stronger in action following and visual quality.

Action following. Open-source models achieve the strongest first-person action scores—Matrix-Game 3.0 (91.34), Lyra 2.0 (90.82), LingBot-World (90.78), and LingBot-World 2.0 (90.74) are within 1 point of one another—and this advantage largely carries over to third-person evaluation among models supporting TPV. Notably, the closed-source models do *not* lead: Genie 3 reaches only 76.31, with per-frame strict accuracy (62.04) in the bottom four, and Happy Oyster 88.22. The strongest overall models therefore need not lead in action following.

Visual quality. Visual quality is comparatively clustered: 10 of 12 models fall within 8.3 points of one another, from SANA-WM-AR (74.40) and SANA-WM (71.42) down to Genie 3 (66.21) and HY-World 1.5 (66.06), with only Matrix-Game 2.0 (60.83) and minWM (58.21) clearly behind. Both SANA-WM variants lead first-person visual quality but rank 7th and 9th overall.

Physics. In contrast, physics exposes a far larger gap: Genie 3 (66.67) and Happy Oyster (61.40) lead the best open-source model, Lyra 2.0 (51.52), by 15.15 and 9.88 points, and the weakest models score in single digits. Mechanics shows a pronounced closed/open-source gap: Genie 3 scores 61.50 and Happy Oyster 45.50, while every open-source model stays below 8. Despite strong camera-command following, open-source models remain weak on these mechanics tests. In third-person view, Happy Oyster is strongest in physics (68.90), narrowly ahead of Genie 3 (67.15).

Memory. Beyond controllability, memory reveals a persistence gap: Genie 3 leads first-person memory (71.40), followed by Lyra 2.0 (66.49) and SANA-WM-AR (64.48), while LingBot-World

Table 2: **Unified quantitative results on WorldRoamBench.** All values $\times 100$. Metrics are rows and models are columns, organized into an FPV table, a TPV table, and an Overall-score ranking leaderboard. Gray columns denote closed-source models. **Bold**/underline = best/2nd-best within each metric. ↑ = higher is better; ↓ = lower is better.

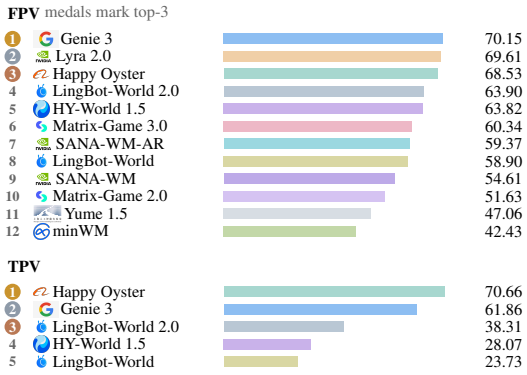
(a) First-person View (FPV)

Metric	Genie 3	Happy Oyster	Lyra 2.0	LingBot-World 2.0	HY-World 1.5	Matrix-Game 3.0	SANA-WM-AR	LingBot-World	SANA-WM	Matrix-Game 2.0	Yume 1.5	minWM
Overall Score ↑	**70.15**	68.53	69.61	63.90	63.82	60.34	59.37	58.90	54.61	51.63	47.06	42.43
Action Score ↑	76.31	88.22	90.82	90.74	87.38	**91.34**	75.73	90.78	70.42	89.67	71.77	60.79
Strict Accuracy ↑	62.04	85.90	90.85	88.95	83.21	88.77	71.56	**91.17**	61.91	88.11	58.45	46.15
Partial Accuracy ↑	78.06	89.96	93.82	**96.30**	88.49	94.45	76.66	94.38	69.77	92.79	76.41	53.18
Trajectory ↑	88.83	88.80	87.78	86.98	90.44	**90.79**	78.96	86.80	79.59	88.10	80.45	83.04
$nATE_t$ ↓	11.48	11.02	9.61	**6.85**	8.63	9.24	23.23	9.92	22.60	11.57	18.91	13.63
$nATE_r$ ↓	10.85	11.37	14.83	19.18	10.49	**9.17**	18.86	16.47	18.22	12.23	20.18	20.28
Visual Score ↑	66.21	66.49	69.61	68.55	66.06	68.70	**74.40**	69.28	71.42	60.83	69.07	58.21
Aesthetic Score ↑	53.41	55.93	59.62	59.93	58.22	56.50	60.08	**60.81**	57.17	53.69	57.97	48.93
Aesthetic Drift ↓	**17.22**	19.15	22.63	24.89	27.06	26.25	18.36	21.20	21.83	32.43	23.91	43.50
Imaging Quality ↑	54.53	49.04	60.82	61.57	55.11	64.77	**69.08**	59.96	67.80	48.52	64.64	51.59
Imaging Drift ↓	25.87	19.87	19.36	22.43	22.04	20.24	**13.20**	22.47	17.48	26.46	22.42	24.20
Physics Score ↑	**66.67**	61.40	51.52	43.52	40.43	25.56	22.88	31.08	17.38	12.21	14.33	7.12
Mechanics ↑	**61.50**	45.50	4.80	3.80	0.00	0.00	6.40	3.80	3.20	0.00	7.40	0.00
Optics ↑	60.60	67.50	**69.10**	56.60	48.80	17.20	41.20	51.20	44.70	9.70	10.30	4.40
3D Consistency ↑	77.90	71.21	**80.65**	70.17	72.50	59.49	21.05	38.25	4.23	26.92	25.30	16.97
Memory Score ↑	**71.40**	58.00	66.49	52.80	61.40	55.76	64.48	44.46	59.23	43.81	33.05	43.60
Retention ↑	**73.21**	55.11	62.25	53.12	58.75	59.24	62.79	39.09	60.09	44.13	30.67	43.82
Anti-Hallucination ↑	76.65	71.29	**79.45**	60.01	70.56	59.78	77.44	62.76	72.51	53.69	48.72	61.09

(b) Third-person View (TPV)

Metric	Happy Oyster	Genie 3	LingBot-World 2.0	HY-World 1.5	LingBot-World
Overall Score ↑	**70.66**	61.86	38.31	28.07	23.73
Control Rate ↑	**96.43**	83.82	53.78	39.71	35.50
Action Score ↑	84.81	84.76	88.61	**91.46**	82.29
Partial Accuracy ↑	81.16	79.83	90.29	**91.88**	81.28
Trajectory ↑	88.45	89.68	86.92	**91.03**	83.30
$nATE_t$ ↓	13.39	10.78	12.57	**9.14**	19.55
$nATE_r$ ↓	9.71	9.86	13.60	**8.80**	13.85
Visual Score ↑	69.84	72.03	**74.46**	71.52	73.47
Aesthetic Score ↑	61.18	60.65	**64.62**	63.25	62.25
Aesthetic Drift ↓	14.45	**12.21**	15.38	17.83	14.49
Imaging Quality ↑	49.10	57.60	**62.96**	56.29	62.37
Imaging Drift ↓	16.47	17.92	**14.35**	15.65	16.27
Physics Score ↑	**68.90**	67.15	58.00	51.35	53.85
Mechanics ↑	52.10	**70.70**	41.00	41.20	35.00
Optics ↑	**85.70**	63.60	75.00	61.50	72.70
Memory Score ↑	69.55	**71.27**	63.88	68.41	57.75
Retention ↑	59.76	66.56	59.46	**72.86**	49.52
Subject Memory ↑	**81.09**	77.99	73.88	67.52	62.22
Anti-Hallucination ↑	69.14	72.19	58.17	**74.90**	70.02

(c) Overall-score Ranking



reaches only 44.46 despite a 90.78 action score. In third-person view Genie 3 leads overall TPV memory (71.27) and Happy Oyster leads subject memory (81.09). Memory is also the dimension that degrades most sharply with rollout horizon, which we analyze in Appendix C.1.4.

4 Discussion

4.1 Key Findings

A high trajectory score does not imply high per-frame correctness. Models with trajectory scores above 85 can still fall below 65% strict per-frame action accuracy because short-term errors can cancel geometrically, motivating keystroke-level evaluation alongside trajectory scoring. **Low visual quality does not necessarily imply poor controllability.** Matrix-Game 2.0 ranks near the bottom in first-person visual quality but 5th in action following, whereas the two SANA-WM variants lead visual quality but rank 9th and 11th in action following, showing that perceptual qual-

Table 3: **First-person scores across horizons.** For Overall, Vision, Action, and Memory: Short/Mid/Long scores ($\times 100$, ↑) and the Short-to-Long decay rate ρ (↓, Section 4.2). Rows ordered by Short-horizon Overall; gray = closed-source; **bold**/underline = best/2nd-best per metric and horizon. In the Mean row, decay rates are averaged across models.

Model	Overall				Vision				Action				Memory			
	Short ↑	Mid ↑	Long ↑	ρ↓	Short ↑	Mid ↑	Long ↑	ρ↓	Short ↑	Mid ↑	Long ↑	ρ↓	Short ↑	Mid ↑	Long ↑	ρ↓
Genie 3	**73.09**	69.71	**67.17**	8.1%	68.04	66.13	64.46	5.3%	77.31	74.41	77.22	0.1%	**80.33**	**71.64**	60.32	24.9%
Lyra 2.0	72.80	**70.63**	64.82	11.0%	73.80	70.47	64.58	12.5%	92.16	91.02	89.26	3.1%	73.72	69.52	53.91	26.9%
Happy Oyster	71.60	69.99	63.28	11.6%	69.21	66.90	63.36	8.5%	88.71	88.94	87.02	1.9%	67.06	62.72	41.32	38.4%
HY-World 1.5	71.00	66.40	53.28	25.0%	72.64	68.32	57.22	21.2%	**93.69**	91.08	77.38	17.4%	77.23	65.75	38.10	50.7%
LingBot-World 2.0	69.52	63.18	58.30	16.1%	72.39	69.16	64.09	11.5%	90.42	91.39	**90.43**	0.0%	71.77	48.65	35.15	51.0%
Matrix-Game 3.0	65.91	61.77	52.17	20.8%	72.79	69.93	63.36	13.0%	92.31	**91.85**	89.85	2.7%	72.96	59.75	29.91	59.0%
LingBot-World	63.52	59.07	53.52	15.7%	72.95	69.87	65.01	10.9%	91.55	90.83	89.98	1.7%	58.51	44.49	28.02	52.1%
SANA-WM-AR	60.36	61.54	55.92	7.4%	**77.40**	**74.48**	**71.33**	7.8%	70.16	84.74	72.29	-3.0%	70.99	64.05	57.18	19.5%
SANA-WM	57.61	54.13	52.25	9.3%	74.39	73.40	66.47	10.6%	67.86	79.27	64.14	5.5%	70.83	46.48	**61.03**	13.8%
Matrix-Game 2.0	56.30	51.75	46.24	17.9%	67.31	60.91	54.27	19.4%	90.32	89.85	88.83	1.6%	55.37	44.04	29.67	46.4%
Yume 1.5	51.14	46.39	43.22	15.5%	74.53	68.98	63.70	14.5%	71.53	69.97	73.81	-3.2%	44.17	32.28	21.02	52.4%
minWM	45.12	43.13	39.15	13.2%	66.84	58.92	48.86	26.9%	64.96	62.11	55.31	14.9%	41.57	44.38	45.32	-9.0%
Mean	63.16	59.81	54.11	14.3%	71.86	68.12	62.23	13.5%	82.58	83.79	79.63	3.6%	65.38	54.48	41.75	35.5%

ity and controllability are distinct capabilities. **Stricter physics adherence may compromise GT action following.** Physically desirable behavior, such as stopping at a wall, can deviate from an unconstrained ground-truth trajectory, motivating separate action-following and interaction-physics metrics. **Memory evaluation is confounded by action-following imprecision.** Imperfect action execution can prevent exact revisits, causing frame-pair metrics to conflate spatial misalignment with memory degradation. Our action-aware transition localization and scene-level 3D reconstruction reduce this confounding and are more robust than GT-indexed frame matching under imperfect control (Appendix M).

4.2 Performance Across Horizons

We compare first-person performance over the Short-, Mid-, and Long-horizon splits defined in Appendix B.3. For each metric $d \in \{\text{Overall}, \text{Vision}, \text{Action}, \text{Memory}\}$, we define its Short-to-Long decay rate as $\rho_d = 1 - S_d^{\text{Long}}/S_d^{\text{Short}}$. A larger ρ_d indicates a greater Short-to-Long decrease and $\rho_d < 0$ an increase. The full-suite Physics score enters every split unchanged, so these rates characterize the horizon sensitivity of Action, Vision, Memory, and the Overall composite, not of Physics itself.

Table 3 shows that horizon sensitivity varies sharply across dimensions and reshuffles the leaderboard. Mean Overall scores fall from 63.16 to 54.11 from Short to Long. Mean per-model decay is largest for Memory (35.5%), followed by Overall (14.3%) and Vision (13.5%), and smallest for Action (3.6%). The top-4 Overall spread grows $4.2\times$ (2.09→8.87 points), while the same 3 models (Genie 3, Lyra 2.0, and Happy Oyster) occupy the top 3 at every horizon. From Short to Long, 11 of 12 models change rank in Memory, compared with 6 in Overall. Decay-rate plots and per-dimension analysis are in Appendix C; third-person horizon results are in Appendix C.2.

Yume 1.5 has slightly higher partial action accuracy in the Long than in the Short split, but its partial accuracy is 20.2 pp lower with 2 switches than with 0 switches (Appendix C.3).

5 Conclusion

WorldRoamBench benchmarks long-horizon interactive world models under continuous WASD/IJKL control, jointly evaluating action following, visual stability, interaction physics, and memory to expose failures largely hidden by short clips and trajectory- or frame-pair-only metrics. Across 12 open- and closed-source models, no model dominates all dimensions, and stratifying action, visual, and memory scores by rollout horizon, with Physics fixed at its full-suite value, shows that short-horizon rankings do not fully characterize long-horizon performance; WorldRoamBench thus provides a diagnostic foundation for world models that are simultaneously action-responsive, physically grounded, visually stable, and memory-faithful. Limitations and future directions are discussed in Appendix P.

AI USE STATEMENT

AI assistants were used for language polishing, grammar and usage checks, and programming assistance. All AI-generated suggestions incorporated into this work were reviewed, revised as needed, and validated by the authors. The authors are responsible for the final manuscript, methods, and reported results.

REPRODUCIBILITY STATEMENT

To support reproducibility, we specify the evaluation metrics and aggregation rules in Section 2. The appendices document test-suite construction (Appendix B), model interfaces and evaluation adapters (Appendix E), and metric algorithms, parameter settings, and VLM prompts (Appendices F–L). Human calibration and agreement studies are described in Appendix O.2. Reproducing closed-source rollouts depends on access to the evaluated product versions. The benchmark data and evaluation code will be publicly released soon.

ETHICS STATEMENT

WorldRoamBench uses Internet-sourced and benchmark images, including generatively edited variants, and expert ratings of generated videos (Appendices B and O.2). Expert raters participated voluntarily and provided informed consent. We will redistribute source images and derived assets only as permitted by the applicable licenses and terms. The benchmark is intended for comparative research; its scores do not certify safety or physical reliability in real-world deployment.

REFERENCES

Alibaba Cloud. Wan 2.7: Image-to-video API. https://www.alibabacloud.com/help/en/model-studio/image-to-video-general-api-reference, 2026.

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

Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, and Yu Yamaguchi. ACT-Bench: Towards action controllable world models for autonomous driving. *arXiv preprint arXiv:2412.05337*, 2024.

Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (VPT): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Wendong Bu, Yang Wu, Qifan Yu, Minghe Gao, Bingchen Miao, Zhenkui Zhang, Kaihang Pan, Yunfei Li, Mengze Li, Wei Ji, Juncheng Li, Siliang Tang, and Yueting Zhuang. What limits virtual agent application? OmniBench: A scalable multi-dimensional benchmark for essential virtual agent capabilities. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 5725–5748, 2025. URL https://proceedings.mlr.press/v267/bu25b.html.

Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 27713–27724, 2025.

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

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

Jess Gallegos and Thomas Iljic. Introducing Veo 3.1 and advanced capabilities in Flow. https://blog.google/innovation-and-ai/products/veo-updates-flow/, 2025.

Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, et al. Do vision-language models have internal world models? towards an atomic evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26170–26195, 2025.

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.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. MineWorld: a real-time and open-source interactive world model on Minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

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

Yue Hu, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Guanghui Ren, and Maoqing Yao. EWMBench: Evaluating scene, motion, and semantic quality in embodied world models. In *British Machine Vision Conference*, 2025. URL https://bmvc2025.bmva.org/proceedings/736/.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. ViPE: Video pose engine for 3D geometric perception. *arXiv preprint arXiv:2508.10934*, 2025a.

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

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b. doi: 10.1109/TPAMI.2025.3633890.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 28991–29017, 2025. URL https://proceedings.mlr.press/v267/kang25g.html.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Kuaishou Technology. Kling AI launches 3.0 model, ushering in an era where everyone can be a director. https://ir.kuaishou.com/node/11216/pdf, 2026.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Zhen Li, Zian Meng, Shuwei Shi, Wenshuo Peng, Yuwei Wu, Bo Zheng, Chuanhao Li, and Kaipeng Zhang. WildWorld: A large-scale dataset for dynamic world modeling with actions and explicit state toward generative ARPG. *arXiv preprint arXiv:2603.23497*, 2026.

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

Ao Liang, Lingdong Kong, Tianyi Yan, Hongsi Liu, Yu Yang, Ziqi Huang, Wei Yin, Jialong Zuo, Yixuan Hu, Dekai Zhu, et al. WorldLens: Full-spectrum evaluations of driving world models in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36385–36399, 2026.

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

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.

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

OpenAI. Sora 2 system card. https://openai.com/index/sora-2-system-card/, 2025.

Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/, 2025.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10106–10116. IEEE, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. arXiv preprint arXiv:2601.20540, 2026. URL https://arxiv.org/abs/2601.20540.

Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *2009 IEEE international conference on robotics and automation*, pp. 3212–3217. IEEE, 2009.

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

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. *AI Open*, 7:208–226, 2026.

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

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

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing video generation for world complexity. arXiv preprint arXiv:2604.14148, 2026. URL https://arxiv.org/abs/2604.14148.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Yunhao Ba, Alex Wong, Celso M de Melo, and Achuta Kadambi. WorldBench: Benchmarking physical understanding of world models by isolating physics concepts. arXiv preprint arXiv:2601.21282, 2026. URL https://arxiv.org/abs/2601.21282.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems*, volume 38, pp. 35928–35959, 2025.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2507.13347.

Zehan Wang, Tengfei Wang, Haiyu Zhang, Xuhui Zuo, Junta Wu, Haoyuan Wang, Wenqiang Sun, Zhenwei Wang, Chenjie Cao, Hengshuang Zhao, et al. WorldCompass: Reinforcement learning for long-horizon world models. *arXiv preprint arXiv:2602.09022*, 2026b.

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

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

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point Transformer V3: Simpler, faster, stronger. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4840–4851. IEEE, 2024.

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

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.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

Min Zhao, Hongzhou Zhu, Bokai Yan, Zihan Zhou, Yimin Chen, Honglie Wang, Wenqiang Sun, Zhengwei Fang, Zizheng Xun, Zihao Li, Kaiwen Zheng, Guande He, Xiao Yang, Chongxuan Li, Fan Bao, and Jun Zhu. minWM: A full-stack open-source framework for real-time interactive video world models. *arXiv preprint arXiv:2605.30263*, 2026a.

Yuke Zhao, Wangbo Zhao, Weijie Wang, Zeyu Zhang, Dakai An, Akide Liu, Yinghao Yu, Jiasheng Tang, Fan Wang, Wei Wang, et al. WorldOlympiad: Can your world model survive a triathlon? *arXiv preprint arXiv:2606.11129*, 2026b.

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

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 Evaluation Pipeline Overview

Figure 4 provides a full overview of the WorldRoamBench evaluation pipeline across its four evaluation dimensions.

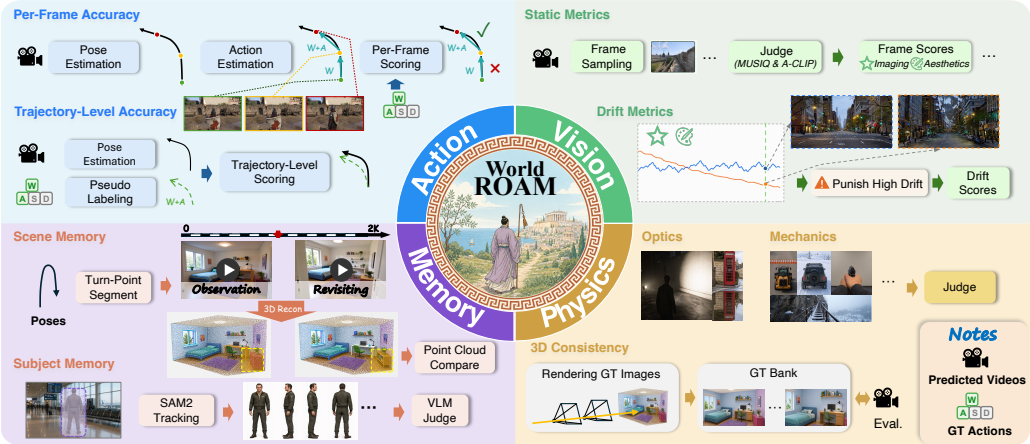


Figure 4: **WorldRoamBench evaluation pipeline.** Given videos generated from shared initial frames and WASD/IJKL action programs, WorldRoamBench evaluates four complementary dimensions: action following through pose-aligned frame scoring, visual quality through frame-level and drift-aware metrics, interaction physics through mechanics, optics, and 3D-consistency tests, and memory through turning-point-aware scene reconstruction and third-person subject memory.

B Test Suite

The WorldRoamBench test suite comprises 1,006 test cases spanning Nature, Urban, and Indoor scenes in first- and third-person views, providing first-frame images and keystroke action sequences for all four dimensions, built with a unified methodology but tailored to each dimension's requirements (representative cases in Appendix B.4).

B.1 Image Suite

Unified construction pipeline. First-frame images spanning first- and third-person views are collected from the open Internet and existing world-model benchmarks (*e.g.*, WorldScore (Duan et al., 2025)), then refined with GPT-Image-2 through targeted edits: resolution and layout adjustments, first-person↔third-person conversion, and caption-and-regenerate variants. This yields high-quality, diverse first frames while aiming to reduce potential training-image overlap that could inflate benchmark scores.

Dimension-specific image selection. All dimensions draw from the same image library, but selection criteria differ by goal: the *action* dimension favors spacious, unobstructed scenes for per-keystroke evaluation with fewer obstacle-induced confounds; the *physics* dimension requires specific affordances (walls, pillars, or furniture for clipping; platforms or staircases for terrain following; long corridors for 3D consistency); and the *memory* dimension spans simple-to-complex scenes paired with simple and multi-step actions to vary revisit difficulty.

B.2 Action Suite

Dimension-specific action design. Action sequences are tailored to each dimension under one horizon definition based on commanded rollout duration: *Short* ($t \leq 15$ s), *Mid* ($15 < t \leq 30$ s), and *Long* ($30 < t \leq 60$ s). The *action* dimension crosses three durations (15/30/60 s) with 0/1/2 switches in a balanced 3×3 design (a single key per segment), supporting separate analyses of duration and switching. Appendix B.5 details the action-control construction. The *memory* dimension

follows symmetric traversals A→B→ inv(B)→ inv(A), where $\mathrm{inv}(\cdot)$ is the inverse action, to enable revisit; and the *physics* dimension directs paths toward walls (collision), platform edges (gravity), and through long corridors at a constant input rate (3D consistency) to trigger targeted interactions. Action and memory paths are routed through open space to limit physics-driven confounds.

B.3 Test Suite Statistics

Figure 5 shows the commanded-key distribution per dimension in *key-seconds* (holding one key for one second is one key-second). Memory is dominated by the symmetric J/L and W/S pairs, reflecting its revisit design; Action covers all six keys and is the most balanced, limiting concentration on any single motion axis; and Physics is strongly skewed toward forward motion (W) toward obstacles, terrain, and corridors. Figure 6 and Figure 7 summarize case counts by dimension, horizon, and physics protocol. Of the 1,006 cases, 648 are action (351 FPV, 297 TPV), 211 memory (103 FPV, 108 TPV), and 147 physics (76 FPV, 71 TPV), with horizons spanning 10–60 s: 271 Short, 298 Mid, and 437 Long, the Long group being largest as it includes every physics case and the longest action and memory rollouts.

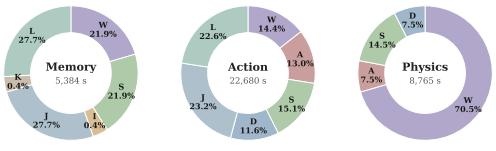


Figure 5: **Action key distribution by dimension.** Donuts show commanded-key proportions and total key-seconds.

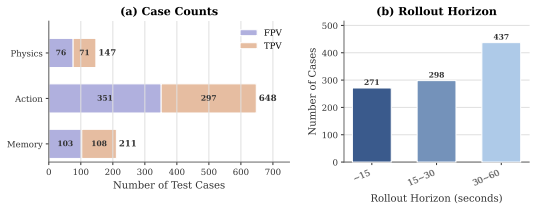


Figure 6: **Test case counts and rollout horizons.** (a) Counts by dimension and view (FPV/TPV). (b) Commanded rollout durations (s).

Action is balanced across the three horizon splits (216/216/216), and Memory spans 55/82/74 across Short, Mid, and Long. Because all physics rollouts last 60 s, horizon-specific evaluations reuse the single full-suite Physics score as a fixed term (Section 4.2).

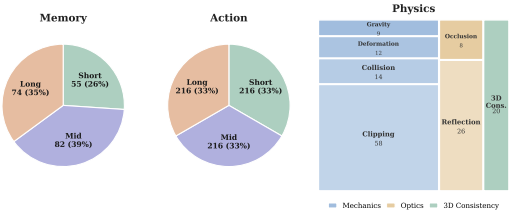


Figure 7: **Rollout-horizon and physics-protocol distribution.** Memory and Action counts by Short/Mid/Long split; 147 Physics cases by protocol.

B.4 Representative Case Gallery

To provide a qualitative overview of the visual and action coverage in WorldRoamBench, we include a gallery of representative test cases in Figure 8. Each subfigure shows the first-frame image together with an action schematic, illustrating both the scene context and the intended WASD/IJKL trajectory. The gallery is organized along two axes: scene category and camera view. Rows correspond to the three major scene families used throughout the benchmark—Indoor, Urban, and Nature—while columns separate first-person and third-person views. This layout highlights that each semantic environment type is evaluated under both egocentric exploration and avatar-centered interaction, enabling the benchmark to probe camera control, physical interaction, scene memory, and subject memory across diverse visual domains.

B.5 Action-Control Test Design

Rollout duration and action switches. Each first-frame image is paired with nine action sequences, crossing three rollout durations (15/30/60 s) with three action-switch counts (0/1/2), as illustrated in Figure 9(a). A rollout contains one, two, or three equal-duration segments, respectively; each segment sustains a single key, and adjacent segments use different keys. The 39 FPV and 33 TPV images yield 351 and 297 cases, respectively. Equal case counts across the nine conditions support comparisons across horizons at a fixed switch count and across switch counts at a fixed horizon.



Figure 8: **Test suite gallery.** Each subfigure shows a representative test case with the first frame, trajectory schematic, and action sequence. The gallery is organized by scene category from top to bottom (Indoor, Urban, Nature) and by view from left to right (FPV, TPV).

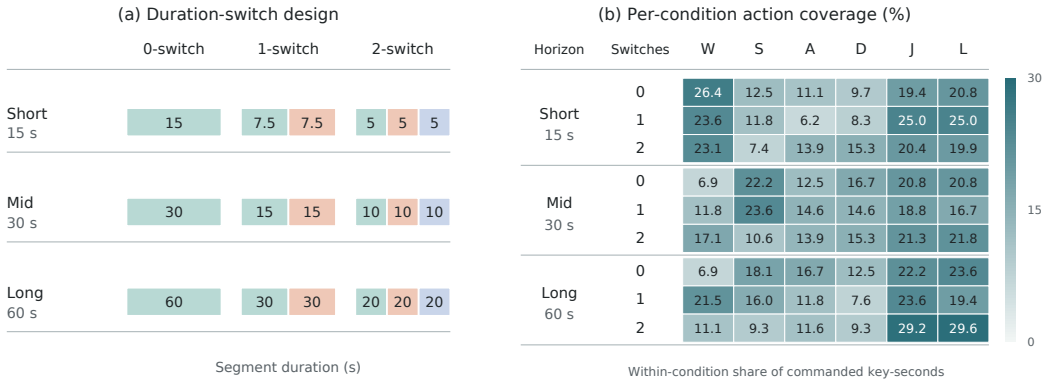


Figure 9: **Action-control test design and coverage.** (a) Nine duration–switch combinations per initial image, with segment durations in seconds. Colors indicate segment positions, not fixed actions. (b) Within-condition action shares (%) measured in commanded key-seconds. All six keys occur in every condition.

Scene-aware sequence construction. Initial images are inspected for forward and lateral clearance, and action identities and templates are adapted to scene layout and rollout duration. Lateral motion is restricted to scenes with sufficient lateral clearance and excluded for TPV vehicle subjects. In sequences containing both translation and yaw rotation, all translation segments precede the rotations. Models can produce different turning angles under the same yaw command; this order-

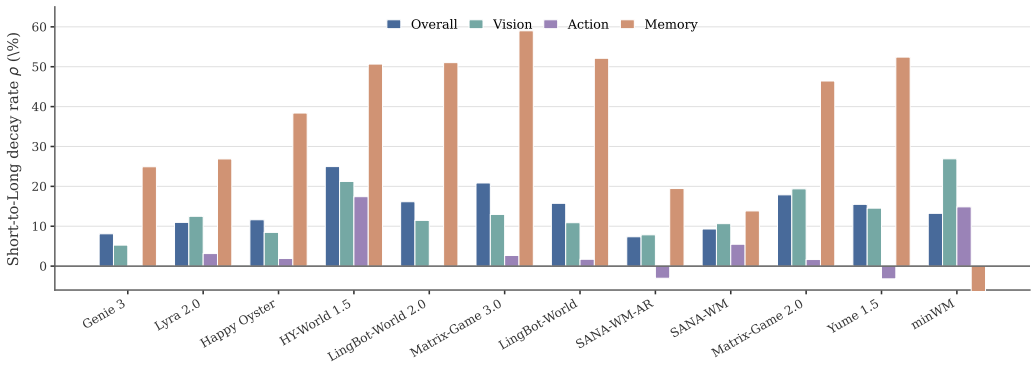


Figure 10: **Short-to-Long horizon decay rates.** Decay rates ρ_d (defined in Section 4.2) for Overall, Vision, Action, and Memory. Positive values denote degradation; negative values denote models that improve at Long horizon.

ing avoids subsequent translation along model-dependent headings, reducing the risk that collision responses are mistaken for action-following errors.

Representative patterns and action coverage. Templates include translation followed by opposite yaw commands (W→J→L), same-axis reversals (A→D→A), and alternating yaw commands (J→L→J). These cover transitions between motion modalities and direction reversals without placing translation after rotation. The suite uses four translation keys (W/S/A/D) and two yaw keys (J/L); left/right variants are alternated across samples to improve directional balance subject to scene and subject constraints. Figure 9(b) reports the within-condition action shares, computed from commanded key-seconds. All six action primitives are represented in every duration–switch condition, supporting per-action evaluation across horizons and switch counts. Figure 5 reports the aggregate distribution.

C Extended Horizon and Behavioral Analysis

This appendix provides decay-rate plots and extended behavioral analyses complementing the per-model horizon results in Section 4.2.

C.1 First-Person Performance Across Horizons

C.1.1 Overall Performance

The mean Overall score decreases from 63.16 (Short) to 59.81 (Mid) and 54.11 (Long) (Table 3 and Figure 10). All 12 models decline, but by markedly different amounts—SANA-WM-AR and Genie 3 lose only 7.4% and 8.1%, whereas HY-World 1.5 and Matrix-Game 3.0 lose 25.0% and 20.8%—so horizon robustness is a complementary axis that a single aggregate score conceals. Long-horizon evaluation is also more discriminative among the leaders: the top-4 spread grows from 2.09 points at Short (73.09–71.00) to 8.87 at Long (67.17–58.30), while the full-leaderboard spread barely changes (27.97→28.02).

Rank migration is concentrated in the mid-field: SANA-WM-AR rises from 8th to 5th and LingBot-World 2.0 from 5th to 4th, while HY-World 1.5 falls from 4th to 7th, Matrix-Game 3.0 from 6th to 9th, and LingBot-World edges up from 7th to 6th, producing 12 positions of total absolute rank migration.

C.1.2 Action Following

Action is the most horizon-stable dimension among those stratified by horizon: the mean score changes only from 82.58 to 83.79 and 79.63 across the three horizons. LingBot-World 2.0 is effectively horizon-invariant and rises from 5th to 1st at Long horizon, while Yume 1.5 and SANA-WM-

Table 4: **Third-person scores across horizons.** Notation follows Table 3: Short/Mid/Long scores ($\times 100$, $\uparrow$) and Short-to-Long decay rate ρ ($\downarrow$, Section 4.2) for Overall, Vision, Action, and Memory; Overall is the plain composite without the control-rate weighting used for the TPV leaderboard. Rows ordered by Short-horizon Overall; gray = closed-source; **bold**/underline = best/2nd-best per metric and horizon. In the Mean row, decay rates are averaged across models.

Model	Overall				Vision				Action				Memory			
	Short ↑	Mid ↑	Long ↑	$\rho\downarrow$	Short ↑	Mid ↑	Long ↑	$\rho\downarrow$	Short ↑	Mid ↑	Long ↑	$\rho\downarrow$	Short ↑	Mid ↑	Long ↑	$\rho\downarrow$
Happy Oyster	**74.65**	74.65	70.85	5.1%	71.86	70.01	67.68	5.8%	84.28	83.46	86.65	-2.8%	73.57	76.21	60.15	18.2%
Genie 3	74.57	**75.24**	**71.75**	3.8%	73.87	71.68	**70.51**	4.5%	83.38	83.50	87.38	-4.8%	**73.87**	**78.64**	**61.98**	16.1%
HY-World 1.5	73.84	72.21	65.72	11.0%	76.12	70.54	65.24	14.3%	**94.80**	**91.95**	85.08	10.3%	73.08	75.02	61.21	16.2%
LingBot-World 2.0	73.60	72.12	68.38	7.1%	**77.42**	**74.50**	70.35	9.1%	88.83	87.75	**89.34**	-0.6%	70.14	68.25	55.82	20.4%
LingBot-World	68.98	67.10	57.51	16.6%	75.55	74.08	68.84	8.9%	82.91	82.43	80.96	2.4%	63.62	58.06	26.38	58.5%
Mean	73.13	72.26	66.84	8.7%	74.96	72.16	68.52	8.5%	86.84	85.82	85.88	0.9%	70.86	71.24	53.11	25.9%

AR even improve relative to Short horizon. HY-World 1.5 and minWM are the main exceptions, decaying by 17.4% and 14.9%; HY-World 1.5 consequently falls from 1st to 7th.

C.1.3 Visual Quality

Mean Visual Score decreases monotonically from 71.86 to 68.12 and 62.23 across the three horizons. Because Visual Score includes range-based drift, cross-horizon differences also depend on the number of evaluated windows and do not isolate temporal degradation. SANA-WM-AR remains the visual leader (77.40→71.33), while Genie 3 has the lowest decay (5.3%) and rises from 10th to 5th. minWM and HY-World 1.5 show the largest relative score decreases (26.9% and 21.2%), and Yume 1.5 falls from 2nd to 7th.

C.1.4 Memory

Memory is the most horizon-sensitive dimension: its mean score falls from 65.38 at Short horizon to 54.48 at Mid horizon and 41.75 at Long horizon, corresponding to 35.5% average decay. A model can therefore retain action responsiveness while progressively losing previously observed scene content. At Long horizon, only SANA-WM (61.03) and Genie 3 (60.32) remain above 60, while 6 models fall below 40.

Model-wise decay varies from -9.0% (minWM, whose memory instead improves at Long horizon) to 59.0% (Matrix-Game 3.0), defining fast ($\rho \geq 50\%$), moderate ($30\% \leq \rho < 50\%$), and slow ($\rho < 30\%$) regimes. Memory decay exceeds 50% for 5 models. Its rank correlation with Short-horizon memory is only $+0.04$, so strong initial memory does not imply robustness: Matrix-Game 3.0 falls from 4th to 9th, whereas SANA-WM rises from 7th to 1st.

Rank migration in Memory is the largest of any dimension: Genie 3 leads at Short and Mid but yields 1st to SANA-WM at Long, and the largest movements are minWM ($12 \rightarrow 5$), SANA-WM ($7 \rightarrow 1$), Matrix-Game 3.0 ($4 \rightarrow 9$), and HY-World 1.5 ($2 \rightarrow 7$).

C.2 Third-Person Performance Across Horizons

For TPV, Action, Visual, Physics, and Memory scores are computed on rollouts passing the controllability gate. For Action, Visual, and Memory, horizon-specific scores average valid rollouts within each split; leaderboard scores pool valid rollouts across splits before rounding. Table 5 reports the valid sample counts.

Table 4 reports horizon splits for the five TPV-capable models. The sensitivity ordering mirrors the first-person case: Memory decays most (25.9% mean) and Action is essentially flat (0.9%), while mean Overall falls from 73.13 to 66.84 (8.7% mean decay, vs. 14.3% first-person). Happy Oyster leads Overall at Short (74.65), Genie 3 leads at both Mid (75.24) and Long (71.75), and it is both the most robust (3.8% decay) and the Memory leader at all three horizons. Unlike the first-person split, mean TPV Memory rises from Short to Mid (70.86→71.24) before dropping to 53.11 at Long; LingBot-World shows the steepest Memory collapse (58.5% decay, 26.38 at Long), consistent with its lowest control rate (Table 5). Action even improves at Long for Genie 3, Happy Oyster, and LingBot-World 2.0 (negative ρ), whereas HY-World 1.5 loses 10.3% and falls from 1st to 4th.

Table 5: **Valid TPV rollouts and control rates.** Counts are valid/total after the controllability gate. Action and Vision share one suite; Memory and Physics use separate suites. Control Rate ($\times 100$) pools all 476 TPV cases (297 action, 71 physics, 108 memory).

Model	Action / Vision			Memory			Physics	Control Rate ↑
	Short	Mid	Long	Short	Mid	Long		
Happy Oyster	95/99	94/99	96/99	19/19	45/45	40/44	70/71	**96.43**
Genie 3	80/99	77/99	79/99	19/19	45/45	41/44	58/71	<u>83.82</u>
HY-World 1.5	49/99	44/99	29/99	12/19	10/45	17/44	28/71	39.71
LingBot-World 2.0	55/99	48/99	40/99	13/19	22/45	22/44	56/71	53.78
LingBot-World	43/99	28/99	23/99	10/19	13/45	2/44	50/71	35.50

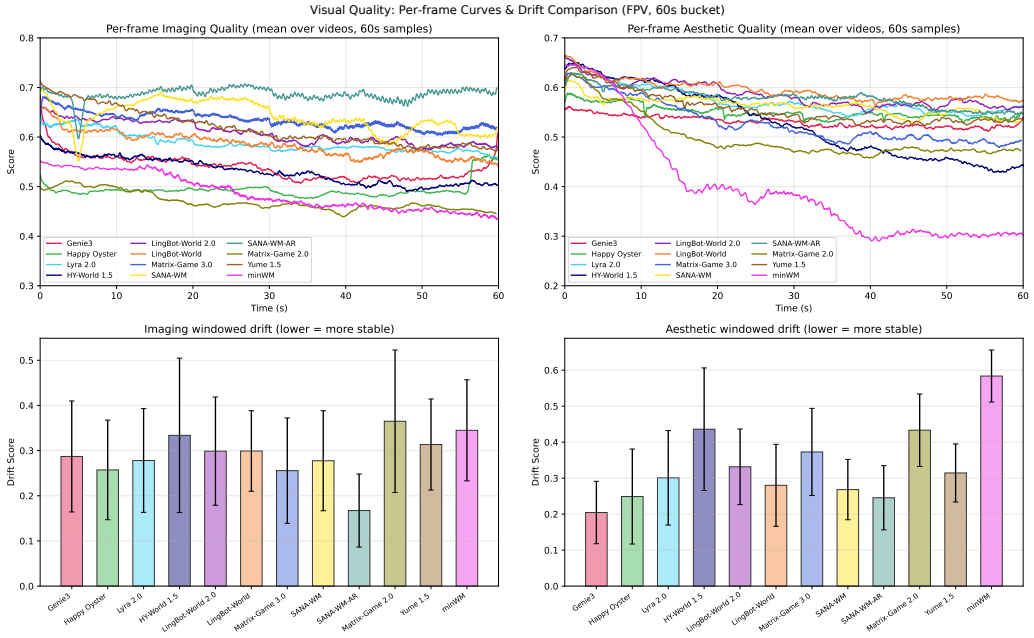


Figure 11: **Temporal visual quality and windowed drift (first-person, 60 s).** Top: mean imaging and aesthetic scores as functions of elapsed time over 117 cases per model; timestamps are derived using each model's native output frame rate (12–24 fps), and each time point averages videos that reach that elapsed time. Bottom: mean per-video relative drop from the best to worst 3 s window (1 s stride; lower is more stable); black error bars denote ± 1 std across videos.

C.3 Other Observations in World Models

Visual quality degradation exhibits model-specific temporal patterns. Figure 11 shows model-specific and axis-dependent degradation over 60 s. SANA-WM-AR is the most stable on imaging (16.73 windowed drift) and Genie 3 on aesthetics (20.45). At the other extreme, minWM exhibits selective collapse—its aesthetic score falls from 0.643 to 0.310, mostly within the first 40 s (largest aesthetic drift, 58.36)—while HY-World 1.5 and Matrix-Game 2.0 degrade progressively on both axes; such imaging-versus-aesthetic asymmetries confirm that the two axes capture distinct failure modes. Happy Oyster and SANA-WM instead show transient dips around 5 s followed by partial recovery, illustrating why best-to-worst windowed drift captures non-monotonic failures that start-to-end comparisons may miss. Absolute quality and temporal stability are also distinct: Happy Oyster has a low full-rollout imaging score (0.490) yet the 3rd-lowest imaging drift. Yume 1.5 starts with the highest imaging score (0.696) yet reaches 31.34 imaging drift. Thus, time-averaged quality alone does not capture stability.

Longer rollouts and action switching expose model-dependent control failures. Figure 12 compares partial action accuracy across Short, Mid, and Long rollouts (15/30/60 s) and 0/1/2 action

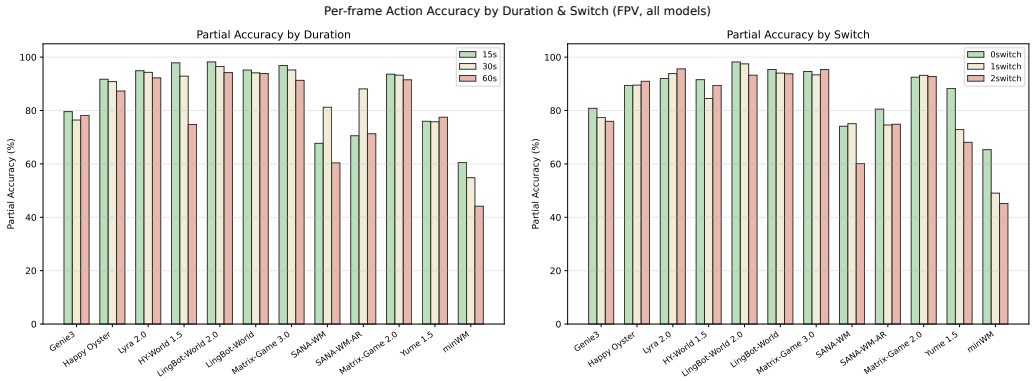


Figure 12: **Per-frame partial action accuracy by rollout duration and action-switch count (first-person).** Bars are grouped by rollout duration (left: 15/30/60 s) and number of ground-truth action switches (right: 0/1/2). The balanced 3×3 design has 39 cases per duration–switch cell and 351 per model; each bar is the unweighted mean over 117 cases. Camera poses use the MegaSaM + DA-v3 pipeline. A frame is a partial match when the predicted and ground-truth atomic-action sets intersect.

switches. Action proportions vary across conditions, so these comparisons characterize performance under the benchmark design rather than isolated causal effects. Mean accuracy is 85.2%, 86.1%, and 79.7% across the three horizons, with a 5.5 pp decrease from Short to Long. Ten of twelve models decline, most notably HY-World 1.5 ($97.9\% \rightarrow 74.7\%$) and minWM ($60.6\% \rightarrow 44.2\%$). LingBot-World 2.0 maintains the highest accuracy at every horizon (98.2/96.5/94.2%). SANA-WM and SANA-WM-AR peak at Mid horizon, while SANA-WM-AR and Yume 1.5 score slightly higher at Long than at Short, showing that longer rollouts do not uniformly reduce action accuracy.

Across 0/1/2 switches, mean accuracy is 86.9%, 82.9%, and 81.3%. Eight models decline from 0 to 2 switches, but only 5 decline monotonically. The largest decreases occur for minWM (20.2 pp), Yume 1.5 (20.2 pp), and SANA-WM (14.0 pp), whereas Lyra 2.0 improves by 3.6 pp and the Matrix-Game models remain nearly flat. Yume 1.5 improves slightly with duration yet drops sharply with more switches, illustrating why robustness to longer rollouts does not necessarily imply robustness to action switching.

Partial accuracy credits shared primitives without requiring an exact action match. For example, Yume 1.5 achieves 88.3% partial versus 71.4% strict accuracy with 0 switches, while LingBot-World 2.0 achieves 97.5% versus 80.3% with 1 switch. These gaps show that high partial accuracy can coexist with substantial errors in the complete action label.

Third-person controllability varies dramatically across models. Control Rate is the fraction of TPV cases passing the controllability gate (Appendix J.1). Pooled over the 476 third-person test cases that enter the action (297), physics (71), and memory (108) dimensions, Happy Oyster achieves the highest control rate (96.43), followed by Genie 3 (83.82), LingBot-World 2.0 (53.78), HY-World 1.5 (39.71), and LingBot-World (35.50), matching the TPV control-rate row in Table 2 and summarized in Table 5. The resulting gap is substantial: Happy Oyster's control rate is roughly $2.7\times$ that of LingBot-World and $2.4\times$ that of HY-World 1.5.

Qualitatively, Genie 3 and Happy Oyster maintain coherent third-person interaction across realistic and game-like scenes—Happy Oyster particularly so, consistent with its high subject memory score in Table 2—plausibly reflecting specialized optimization for third-person or avatar-centric interaction. By contrast, HY-World 1.5 shows third-person control mainly in game-style domains (Unity-, FPS-, or Genshin-like scenes). LingBot-World, LingBot-World 2.0, and HY-World 1.5 expose no explicit FPV/TPV switch, suggesting their third-person ability emerges from training data and the base model rather than an explicit view-selection mechanism, whereas the strongest closed-source models may benefit from more deliberate third-person design.

First-person physics depends on whether the camera is embodied by a held object. In the physics dimension, the same first-person scene can yield different physical behavior depending on

whether the view contains a visible held object (*e.g.*, a weapon or a cup in hand) or is a clean camera-only view. Within the same scene, Genie 3 is less likely to clip through walls when the view is visually grounded by a held object, and textual prompts describing the subject as human do not fully remove this gap. This suggests that even models with strong overall physics, such as Genie 3 and Happy Oyster, may rely partly on data-derived visual priors rather than fully internalized physical understanding.

D Related Work Details

D.1 Interactive Video World Models

Unlike passive video generation models such as Wan 2.7 (Alibaba Cloud, 2026), Kling 3.0 (Kuaishou Technology, 2026), Sora 2 (OpenAI, 2025), Veo 3.1 (Gallegos & Iljic, 2025), and Seedance 2.0 (Team Seedance et al., 2026), which synthesize videos from text or image prompts, interactive world models generate visual environments under real-time user control. Genie (Bruce et al., 2024) pioneered this paradigm by learning latent action spaces from unlabeled video. Recent models explore three overlapping control modalities. **Discrete-action control**, used by Matrix-Game 2.0/3.0 (He et al., 2025; Wang et al., 2026c) and HY-World 1.5 (Sun et al., 2025), conditions generation on keyboard commands and can be combined with camera-pose inputs. **Camera-pose control**, adopted by LingBot-World/2.0 (Robbyant Team et al., 2026; Gao et al., 2026), SANA-WM/AR (Zhu et al., 2026), Lyra 2.0 (Shen et al., 2026), and minWM (Zhao et al., 2026a), specifies motion through continuous camera trajectories. **Language-based control**, exemplified by Yume 1.5 (Mao et al., 2025), uses text instructions to guide motion and world events. Closed-source models such as Genie 3 (Parker-Holder & Fruchter, 2025) and Happy Oyster (Alibaba Group, 2026) further push visual quality and temporal coherence. For cross-interface comparison, models receive an initial image and discrete actions (WASD/IJKL or idle), with actions passed directly or translated by deterministic adapters, and produce a video rollout.

D.2 Benchmarks for World Models

Evaluation benchmarks span several paradigms. **Video generation** benchmarks (VBench (Huang et al., 2024; 2025b; Zheng et al., 2025), WorldScore (Duan et al., 2025), WorldModelBench (Li et al., 2025a)) assess visual quality, text alignment, or physical commonsense without closed-loop interactive control. **Domain-specific** benchmarks target narrow scenarios: ACT-Bench (Arai et al., 2024) and WorldLens (Liang et al., 2026) for autonomous driving; EWMBench (Hu et al., 2025) and WorldArena (Shang et al., 2026) for embodied robotic manipulation. **Physics and interaction fidelity** benchmarks (PhyGenBench (Meng et al., 2024), WorldBench (Upadhyay et al., 2026), Omni-WorldBench (Wu et al., 2026)) evaluate physical plausibility or causal interaction correctness on prompt-conditioned generation without discrete-keystroke control. **Interactive world model** benchmarks are most relevant to our setting. MIND (Ye et al., 2026) introduced memory via closed-loop revisit tests; WorldMark (Xu et al., 2026) unified WASD actions with three difficulty tiers; iWorld-Bench (Fang et al., 2026) proposed a control quadruple with six levels; WBench (Ying et al., 2026) unified multi-modal inputs for multi-turn evaluation; WorldOlympiad (Zhao et al., 2026b) jointly assessed physics, geometry, and interaction across chunk-by-chunk rollouts; and WildBench, introduced with WildWorld (Li et al., 2026), evaluated segment-level action following and state alignment in a single game.

These protocols differ in evaluation unit and temporal scope (Table 1), using clips, windows, segments, turns, or chunks; WorldMark evaluates 20–60 s rollouts (Xu et al., 2026). We jointly evaluate action following, visual drift, interaction physics, and scene/subject memory in open-world, long-horizon (10–60s) interactions. Our action suite crosses three rollout durations with three switch counts for separate comparisons of duration and switching.

D.3 Action Evaluation Methodologies

Action evaluation in world models generally follows four paradigms. **IDM-based** approaches (VPT (Baker et al., 2022), MineWorld (Guo et al., 2025)) train an Inverse Dynamics Model to classify per-frame actions from video, achieving high accuracy in Minecraft but limited to that domain.

Trajectory-based approaches (MIND (Ye et al., 2026), iWorld-Bench (Fang et al., 2026)) recover 3D camera poses via SLAM or ViPE and compute geometric errors. **Flow-based** approaches, exemplified by WorldMark (Xu et al., 2026), measure continuous direction accuracy and purity from optical flow, along with response latency and motion stability. These scores differ from strict/partial matching of decoded action labels. **VLM-based** approaches (WildBench (Li et al., 2026)) use vision-language models to judge action consistency at the segment level. WorldRoamBench combines pose-decoded strict and partial action accuracy with TrajScore to assess command agreement and geometric fidelity.

D.4 Physics Evaluation Methodologies

Physics evaluation spans video-generation and world-model benchmarks. General video benchmarks such as VBench (Huang et al., 2024), VBench-2.0 (Zheng et al., 2025), and OmniBench (Bu et al., 2025) include physical plausibility alongside visual and temporal quality. Physics-focused benchmarks assess physical reasoning: PhyGenBench (Meng et al., 2024) covers diverse physical laws, WorldModelBench (Li et al., 2025a) evaluates violations such as gravity, penetration, fluid dynamics, and conservation, and WM-ABench (Gao et al., 2025) studies mechanical interactions through controlled simulation. World-model benchmarks assess complementary aspects: WorldScore (Duan et al., 2025) includes 3D and photometric consistency, WorldBench (Upadhyay et al., 2026) evaluates physical plausibility in generated worlds, WorldLens (Liang et al., 2026) studies closed-loop degradation, and Omni-WorldBench (Wu et al., 2026) introduces multi-level causal interaction scenarios.

However, existing physics tests are still mostly designed for passive observations or coarse semantic judgments. They rarely ask whether a user's action should trigger a specific physical consequence, and they do not provide case-specific metrics for contact, occlusion, reflection, shadows, or traversability. WorldRoamBench complements prior benchmarks by evaluating fine-grained interaction physics in 3D-grounded scenes under explicit WASD/IJKL control.

D.5 Memory Evaluation Methodologies

Memory evaluation commonly relies on frame-pair similarity. MIND (Ye et al., 2026) constructs symmetric action sequences and compares temporally symmetric frames using pixel-level metrics such as LPIPS. This formulation implicitly assumes perfect action execution and uniform action magnitude, such that symmetric temporal indices correspond to the same spatial location. However, interactive models may drift, under- or over-execute commands, and accumulate timing offsets, causing frame-pair metrics to conflate control error with memory degradation. WBench (Ying et al., 2026) partially addresses temporal misalignment by estimating poses and selecting the revisit frame closest to the initial frame, but the resulting metric remains frame-level and first-frame-centric, rather than assessing preservation of the broader visited scene.

A related line of work evaluates short-term geometric consistency rather than memory. WBench's spatial consistency, WorldMark's geometry consistency (Xu et al., 2026), and WorldOlympiad's geometric evaluation (Zhao et al., 2026b) measure whether nearby generated views remain 3D-coherent. These metrics do not assess retention of previously generated or observed content after a long interactive loop. WorldRoamBench therefore formulates memory as revisit evaluation under imperfect control, combining action-aware transition localization with scene-level 3D reconstruction to assess preservation beyond a single frame pair.

E Evaluated Model Details

WorldRoamBench evaluates models as black-box or gray-box interactive generators: each model receives the same initial frame and an action program derived from the WASD/IJKL vocabulary, then generates a video rollout scored by the same action, visual-quality, physics, and memory pipelines. No model is fine-tuned on WorldRoamBench. For models with non-keyboard interfaces, deterministic adapters translate the benchmark action program into each model's supported control format. Scores therefore characterize performance under the evaluated configurations and adapters. Table 7 summarizes the adapter used for each evaluated model.

Table 7: **Model evaluation adapters.** Each model receives the benchmark's WASD/IJKL actions through a deterministic adapter that converts them to the model's native input format.

Model	Evaluation Adapter
Genie 3	Native WASD/IJKL support. Actions are injected via frontend automation at 1/20 s per action, synchronized with the model's 20 fps output.
Happy Oyster	Native WASD/IJKL support. Actions are injected via frontend automation at 1/24 s per action, synchronized with the model's 24 fps output.
LingBot-World	Autoregressive (Fast) checkpoint; benchmark actions are supplied through camera-pose conditioning. Actions are converted to camera poses at a translation speed of 0.02 per frame and a rotation speed of 0.75° per frame.
LingBot-World 2.0	Benchmark WASD/IJKL commands are converted into camera-pose sequences for the evaluated configuration.
HY-World 1.5	Native keyboard support, but the model accepts one action per latent frame. Four consecutive frame-level actions are merged into a single latent-frame action before being fed to the model.
Lyra 2.0	Actions are converted to camera poses at a translation speed of 0.02 per frame and a rotation speed of 0.75° per frame, preserving action order.
SANA-WM	WASD/IJKL are mapped to local camera increments: A/D for translation and J/L for rotation.
SANA-WM-AR	Chunk-causal configuration; benchmark WASD/IJKL commands are converted into camera-pose conditioning, with 24 generated video frames per chunk.
Matrix-Game 2.0	Native per-frame action support; each benchmark action is passed directly as a per-frame action token.
Matrix-Game 3.0	Native per-frame action support; an external adapter converts benchmark keyboard actions into per-frame action conditions and camera-pose conditioning.
Yume 1.5	In the evaluated configuration, motion control is expressed through text rather than pose or action tensors; the initial image is also provided. Every 32 frame-level actions are aggregated by taking the dominant WASD and IJKL keys within the segment, then rendered into a fixed-template motion caption (e.g., "Person moves forward; Camera turns left"). Each caption controls one autoregressive chunk of 32 pixel frames.
minWM	Benchmark actions are converted into camera-pose sequences and supplied to the model together with the initial image.

Table 6: Evaluated interactive world models. Input: 🎮 = keyboard actions, 🎥 = 6-DoF camera poses, 📝 = text; → denotes conversion from benchmark actions.

Model	Type	Params	Year	Views	Input	Resolution	FPS	Chunk (frames)	Inf. Speed (Hz)	MPPS	Source
Genie 3	Closed	—	2025	FPV/TPV	🎮	1280×704	20	—	—	—	DeepMind
Happy Oyster	Closed	—	2026	FPV/TPV	🎮	1296×720	24	—	—	—	Alibaba
LingBot-World	Open	14B	2026	FPV/TPV	🎮 → 🎥	832×480	16	12	1.79	0.72	Ant Group
LingBot-World 2.0	Open	14B	2026	FPV/TPV	🎮 → 🎥	832×480	16	16	1.75	0.70	Ant Group
HY-World 1.5	Open	8B	2025	FPV/TPV	🎮 → 🎮 + 🎥	832×480	24	16	1.30	0.52	Tencent
Lyra 2.0	Open	14B	2026	FPV	🎮 → 🎥	832×480	16	80	2.31	0.92	NVIDIA
SANA-WM	Open	2.6B	2026	FPV	🎮 → 🎥	1280×704	16	56	0.81	0.73	NVIDIA
SANA-WM-AR	Open	2.6B	2026	FPV	🎮 → 🎥	1280×704	16	24	0.93	0.84	NVIDIA
Matrix-Game 3.0	Open	5B	2026	FPV	🎮 → 🎮 + 🎥	1280×704	17	40	5.34	4.82	Skywork
Matrix-Game 2.0	Open	1.8B	2025	FPV	🎮	640×352	12	12	6.26	1.41	Skywork
Yume 1.5	Open	5B	2025	FPV	🎮 → 📝	1280×704	16	32	6.49	5.85	Shanghai AI Lab
minWM	Open	8B	2026	FPV	🎮 → 🎥	832×480	16	16	3.97	1.58	ShengShu

E.1 Control-Interface Families and Inference Efficiency

This model set is intentionally heterogeneous and spans four control-interface families. Genie 3, Happy Oyster, and Matrix-Game 2.0 natively accept keyboard input with no conversion and are closest to the target WASD/IJKL interaction protocol. LingBot-World, LingBot-World 2.0, Lyra 2.0, SANA-WM, SANA-WM-AR, and minWM are conditioned on continuous camera-pose sequences, testing whether pose-driven models can follow discrete keystroke programs after action-to-pose conversion. HY-World 1.5 and Matrix-Game 3.0 are conditioned on both discrete keyboard actions and converted camera poses, evaluating a hybrid action-plus-pose interface. Yume 1.5 represents a caption- or instruction-driven interface, with benchmark keyboard commands aggregated over 32-frame windows and converted into motion captions.

Table 6 reports native output FPS, frames per autoregressive chunk, and generation throughput. For open-source models, inference speed (Hz) is chunk frames divided by wall-clock chunk time, averaged over 20 runs on a single NVIDIA H20 GPU; MPPS (million pixels generated per second) equals resolution width $\times$ height $\times$ inference speed divided by 10^6. The bidirectional SANA-WM is evaluated with 56 frames per chunk. Open-source models vary widely in efficiency: Yume 1.5 is fastest at 6.49 Hz, followed by Matrix-Game 2.0 (6.26 Hz) and Matrix-Game 3.0 (5.34 Hz), while SANA-WM is slowest at 0.81 Hz, with HY-World 1.5 (1.30 Hz) and LingBot-World (1.79 Hz) similarly throughput-limited. Normalized by resolution, Yume 1.5 and Matrix-Game 3.0 still lead in MPPS (5.85 and 4.82), reflecting efficient high-resolution generation, whereas HY-World 1.5, LingBot-World, and SANA-WM stay below 1 MPPS (0.52, 0.72, 0.73). These gaps highlight the trade-off between scale, visual fidelity, and real-time deployability in interactive world generation.

E.2 Model Descriptions

Genie 3 (Parker-Holder & Fruchter, 2025). Genie 3 is Google DeepMind's closed-source interactive world model. It generates subsequent video frames in response to user actions, enabling interactive exploration of generated environments. Public demonstrations also show the persistence of previously observed scene content and changes to the environment through text prompts. We evaluate Genie 3 through its web interface, using browser automation to issue the benchmark keyboard commands after providing the initial image and associated scene and character descriptions. We treat the model as a black box, without access to its internal architecture or sampling configuration. Genie 3 ranks 1st in first-person Overall, Physics, and Memory Scores, but 8th in Action Score and 9th in Visual Score.

Happy Oyster (Alibaba Group, 2026). Happy Oyster is Alibaba's closed-source interactive world model. It supports exploration of generated environments and scene editing through text, voice, and image inputs. We evaluate its responses to the benchmark keyboard commands through its web interface using browser automation. As with Genie 3, we treat the model as a black box. Happy Oyster ranks 1st in third-person Overall and Physics Scores and in the Subject Memory submetric, but 3rd in Action Score and last in Visual Score among the 5 third-person models.

LingBot-World (Robbyant Team et al., 2026). LingBot-World is Ant Group's open-source world-model framework built on a Wan2.2 image-to-video backbone. Its training introduces action conditioning and progressively increases the video duration to learn control responses and long-term temporal dependencies. The Fast variant is obtained through block-causal adaptation and few-step distillation of the bidirectional model, allowing video chunks to be generated sequentially from past visual context and control inputs. We evaluate the 14B `LingBot-World-Fast` checkpoint, converting benchmark navigation and view-control commands into camera-pose sequences. Its scene memory is learned from video context rather than maintained in an explicit 3D scene representation. LingBot-World ranks 3rd in first-person Action Score but 9th in Memory Score.

LingBot-World 2.0 (Gao et al., 2026). LingBot-World 2.0 is Ant Group's open-source interactive world model with camera-pose and text conditioning. Compared with LingBot-World, LingBot-World 2.0 introduces causal world-model pretraining before distillation, learning to generate subsequent video chunks from historical visual context and control inputs. Camera poses control viewpoint motion, while per-chunk text prompts specify character actions and environmental events. Post-training combines distillation with optimization over long rollouts to improve generation efficiency and reduce accumulated errors. We evaluate the `lingbot-world-v2-14b-causal-fast` checkpoint with 16 generated video frames per chunk, converting benchmark navigation and view-control commands into camera-pose sequences. LingBot-World 2.0 ranks 1st in third-person Visual Score but 4th of 5 models in third-person Memory Score.

HY-World 1.5 (Sun et al., 2025). HY-World 1.5 is Tencent Hunyuan's open-source interactive world model built on HunyuanVideo 1.5. It uses a dual action representation that combines discrete action signals with continuous camera information. Its reconstituted context memory retrieves relevant past frames and adjusts their temporal encoding to retain access to observations from earlier parts of the rollout. We evaluate the 8B configuration, mapping benchmark navigation and view-

control commands to the model's action-conditioning inputs. HY-World 1.5 ranks 1st in third-person Action Score but last in third-person Physics Score among the 5 evaluated models.

Lyra 2.0 (Shen et al., 2026). Lyra 2.0 is NVIDIA's open-source framework for generating explorable 3D environments from camera-controlled video. Its video generator uses per-frame geometry to retrieve relevant past frames and establish dense correspondences with target viewpoints. The retrieved information guides appearance synthesis by the generative model. During training, self-augmented histories expose the model to its own degraded outputs so that it learns to correct rather than propagate generation errors. We evaluate the 14B video generator using continuous camera-pose sequences derived from benchmark navigation and view-control commands. The evaluation scores the generated video, not the framework's downstream 3D reconstruction output. Lyra 2.0 ranks 2nd in first-person Overall Score and 1st in the first-person Optics and 3D Consistency sub-metrics, but ranks 5th in Mechanics.

SANA-WM (Zhu et al., 2026). SANA-WM is NVIDIA's open-source framework for camera-controlled video generation. Its hybrid attention architecture combines Gated DeltaNet with softmax attention to model long video sequences, while a dual-branch conditioning mechanism incorporates 6-DoF camera poses. We evaluate its 2.6B-parameter bidirectional configuration, whose attention allows information exchange across temporal positions within the generation window. Benchmark navigation and view-control commands are converted into camera-pose sequences and supplied as conditioning inputs. We report this configuration separately from SANA-WM-AR because the two use different temporal attention structures. SANA-WM ranks 2nd in first-person Visual Score but 11th in Action Score and 9th in Physics Score.

SANA-WM-AR (Zhu et al., 2026). SANA-WM-AR denotes the chunk-causal SANA-WM configuration evaluated in WorldRoamBench. Its generative backbone processes video in temporal chunks, with each chunk conditioned on preceding visual context rather than later chunks. This differs from the bidirectional temporal attention used by SANA-WM above. Our evaluation uses a 2.6B-parameter generative model with 24 generated video frames per chunk. Benchmark navigation and view-control commands are converted into camera-pose sequences and supplied as conditioning inputs. SANA-WM-AR ranks 1st in first-person Visual Score and 3rd in Memory Score, but 9th in Action Score and 8th in Physics Score.

Matrix-Game 2.0 (He et al., 2025). Matrix-Game 2.0 is Skywork's open-source 1.8B interactive world model for streaming video generation. Its action-injection module incorporates keyboard and mouse signals into a causal video generator, and few-step distillation reduces the number of denoising steps required to generate each video chunk. For WorldRoamBench, benchmark WASD/IJKL commands are converted into the model's action-conditioning inputs. Matrix-Game 2.0 ranks 5th in first-person Action Score but 11th in both Visual and Physics Scores.

Matrix-Game 3.0 (Wang et al., 2026c). Matrix-Game 3.0 is Skywork's open-source interactive world model with a memory mechanism for long video rollouts. It uses camera information to retrieve and incorporate relevant historical observations, while training on imperfect generated histories teaches the model to correct accumulated errors. Multi-segment autoregressive distillation reduces the number of denoising steps required for streaming generation. We evaluate the 5B configuration, mapping benchmark commands to its action-conditioning inputs, which support discrete action and continuous camera-pose representations. Matrix-Game 3.0 ranks 1st in first-person Action Score but 7th in both Physics and Memory Scores.

Yume 1.5 (Mao et al., 2025). Yume 1.5 is Shanghai AI Lab's open-source 5B world-generation model with text-based control. It combines context compression and linear attention to process video history, and uses bidirectional attention distillation to accelerate streaming generation. In our evaluated configuration, the model receives the benchmark initial image and motion instructions expressed in natural language. The adapter summarizes the dominant navigation and view-control commands in each 32-frame window using a fixed text template. The video is evaluated using the same action-following metrics as the other models. This temporal aggregation may omit brief actions or action transitions within a window. Yume 1.5 ranks 5th in first-person Visual Score, but 10th in Action Score and last in Memory Score.

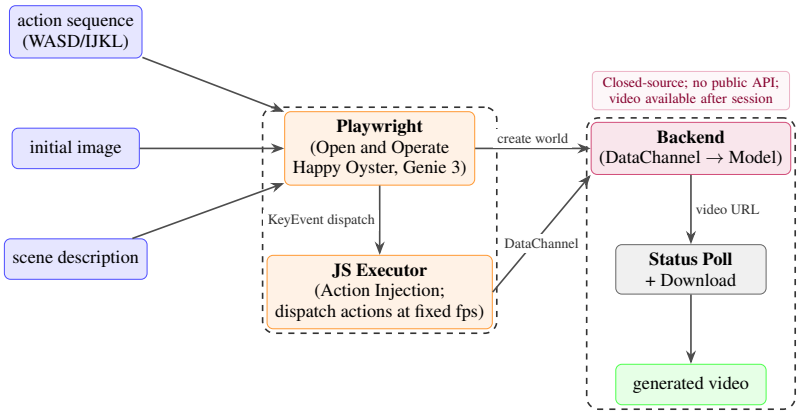


Figure 13: Closed-source automated interaction pipeline for Happy Oyster and Genie 3. The system reads benchmark test cases (action sequence, first frame, and scene description), then uses Playwright to drive the target platform's web interface. The entire action list is injected into the browser's V8 engine via a single `page.evaluate()` call, where actions are dispatched at intervals matching the model's native frame rate (e.g., $\frac{1}{24}$ s for Happy Oyster, $\frac{1}{20}$ s for Genie 3). The frontend encodes the keyboard commands and sends them to the backend via a WebRTC DataChannel. After the session completes, the backend-generated video is downloaded via status API polling.

minWM (Zhao et al., 2026a). minWM is ShengShu's open-source framework for converting pretrained video diffusion models into camera-controlled autoregressive generators. Its training pipeline first adds camera conditioning to a bidirectional backbone, then applies autoregressive training and distillation to obtain a generator that requires few denoising steps. We use the 8B `HY15/Action2V/dmd` checkpoint downloaded from the official Hugging Face repository (`MIN-Lab/minWM`). For WorldRoamBench, benchmark actions are converted into camera-pose sequences and supplied to the model together with the initial image. In our evaluation, minWM ranks last in first-person Overall, Action, Visual, and Physics Scores.

E.3 Closed-Source Model Evaluation

E.3.1 Closed-Source Model Data Acquisition

Evaluating closed-source interactive world models such as Genie 3 and Happy Oyster poses a practical challenge: neither downloadable model weights nor programmatic APIs for batch inference were available to us during the evaluation. We therefore access these models through their product interfaces using browser automation. A frontend automation pipeline drives the interactive session by injecting WASD/IJKL keystrokes at the model's native frame rate. For Genie 3, which outputs video at 20 fps, each benchmark action is held for $\frac{1}{20}$ s before the next action is issued; for Happy Oyster, which runs at 24 fps, each action is held for $\frac{1}{24}$ s. The automation layer injects keyboard events according to a predefined command sequence and timing schedule. Figure 13 illustrates the end-to-end pipeline: benchmark test cases are fed into a Playwright-driven browser session, where the full action list is injected and dispatched at the model's native frame rate, and the backend-generated video is retrieved once the session completes.

E.3.2 RAFT Alignment for Closed-Source Videos

Unlike open-source models with end-to-end control over the generation pipeline and known frame-to-action correspondence, closed-source models return a video file downloaded from the backend after the interactive session completes. The downloaded video often contains extra frames before the first effective action takes effect (e.g., loading screens, transition animations, or buffering delays), and may have a different total frame count than the benchmark action sequence.

To align the downloaded video with the ground-truth action program, we use RAFT (Recurrent All-Pairs Field Transforms) (Teed & Deng, 2020) optical flow to detect the onset of motion. Specifically, we compute the mean optical flow magnitude between consecutive frames. The first frame whose

optical-flow magnitude exceeds the threshold provides an estimate of the effective start frame. From this start frame, we crop the video to the required number of frames that matches the benchmark action schedule. The cropped videos are then used to evaluate action following, visual quality, physics, and memory over matched interaction windows for closed- and open-source models.

F Pose Estimation Pipeline Comparison

Comparison protocol. We compare pose pipelines by their suitability for action decoding. For each video, the *oracle ceiling* is the maximum strict accuracy over a grid of 31 translation and 8 rotation thresholds. This reduces dependence on a particular operating point, but measures action recoverability rather than pose error against ground-truth geometry. The evaluation covers 12 models and 351 cases per model. This full-grid diagnostic is distinct from the four-candidate selection used in the benchmark.

We compare six variants. Three use ViPE (Huang et al., 2025a) with a UniDepth-L (Piccinelli et al., 2024), Depth-Anything-v3 (Lin et al., 2025) (DA-v3), or Pi3X (Wang et al., 2026a)+MoGe-2 (Wang et al., 2025) depth prior; two use MegaSaM (Li et al., 2025b) with a DA-v1+UniDepth-L or DA-v3 prior; the sixth uses DA-v3 as a standalone feed-forward pose estimator. The ViPE+Pi3X+MoGe-2 and MegaSaM+DA-v3 variants are configurations that we assembled and implemented on top of the original ViPE and MegaSaM frameworks, respectively.

Configuration differences. The five SLAM-based variants share a DROID-SLAM (Teed & Deng, 2021) backbone but differ in their use of depth and dynamic-scene cues. MegaSaM initializes disparity from aligned monocular depth, retains a depth prior during bundle adjustment, and weights flow confidence using a learned motion probability. ViPE uses constant disparity initialization, depth regularization, and instance masks to suppress dynamic regions.

The comparison also varies focal handling and depth-model capacity. The upstream MegaSaM configuration refines a shared focal length when calibration is well-conditioned; our DA-v3 variant fixes the DA-v3 field-of-view estimate. ViPE uses DA-v3 Metric-L, whereas MegaSaM+DA-v3 and the standalone baseline use DA-v3 Giant. The latter predicts poses directly without bundle adjustment. ViPE's depth-only post-processing leaves camera poses unchanged and is excluded from this comparison. These configurations compare complete pipelines rather than isolate the effects of individual components.

Table 8: Six pose-estimation pipelines, grouped by SLAM framework. "BA": global bundle adjustment. "Focal": bundle-adjusted focal refinement. "Ceil.": mean per-sample oracle ceiling accuracy. † marks configurations assembled and implemented by us on top of the corresponding original framework.

Tag	Depth prior	BA	Focal	Ceil.
MegaSaM (DROID-SLAM based)				
`megasam_dav3`†	DA-v3 Giant	✓	×	**0.795**
`megasam`	DA-v1+UniDepth-L	✓	✓	0.770
ViPE (DROID-SLAM based)				
`default`	UniDepth-L	✓	✓	0.760
`dav3`	DA-v3 Metric-L	✓	✓	0.759
`pi3x_moge2`†	Pi3X+MoGe-2	✓	✓	0.752
Feed-forward pose				
`da3_direct`	DA-v3 Giant	×	×	0.689

Results and default choice. The mean ceiling accuracies follow the ranking `megasam_dav3` > `megasam` > `default` ≈ `dav3` > `pi3x_moge2` > `da3_direct` (Table 8 and Figure 14). `megasam_dav3` achieves the highest mean ceiling accuracy (0.795) and leads on 11 of 12 models, with `dav3` leading on LingBot-World 2.0. Its advantage therefore extends across most evaluated models.

Relative to upstream MegaSaM, its sample-level win and loss rates are nearly balanced (25.0% wins, 25.5% losses, and 49.5% ties), yet its mean ceiling is higher by 2.5 percentage points. This

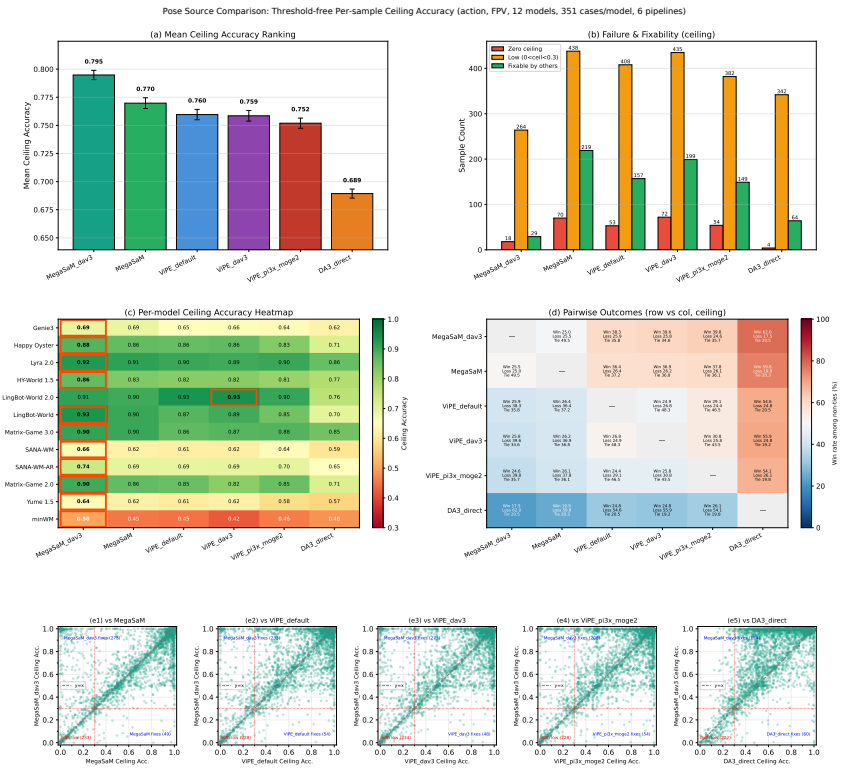


Figure 14: **Action-decoding ceiling across six pose pipelines.** Mean ceiling accuracy and low-accuracy sample rates (top), breakdowns by model and action (middle), and pairwise comparisons (bottom). In the scatter plots, points above the diagonal favor the y-axis pipeline.

higher mean is accompanied by fewer low-accuracy samples: 6.7% have ceiling accuracy below 0.3, compared with 12.1% for upstream MegaSaM. The ViPE variants span 0.752–0.760, while standalone DA-v3 reaches 0.689. MegaSaM+DA-v3 is used as the default action-decoding pipeline across models.

G Latent-Stride Action Discretization Details

This appendix specifies the windowed pose-to-action decoder, threshold selection, and stride comparison in Section 2.2.1. The decoder extends WorldCompass's CompassReward implementation (Wang et al., 2026b) with motion aggregation over K consecutive action intervals.

G.1 Motivation and Windowed Decoding

Chunk-wise world models such as Matrix-Game 2.0 (He et al., 2025) jointly denoise blocks of latent frames and use a video VAE with a temporal compression factor of four. Our benchmark uses sustained command segments, so most four-frame windows share a command. Together, these motivate decoding one action over K frames. Aggregating motion over this window can strengthen the motion signal relative to pose-estimation noise, whereas per-frame discretization can produce spurious labels when small inter-frame displacements are comparable to the noise floor.

For each complete window $c = 0, \ldots, \lfloor T/K \rfloor - 1$, we compute the relative pose as

$$\Delta P_c = \hat{P}_{cK}^{-1} \hat{P}_{(c+1)K}. \tag{11}$$

Its translation is expressed in the camera coordinates at the start of the window. The decoded action is assigned to steps $cK+1, \ldots, (c+1)K$ and compared with their original frame-level commands. Setting $K = 1$ yields frame-by-frame action decoding.

To assess the choice of window size, we compare $K \in \{1, 2, 4, 8\}$ on all 4,212 first-person action samples from 12 models. Translation and rotation thresholds are recalibrated independently for each K. Figure 15 reports three strict-accuracy diagnostics. "Per-sample ceiling" is the per-sample oracle over the full threshold grid; "Four-threshold array" uses four global translation candidates and one rotation threshold, with the best candidate selected per sample; "Global single threshold" uses the best global threshold pair on the same samples.

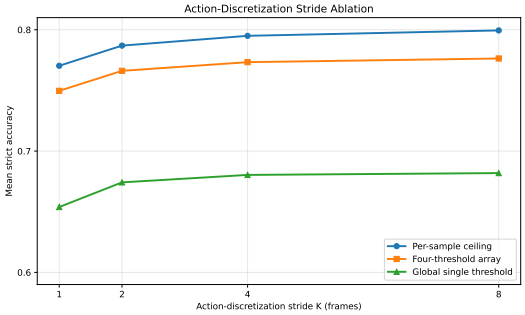


Figure 15: Action-discretization stride ablation.

Across all three strategies, mean strict accuracy improves as K increases from 1 to 4, with diminishing gains beyond $K{=}4$. This trend is consistent with temporal motion aggregation improving robustness to inter-frame pose noise. Increasing the window to $K{=}8$ yields only marginal gains while halving temporal resolution; assigning a shared label to a longer window can introduce errors over multiple frames near action switches. We therefore use $K{=}4$ by default to balance decoding accuracy and temporal resolution.

G.2 Threshold Scaling and Calibration

The same action can produce different translation magnitudes across models and scenes, making a single motion-detection threshold unsuitable for all rollouts. A high threshold can miss low-amplitude motion, whereas a low threshold can admit small pose-estimation fluctuations. WorldCompass (Wang et al., 2026b) similarly motivates multiple translation thresholds by scene-dependent pose scale; its released CompassReward implementation uses three fixed candidates.[2] We extend this design to four candidates calibrated across the evaluated models and scenes to accommodate their motion-scale variation. Calibration is performed separately for each pose source to account for differences in coordinate scale; within each source, all models and scenes share the same four candidates.

For the final $K{=}4$ decoder, the effective rotation threshold is $\tau_{\text{rot}} = 1.0°$, corresponding to a base threshold of $0.25°$ per action interval. Table 9 lists effective translation candidates in each pose source's reconstructed coordinate units; they are used directly without further stride scaling. The stride ablation independently recalibrates both threshold types for each K.

The candidate set Θ is fixed for each pose source and shared across models and scenes. To accommodate variation in reconstructed motion scale, we adaptively select one candidate per video. Let $\hat{a}_t^{(\tau)}$ be the action at step t decoded using translation threshold τ and the fixed rotation threshold. We select the candidate with the highest action accuracy, measured as exact agreement with the input commands:

$$\tau^* = \arg\max_{\tau \in \Theta} \frac{1}{T} \sum_{t=1}^{T} [\![\hat{a}_t^{(\tau)} = a_t]\!]. \tag{12}$$

The selected threshold is held fixed throughout the video, and its decoded sequence $\hat{\mathcal{A}}^{(\tau^*)}$ is used for all action-accuracy metrics.

G.3 Discretization Algorithm

The command space $\mathbb{A}$ comprises eight primitives: `forward` (W), `backward` (S), `left` (A), `right` (D), `look_left` (J, yaw$-$), `look_right` (L, yaw$+$), `look_up` (I, pitch$+$), and `look_down` (K, pitch$-$), together with valid combinations of these primitives and `no_op`.

Algorithm 1 decodes a window into a label from $\mathbb{A}$, using its translation magnitude and direction, with signed pitch and yaw. Translation primitives are activated only above τ_{move} and within $60°$

[2] Official CompassReward code uses thresholds $\{0.002, 0.005, 0.01\}$ and selects the decoded sequence with the highest exact command agreement for the evaluated clip.

Table 9: Effective translation thresholds (already scaled by $K{=}4$), calibrated per pose source. The rotation threshold is $0.25°$ before scaling, i.e., $\tau_{\text{rot}}{=}1.0°$, for every source.

Pose source	Candidate τ_{move}			
`megasam_dav3`	0.0001	0.015	0.05	0.3
`megasam`	0.0001	0.008	0.03	0.3
`default`	0.0001	0.04	0.1	0.3
`dav3`	0.0001	0.03	0.1	0.3
`pi3x_moge2`	0.0001	0.04	0.12	0.3
`da3_direct`	0.0001	0.006	0.015	0.3

Input: $\Delta P \in \mathrm{SE}(3)$ ▷ relative camera pose
$\tau_{\text{move}} \in \mathbb{R}^+$ ▷ overall translation threshold
$\tau_{\text{rot}} \in \mathbb{R}^+$ ▷ rotation threshold (degrees)
Output: $\hat{a} \in \mathbb{A}$ ▷ predicted action label

$\Delta\mathbf{t} \leftarrow \Delta P[:3,3]$; $\nu \leftarrow \|\Delta\mathbf{t}\|_2$
$\mathbf{o}_t \leftarrow (0,0,0,0)$ ▷ (fwd, bwd, right, left)
if $\nu > \tau_{\text{move}}$ **then**
  $\hat{\mathbf{n}} \leftarrow \mathrm{clip}(\Delta\mathbf{t}/\nu, -1, 1)$
  $\boldsymbol{\theta} \leftarrow \arccos(\hat{\mathbf{n}}) \cdot 180/\pi$ ▷ angle to each axis
  if $\theta_z < 60°$: $\mathbf{o}_t[1] \leftarrow 1$ ▷ forward
  if $\theta_z > 120°$: $\mathbf{o}_t[2] \leftarrow 1$ ▷ backward
  if $\theta_x < 60°$: $\mathbf{o}_t[3] \leftarrow 1$ ▷ right
  if $\theta_x > 120°$: $\mathbf{o}_t[4] \leftarrow 1$ ▷ left
end if
$R \leftarrow \Delta P[:3,:3]$; $s_y \leftarrow \sqrt{R_{00}^2 + R_{10}^2}$
$(\phi, \psi) \leftarrow \mathrm{EulerPitchYaw}(R, s_y)$ ▷ gimbal-safe
$\mathbf{o}_r \leftarrow (0,0,0,0)$ ▷ (yaw+, yaw−, pitch+, pitch−)
if $\psi > +\tau_{\text{rot}}$: $\mathbf{o}_r[1] \leftarrow 1$ **else if** $\psi < -\tau_{\text{rot}}$: $\mathbf{o}_r[2] \leftarrow 1$
if $\phi > +\tau_{\text{rot}}$: $\mathbf{o}_r[3] \leftarrow 1$ **else if** $\phi < -\tau_{\text{rot}}$: $\mathbf{o}_r[4] \leftarrow 1$
$\ell_t \leftarrow \mathrm{OneHotToLabel}(\mathbf{o}_t)$ ▷ translation label
$\ell_r \leftarrow \mathrm{OneHotToLabel}(\mathbf{o}_r)$ ▷ rotation label
$\mathcal{S} \leftarrow \mathrm{TransParts}(\ell_t) \cup \mathrm{RotateParts}(\ell_r)$
$\hat{a} \leftarrow \bigoplus_{p \in \mathrm{sort}(\mathcal{S})} p$ ($\oplus$ = "+"); $\hat{a} \leftarrow$ `no_op` if $\mathcal{S} = \emptyset$
return $\hat{a}$

Algorithm 1: Window-level pose-to-action decoding.

of the corresponding signed camera axis; multiple active primitives form a compound label. Look primitives are activated when the associated angle exceeds τ_{rot} in magnitude. If no primitive is active, the label is `no_op`.

For rotation matrix R, the decoder uses $s_y = \sqrt{R_{00}^2 + R_{10}^2}$, yaw $\psi = \mathrm{atan2}(-R_{20}, s_y)$, and pitch $\phi = \mathrm{atan2}(R_{21}, R_{22})$. When $s_y \le 10^{-6}$, pitch instead uses $\mathrm{atan2}(-R_{12}, R_{11})$. Angles are converted to degrees; positive yaw denotes `look_right` and positive pitch denotes `look_up`.

Algorithm 2 combines windowed decoding with threshold selection to produce the frame-level prediction sequence. For Equation (1), we use $\mathcal{P}(\texttt{no_op}) = \{\texttt{no_op}\}$ so that two `no_op` labels also count as a partial match.

H TrajScore: Detailed Computation

Core idea. Above minimum-motion safeguards, the reference retains the model's displacement or rotation magnitude while imposing the ideal direction and shape prescribed by the action. It therefore answers: *"Given that the model moved this far, what is the ideal trajectory shape?"* For a `forward` action, the reference is a straight line in the commanded direction with the same endpoint displacement magnitude as the prediction, but no lateral drift. This evaluates geometric deviations without requiring a fixed movement distance or turning angle across models.

Input: $\hat{\mathcal{T}} = (\hat{P}_0, \ldots, \hat{P}_T)$ ▷ $T{+}1$ aligned camera poses
$\mathcal{A} = (a_1, \ldots, a_T)$ ▷ T GT action commands
$K = 4$ ▷ action-window size
Θ ▷ 4 fixed effective move candidates
$\tau_{\text{rot}} = 1.0^\circ$ ▷ effective rotation threshold
Output: $\hat{\mathcal{A}} = (\hat{a}_1, \ldots, \hat{a}_T)$ ▷ per-action predictions
τ^*, π ▷ selected threshold & sweep table

if $T < K$ **then return** $\langle \texttt{no_op} \rangle^T$, $\varnothing$, $\varnothing$
$C \leftarrow \lfloor T/K \rfloor$ ▷ number of complete windows
for $c = 0, \ldots, C{-}1$ **do**
  $s \leftarrow c \cdot K$; $e \leftarrow s{+}K$
  $\Delta P_c \leftarrow \hat{P}_s^{-1} \cdot \hat{P}_e$ ▷ relative pose across window c
end for
$(\hat{\mathcal{A}}^*, \alpha^*) \leftarrow (\emptyset, -\infty)$
for each $\tau_{\text{move}} \in \Theta$ **do**
  for $c = 0, \ldots, C{-}1$ **do**
    $\hat{b}_c \leftarrow \textsc{DiscretizeWindow}(\Delta P_c, \tau_{\text{move}}, \tau_{\text{rot}})$
  end for
  $\hat{\mathcal{A}} \leftarrow \text{RepeatEach}(\langle \hat{b}_c \rangle, K)$
  if $|\hat{\mathcal{A}}| < T$ **then** $\hat{\mathcal{A}} \leftarrow \hat{\mathcal{A}} \,\|\, \langle \hat{b}_{C-1} \rangle^{T-|\hat{\mathcal{A}}|}$
  $\alpha \leftarrow \frac{1}{T} \sum_{i=1}^{T} [\![\hat{a}_i = a_i]\!]$ ▷ exact-match accuracy
  $\pi[\tau_{\text{move}}] \leftarrow \alpha$
  if $\alpha > \alpha^*$: $(\hat{\mathcal{A}}^*, \alpha^*, \tau^*) \leftarrow (\hat{\mathcal{A}}, \alpha, \tau_{\text{move}})$
end for
$\text{Acc}_{\text{strict}} \leftarrow \alpha^*$
$\text{Acc}_{\text{partial}} \leftarrow \frac{1}{T} \sum_{i=1}^{T} [\![\mathcal{P}(\hat{a}_i^*) \cap \mathcal{P}(a_i) \neq \emptyset]\!]$
return $\hat{\mathcal{A}}^*$, τ^*, π

Algorithm 2: Video-level action decoding and threshold selection.

H.1 Adaptive Reference Construction

Action changes partition $\mathcal{A}$ into constant-command segments. The trajectory evaluator uses half-open slices: commands $a_{s+1}, \ldots, a_e$ define a segment whose scored poses are $\hat{P}_s, \ldots, \hat{P}_{e-1}$. Write $\hat{P}_j = (\hat{R}_j, \hat{\mathbf{p}}_j)$ for camera-to-world rotation and position.

Translation. For a pure movement command, let $\mathbf{u}_a$ be its unit direction in local camera coordinates and R_{world} the running reference orientation, initialized from the first scored pose. The reference is a straight line with direction $\mathbf{d}_a = R_{\text{world}}\mathbf{u}_a$ and displacement

$$d_{\text{ref}} = \begin{cases} \|\hat{\mathbf{p}}_{e-1} - \hat{\mathbf{p}}_s\|_2, & \|\hat{\mathbf{p}}_{e-1} - \hat{\mathbf{p}}_s\|_2 > 0.1, \\ 1.0, & \text{otherwise.} \end{cases} \tag{13}$$

For an isolated segment, its reference positions are

$$\mathbf{p}_j^{\text{ref}} = \hat{\mathbf{p}}_s + \frac{j - s}{\max(e - s - 1, 1)} \, d_{\text{ref}} \mathbf{d}_a, \quad j = s, \ldots, e - 1. \tag{14}$$

The orientation remains fixed. The fallback displacement prevents a nearly static prediction from defining a zero-length target.

Rotation and compound actions. Look commands define in-place rotation in FPV and an orbit about the subject in TPV; compound commands combine this rotation with the prescribed movement. For example, `forward+look_right` in FPV defines a curved path that advances while turning right. Reference rotation has the commanded direction and uses the predicted magnitude with a minimum of 30°. For TPV, subject depth determines the orbit radius: it is estimated from the predicted trajectory when predicted accumulated rotation is at least 5°, otherwise set to 1.0, with estimated depths also bounded below by 1.0.

Rotating reference segments are aligned to the initial predicted heading through $R_{\text{align}} = \hat{R}_s(R_s^{\text{ref}})^\top$. Their extent is matched using the ratio of mean radial extents, $\bar{r}_{\text{pred}}/\bar{r}_{\text{ref}}$; this scale is reset to 1 when predicted displacement or rotation falls below the minimum-motion thresholds.

Segment composition. Reference segments are translated to begin at the preceding reference endpoint. The running reference orientation is advanced after segments containing rotation. This constructs a continuous multi-action reference while preserving the commanded sequence.

H.2 Segment-wise Alignment

Two paths with the same geometry may traverse it at different speeds, producing apparent errors under frame-wise comparison. For segments with nonzero translational path length, we therefore match samples at equal fractions of the predicted and reference path lengths. Positions are linearly interpolated along cumulative arc length and orientations use spherical linear interpolation (Slerp).

For zero-length paths, including pure in-place rotation, we instead sample at uniformly spaced frame indices, retaining rotational timing.

We target 60 evaluation pairs across N_{seg} scored segments. Each segment with more than two poses receives

$$n_{\text{per}} = \max\left(\left\lfloor \frac{60}{N_{\text{seg}}} \right\rfloor, 4\right) \tag{15}$$

paired samples; shorter segments retain their available poses. This allocation normally yields $M = N_{\text{seg}} n_{\text{per}}$ pairs with equal segment weighting.

H.3 Geometric Errors and Normalization

For paired samples $(\hat{\mathbf{p}}_i, \hat{R}_i)$ and $(\mathbf{p}_i^{\text{ref}}, R_i^{\text{ref}})$, define

$$e_t^{(i)} = \|\hat{\mathbf{p}}_i - \mathbf{p}_i^{\text{ref}}\|_2, \qquad e_r^{(i)} = \angle(\hat{R}_i, R_i^{\text{ref}}). \tag{16}$$

The angle is the geodesic rotation difference in degrees. Averaging over all M pairs gives ATE_t and ATE_r.

For the same absolute error, a longer predicted path or greater accumulated rotation can yield a smaller normalized error. We therefore normalize by the predicted total path length L_{path} and accumulated rotation Φ_{total}, as in the main text, with denominator floors $\ell_{\min} = 0.5$ and $\phi_{\min} = 10^\circ$. Each normalized error is clipped to $[0, 1]$ before their average is subtracted from 1.

The denominator floors stabilize low-motion normalization, while reference-motion safeguards prevent trivial targets.

H.4 Numerical Example

Consider 300 consecutive `forward` action intervals, forming one segment. The path advances mainly along z, with lateral and vertical drift. Its start and end positions are

$$\hat{\mathbf{p}}_{\text{start}} = (-0.001,\ -0.000,\ -0.000), \qquad \hat{\mathbf{p}}_{\text{end}} = (0.002,\ -0.245,\ 6.509). \tag{17}$$

The endpoint displacement is 6.513, the path length is 6.603, and accumulated rotation is 2.72°. Since the displacement exceeds 0.1, the reference is a straight line of length 6.513 along the initial forward direction, approximately $(0, 0, 1)^\top$.

Resampling produces 60 pairs, with mean errors $\text{ATE}_t = 0.090$ and $\text{ATE}_r = 0.41^\circ$. Therefore,

$$\text{nATE}_t = \frac{0.090}{\max(6.603, 0.5)} \approx 0.014, \qquad \text{nATE}_r = \frac{0.41}{\max(2.72, 10)} = 0.041, \tag{18}$$

$$\text{TrajScore} = 1 - \frac{0.014 + 0.041}{2} \approx 0.973. \tag{19}$$

The high score reflects a nearly straight forward path with small position and heading deviations. It does not imply exact timing or speed agreement. Temporal command agreement is assessed separately by the discrete action metrics.

I Visual Quality Evaluation Details

Frame-level quality predictors. Aesthetic quality uses a CLIP-based linear predictor (Schuhmann et al., 2022): frozen CLIP ViT-L/14 features $\mathbf{z}_t = \mathrm{norm}(\mathrm{CLIP}(I_t))$ feed a head $g : \mathbb{R}^{768} \to \mathbb{R}$ trained on LAION-Aesthetics ratings, yielding $s_t^{\mathrm{aes}} = g(\mathbf{z}_t)/10$. MUSIQ (Ke et al., 2021), a reference-free multi-scale transformer with the SPAQ-pretrained checkpoint, assesses low-level imaging quality (e.g., sharpness, noise, and compression artifacts), yielding a per-frame score $s_t^{\mathrm{img}} = \mathrm{MUSIQ}(I_t)/100$. We average the normalized scores over all T frames.

Window construction. For per-frame aesthetic or imaging scores $\{s_t\}_{t=1}^{T}$ at frame rate f, a W-second window with an R-second stride contains $w = \mathrm{round}(fW)$ frames and advances by $r = \mathrm{round}(fR)$ frames. For the frame-index set $\mathcal{T}_i$ of window i, let $\bar{s}_i = |\mathcal{T}_i|^{-1} \sum_{t \in \mathcal{T}_i} s_t$. To cover the tail, the last window is $\mathcal{T}_{\mathrm{last}} = \{t : T - w < t \le T\}$; clips shorter than W use one full-clip window. We use $W = 3\,\mathrm{s}$ and $R = 1\,\mathrm{s}$ throughout, so $D = 0.05$ means the worst 3-second window scores 5% below the best.

Window and stride selection. The drift metric of Equation (5) has two free parameters, the window duration W and the stride R. This appendix reports the sweep that fixes $W = 3\,\mathrm{s}$ and $R = 1\,\mathrm{s}$. All numbers are measured on 9,460 evaluated rollouts covering the full duration range of our internal runs (from 15 s up to 3600 s, i.e. including rollouts far longer than the released 15/30/60 s buckets), so that both ends of the temporal scale are stressed. As a sensitivity reference we use the Spearman rank correlation η between the windowed drift and the earlier fixed-10-segment variant on the 15 s bucket, where the two definitions are expected to agree most closely. With ten equal segments, segment duration increases from 1.5 s for a 15 s rollout to 6 s for a 60 s rollout, potentially averaging away brief collapses in longer rollouts. On the 15 s bucket, a window near 1.5 s closely reproduces the segment-based ranking; this correlation measures agreement between the two definitions.

Window duration. Sweeping $W \in \{1.5, 2, 3, 5\}\,\mathrm{s}$ gives $\eta = 0.992$, 0.991, 0.977 and 0.932 respectively on the 15 s bucket. A 5 s window can dilute 1–2 s collapses through temporal averaging and yields lower rank agreement with the segment-based reference. At $W = 3\,\mathrm{s}$, the correlation is 0.977 and the mean drift is 0.142, compared with 0.165 for the reference. At the long end the choice matters much less—on the 3600 s bucket $W = 3\,\mathrm{s}$ yields 0.587 against 0.554 for $W = 5\,\mathrm{s}$, a $\sim$6% difference with no qualitative change—so the decision is driven by the short-clip regime. We prefer 3 s over 2 s despite the latter's higher correlation ($\eta = 0.991$): at 16 fps, a 3 s window averages 48 frames, reducing the contribution of isolated frame-level fluctuations while emphasizing *sustained* degradation. $W = 3\,\mathrm{s}$ is the compromise between temporal resolution and robustness.

Stride. With W at 3 s, disjoint windows ($R = W = 3\,\mathrm{s}$; 5 windows on a 15 s rollout) and sliding windows ($R = 1\,\mathrm{s}$) agree closely in aggregate: mean drift differs by only 3–7% and η between the two is 0.987–0.991. Despite this aggregate agreement, overlapping windows reduce sensitivity to window placement by mitigating the dilution of collapses split across disjoint-window boundaries. For example, on the most divergent sample the disjoint variant reports 0.468 while the sliding variant reports 0.674, a gap of 0.21 caused purely by a collapse straddling a window boundary and being split across two neighbouring means. The effect is stronger for larger windows—in an earlier comparison at $W = 5\,\mathrm{s}$, switching to a 1 s stride raised per-model mean drift by 6–27% and lifted one boundary-straddling case from 0.0038 to 0.0853. Since the sliding window set is a superset of the disjoint one, the sliding drift is by construction never smaller. The cost is negligible: a 3600 s rollout produces $\sim$3600 windows, each a mean over a list of scalars, evaluated in milliseconds. We therefore keep both choices, as they address orthogonal problems: reducing W from 5 s to 3 s fixes the temporal scale being too coarse for short rollouts, while overlapping windows prevent boundary-aligned collapses from being diluted in individual cases.

J Interaction Physics Evaluation Details

This appendix provides the implementation details omitted from the main interaction-physics section. The high-level goal remains to evaluate whether a generated rollout produces physically plausible consequences under user control. We first apply a validity gate to remove static or uncontrolled

generations, and then evaluate valid rollouts along three complementary domains: mechanics, optics, and 3D consistency. Unless otherwise noted, all VLM queries in this appendix are instantiated with Qwen3-VL-PLUS. Figure 16 shows the complete implementation pipeline.

J.1 Validity Gate Details

Motivation. Before physics is evaluated, a validity gate checks whether the rollout provides usable motion evidence for the corresponding view. Invalid cases receive a score of -1 and are excluded from aggregate physics statistics.

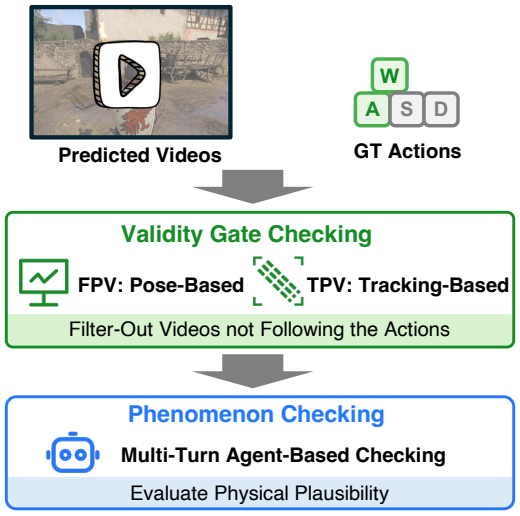


Figure 16: **Detailed interaction-physics evaluation pipeline for mechanics and optics.** The full pipeline includes action-responsiveness validation, mechanics protocols, optics protocols, Qwen3-VL-PLUS-based VLM scoring, fallback queries, and domain-level aggregation.

FPV validity. For FPV videos, we use a pose-based pre-filter. The camera trajectory is reconstructed with the benchmark pose pipeline, and for each prescribed translation segment we measure the forward displacement of the camera in its local frame; a case is marked invalid when this displacement falls below a per-pose-source calibrated threshold, i.e., the camera fails to respond to the commanded motion. Compared with an optical-flow direction check, the pose-based gate remains reliable in texture-poor scenes where flow estimates degrade.

TPV validity. For TPV videos, optical flow cannot reliably distinguish camera motion from subject motion. We therefore adopt a subject-centric validity check based on SAM 2 (Ravi et al., 2025), defined on tracking *continuity* rather than on displacement from the initial pose: a correctly controlled protagonist is expected to move away from where it started, so we penalize only discontinuities such as disappearance, teleportation, tracking loss, and abrupt scale change. This tracking-continuity criterion provides usable subject evidence rather than, by itself, certifying action fidelity. It is also used for subject memory (Appendix K.2), and the resulting video-level decisions are aggregated into R_{ctrl} for TPV leaderboard weighting (Section 2.6).

Concretely, we track and segment the subject every $\Delta_{\text{ctrl}} = 20$ frames and extract an axis-aligned bounding box B_{t_i} for each sampled frame. For each pair of adjacent sampled frames, we compute

$$\text{IoU}_{t_i} = \frac{|B_{t_{i-1}} \cap B_{t_i}|}{|B_{t_{i-1}} \cup B_{t_i}|}, \qquad i = 2, \ldots, N. \tag{20}$$

A sampled frame in which no subject mask is found contributes zero IoU. For a video V, the tracking-continuity score $c(V)$ is the fraction of adjacent sampled-frame transitions satisfying $\text{IoU}_{t_i} \geq \tau_{\text{iou}}$:

$$c(V) = \frac{1}{N-1} \sum_{i=2}^{N} \mathbf{1}[\text{IoU}_{t_i} \geq \tau_{\text{iou}}]. \tag{21}$$

A video passes validity when $c(V) \geq \tau_{\text{ctrl}}$. We set $\tau_{\text{iou}} = 0.4$ and $\tau_{\text{ctrl}} = 0.9$ throughout. In principle, continuity and commanded motion need not coincide; however, a per-segment speed audit with the pose-pipeline maintainer found no such divergence among the evaluated rollouts. When masks are unavailable, a lightweight Qwen3-VL-PLUS query over three frames checks the expected motion direction as a substitute action check, rather than as a continuity estimate.

J.2 Physics Protocol Details

J.2.1 Mechanics Details

Mechanics is evaluated through four protocol families: collision, clipping, deformation, and gravity. Ground-contact adherence on slopes and steps (terrain following) is scored within the gravity family but retains its own rubric (Appendix J.13), matching the protocol taxonomy of Figure 7. Unless otherwise specified, each protocol returns a binary score, where 1 indicates a physically plausible outcome and 0 indicates failure.

Two evidence regimes. FPV and TPV rollouts expose different evidence. In FPV, the reconstructed camera trajectory is the primary signal: after the validity pre-filter, a bird's-eye-view protocol checks whether the trajectory exhibits the environmental influence expected under the prescribed action—lateral deviation toward the expected side, contact with a lateral barrier, or disturbance on either side—and pass candidates are re-checked by Qwen3-VL-PLUS for scene degradation before being accepted. In TPV, trajectory signals cannot localize subject–object events, so each criterion is judged by an evidence-gathering agent loop.

TPV agent loop. The agent receives the case objective (criterion plus scene description) together with a coarse overview timeline of the rollout. It then iterates for at most four rounds: localize the critical event window, call an inspection tool, and either gather more evidence or emit a structured verdict. Two tools are exposed: `inspect_interval(start, end, fps, crop)`, which densely samples a time window at 1–8 fps, and `inspect_frames(timestamps, crop)`, which retrieves exact frames; the crop argument (`full`, `center`, `lower`, `upper`) zooms into the interaction region at up to 768 px. The verdict is a structured object containing the label, a confidence, the event interval, and timestamped evidence observations, which makes every score auditable. Across TPV mechanics protocols, cases assigned a final `INVALID` verdict are excluded from aggregation, with averages computed over the remaining valid cases.

Per-criterion rubrics. Each criterion instantiates the same loop with a one-line evidence requirement: collision compares the target's position and pose before and after first contact; clipping tests whether the subject ever traverses a solid obstacle (body–obstacle interpenetration, the obstacle vanishing at the subject's position, or side-switching without a passage that visibly pre-exists contact); deformation compares the contacted surface before and after passage (persistent traces) or during contact (real-time ripples and crop parting); gravity requires a continuous descent or submersion response at edges and water entries without demanding free-fall signatures, and, for its terrain-adherence subfamily, continuous height and foot contact over slopes and steps.

In-loop evidence audit. Before a verdict is committed, the loop applies a final audit step: clipping pass verdicts are re-examined by an independent crossing reviewer that checks obstacle-topology continuity around the contact window, and fail verdicts on interaction criteria receive one high-recall positive-evidence pass that searches for subtle missed responses (faint traces, small splashes, partial displacement). The audit is part of the loop rather than a separate pipeline stage, so its evidence enters the same structured verdict.

J.2.2 Optics Details

Optics is evaluated through two protocols: reflection and occlusion, where occlusion is instantiated as shadow correctness. These protocols differ in score granularity.

Reflection. Reflection uses a two-tier evaluation. Tier 1 is a coarse global check. We uniformly sample $N = 12$ frames at 384 px resolution and ask the Qwen3-VL-PLUS VLM whether reflections are globally plausible: consistent in shape, approximately symmetric about the reflecting surface, and changing naturally with camera or subject motion. A failure at Tier 1 yields a score of 0.

If Tier 1 passes, Tier 2 performs fine-grained per-frame scoring. We sample $N = 8$ frames and pair each sampled frame with the first frame from the test case as reference. The Qwen3-VL-PLUS VLM assigns each frame an integer score:

$$s_t \in \{-1,\ 0,\ 1,\ 2,\ \mathrm{N/A}\}, \tag{22}$$

where 2 denotes a physically correct reflection, 1 indicates partial errors, 0 means the reflecting surface is visible but the reflection is missing or wrong, -1 marks a collapsed scene, and N/A indicates that the reflecting surface has left the field of view. Collapsed (-1) frames are excluded; N/A frames are instead scored by a surface-agnostic optical-quality query on the same 0–2 scale (normal $= 2$, degraded $= 0$) and enter $\mathcal{V}$, so surface absence alone is not treated as a reflection failure. The reflection score is

$$S_{\text{refl}} = \frac{1}{2|\mathcal{V}|} \sum_{t \in \mathcal{V}} s_t, \qquad \mathcal{V} = \{t : s_t \geq 0\}, \tag{23}$$

yielding a continuous score in $[0, 1]$. A per-case degradation re-check samples start, mid, and end frames and asks Qwen3-VL-PLUS whether the rollout has visually collapsed; a degraded verdict overrides the reflection score to 0 irrespective of the tier outcomes.

Occlusion and shadow correctness. The occlusion protocol evaluates whether shadows appear and behave consistently with the scene lighting. All reported shadow cases are TPV and use $N = 10$ samples from the 10–100% range; the FPV configuration samples six frames from 40–100% and prepends the first frame as a reference. The Qwen3-VL-PLUS VLM judges whether the expected shadow is present and whether its direction and shape are physically consistent with the light source. Scoring is binary.

J.2.3 3D Consistency Details

3D consistency evaluates whether the generated scene preserves rigid spatial structure under controlled camera motion. In the current benchmark, these cases are primarily instantiated in FPV settings with long, unobstructed trajectories and a constant input rate. This design makes geometric distortions, inconsistent depth ordering, and non-rigid scene deformation easier to detect with reduced confounding effects from collisions, deformation, or terrain changes. In implementation, propagated landmarks define per-region interest masks on reference crops; each sampled predicted frame is compared with its matched reference by a perceptual (LPIPS) distance within each region. A two-threshold calibration then maps per-region distances to scores: distances below a rejection threshold indicate retrievable mild distortion and are stretched into the upper score range, while distances in an intermediate collapse band score 0. Regions above the upper threshold are excluded when correspondence to the reference is unreliable (e.g., reference content is no longer visible or appearance divergence is extreme), reducing the frame denominator rather than contributing zero. This exclusion is heuristic: high perceptual distance may also reflect disappearance or severe distortion, and validating it against ground-truth visibility remains future work. Matching uses rendered landmark crops and perceptual similarity rather than model trajectory scale; views too far from every reference enter the unreliable-correspondence regime. The resulting per-case scores are treated as a third physics domain when available.

J.3 Prompt Design Principles

Across all protocols, prompt construction follows four design principles derived from iterative calibration against human judgments.

Scene-grounded context. Each agent run begins from the scene description in the test-case metadata, giving the judge the expected scenario without imposing rigid subject assumptions (the subject may be a person, vehicle, or animal).

Evidence before verdict. Verdicts must cite timestamped observations collected through the inspection tools; a pass without positive visual evidence, or a fail without a violated-constraint observation, is not accepted. This replaces one-shot whole-video questions, which are sensitive to temporal aliasing in sparse uniform samples.

Adaptive resolution and windowing. The agent chooses both the temporal window and the spatial crop per query: coarse full-frame overviews for localization, dense cropped samples (up to 768 px) for the interaction region. Fine-grained criteria such as deformation therefore receive the resolution their evidence requires without paying it globally.

In-loop audit instead of retries. Rather than re-issuing the same query on a different frame range, the loop ends with a targeted audit (crossing review for clipping passes, high-recall positive-evidence search for interaction fails). The audit consumes the same structured evidence and its outcome is recorded in the verdict trace.

J.4 Physics Score Details

Per-case scores are aggregated into a physics composite. Because validity filtering is model dependent, the resulting domain scores are conditional on different accepted subsets; TPV Control Rate is a complementary overall coverage indicator, not a replacement for identical per-domain coverage. Let $\mathcal{C}_{\text{mech}}$, $\mathcal{C}_{\text{opt}}$, and $\mathcal{C}_{\text{3d}}$ denote the sets of valid mechanics, optics, and 3D-consistency cases, respectively. The domain-level scores are

$$S_{\text{mech}} = \frac{1}{|\mathcal{C}_{\text{mech}}|} \sum_{c \in \mathcal{C}_{\text{mech}}} s_c, \qquad S_{\text{opt}} = \frac{1}{|\mathcal{C}_{\text{opt}}|} \sum_{c \in \mathcal{C}_{\text{opt}}} s_c, \tag{24}$$

with S_{3d} defined analogously when $\mathcal{C}_{\text{3d}}$ is available. Here, $s_c \in [0, 1]$ is the per-case score. The overall physics score averages all available physics domains:

$$S_{\text{physics}} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} S_d, \qquad \mathcal{D} \subseteq \{\text{mech}, \text{opt}, \text{3d}\}. \tag{25}$$

The available domains are weighted equally because they test complementary aspects of physical plausibility: mechanics evaluates interaction-level consequences, optics evaluates light-level consistency, and 3D consistency evaluates spatial coherence. FPV averages mechanics, optics, and 3D consistency, whereas TPV averages mechanics and optics only. Physics scores are therefore compared within each view.

J.5 Collision Evaluation Details

Evidence-localized verification. The agent localizes first contact between the subject and the target object, then compares the target's position and pose before and after that moment. Target identity and expected response come from the test-case metadata. The design deliberately avoids requiring the judge to witness the exact contact frame: displacement, tumbling, or deformation of the target after the localized contact window is sufficient evidence, which suppresses both false positives from rollouts that never reach the target and false negatives from brief responses.

Scoring. Binary: $\text{score}{=}1$ if the target shows a physically plausible response and remains observable, else 0; validity-gate failures are excluded with $\text{score}{=}{-}1$. Cases assigned an `INVALID` verdict by the agent are excluded from score aggregation.

J.6 Collision Evaluation Prompt

The collision agent instantiates the shared rubric, tool protocol, and structured verdict format of Figure 24 with the criterion-specific evidence requirement described above.

J.7 Clipping Evaluation Details

Crossing-centric detection. Clipping (pass-through of solid obstacles) is scored inversely: detecting a crossing yields 0. The rubric defines a crossing as any of: body–obstacle interpenetration; the obstacle vanishing or being replaced at the subject's position; or the subject appearing on the far side of the obstacle without a passage that was visibly present before contact. Static subjects and subjects blocked at the obstacle are passes, which keeps the criterion purely geometric and independent of action responsiveness (handled by the validity gate).

In-loop crossing audit. Every pass verdict triggers the loop's final audit: an independent reviewer re-inspects the contact window and the post-contact segment for topology-change artifacts (obstacle disappearance, sudden side switches). Only audits that confirm a pre-existing passage or continuous same-side evidence keep the pass.

Scoring. score=0 if a crossing is confirmed, else 1; validity-gate failures yield score=−1.

J.8 Clipping Evaluation Prompt

The clipping agent instantiates the shared rubric, tool protocol, and structured verdict format of Figure 24 with the criterion-specific evidence requirement described above.

J.9 Deformation Evaluation Details

Before/after surface comparison. Deformation requires the right evidence window rather than a fixed sampling rule: the agent localizes when the subject contacts the deformable surface, then compares that surface before contact (reference frames) with frames after passage for persistent traces (footprints, tire tracks), or during contact for real-time responses (ripples, splashes, crop parting). Crops of the contact region at up to 768 px preserve the fine detail these effects live in.

Leniency with guards. Faint, partial, or short-lived responses count as passes, but illumination changes, background texture, and camera motion are explicitly excluded as evidence, which prevents texture hallucination from scoring as deformation.

Scoring. Binary: score=1 if a trace or real-time response is evidenced, else 0.

J.10 Deformation Evaluation Prompt

The deformation agent instantiates the shared rubric, tool protocol, and structured verdict format of Figure 24 with the criterion-specific evidence requirement described above.

J.11 Gravity Evaluation Details

Descent-evidence rubric. The terrain-adherence subfamily (slopes, steps) is detailed in Appendix J.13.

Gravity cases vary (falls from edges, sinking into water, projectiles), so the rubric states the expected descent signature per scenario and accepts continuous height decrease, progressive submersion, or arc-like motion as evidence; free-fall acceleration is not required, since third-person cameras rarely resolve it. Absence of any descent or submersion over the localized window yields failure.

Scoring. Binary: score=1 if the descent response is evidenced, else 0.

J.12 Gravity Evaluation Prompt

The gravity agent instantiates the shared rubric, tool protocol, and structured verdict format of Figure 24 with the criterion-specific evidence requirement described above.

J.13 Terrain Following Evaluation Details

Terrain following is reported under the gravity family in Figure 7; we retain a separate rubric because its evidence (continuous ground contact over elevation changes) differs from free descent.

Ground-contact continuity. The agent localizes slope, step, or uneven segments and verifies that the subject's height and foot contact change continuously with the terrain, rather than floating above or sinking into the surface. Elevation changes themselves are natural and are not penalized; the failure signatures are hovering, ground-level camera collapse, and view-through-surface artifacts.

Scoring. Binary: score=1 if ground contact remains plausible throughout the localized segment, else 0.

J.14 Terrain Following Evaluation Prompt

The terrain agent instantiates the shared rubric, tool protocol, and structured verdict format of Figure 24 with the criterion-specific evidence requirement described above.

J.15 Reflection Evaluation Details

Two-tier coarse-to-fine evaluation. Reflection is the only VLM-based physics protocol producing a continuous score. The evaluation proceeds in two tiers:

- **Tier 1 (coarse)**: 12 uniformly sampled frames are presented to the Qwen3-VL-PLUS VLM in a single call. It judges global reflection plausibility: are reflections consistent in shape, approximately symmetric about the reflecting surface, and changing naturally with camera movement? A "no" verdict short-circuits to a score of 0.
- **Tier 2 (fine)**: 8 frames are sampled and each is scored independently (parallelized with 4 workers) against the first-frame reference. The Qwen3-VL-PLUS VLM assigns a per-frame score of -1 (invalid), 0 (missing/wrong), 1 (physics errors), 2 (correct), or N/A (surface not visible).

Per-frame scoring rubric. The Tier 2 prompt assigns N/A when the reflecting surface has moved out of view, rather than 0 for an incorrect reflection on a visible surface. Before aggregation, N/A frames receive surface-agnostic optical-quality scores on the same 0–2 scale (normal $= 2$, degraded $= 0$) and are included in $\mathcal{V}$.

Aggregation. The final score is:

$$S_{\text{refl}} = \frac{1}{2 \cdot |\mathcal{V}|} \sum_{i \in \mathcal{V}} s_i, \qquad \mathcal{V} = \{i : s_i \geq 0\}, \tag{26}$$

where $s_i \in \{0, 1, 2\}$ are per-frame scores; collapsed (-1) frames are excluded, while N/A frames contribute surface-agnostic optical-quality scores on the same scale. The result is a continuous value in $[0, 1]$.

Degradation re-check. A per-case re-check samples start, mid, and end frames and asks Qwen3-VL-PLUS whether the rollout has visually collapsed (black frames, heavy artifacts, nonsensical imagery). A degraded verdict overrides the tiered reflection score to 0, preventing collapsed rollouts from collecting partial credit on the few frames that remain interpretable.

Surface-aware prompting. The reflecting-surface type (`water_puddle`, `water`, `canal`, `mirror`, `glass`) is injected into both tiers so that the judge applies the specular-reflection expectation appropriate to that surface.

J.16 Reflection Evaluation Prompt

The reflection prompts implement the coarse global plausibility check and the per-frame scoring rubric; both are shown in Figure 25.

J.17 Occlusion / Shadow Evaluation Details

Temporal sampling with reference frame. Shadow evaluation uses view-specific temporal ranges: FPV samples six frames from 40–100% and prepends the first frame as a visual reference, whereas TPV samples ten frames from 10–100%. All shadow cases reported here are TPV and use the latter configuration.

Strict shadow recognition criteria. The prompt enforces strict criteria to distinguish real shadows from ambient darkening:

- A real shadow must have a *recognizable silhouette* matching the casting object (e.g., a person shadow looks like a person outline).

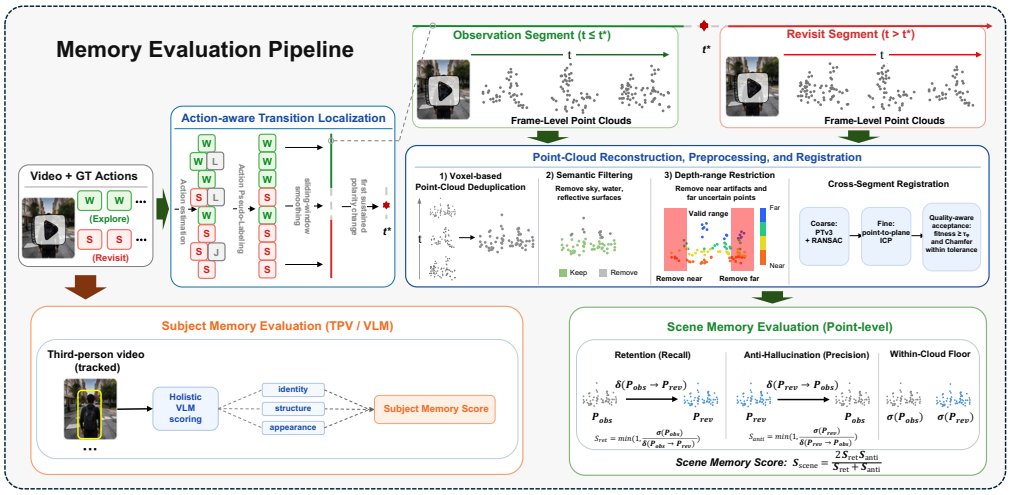


Figure 17: **Detailed memory evaluation pipeline.** The full pipeline includes action-aware transition localization, point-cloud reconstruction, registration, filtering, scene-memory scoring, subject tracking, controllability gating, and Qwen3-VL-PLUS-based VLM subject-memory evaluation.

- The shadow must have a *clear boundary* separating shadowed from lit regions—vague dark patches, ambient shading, surface texture, or lighting gradients do *not* count.
- The shadow direction must be physically consistent with visible light sources.

The Qwen3-VL-PLUS VLM is instructed to answer "yes" *only* if confident the shadow is real and correct. This deliberately strict, visibility-oriented rubric favors unambiguous attributable shadows; physically plausible soft shadows from extended or scattered illumination may not satisfy it, so these criteria are evaluation anchors rather than universal requirements on shadow appearance.

Expected-shadow injection. When `eval_protocol_args.expected_shadow` is provided (e.g., "wall shadow," "person shadow on the ground"), it is injected into the prompt to specify what the Qwen3-VL-PLUS VLM should look for. Without it, a generic shadow detection prompt is used.

Scoring. Binary: score=1 if the expected shadow is present and behaves correctly, else 0.

J.18 Occlusion / Shadow Evaluation Prompt

The occlusion/shadow prompts include expected-shadow checks, third-person fallback checks, and a generic fallback for scenes without explicit shadow metadata, as shown in Figure 26.

K Memory Evaluation Details

This appendix provides the implementation details omitted from the main memory section. Scene memory aggregates geometry over observation and revisit segments, while subject memory evaluates whether the protagonist is preserved in controllable TPV rollouts. Figure 17 shows the complete implementation pipeline.

K.1 Scene Memory Details

Action-aware transition localization. We localize the observation–revisit transition from the generated video rather than using the symmetric GT index. Given the video and its GT action sequence, CompassReward estimates a frame-level action for every generated frame. Since the estimator may output compound actions, such as `W+L`, we convert predictions into action pseudo-labels in the GT action space. For each predicted action, we search the set of GT actions occurring in the rollout and assign the first GT action whose components are all contained in the prediction;

unmatched predictions receive no pseudo-label. The resulting pseudo-label sequence is temporally smoothed, and the transition point t^* is identified as the first stable change point from the observation action to the revisit action. This procedure aligns the memory split with the model's executed action transition and reduces contamination from compound or noisy action predictions.

Point-cloud reconstruction and preprocessing. From the observation segment $\{I_t\}_{t\le t^*}$ and revisit segment $\{I_t\}_{t>t^*}$, subsampled at a fixed frame stride, we reconstruct point clouds $\mathcal{P}_{\text{obs}}$ and $\mathcal{P}_{\text{rev}}$ using monocular depth estimates and camera intrinsics. To reduce depth-estimation noise, particularly in distant and textureless regions, we apply the following preprocessing steps.

Voxel-based deduplication. We average points within each voxel of an adaptive grid, reducing redundancy while preserving scene structure.

Semantic filtering. We use a pretrained SegFormer to remove categories with unreliable depth, such as sky, water, and reflective surfaces.

Depth-range restriction. We retain only points whose estimated depth falls inside a scene-dependent absolute range, discarding both near-field artifacts and distant regions with high depth uncertainty.

Subject removal (TPV only). In third-person rollouts the protagonist is itself a large moving object, so its geometry would appear in one segment and not the other and be charged as scene inconsistency. We therefore propagate the annotated subject mask with SAM 2 and exclude the tracked subject from both segment clouds, leaving only the static scene. This is the sole difference between the FPV and TPV scene-memory pipelines.

Cases with failed 3D reconstruction or insufficient points remaining after the above filtering are excluded from scene-memory scoring. Table 10 reports the valid FPV scene-memory counts.

Table 10: **FPV scene-memory sample counts.** Counts are valid/total after the above exclusions. Horizon-specific scores average valid cases within each split; leaderboard scores pool the same valid cases across splits before rounding.

Model	Short	Mid	Long	All
Genie 3	35/36	37/37	29/30	101/103
Happy Oyster	36/36	37/37	30/30	103/103
Lyra 2.0	35/36	37/37	29/30	101/103
HY-World 1.5	34/36	37/37	30/30	101/103
LingBot-World 2.0	36/36	37/37	30/30	103/103
Matrix-Game 3.0	35/36	37/37	29/30	101/103
SANA-WM-AR	36/36	37/37	30/30	103/103
LingBot-World	35/36	37/37	30/30	102/103
SANA-WM	36/36	37/37	30/30	103/103
Matrix-Game 2.0	36/36	36/37	30/30	102/103
Yume 1.5	35/36	37/37	30/30	102/103
minWM	36/36	37/37	26/30	99/103

Cross-segment point-cloud registration. The coordinates of $\mathcal{P}_{\text{obs}}$ and $\mathcal{P}_{\text{rev}}$ are determined by the estimated camera-pose chain. Pose drift can introduce a rigid displacement unrelated to memory, artificially increasing apparent forgetting and hallucination. We therefore register $\mathcal{P}_{\text{rev}}$ to the reference $\mathcal{P}_{\text{obs}}$ using coarse-to-fine alignment. We estimate the transformation once from semantically filtered, scene-range-restricted, voxel-deduplicated clouds and reuse it for the scoring clouds.

Coarse registration. When the initial misalignment, measured as the maximum of the inter-centroid displacement and symmetric Chamfer distance, exceeds τ_c, we perform global registration. We compute per-point descriptors using Point Transformer V3 (Wu et al., 2024) features. Nearest-neighbor matching in descriptor space produces candidate correspondences, from which RANSAC estimates a geometrically consistent rigid transformation.

Fine registration. Starting from the coarse estimate, or the identity transformation when global registration is unnecessary, we refine the alignment using point-to-plane Iterative Closest Point (ICP) (Chen & Medioni, 1992).

Quality-aware acceptance. To reject degenerate registrations, we apply a dual acceptance criterion. Let T^* be the estimated transformation, $\phi(T)$ the ICP fitness, and $\mathrm{CD}(T)$ the symmetric Chamfer distance after applying T. We use

$$T = \begin{cases} T^*, & \text{if } \phi(T^*) \ge \tau_\phi \text{ and } \mathrm{CD}(T^*) \le (1+\epsilon)\,\mathrm{CD}(I), \\ I, & \text{otherwise,} \end{cases} \tag{27}$$

where τ_ϕ is the minimum acceptable fitness and I is the 4×4 identity. Thus, registration is accepted only if it achieves sufficient fitness without increasing the Chamfer distance by more than a small relative tolerance ϵ. All subsequent metrics are computed on the aligned pair $(\mathcal{P}_{\text{obs}},\ T \circ \mathcal{P}_{\text{rev}})$; for brevity, we use $\mathcal{P}_{\text{rev}}$ to denote the aligned revisit cloud below.

Density-normalized retention and hallucination. We quantify geometric memory from nearest-neighbor distances between the aligned point clouds, normalized by the sampling resolution of the clouds themselves rather than by an absolute distance threshold. Writing $d(\mathbf{p}, \mathcal{Y})$ for the distance from $\mathbf{p}$ to its nearest neighbor in $\mathcal{Y}$, we define the sampling floor of a cloud $\mathcal{X}$ and the directed cross-segment discrepancy from $\mathcal{X}$ to $\mathcal{Y}$ as

$$\sigma(\mathcal{X}) = \frac{1}{|\mathcal{X}_1|} \sum_{\mathbf{p}\in\mathcal{X}_1} d(\mathbf{p},\, \mathcal{X}_2), \qquad \delta(\mathcal{X}\to\mathcal{Y}) = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{p}\in\mathcal{X}} d(\mathbf{p},\, \mathcal{Y}). \tag{28}$$

The sampling floor $\sigma(\mathcal{X})$ is estimated by drawing two *disjoint* random subsets $\mathcal{X}_1, \mathcal{X}_2 \subset \mathcal{X}$ of at most N_{split} points each under a fixed seed, and averaging $d(\mathbf{p}, \mathcal{X}_2)$ over $\mathbf{p} \in \mathcal{X}_1$. Because the subsets are disjoint, no point can match itself, so σ measures the characteristic spacing between independent samples of the same surface. The directed cross-segment term δ of Equation (28) is computed on subsamples of at most $N_{\max}$ points. The two scores are the clamped ratios of Equation (8), combined by the harmonic mean defined in the scene-memory protocol. By analogy with precision and recall, the direction $\mathcal{P}_{\text{obs}} \to \mathcal{P}_{\text{rev}}$ asks how closely the geometry that should have been remembered is recovered and hence yields the recall term S_{ret}, while the direction $\mathcal{P}_{\text{rev}} \to \mathcal{P}_{\text{obs}}$ asks how consistently the produced geometry is supported by prior observation and hence yields the precision term S_{anti}. The results table reports S_{anti} directly, so that all three memory columns share the same orientation and the reported sub-scores are exactly the terms entering the harmonic mean. A clamped score of 1 indicates that the cross-segment discrepancy is at or below the estimated sampling floor; it does not imply identical point clouds. 20.2% of first-person cases saturate.

K.2 Subject Memory Details

Third-person controllability filtering. Before evaluating subject memory, we filter out test cases with insufficient subject-tracking continuity. When control degenerates, the camera collapses toward a first-person view and produces characteristic artifacts: the subject grows or shrinks under translation actions (W/S) and disappears then reappears under yaw rotation actions (J/L).

To assess tracking continuity, we use SAM 2 (Ravi et al., 2025) to track and segment the subject every $\Delta_{\text{ctrl}} = 20$ frames and extract an axis-aligned bounding box B_{t_i} for each sampled frame. For each pair of adjacent sampled frames, we compute

$$\text{IoU}_{t_i} = \frac{|B_{t_{i-1}} \cap B_{t_i}|}{|B_{t_{i-1}} \cup B_{t_i}|}, \qquad i = 2, \ldots, N. \tag{29}$$

A sampled frame in which no subject mask is found contributes zero IoU. A rollout passes the gate when at least 90% of the $N-1$ adjacent sampled-frame transitions satisfy $\text{IoU}_{t_i} \geq \tau_{\text{iou}} = 0.4$; cases that fail are excluded from subject-memory evaluation.

Failure modes. We characterize subject memory failure using five failure modes and one benign factor: identity change, structural distortion, appearance drift, subject disappearance, quality degradation, and benign viewpoint change. The evaluator must distinguish true subject-memory failure from benign variation caused by viewpoint, illumination, or shadow.

Sampling and subject extraction. Given a generated third-person video $V = \{I_1, I_2, \ldots, I_T\}$, we sample frames uniformly with stride $\Delta = 20$, yielding indices $\mathcal{S} = \{t_1, t_2, \ldots, t_N\}$, where $t_1 = 1$. Each test case provides a first-frame subject mask, which initializes SAM 2 (Ravi et al., 2025) for mask propagation over the sampled sequence. Frames without valid subject tracking are excluded. We extract each retained subject observation as

$$O_{t_i} = \text{crop}(I_{t_i} \odot M_{t_i}), \qquad t_i \in \mathcal{S}, \tag{30}$$

where $\odot$ denotes element-wise masking and $\text{crop}(\cdot)$ isolates the masked subject region. Suppressing the background prevents the judge from using scene content as an identity cue.

Holistic VLM evaluation. We present the temporally ordered sequence $\mathcal{O} = \{O_{t_1}, \ldots, O_{t_N}\}$ to the Qwen3-VL-PLUS VLM in a *single* deterministic request, with the first retained crop acting as the reference and all remaining crops following in temporal order. The judge evaluates the sequence holistically rather than through pairwise frame comparisons, producing an integer subject memory score $S_{\text{VLM}} \in \{1, \ldots, 10\}$ with binary diagnostic flags. Appendix L provides the full prompt.

We linearly normalize the integer rating to $[0, 1]$:

$$S_{\text{sub}} = \frac{S_{\text{VLM}} - 1}{9}. \tag{31}$$

Thus, a rating of 1 maps to 0, and a rating of 10 maps to 1. Holistic scoring accommodates smooth viewpoint and illumination changes that can confound frame-level comparisons, while producing a single benchmark-compatible scalar without requiring an external reference-feature library.

L Subject Memory Prompt

For third-person memory evaluation, the Subject Memory Evaluation Prompt in Figure 27 scores video-level subject memory from temporally ordered subject crops. The prompt emphasizes holistic temporal judgment, separates viewpoint and lighting changes from true identity drift, and explicitly penalizes structural collapse, disappearance, and severe quality degradation.

M Why Not Frame-Pair Memory?

A natural alternative to scene-level memory is to directly compare observation and revisit frames at corresponding spatial locations. We instantiate this frame-pair approach by first reusing the action-aware transition localization step from the scene-memory protocol to obtain the transition frame t^*, and then constructing geometrically corresponding pairs around t^* via *pose-aware frame pair construction*, described below. Even with t^* given by the same procedure that anchors our scene-level pipeline, frame-pair evaluation remains insufficient for the reasons we explain at the end of this appendix. Figure 18 visualizes a representative failure case.



Figure 18: **Failure mode of frame-pair memory evaluation.** Frame-pair metrics assume observation and revisit frames depict the same location. In long-horizon rollouts, imperfect action execution can shift the revisit trajectory, so paired frames correspond to different viewpoints or scene regions. Image-level discrepancies then mix memory degradation, action-following error, and pose-estimation noise, making frame-pair scores unreliable indicators of scene memory.

Pose-aware frame pair construction. Rather than pairing frames by symmetric index around t^*, camera poses estimated by the pose estimation pipeline are used to identify geometrically corresponding revisit frames. For each observation frame f_i, let $\hat{f}_i$ denote its temporally symmetric counterpart with respect to t^*. The revisit frame is selected by minimizing the geodesic distance to the pose of f_i within a local temporal window:

$$f_i^* = \arg \min_{f_j \in \mathcal{N}(\hat{f}_i)} d\big(P(f_i), P(f_j)\big), \tag{32}$$

where $P(\cdot)$ denotes the estimated camera pose and $d(\cdot, \cdot)$ is a pose distance metric combining translational and rotational components.

This construction offers two advantages: (i) evaluating memory on the most geometrically aligned pair minimizes contamination from incomplete revisits, and (ii) restricting the search

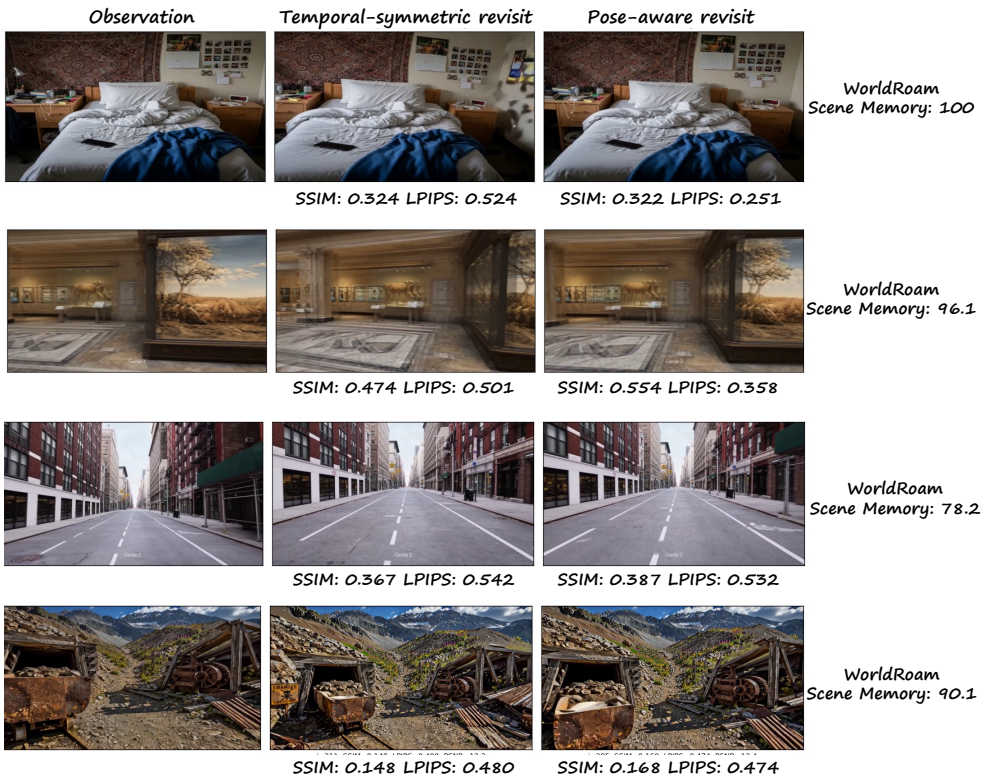


Figure 19: **Frame-pair similarity can underestimate preserved scene memory.** Each row shows the observation frame, its temporally symmetric revisit $\hat{f}_i$, and the pose-aware revisit f_i^*, with SSIM (↑) / LPIPS (↓) below and S_{scene} at right. Pose-aware pairing generally improves over symmetric pairing in these examples, illustrating how spatial misalignment affects frame-pair scores; yet substantial image-level differences remain (SSIM ≤ 0.554, LPIPS ≥ 0.251) despite visibly preserved content. These examples suggest frame-pair discrepancies need not indicate scene-memory loss; the relatively high S_{scene} values are consistent with visibly preserved scene structure.

to a neighborhood of the symmetric index preserves the trajectory prior and mitigates the effect of noisy pose estimates.

Given the constructed pairs $\{(f_i, f_i^*)\}$, standard image-level fidelity metrics are reported: MSE, PSNR, SSIM, and LPIPS.

Limitations of frame-pair memory evaluation. Despite the above mitigations, frame-pair metrics remain unreliable indicators of memory capability under the action-following quality exhibited by current world models. Three fundamental issues persist.

Systematic spatial offset from inconsistent action magnitudes. Even when the transition frame t^* is correctly identified, forward and backward actions often traverse different distances. This creates a systematic spatial displacement between the observation and revisit segments that no temporal re-indexing can correct. Pose-aware pairing alleviates this issue to some degree, but the search is limited to a local neighborhood and cannot compensate for large cumulative offsets.

Trajectory deviation from spurious motion. Certain models exhibit unprompted camera movements that cause the generated trajectory to diverge substantially from the prescribed path. In such cases, the revisit segment may explore entirely different regions of the scene, rendering frame-level comparison unreliable regardless of the pairing strategy.

Compounding of control jitter and pose estimation noise. For closed-source models accessed through web interfaces, network latency introduces temporal uncertainty in action execution, while



Figure 20: **Action gap in closed-source models. Top:** Genie 3. When the D key (rightward translation) is issued, the model rotates the camera to the right instead of translating the character rightward within a stable scene. The character remains centered while the entire scene rotates, making controlled navigation impossible. **Bottom:** Happy Oyster. When the D key (rightward translation) is issued, the model faithfully enforces collision with a newly generated obstacle, preventing the character from moving rightward. This strong adherence to physics results in lower action-following scores. Each row shows frames from left to right in temporal order.

the pose estimator itself contributes additional noise. These two error sources compound: noisy poses lead to suboptimal frame pair selection, and control jitter ensures that even correctly selected pairs depict slightly different viewpoints, both inflating the measured dissimilarity.

As a result, models that visibly preserve scene content upon human inspection can receive misleadingly poor frame-pair memory scores, not because they fail to remember, but because the evaluation framework conflates action-following deficiency with memory degradation. For these reasons, WorldRoamBench does not adopt frame-pair memory and instead relies on the scene-level memory evaluation described in the main text, which aggregates geometric information across frames and is substantially more robust to per-frame action imprecision. Figure 19 illustrates two limitations: temporally symmetric pairs can differ in viewpoint or location, and even after pose-aware pairing reduces some spatial offset, SSIM and LPIPS still indicate substantial image-level differences in examples with visibly preserved scene content.

N Action Gap in Closed-Source Models

Table 2 shows that Genie 3 and Happy Oyster trail the strongest open models in action-following scores despite their strong overall performance. We also present representative failure cases from both closed-source models to illustrate the underlying causes.

N.1 Genie 3: Lateral-Action Camera Drift

When the A or D key is pressed to move the character laterally, Genie 3 interprets the input as a camera rotation command rather than producing a lateral translation. As shown in Figure 20 (top), instead of the character shifting left or right within a stable scene, the entire camera view rotates, leaving the character centered but drastically altering the visible environment. This spurious rotation makes precise steering impossible and directly illustrates a case-level action-following failure.

N.2 Happy Oyster: Physics-Induced Action Failure

In the illustrated rollout, Happy Oyster prioritizes collision response over commanded motion, which hurts its action-following score. As shown in Figure 20 (bottom), when the autoregressive generation produces new obstacles along the navigation path, the model simulates collision rather than allowing the character to pass through. Once the character collides with a generated obstacle, subsequent movement actions cannot dislodge it because the collision response overrides the commanded input. This case illustrates a local trade-off between physical coherence and action following: the character remains trapped even as movement commands continue.

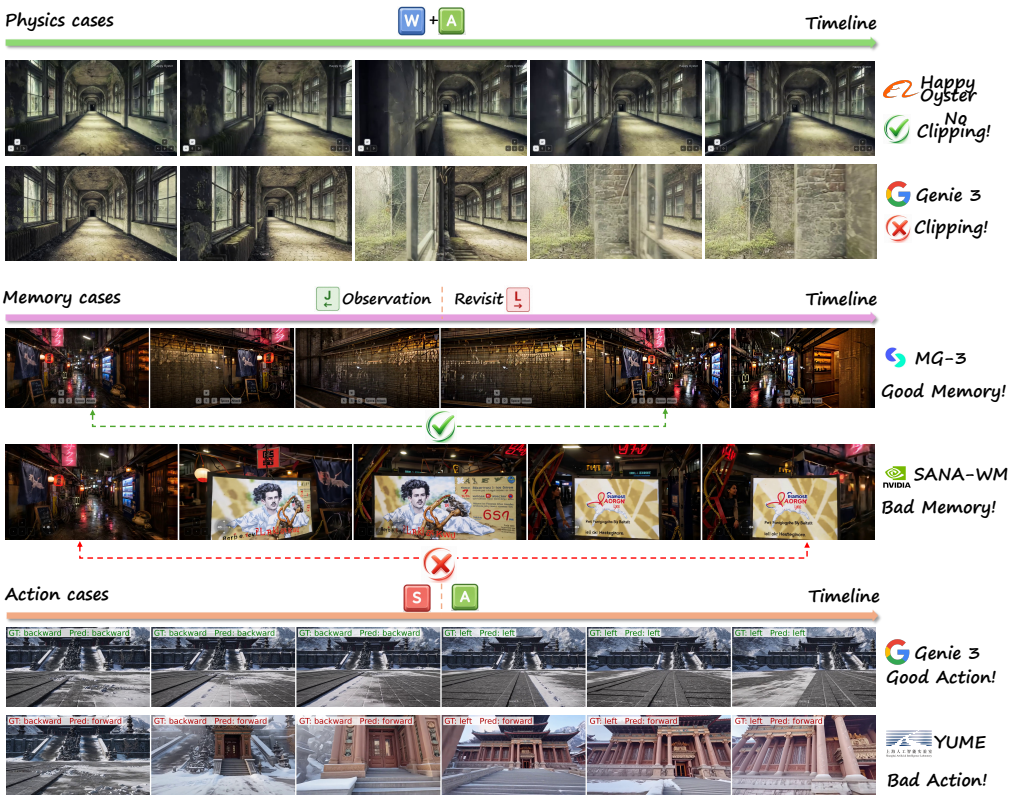


Figure 21: **Qualitative examples across physics, memory, and action.** Each block contrasts a positive and a negative rollout that share the same initial frame and action schedule. **Physics:** `W+A` in a corridor—Happy Oyster (*No Clipping*) vs. Genie 3 (*Clipping*). **Memory:** `J` → `L` Observation/Revisit in an alley—Matrix-Game 3.0 (*Good Memory*) vs. SANA-WM (*Bad Memory*). **Action:** `S` → `A` (`backward` → `left`)—Genie 3 (*Good Action*) vs. Yume 1.5 (*Bad Action*). Frames are shown in temporal order, left to right.

O Qualitative Study and Human Agreement Analysis

This appendix complements the quantitative leaderboard with qualitative examples, mechanics-judge calibration results, and a human-agreement analysis of subject-memory ratings.

O.1 Qualitative Study

Figure 21 contrasts a positive and a negative rollout for each of the three dimensions—interaction physics, memory, and action following; the two rollouts in each block share both the initial frame and the keyboard action schedule, supporting comparison within the evaluated dimension.

Physics. Under a sustained `W+A` in a narrow corridor, Happy Oyster keeps the avatar inside the geometry (*No Clipping*), whereas Genie 3 penetrates the wall into vegetation (*Clipping*). **Memory.** Under `J` (Observation) → `L` (Revisit) along a neon alley, Matrix-Game 3.0 reproduces the same storefronts on revisit (*Good Memory*), while SANA-WM regenerates different murals and shopfronts (*Bad Memory*). **Action.** Under an `S` → `A` schedule (`backward` → `left`), Genie 3's per-frame predictions match the ground truth throughout (*Good Action*), while Yume 1.5 outputs `forward` in every segment (*Bad Action*).

O.2 Human Calibration and Agreement Studies

The interaction-physics pipeline (Appendix J) and the subject-memory pipeline (Appendix K.2) use Qwen3-VL-PLUS for perceptual judgments. We report two human studies: calibration of the binary mechanics judge and agreement analysis for ordinal subject-memory ratings.

O.2.1 Mechanics VLM Calibration

From a calibration subset of 630 mechanics rollouts spanning 63 cases and 10 models, we randomly sample 100 videos for human evaluation. Twenty experts label each sampled video; the per-video majority vote serves as ground truth, and accuracy is the fraction of videos whose VLM verdict matches it, averaged over the 10 models of the calibration subset. Iterative refinement of the mechanics prompts and the evidence-gathering harness increases accuracy against the human majority-vote ground truth from **52.5%** (fixed template) to **73.2%** on this calibration subset. Both percentages are computed on the same 100-video set used for calibration, rather than on held-out validation data.

Figure 22 visualizes a representative sampled rollout in which the agent successfully zooms in on the correct time interval and produces a score matching the human ground truth.

O.2.2 Subject Memory Rating Agreement

We randomly sample 60 TPV rollouts from Happy Oyster, Genie 3, LingBot-World, and HY-World 1.5, and ask 20 human experts to score each rollout on the same 1–10 scale and rubric as Qwen3-VL-PLUS (Figure 27), blind to the VLM score. Defining a (rollout, expert) pair as concordant when $|S_{\mathrm{VLM}} - S_{\mathrm{human}}| \leq 1$, the concordance rate over the $60 \times 20 = 1{,}200$ pairs reaches **88.3%**, with a mean absolute difference of **0.83**, indicating that the judge tracks expert ratings closely under this tolerance.

Figure 23 visualizes representative sampled rollouts as temporally ordered subject crops together with their VLM and human scores.

The three rows track subject fidelity over time: the high-scoring subject keeps facial and clothing details consistent across multiple viewpoint changes; the mid-scoring subject keeps overall body proportions but drifts in garment attributes (a tie appears in later crops); and the low-scoring subject changes in both body shape and accessories (a backpack persists after the subject turns around). In each case the VLM score lies within one point of the displayed human scores, illustrating the concordance criterion reported above.

P Limitations and Future Work

Although WorldRoamBench comprehensively evaluates action following, visual quality, interaction physics, and memory, several limitations remain. The benchmark covers rollouts of up to 60 seconds over a finite set of scenes and an action space restricted mainly to camera movement and navigation, so it does not capture capabilities such as attacking, jumping, and object manipulation. Several components rely on estimated camera poses, monocular depth, segmentation, and VLM judgments, whose errors may propagate into the scores, and differences in native resolution, frame rate, control interface, and access conditions prevent perfectly standardized inference. Finally, all physics rollouts last 60 seconds, and horizon-specific Overall scores reuse the full-suite Physics score; physics-specific horizon sensitivity is therefore not assessed.

Future extensions should evaluate longer interaction horizons and broaden the diversity of environments, visual styles, and interaction contexts. The action suite should grow beyond navigation to richer embodied and compositional interactions, and the physics suite beyond mechanics and optics to electricity, thermodynamics, and other energy transfer; constructing physics cases at multiple durations would further enable a duration ablation of physical plausibility. Future work should also improve the robustness, calibration, and uncertainty estimation of the pipeline and develop more unified inference and control protocols for cross-model comparison.

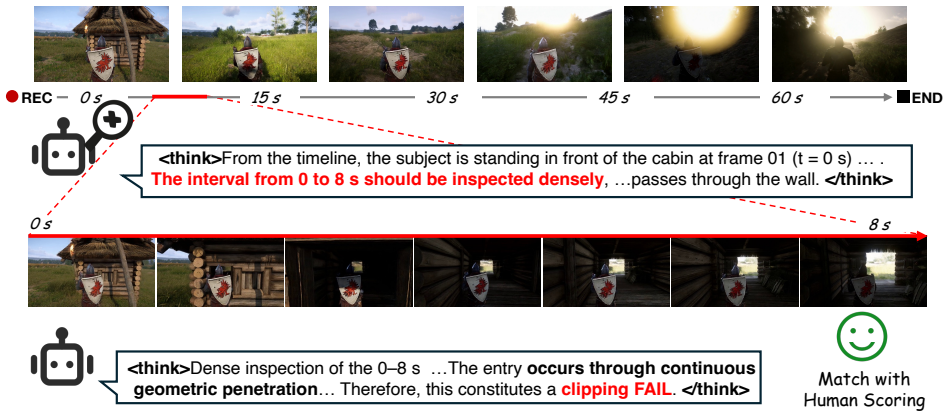


Figure 22: **Mechanics VLM Calibration Example.** The agent follows a coarse-to-fine scoring process: it first scans the full video timeline, progressively zooms in on the suspicious temporal interval for dense frame-level inspection, localizes the physical event, and assigns the final score.



Figure 23: **Subject-memory rating agreement examples.** Each row shows the temporally ordered subject crops fed to Qwen3-VL-PLUS for one sampled TPV rollout, together with the VLM score and the human expert scores on the same 1–10 scale.

TPV Mechanics Agent — System Rubric (excerpt)

```
You are a TPV mechanics evaluation agent.  You judge ONE physical criterion per case
(collision / clipping / deformation / gravity, including terrain adherence).
Loop (at most 4 rounds):  (1) study the overview timeline; (2) call a tool to localize
and densely inspect the critical event window; (3) either gather more evidence or emit a
structured verdict.
Tools:  inspect_interval(start, end, fps, crop) --- dense frames from a time window at
1--8 fps; inspect_frames(timestamps, crop) --- exact frames; crop in {full, center, lower,
upper} at up to 768 px.
Verdict rules:  PASS requires positive visual evidence; FAIL requires evidence of the
violated constraint; INVALID only if the subject or target is unobservable.
Clipping:  PASS iff the subject never traverses the obstacle (static, blocked, or a
passage that visibly pre-exists contact); FAIL on body--obstacle interpenetration, the
obstacle vanishing at the subject, or side-switching without a pre-existing passage.
The loop ends with an in-loop evidence audit:  clipping PASS verdicts are re-examined
for obstacle-topology artifacts; FAIL interaction verdicts get one high-recall
positive-evidence pass before being committed.
```

Agent Tool Call (example)

```
{"thought":  "Subject contacts the railing at t=16.6s; verify whether it stops or
traverses.", "tool_call":  {"name":  "inspect_interval", "args":  {"start":  16.0, "end":
24.0, "fps":  4.0, "crop":  "center"}}}
```

Structured Verdict (example)

```
{"verdict":  "FAIL", "confidence":  0.95, "event_interval":  [16.6, 21.5],
"subject_crossed":  true,
"evidence":  [{"timestamp":  18.2, "observation":  "railing absent at subject position"},
{"timestamp":  21.5, "observation":  "subject on far side of railing"}],
"reason":  "Subject switches sides while the railing disappears at its position; no
pre-existing passage."}
```

Figure 24: **TPV mechanics agent prompts.** Excerpt of the agent system rubric, an example tool call, and the structured verdict format. Per-criterion rubrics instantiate the same loop with criterion-specific evidence requirements.

Reflection Evaluation Prompt — Tier 1: Coarse Check (FPV)

```
There is {target_object} reflected in {surface_desc}.  Looking at the overall scene across
all these frames, are the reflections physically plausible | i.e., consistent in shape,
approximately symmetric about the reflecting surface, and changing naturally with the
camera movement?

After your reasoning, conclude with exactly one line:  Answer:  yes or Answer:  no
```

Reflection Evaluation Prompt — Tier 1: Coarse Check (TPV)

```
In this third-person view video, watch {target_object} and its reflection in
{surface_desc}.  As the subject moves, does the reflection behave physically correctly
| i.e., consistent in shape, approximately symmetric about the reflecting surface, and
moving consistently with {target_object}?

After your reasoning, conclude with exactly one line:  Answer:  yes or Answer:  no
```

Reflection Evaluation Prompt — Tier 2: Per-Frame Scoring

```
I show you two images.  Image 1 is the REFERENCE first frame showing the original scene.
Image 2 is a frame from the generated video at timestamp {timestamp}s.

In the reference frame, {target_object} is reflected in {surface_desc}.

FIRST, check whether {surface_desc} is clearly visible in Image 2.  If the camera has moved
so that {surface_desc} is no longer in view, or is only a tiny sliver at the edge, score
N/A | do NOT score 0 just because you cannot see a reflection.  Score 0 means the surface
IS visible but the reflection is wrong.

Score the reflection quality in Image 2:
- 2:  Reflection is physically correct | properly flipped, content matches the real scene
above, proportions and perspective are consistent with the viewing angle
- 1:  Reflection exists but has physics errors | wrong flip direction, noticeable
asymmetry, content partially mismatched, or unnatural distortion
- 0:  The reflecting surface IS visible and occupies a reasonable area, but the reflection
is missing, completely wrong, or shows unrelated content
- -1:  INVALID | the scene has collapsed or deviated so far from the reference that
reflection evaluation is not meaningful (e.g., completely different environment)
- N/A: The reflecting surface ({surface_desc}) is NOT visible or barely visible in this
frame | camera has moved away from it

After your reasoning, output exactly one line:  Score:  -1, Score:  0, Score:  1, Score:
2, or Score:  N/A
```

Figure 25: **Reflection evaluation prompts.** Full VLM prompts for coarse reflection plausibility checks and per-frame reflection scoring.

Shadow Evaluation Prompt — FPV with Reference Frame and Expected Shadow

```
The FIRST image is the REFERENCE first frame showing the initial scene.
The remaining images are frames from the generated video.

The expected physical phenomenon is: {expected_shadow}.

Compare the video frames against the reference. Does the generated video correctly show
{expected_shadow}?

Judgment criteria:
- A real shadow must have a recognizable silhouette shape matching the object that casts
it (e.g., a person shadow looks like a person outline, a pillar shadow looks like an
elongated rectangle)
- The shadow must have a clear boundary | vague dark patches, ambient shading, surface
texture, or lighting gradients do NOT count
- The shadow direction must be physically consistent with visible light sources in the
scene
- Answer 'yes' ONLY if you are confident the shadow is real and correct

After your reasoning, conclude with exactly one line: Answer: yes or Answer: no
```

Shadow Evaluation Prompt — TPV with Expected Shadow

```
In this third-person view video, {expected_shadow} should be visible. Looking at these
frames, does {expected_shadow} appear and behave in a physically correct manner (consistent
direction relative to the light source, moving with the subject)?

After your reasoning, conclude with exactly one line: Answer: yes or Answer: no
```

Shadow Evaluation Prompt — TPV Fallback (No Expected Shadow)

```
In this third-person view video, as the subject moves forward, does the subject cast a
visible shadow on the ground or nearby surfaces (walls, floors, etc.)? If so, is the
shadow shape recognizable and its direction consistent with the lighting in the scene?

After your reasoning, conclude with exactly one line: Answer: yes or Answer: no
```

Shadow Evaluation Prompt — Generic Fallback

```
In this scene there is a light source illuminating a wall or floor. As the camera moves,
a shadow should appear or change on the surface. Looking at these frames from the latter
portion of the video:

1. Is there a shadow visible on the wall or floor?
2. If yes, does the shadow move or change in a way that is physically consistent with the
camera movement and the light source position?

Answer 'yes' if the shadow appears and behaves correctly, 'no' if the shadow is missing or
behaves incorrectly.

After your reasoning, conclude with exactly one line: Answer: yes or Answer: no
```

Figure 26: **Occlusion and shadow evaluation prompts.** Full VLM prompts used for expected-shadow, fallback shadow, and generic shadow checks.

Subject Memory Evaluation Prompt

```
You are an expert evaluator of third-person video subject memory.  The input is a
temporally ordered sequence of subject crops from one generated video.  The first crop
is the reference subject; the remaining crops show the subject at later time points.
Evaluate whether the subject remains the same individual across the whole sequence, and
output a video-level subject memory score from 1 to 10.

Assess five failure types:
1.  Identity change:  the subject becomes a different individual or category.
2.  Structural distortion:  the body, anatomy, proportions, limbs, head, face, or key
parts become deformed or implausible.
3.  Appearance drift:  color, texture, clothing, hair, material, or style is truly
rewritten.
4.  Subject disappearance:  the subject becomes partly or fully invisible.
5.  Quality degradation:  severe blur, low resolution, diffused edges, or loss of key
details makes the subject hard to identify.

Important calibration rules:
- Viewpoint change alone is not inconsistency.  Back-to-front, front-to-back,
side-to-front, or far-to-close changes are acceptable if the subject can reasonably be
the same individual under the new view.
- New frontal details, such as face or front clothing, should not be treated as
contradictions when the reference was a back view.
- Impossible orientation conflicts are structural distortion.  If the body is still
back-facing but the face suddenly looks toward the camera, the score should be at most
5; if this persists or the head-body relation collapses, score 1--3.
- Same category does not imply consistency.  If a person, animal, or vehicle remains in
the same broad class but gradually becomes a blob, oval mass, fused-limb shape, or loses a
clear part structure, assign a low score.
- Lighting and shadow changes should be treated leniently.  If the subject becomes
darker, bluish-gray, lower contrast, or slightly softer because of shadow, but structure,
proportions, clothing shape, and identity cues remain stable, this is usually 8--9.
- Do not mark appearance drift unless color, texture, clothing, fur, or material changes
in a way not explainable by lighting, exposure, or white balance.
- Do not mark structural distortion only because shadows, low contrast, or mild blur
soften boundaries.  Mark it only when real geometry or anatomy changes.
- Mild blur or reduced texture contrast can be ignored; severe blur, low resolution, or
missing key details should reduce the score.
- Judge the full sequence holistically.  Progressive deterioration is more serious than an
isolated artifact.
- Hallucinated non-subject objects, such as a handheld item, bag, stick, or small prop,
are not identity changes if the subject body and main appearance remain consistent.

Scoring guide:
10:  Highly consistent identity, structure, proportions, appearance, and details.
8--9:  Mostly consistent; only minor appearance variation, mild blur, normal viewpoint
change, or lighting/shadow change.
7:  Same subject with stable structure and main appearance, but noticeable darkening,
mild-to-moderate blur, lower texture contrast, or edge softening.
6:  Roughly consistent, but clear non-lighting appearance drift or persistent detail
degradation begins to affect identity details.
4--5:  Poor consistency, with real structural distortion, obvious appearance rewriting,
partial disappearance, or severe blur.
2--3:  Severe degradation, such as fused limbs, unclear head-body relation, blob/mass/oval
collapse, or large-scale failure of key structure.
1:  The subject clearly becomes a different object, disappears for most of the sequence,
or can no longer be matched to the reference.

Output JSON only.  Do not output anything else:
{
"subject_memory_score":  [integer from 1 to 10],
"identity_change":  [true or false],
"structural_distortion":  [true or false],
"appearance_drift":  [true or false],
"subject_disappearance":  [true or false],
"quality_degradation":  [true or false],
"viewpoint_change":  [true or false],
"lighting_change":  [true or false],
"reason":  [brief explanation of the scoring rationale]
}
```

Figure 27: **Subject memory evaluation prompt.** Full VLM prompt used for holistic third-person subject-memory scoring and diagnostic flag extraction.