REASONING-STRUCTURED VIDEOS: A STRATIFIED DIAGNOSTIC SUITE FOR COMPOSITIONAL CONSISTENCY IN WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing evaluations of video world models measure visual fidelity, control adherence, and revisit consistency. We extend this perspective from within-rollout revisitation to agreement across distinct histories reaching the same reference state. We introduce Reasoning-Structured Videos, an Unreal Engine 5 benchmark for evaluating these relationships in static scenes. The dataset comprises 11,377 exploration and relation-structured videos with action and camera-pose annotations. Its relation-structured trajectories form rooted graphs with shared initial states and annotated endpoint relationships. Inverse tests recovery after retracing a path, Loop tests return after an exploratory detour, and Equivalence tests agreement between distinct paths reaching the same state. Our evaluation separates self-consistency from reference fidelity and complements image-based metrics with layout and geometry readouts. Across nine world-model configurations, return-consistency rankings broadly align with distributional quality, whereas cross-path agreement yields different rankings. Strong layout agreement can coexist with poor geometric fidelity, and revisit consistency deteriorates over longer intervals. Together, these findings show why cross-history agreement, reference fidelity, and long-range retention should be evaluated separately.

1 INTRODUCTION

Video world models generate future frames from visual context and controls, supporting planning (Chua et al., 2018; Hafner et al., 2019), interactive environments (Bruce et al., 2024), and robotic or driving simulation (Yang et al., 2023). These uses require coherent predictions as controls are composed: a world should preserve a scene through a return or an alternative route. Existing evaluations measure visual fidelity, camera control (Duan et al., 2025), decision quality, physical plausibility, and revisit memory (Li et al., 2026a; Ye et al., 2026). These are essential checks, but plausible frames and faithful controls alone do not establish whether different histories reaching one state produce compatible observations. Evaluating such relationships requires explicitly linked trajectories with known endpoint relations.

We study this requirement as *compositional consistency*. **Inverse** tests recovery after retracing a path, **Loop** tests return after exploration, and **Equivalence** tests agreement between distinct paths reaching the same state, extending evaluation from within-rollout revisitation to cross-rollout path dependence. We instantiate these relations in *Reasoning-Structured Videos*, an Unreal Engine 5 benchmark of rooted trajectory graphs with discrete actions and frame-level camera poses (Figure 1). Two complementary tiers distinguish *GT-anchored* fidelity to reference frames from *self-consistency* between relation-defined observations. This distinction matters because agreement can coexist with a shared incorrect prediction.

Equivalence concerns task-relevant state compatibility, allowing stochastic variation in texture and other nuisance appearance. Alternative histories should preserve the state information needed for uses such as navigation or goal-conditioned planning; visual-representation planners motivate this perspective (Zhou et al., 2024; Terver et al., 2025; Assran et al., 2025). We therefore complement image distances with frozen layout and geometry readouts as auxiliary diagnostics. Explicit relation annotations make these comparisons testable: independent clips do not identify which endpoints

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

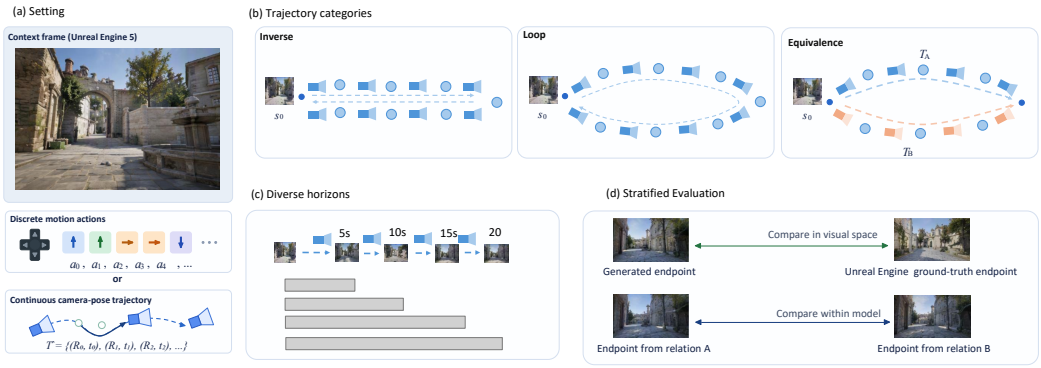

Figure 1: Overview of Reasoning-Structured Videos. (a) Models generate video from visual context and discrete actions or a continuous camera-pose trajectory. (b) Inverse, Loop, and Equivalence test reversal recovery, return after exploration, and agreement between alternative paths to the same state. (c) Evaluation at multiple horizons reveals how errors accumulate; the illustrated checkpoints are schematic. (d) GT-anchored evaluation measures reference fidelity (top), while self-consistency measures agreement between relation-defined observations (bottom, illustrated by paired paths). Together, these dimensions yield capability profiles resolved by relation, horizon, and evaluation tier.

should be compared. The benchmark therefore evaluates observable consistency without making claims about latent-state recovery or relation-aware training.

Findings. Across nine world-model configurations, we find that revisit construction alone does not isolate revisit ability. When Inverse and Loop returns are scored with LPIPS, PSNR, FID, or the recent NMR protocol (Gu et al., 2026b), model rankings largely follow FVD, because generation quality dominates these scores. Separating the two relations is still informative. Inverse targets already appear in the context, and every model recovers them better than Loop endpoints, but the size of this gap varies widely, revealing different degrees of history reuse. Equivalence is the relation least aligned with FVD, since comparing two rollouts of the same model cancels much of the quality component. Agreement, however, is not correctness: some models converge to consistent but unfaithful endpoints, which only the GT-anchored tier and structural SAM/depth comparisons expose. Compositional consistency therefore calls for relation- and tier-resolved evaluation beyond aggregate quality scores.

Our contributions are threefold: **(1)** a trajectory-graph benchmark covering Inverse, Loop, and cross-history Equivalence; **(2)** complementary GT-anchored and self-consistency protocols separating agreement from reference correctness; and **(3)** a multi-model study of relational, layout, and geometric consistency alongside distributional quality and revisit diagnostics.

2 Related Work

Evaluation of controllable video world models. Video world models are evaluated by visual quality, control adherence, and scene persistence. The Matrix combines distributional quality with movement-conditioned frame fidelity (Feng et al., 2026), while Infinite-World includes human judgments of memory and action following (Wu et al., 2026). WorldPlay and SANA-WM further assess reconstruction, camera accuracy, and revisitation (Sun et al., 2025; Zhu et al., 2026); WorldScore organizes evaluation around controllability, quality, and dynamics (Duan et al., 2025). These protocols motivate evaluating multiple capabilities. Our focus is how generated observations preserve known state relations within and across trajectories.

Distributional quality and reference fidelity. FVD compares video-feature distributions (Unterthiner et al., 2018), but does not explicitly test correspondence between observations of the same state. Its sensitivity also depends on the representation: Ge et al. (2024) identify a bias toward frame content over temporal realism. PSNR, SSIM, and LPIPS measure pairwise fidelity, although reconstruction error can reflect appearance, viewpoint, or scene-identity differences. Endpoint FID remains distributional: permuting endpoint pairings leaves it unchanged. These metrics therefore complement relation-defined comparisons rather than directly measure the same property.

Camera control and state consistency. WorldPlay measures rotation and translation errors, while SANA-WM evaluates aligned camera trajectories recovered with Pi3X (Sun et al., 2025; Zhu et al., 2026). Accurate motion does not establish that scene content is recovered at a revisited pose; conversely, an appearance mismatch may arise from control error. These evaluations address complementary questions: distributional quality measures resemblance to reference videos, trajectory adherence measures execution of the requested motion, and relational consistency tests agreement between observations associated with the same state within or across rollouts.

Revisit evaluation and metric specificity. WorldPlay and SANA-WM compare observations at revisited viewpoints, and MIND evaluates memory consistency alongside action control (Sun et al., 2025; Zhu et al., 2026; Ye et al., 2026). R2M-Bench calibrates revisit similarity against within-rollout temporal controls to reduce the influence of generic stability and limited motion (Gu et al., 2026a). We treat NMR as a valuable contemporaneous baseline for the question it addresses: whether a revisit has more appearance similarity than ordinary temporal continuity would predict. Our trajectory graphs add a different comparison target, Equivalence across distinct histories reaching the same state, and also separate self-consistency from GT-anchor correctness. The resulting profiles need not be orthogonal to NMR or FVD; they provide another relation-conditioned view of model behavior, including path dependence and cross-history state compatibility.

Task-relevant state equivalence. Latent world-model planners use frozen visual representations to compare predicted outcomes with image goals (Zhou et al., 2024; Terver et al., 2025; Assran et al., 2025). This motivates evaluating whether alternative histories support compatible state-dependent decisions, beyond pixel agreement. We therefore complement Equivalence-SC with task-relevant layout and geometry readouts, allowing stochastic appearance variation without equating pixel agreement with state correctness.

A detailed comparison with representative video datasets and evaluation suites is in Appendix A.

3 Reasoning-Structured Videos

3.1 Problem Formulation and Benchmark Scope

Given visual context and a control signal, a video world model generates future video frames. The signal can be a sequence of discrete actions or a continuous camera-pose trajectory; our benchmark provides both for each recorded trajectory. Our goal is to evaluate compositional consistency: whether the generated world remains coherent as controls are composed within and across trajectories, beyond the visual quality of individual frames.

To make this consistency testable, we construct reference trajectories in Unreal Engine 5 with known scene states $s_t \in \mathcal{S}$. A feasible control increment u_t induces the transition $s_{t+1} = T_{u_t}(s_t)$, and the renderer produces a frame $x_t = g(s_t)$. For a finite trajectory $A = (u_1, \dots, u_k)$, let $T_A = T_{u_k} \circ \cdots \circ T_{u_1}$, so that $T_A(s_0)$ is the reached state. This reference structure lets us group trajectories by their reached states and test whether generated videos preserve the annotated relationships.

We focus on single-agent camera navigation in scenes with fixed geometry and stabilized rendering. Relation instances are validated against simulator states and recorded camera poses, with construction and tolerance rules in Appendix D. This setting isolates trajectory-level consistency from uncontrolled scene changes. It complements evaluations of visual quality and physical dynamics; dynamic subjects, physical interactions, and stochastic or multi-agent environments are outside the present scope.

3.2 Three Path-Level Relations

A *reasoning-structured* video dataset groups trajectories that share an initial state and annotates how their endpoint states are related. This structure turns compositional properties of a learned transition into explicit evaluation targets. Let s_0 denote the initial state, and let $T_A(s_0)$ denote the state reached by trajectory A. Writing A and B for trajectories, we specify three relations:

- **Inverse:** $T_A^{-1}(T_A(s_0)) = s_0$, i.e., a trajectory followed by its inverse transformation returns to the starting state. The return segment reverses a specified transformation, so part of its target may be represented in the preceding context when the model retains the relevant history;

- **Loop:** $T_A(s_0) = s_0$, i.e., A forms a closed route. The model must preserve the scene state while composing a detour and a return, rather than extend an observed continuation;
- **Equivalence:** $T_A(s_0) = T_B(s_0)$ with $A \neq B$, i.e., two distinct sequences reach the same state. The histories share origin and endpoint but differ in intermediate transitions. This annotation tests whether their generated outcomes support compatible state-dependent decisions.

Together, the relations test cancellation, state preservation over a detour, and consistency across alternative histories. Inverse is a constrained loop whose return follows the inverse of the outbound path; Loop permits an exploratory path followed by a separately constructed return. Equivalence removes the requirement that the histories share intermediate controls. The equations specify identities in the reference environment, not pixel-equality constraints on independently sampled generations. Construction and tolerance rules are given in Appendix D; rendered examples are in Appendix B.

What should be equivalent? For a task-relevant state variable $z(s)$ and a fixed visual readout r, the intended capability is that $r(\hat{x}_A)$ and $r(\hat{x}_B)$ support the same state judgment, and that this judgment agrees with $z(T_A(s_0))$ when reference correctness is required. Texture and other task-irrelevant appearance may differ. Agreement alone permits a shared error, so we retain correctness as a separate axis. Under partial observability, reaching the same physical state need not imply the same belief or access to the same evidence. We therefore do not require unseen scene details to match across stochastic samples, nor do we infer a memory mechanism from endpoint disagreement. Our reference-based diagnostic measures utility for the recorded world; it does not establish that this world was uniquely identifiable from each model's input.

Downstream grounding. Reversible return and loop closure support navigation and mapping (Furgale & Barfoot, 2010; Mur-Artal et al., 2015). For Equivalence, a concrete concern is that alternative histories of one endpoint may give conflicting goal-state judgments. DINO-WM and JEPA-based planners use visual goal representations to evaluate predicted outcomes (Zhou et al., 2024; Terver et al., 2025; Assran et al., 2025). These systems motivate evaluating task-relevant state representations, rather than relying only on pixel similarity or claiming a closed-loop planning result.

3.3 Two-Tier Evaluation Protocol: Ground-Truth-Anchored and Self-Consistency

We distinguish reference fidelity from internal coherence through two complementary tiers: the *GT-anchored tier* conditions each continuation on the model-specific reference context and compares predictions with time-aligned reference frames, whereas the *self-consistency (SC) tier* runs the free-running relation protocol and compares predefined endpoints within or across rollouts. For Equivalence, cross-history agreement $d(\hat{x}_A, \hat{x}_B)$ and reference fidelity $\frac{1}{2}[d(\hat{x}_A, x_A^\star) + d(\hat{x}_B, x_B^\star)]$ are distinct quantities, so SC outputs can and should be compared with GT directly; GT-anchor rollouts are a second inference protocol, not the sole source of fidelity measurements. SC agreement alone is insufficient, since two poor but similar endpoints—including collapsed or shared-wrong states—can earn a favorable score, and we therefore report it alongside GT-anchor fidelity using LPIPS/PSNR and auxiliary layout/geometry readouts. As the tiers use different conditioning, a cross-tier comparison cannot by itself establish that a particular SC output is incorrect; that requires comparing the same output with GT. Formal metric definitions are given in Appendix F.

4 Dataset Construction

4.1 Environment, Scenes, and Capture Pipeline

We build the benchmark in Unreal Engine 5 across authored indoor, outdoor, and mixed scenes, including apartments, offices, warehouses, parks, streets, forest trails, and building entrances. All scenes use the same capture and rendering pipeline, which reduces rendering and capture confounds when comparing models across scenes. Figure 1(a) illustrates the visual context and control signals; additional scene examples are provided in Appendix B.

Stabilized reference rendering. To support the controlled setting in Section 3.1, we use a fixed capsule embodiment and camera configuration, rendering frames at 1280×720. Lighting and

exposure are fixed, warm-up frames are discarded after teleports, and temporally unstable rendering effects are disabled. These controls reduce appearance variation unrelated to camera motion, making reference comparisons interpretable. Exact embodiment and rendering parameters are in Appendix C.

4.2 Trajectory Graph Construction

Data is organized as *trajectory graphs* $G = (V, E; s_0)$ rooted at shared initial states, with a fixed horizon of $T_{\text{steps}} = 40$ control steps (360 frames, 22.5 s at 16 fps). The graph links recorded camera-pose trajectories through their shared root and annotated endpoint relations.

Graph instantiation. We construct trajectories to satisfy the relations defined in §3.2, assigning endpoint identities from simulator states and trajectory composition rather than image similarity. Construction controls, endpoint identifiers, and validation residuals are stored with each graph. Sampling ranges, fallback rules, and closure tolerances are provided in Appendix D.

Validation and split assignment. We validate recorded poses against each construction. The dataset contains 11,377 videos, each with 360 frames at 16 fps (22.5 s), totaling 4,095,720 frames and 71.11 hours. Its three non-overlapping subsets are the **Exploration Corpus**, containing 10,000 random-trajectory videos for broad scene and control coverage; the **Relation Evaluation Set**, containing 200 trajectory videos used in this paper; and the **Relation Development Pool**, containing the remaining 1,177 relation videos. The latter two together form the 1,377-video relation-structured collection of valid Inverse, Loop, and Equivalence constructions. The development pool can be used for training, validation, or additional testing according to the user's protocol. All three subsets use the same rendering, embodiment, horizon, and annotation pipeline and include per-frame actions and camera poses. Construction-branch labels and terminal residuals remain in the metadata for quality-control and tolerance-sensitivity analyses.

Each graph is annotated with RGB frames, both control representations, camera poses, relation labels, and validation metadata; the complete annotation schema is described in Appendix C.

5 Diagnostic Experiments

5.1 Experimental Setup

Models and control signals. We study nine configurations: HY-WorldPlay (Sun et al., 2025), minWM (Zhao et al., 2026a), MagicWorld-Fast (Li et al., 2025), MiniWorld-0.5B and MiniWorld-1B (Zhao et al., 2026b), LingBot-World-v2 Light (Team et al., 2026), SANA-WM (Zhu et al., 2026), our Chunk-AR reference implementation, and Matrix-Game 2.0 (He et al., 2025). Chunk-AR is our in-domain chunk-autoregressive baseline, trained on the Exploration Corpus and described in App. E. The first seven receive continuous camera-pose trajectories; the remaining two receive discrete action sequences. External models are evaluated without fine-tuning, using their supported visual context and generation resolution; camera poses undergo fixed coordinate and unit conversion. For Relation Evaluation Set LPIPS/PSNR, we align frames using recorded timestamps and relation endpoints and resize them to 512×288; conditioning settings and scoring windows are listed in App. G.1.

Relation Evaluation Set and conditioning. The fixed Relation Evaluation Set contains 50 validated graph units for each relation, or 150 graph units and 200 trajectory videos in total because an Equivalence graph contributes paired A and B branches. Each model produces one rollout per trajectory and tier. In the SC tier, each branch is generated from the initial visual context and its recorded control signal without inserting future ground-truth frames. In the GT-anchor tier, a separate rollout is conditioned on the model-specific ground-truth context supported by its implementation before the evaluated continuation; for example, minWM consumes a 180-frame ground-truth prefix, whereas HY-WorldPlay consumes one reference image. The tiers use distinct comparison targets, and their conditioning can also differ: SC compares relation-defined generated endpoints, while GT-anchor compares a generated continuation with the aligned reference frames. Equivalence A and B have the same 40-step horizon and use the same graph-derived generation seed within a frozen model configuration; seed sensitivity is analyzed in App. G.7. For clarity, the two tiers are separate inference protocols, not two distances computed on one rollout. In SC, Inverse and Loop compare

Table 1: Capability profiles on Relation Evaluation Set. Blue and green identify SC and GT-anchor relation blocks; FID compares generated and reference endpoint distributions. For Equivalence, SAM and Depth report A/B agreement for SC and generated-to-GT fidelity for GT-anchor. I3D-FVD compares matched local video-clip distributions. LPIPS-NMR reports normalized revisit gains on pooled Inverse/Loop SC rollouts. Darker shading marks top 4 point estimates within each row, with the best in bold; shading does not indicate statistical significance, and "–" denotes unavailable results.

Metric	HY-WorldPlay	minWM	MagicWorld-Fast	MiniWorld-0.5B	MiniWorld-1B	LingBot Light	SANA-WM	Chunk-AR	Matrix-Game 2.0
Self-consistency (SC)									
Inverse									
FID↓	**118.25**	242.46	210.22	380.30	406.77	207.04	285.05	–	134.94
LPIPS↓	0.460	0.712	0.657	0.841	0.880	0.640	0.737	0.651	0.704
PSNR↑	**15.07**	10.46	13.31	10.96	10.69	11.82	8.55	12.76	9.96
Loop									
FID↓	**138.23**	249.40	221.93	364.47	378.73	241.32	262.50	–	141.33
LPIPS↓	**0.494**	0.744	0.698	0.827	0.826	0.662	0.758	0.699	0.740
PSNR↑	**15.50**	10.85	12.46	11.59	11.31	12.23	8.50	13.00	10.05
Equivalence									
FID↓	**156.76**	240.23	222.02	368.04	371.17	196.05	286.05	–	145.55
LPIPS↓	0.372	0.444	0.433	0.303	0.342	0.499	0.587	**0.299**	0.559
PSNR↑	16.94	14.05	18.32	17.00	**20.05**	15.72	9.08	19.71	12.32
SAM layout cosine↑	0.830	0.826	0.853	0.954	0.940	0.806	0.821	0.901	–
Depth corr.↑	0.857	0.849	0.785	0.502	0.444	0.739	0.676	0.778	–
GT-anchor									
Inverse									
FID↓	**108.08**	207.92	183.83	327.77	351.04	169.53	244.62	–	–
LPIPS↓	0.521	0.571	0.517	0.608	0.677	0.461	0.589	**0.393**	0.487
PSNR↑	14.15	12.93	14.49	14.75	14.47	15.97	11.82	**18.09**	13.82
Loop									
FID↓	**145.68**	219.48	201.48	311.21	338.10	196.98	233.06	–	–
LPIPS↓	0.531	0.620	0.572	0.712	0.762	0.550	0.667	**0.513**	0.592
PSNR↑	14.95	12.67	14.54	13.92	13.96	14.07	10.47	**16.62**	13.21
Equivalence									
FID↓	**158.47**	193.49	170.67	326.18	339.79	174.52	253.23	–	–
LPIPS↓	0.582	0.582	0.551	0.731	0.744	0.544	0.654	**0.369**	0.569
PSNR↑	13.73	13.59	14.57	14.11	14.45	14.27	11.21	**18.61**	12.46
SAM to-GT cosine↑	0.753	0.741	0.758	0.749	0.754	0.760	0.743	0.822	–
Depth to-GT corr.↑	0.521	0.535	0.541	0.175	0.184	0.469	0.373	0.683	–
FVD↓ (I3D)	**200.8**	471.2	275.6	969.1	1097.3	243.8	707.7	234.2	–
LPIPS-NMR↑	**0.188**	0.022	−0.038	−0.102	−0.119	0.034	0.036	0.153	–

the generated terminal frame with the retained start frame, whereas Equivalence compares the two generated terminal frames from its paired paths. GT-anchor LPIPS/PSNR compare generated frames with their time-aligned references over the evaluated continuation; Endpoint FID uses only final frames. Figure 1(d) illustrates these comparison targets, and model-specific contexts are listed in App. G.1.

Metrics and statistics. We report LPIPS ↓/PSNR ↑ for both tiers, Endpoint FID ↓ for endpoint distributions, Equivalence SAM layout cosine ↑ and Depth-Anything relative-depth correlation ↑ as task-relevant auxiliary readouts for both tiers, and I3D-FVD ↓ for video-clip distributions. We estimate 95% graph-bootstrap intervals with 10,000 draws for LPIPS/PSNR; the endpoint-distribution sensitivity analysis uses 1,000 draws. Both Equivalence branches are resampled together as one graph-level unit. The intervals quantify variation across the fixed Relation Evaluation Set; a separate seed analysis quantifies generation-noise variation.

5.2 Relation-Level Capability Profiles

Do revisit constructions isolate revisit ability? Inverse and Loop are designed to test whether a model can return to content it has seen. Yet when their returns are scored with LPIPS, PSNR, or FID, the resulting model ranking closely follows the ranking by I3D-FVD. We attribute this to how these scores are formed: each compares a generated return frame with a reference or context frame, so its

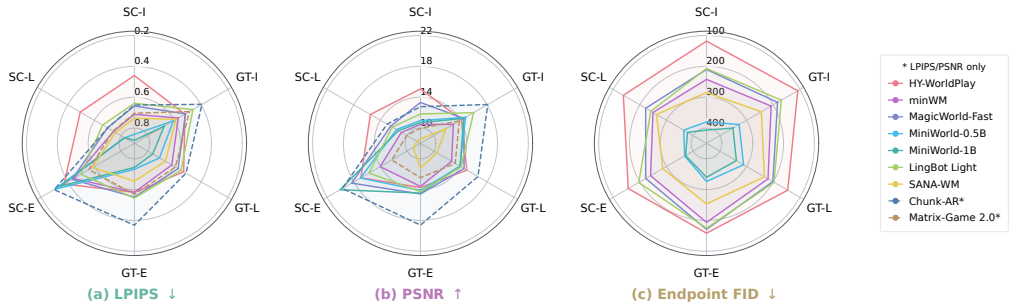

Figure 2: Capability profiles across the six tier–relation settings on the Relation Evaluation Set. Each radar uses a shared linear scale for one metric, with lower LPIPS/FID and higher PSNR farther from the center. I, L, and E denote Inverse, Loop, and Equivalence; SC and GT denote the two evaluation tiers. Solid and dashed lines identify camera-pose and discrete-action configurations. FID profiles include only configurations with complete endpoint results.

value reflects overall generation quality—blur, texture loss, and color drift—as much as whether the correct content was recovered. Across current models, differences in general generation quality are large enough to dominate the revisit-specific component. The same holds for NMR, which normalizes revisit error to isolate revisit gains: when instantiated with LPIPS, its model ordering stays close to that of FVD. Revisit construction thus defines the right test, but endpoint scores computed in pixel or perceptual space do not by themselves separate revisit ability from generation quality.

Inverse and Loop separate history reuse from state preservation. Rather than treating revisiting as a single capability, we split it into two relations that differ in what the model's context already contains. In Inverse, the GT-anchored context covers the outbound half of the trajectory, and the return controls invert it. Every view to be predicted on the way back has therefore been observed, so a model that retains and reuses its context should track the reference closely. In Loop, the model reaches its start only after a detour through unobserved views, and must preserve the scene state rather than retrieve it step by step along the path. An ideal model should therefore perform better on Inverse than on Loop in the GT-anchored tier. All evaluated models follow this ordering, but the size of the gap varies substantially across models. Because both relations are scored for the same model with the same metric, this within-model contrast partly factors out generation quality and indicates how much a model benefits when its target is available in its history. We therefore treat Inverse and Loop as relation-conditioned diagnostics of *how* a model fails to return, rather than as standalone scores of reasoning ability.

Equivalence measures internal consistency, not quality. Unlike Inverse and Loop, Equivalence-SC shows little rank agreement with FVD. It compares two rollouts of the same model that reach the same state along different paths. Both endpoints share the model's rendering characteristics, so much of the quality component cancels, and the score instead reflects how strongly the predicted endpoint depends on the route taken. This makes Equivalence-SC the relation most directly tied to internal coherence, but agreement is not correctness: a model can reach the same wrong or degraded endpoint along both paths. The MiniWorld configurations illustrate this case. They achieve some of the strongest Equivalence-SC agreement despite weaker FVD and poor GT-anchored endpoint fidelity. Equivalence-SC should therefore be read jointly with the Equivalence GT-anchor tier, which tests whether the endpoint that both branches agree on also matches the reference environment.

Structural agreement across histories. Pixel-level agreement is also stricter than the relation requires. Two distinct histories may legitimately differ in fine appearance, whereas downstream uses such as navigation and planning depend mainly on whether the endpoint preserves the same semantic layout and geometry. We therefore also compare SAM features and DepthAnything depth between branches (SC) and against the reference (GT-anchor). These structural views sharpen the distinction above. MiniWorld keeps a highly consistent semantic layout across branches, yet its depth agreement and depth fidelity are among the weakest, indicating endpoints that are consistent but geometrically unfaithful. HY-WorldPlay instead shows the strongest cross-history depth agreement while retaining competitive geometric fidelity. This suggests an advantage in navigation-oriented settings, where

Table 2: Inverse self-consistency at control-step gaps of 10–40. LPIPS and PSNR are grouped separately, and Δ_{40-10} reports the paired 95% graph-bootstrap interval for the change from gap 10 to gap 40. Darker shading marks the top four models within each metric column, with Delta shading based on the corresponding paired point estimate; bold is best. All eight configurations and the GT baseline use the same 50 pose-certified graph units from the Relation Evaluation Set.

Model	LPIPS↓					PSNR↑ (dB)				
	10	20	30	40	Δ_{40-10}	10	20	30	40	Δ_{40-10}
HY-WorldPlay	0.330	**0.412**	**0.423**	**0.453**	[0.071, 0.174]	17.75	15.54	**15.41**	**15.14**	[−4.02, −1.36]
minWM	0.415	0.594	0.669	0.710	[0.256, 0.335]	14.24	12.41	11.19	10.49	[−4.41, −3.15]
MagicWorld-Fast	0.456	0.569	0.618	0.654	[0.169, 0.229]	16.16	14.38	13.76	13.33	[−3.50, −2.18]
MiniWorld-0.5B	0.435	0.635	0.770	0.840	[0.342, 0.466]	17.75	14.61	12.78	10.96	[−7.94, −5.65]
MiniWorld-1B	0.382	0.630	0.770	0.880	[0.443, 0.552]	19.58	**15.95**	13.87	10.64	[−10.32, −7.44]
LingBot Light	0.435	0.562	0.627	0.641	[0.162, 0.254]	16.33	13.66	12.29	11.89	[−5.64, −3.30]
SANA-WM	0.555	0.669	0.706	0.725	[0.143, 0.199]	9.63	8.66	8.25	8.45	[−1.61, −0.78]
Chunk-AR	**0.322**	0.499	0.584	0.631	[0.266, 0.354]	**19.66**	15.40	13.69	12.80	[−8.48, −5.41]
GT	0.024	0.024	0.027	0.026	[−0.007, 0.013]	33.64	32.82	33.45	34.33	[−0.93, 2.39]

route-independent geometry matters more than pixel-exact appearance. Chunk-AR attains the highest structural fidelity to the reference, so no single model dominates both axes.

5.3 Consistency Across Horizon

We express horizon in reference control steps, each spanning nine frames on the recorded timeline, rather than in model-specific inference steps. Table 2 compares generated frames at two visits to the same reference pose, separated by 10, 20, 30, or 40 control steps. The comparable temporal profiles generally lose reference fidelity over the evaluated horizon, but the magnitude is relation dependent. Equivalence has the largest increase for MagicWorld-Fast, LingBot, SANA-WM, and Chunk-AR, even though it has the lowest terminal SC-LPIPS in the relation table. This dissociation shows why a low endpoint agreement score cannot be read as a long-horizon guarantee. MiniWorld-1B has the shallowest Loop increase, whereas Chunk-AR has the steepest Equivalence increase.

Short-range consistency does not predict long-range retention. All eight configurations evaluated on the same 50 revisit graphs have positive 10-to-40-step LPIPS changes with 95% intervals excluding zero. MiniWorld-1B moves from relatively low short-separation error (0.382) to the highest long-separation error (0.880). SANA-WM starts with high error (0.555) but degrades comparatively little, while HY-WorldPlay combines low initial error with the smallest increase. The GT baseline changes by only $+0.002$ LPIPS $[-0.007, 0.013]$, ruling out a comparable rise in reference-image discrepancy. Initial reconstruction and long-range retention are distinct capabilities, and a small drift alone does not imply accurate generation.

5.4 Calibration and Generation Stability

Controlled pose perturbations increase LPIPS and decrease PSNR, while exact-pose and appearance-nuisance controls yield smaller perceptual errors (App. G.3). For minWM, relation-level motion ratios exceed the static-output threshold. These checks support metric sensitivity and argue against relation-wide static copying, without establishing state recovery or excluding individual static rollouts.

Generation-seed sensitivity. We repeat the Matrix-Game 2.0 evaluation across generation seeds with the checkpoint, graphs, and conditioning fixed. Seed variation is small relative to graph-bootstrap uncertainty. Equivalence consistently has the lowest SC-LPIPS, while the nearly tied Inverse and Loop means change order. These results support aggregate stability for this configuration, but do not resolve close rankings or establish stability for other models. Full results appear in App. G.7.

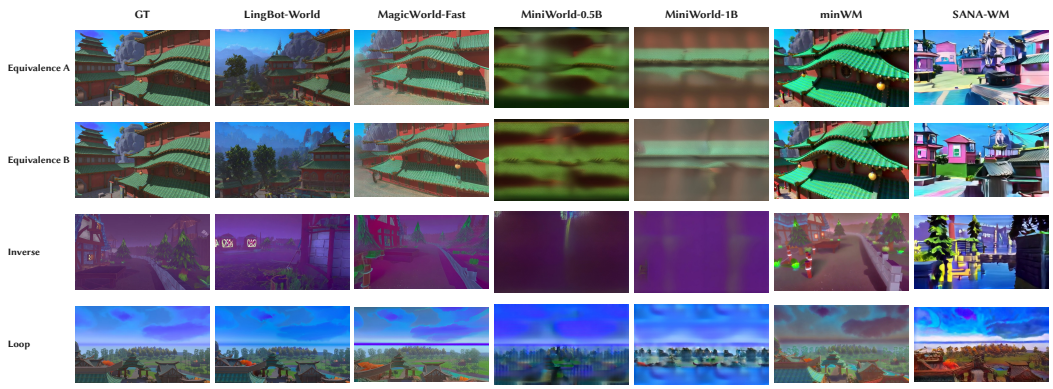

Figure 3: Qualitative comparison on a shared Equivalence graph. LingBot-World produces the richest details and most natural colors, while SANA-WM retains substantial detail but exhibits color shifts. MagicWorld-Fast and minWM better preserve GT geometry and semantic layout. MiniWorld shows lower visual quality but stronger Equivalence SC, illustrating that perceptual quality and internal consistency need not coincide.

5.5 Qualitative Diagnostics

Figure 3 highlights distinct patterns in perceptual quality, reference fidelity, and internal consistency across models on a shared Equivalence graph. LingBot-World achieves the strongest visual quality, producing rich details and natural colors. SANA-WM also generates detailed scenes, but exhibits noticeable color shifts. MagicWorld-Fast and minWM more closely preserve the geometry and semantic layout of the ground-truth (GT) frames. By contrast, MiniWorld produces lower-quality images, yet exhibits stronger consistency under Equivalence SC. These observations underscore that visually compelling outputs, faithful scene reconstruction, and internally consistent predictions are distinct properties: improvements in one do not necessarily imply improvements in the others.

Differences in visual quality may partly reflect differences in the models' training data, whereas our primary focus here is their internal consistency. The qualitative patterns align with the quantitative results in Table 1, supporting the ability of our evaluation to distinguish complementary dimensions of model behavior rather than reduce performance to a single notion of generation quality. Our framework thus complements existing evaluations by providing a more fine-grained diagnosis of internal consistency. Additional visualizations are provided in Appendix H.

6 Limitations

Our benchmark focuses on static, human-scale, single-agent scenes; dynamic interactions, irreversible actions, and stochastic or multi-agent settings require additional state definitions. GT-anchor values depend on calibration, field of view, appearance, and model-specific conditioning, while LPIPS/PSNR remain appearance-sensitive and Endpoint FID is sample-size sensitive. SC comparisons are valid only when the supplied controls preserve the annotated relations. Results use a fixed graph set and frozen generation configuration, so the intervals describe this evaluation setting rather than universal rankings or every model's noise distribution. The benchmark is an output-level diagnostic and does not establish latent-state recovery.

7 Conclusion

We introduced *Reasoning-Structured Videos* as a relational complement to video-quality, control, and revisit metrics. The trajectory graph makes explicit which states should agree within a rollout and across alternative histories. Our results show that FVD, NMR, Inverse, and Loop can share general-quality effects, while Equivalence-SC can reward shared-wrong outputs; GT-anchor fidelity together with SAM/Depth readouts exposes this failure. The benchmark therefore provides a relation-, tier-, and readout-specific capability profile without requiring metric orthogonality. Future work will extend these tests to dynamic and interactive environments.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, code development and debugging. The authors are responsible for verifying source references, experimental procedures, numerical results, and interpretations, and for the final content of the paper.

ETHICS STATEMENT

This work introduces a diagnostic benchmark for compositional consistency in video world models. All visual data is rendered synthetically in Unreal Engine 5; no human subjects, personally identifiable information, or user-generated content are involved, and no human annotation is used. The evaluated models are all publicly released systems, and our protocol reports frozen-subset comparisons rather than universal rankings. The main foreseeable risk is misuse of the reported numbers as a general quality leaderboard despite the scoped, static single-agent setting documented in §3.1 and §6; we therefore state the supported claims, assumptions, and out-of-scope axes explicitly.

REPRODUCIBILITY STATEMENT

The benchmark construction, capture pipeline, and evaluation protocol are described in §3–§5, with the trajectory-sampling, horizon, and residual-tolerance details in App. D, the SC metric definitions in App. F, calibration and validation in App. G. The relation-structured dataset is publicly available at `https://huggingface.co/datasets/VideoWorldmodel/ReasoningStructureTestset`. Evaluation code is provided in the supplementary materials, including model-specific conversion scripts and SC/GT metric computation. The dataset composition and the fixed evaluation subset are described in App. C–D; metric recomputation from saved model rollouts is detailed in App. G.8. Regenerating the minWM and HY-WorldPlay rollouts requires their official repositories and evaluated checkpoints.

REFERENCES

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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

Kaijin Chen, Dingkang Liang, Xin Zhou, Yikang Ding, Xiaoqiang Liu, Pengfei Wan, and Xiang Bai. Out of sight but not out of mind: Hybrid memory for dynamic video world models. *arXiv preprint arXiv:2603.25716*, 2026.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

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.

Ruili Feng, Han Zhang, Zhilei Shu, Zhantao Yang, Longxiang Tang, Zhicai Wang, Andy Zheng, Jie Xiao, Zhiheng Liu, Ruihang Chu, et al. The matrix: Infinite-horizon world generation with real-time moving control. *Advances in Neural Information Processing Systems*, 38:87318–87344, 2026.

Paul Furgale and Timothy D Barfoot. Visual teach and repeat for long-range rover autonomy. *Journal of field robotics*, 27(5):534–560, 2010.

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7277–7288. IEEE, 2024.

Qiwen Gu, Bingjie Gao, Rui Chen, Geng Li, Jifan Li, Qishuai Wen, Li Niu, Jing Tang, Xiangxiang Chu, and Junqiao Zhao. R2m-bench: Evaluating revisit memory via relative consistency in interactive video world models. *arXiv preprint arXiv:2608.27328*, 2026a.

Qiwen Gu, Bingjie Gao, Rui Chen, Geng Li, Jifan Li, Qishuai Wen, Li Niu, Jing Tang, Xiangxiang Chu, and Junqiao Zhao. R2m-bench: Evaluating revisit memory via relative consistency in interactive video world models, 2026b. URL https://arxiv.org/abs/2608.27328.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

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.

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.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *Advances in Neural Information Processing Systems*, 38, 2026a.

Guangyuan Li, Bo Li, Jinwei Chen, Xiaobin Hu, Lei Zhao, and Peng-Tao Jiang. Magicworld: Towards long-horizon stability for interactive video world exploration. *arXiv preprint arXiv:2511.18886*, 2025.

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*, 2026b.

Kewei Lian, Shaofei Cai, Yitao Liang, and Anji Liu. Loopnav: Benchmarking spatial consistency in world models. *arXiv preprint arXiv:2505.22976*, 2025.

Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346. IEEE, 2019.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3616–3626. IEEE, 2022.

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.

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

Basile Terver, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann LeCun. What drives success in physical planning with joint-embedding predictive world models? *arXiv preprint arXiv:2512.24497*, 2025.

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

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

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

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

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.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Min Zhao, Hongzhou Zhu, Bokai Yan, Zihan Zhou, Yimin Chen, Wenqiang Sun, Kaiwen Zheng, Guande He, Xiao Yang, Chongxuan Li, et al. minwm: A full-stack open-source framework for real-time interactive video world models. *arXiv preprint arXiv:2605.30263*, 2026a.

Yian Zhao, Ruochong Zheng, Hongcan Guo, Yu Yan, Jian Zhang, and Jie Chen. Miniworld: Democratizing the training of video world models from scratch. *arXiv preprint arXiv:2608.01127*, 2026b.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

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

Table 3: Comparison with representative video datasets and evaluation suites for world models. Paired rel. denotes explicit trajectory pairing under an annotated compositional state relation, rather than the presence of temporal or revisit information alone. Scale uses each resource's reported unit; our row gives the full dataset size and duration, with the fixed evaluation subset identified separately.

Dataset	Primary use	Scale	Engine	Paired rel.
Evaluation suites				
VBench (Huang et al., 2024)	T2V quality	800 prompts	–	✗
WorldModelBench (Li et al., 2026a)	WM QA	350 prompts	–	✗
WorldScore (Duan et al., 2025)	camera traj. adherence	3,000 scenes	mixed	✗
CLEVRER (Yi et al., 2019)	counterfactual QA	20k videos	physics sim	✗
Concurrent world-model benchmarks				
HM-World (Chen et al., 2026)	subject re-entry	59,225 clips	UE5	✗
MIND (Ye et al., 2026)	memory+control	250 videos	UE5	✗
LoopNav (Lian et al., 2025)	spatial consistency	250 h	Minecraft	✗
WildWorld (Li et al., 2026b)	symbolic state fidelity	108 M frames	AAA ARPG	✗
Ours (Reasoning-Structured Videos)	**compositional consistency**	11,377 videos / 71.11 h; 200 evaluation videos	**UE5**	✓

A Comparison with Existing Datasets and Evaluation Suites

Table 3 places Reasoning-Structured Videos alongside video datasets and world-model evaluation suites. The comparison concerns the evaluation target and the structure of the observations available for scoring. Dataset size provides context, but prompts, scenes, frames, and hours are different units and should not be interpreted as a common coverage scale.

Evaluation target and sampling unit. The resources in the table cover video quality, question answering, camera control, revisitation, and state fidelity. Our evaluation unit is a rooted trajectory graph: Inverse and Loop define a start–return comparison within one trajectory, while Equivalence links two trajectories through a common origin and terminal state. The fixed evaluation set contains 150 graph units and 200 videos because each of its 50 Equivalence graphs contributes two branches. These branches remain paired during scoring and bootstrap resampling. The larger Exploration Corpus supports scene and control coverage; its size is separate from the number of relation tests.

What explicit pairing contributes. The paired-relation column concerns annotations specifying which trajectories or observations should be compared under a known state relation. A revisit tests recovery of an earlier observation, whereas Equivalence compares outcomes reached through different histories. This pairing also distinguishes the test from an endpoint-distribution comparison: permuting generated endpoints leaves their empirical distribution unchanged but can change which annotated pairs agree. We therefore retain both relation-defined distances and distributional quality measures. As discussed in §2, revisit diagnostics and cross-history comparisons address complementary aspects of consistency.

Reference support and scope. The graph annotations identify the comparison targets before model inference. For a generated Equivalence pair, A/B agreement and the distance of each branch to its reference answer separate questions; a common incorrect scene can satisfy the first without satisfying the second. The GT-anchor tier additionally evaluates reference-conditioned continuations. Our contribution is this controlled relational diagnostic in static camera-navigation scenes. The comparison does not imply coverage of the dynamic interactions or question-answering tasks addressed by other resources.

B Visual Dataset and Relation Examples

Figures 4 and 5 connect the trajectory-graph definition to the rendered reference data. The first shows the range of scene appearances; the second shows how observations change along paths with an annotated return or shared endpoint. Both figures contain simulator-rendered reference frames. Model-generated comparisons are shown separately in Appendix H.

Scene appearance and visibility. Figure 4 includes urban streets, industrial interiors, wooded and snowy landscapes, and stylized architectural environments. The examples span differences in texture, spatial enclosure, and landmark visibility: factory views contain repeated structural elements, while trees and buildings can obscure parts of an outdoor scene as the camera moves. These examples illustrate the visual conditions under which a returned view must be recovered. They are not a scene-frequency histogram; the number of tiles should not be read as the number of distinct scenes or their sampling weights in the evaluation set.

The visual diversity is across scenes. Within each recorded trajectory, geometry and rendering are stabilized so that a change in observation can be interpreted relative to the recorded camera motion. Lighting differences between tiles do not indicate a time-varying lighting task. Camera settings, rendering controls, and released modalities are specified in Appendix C.



Figure 4: Reference scene examples. The montage includes urban streets, industrial interiors, wooded and snowy landscapes, and stylized architecture. Appearance varies across scenes, while rendering is stabilized within each trajectory. Tiles illustrate visual coverage rather than scene counts or evaluation frequencies.

Reading the relation examples. Figure 5 is organized from top to bottom as Loop, Inverse, and Equivalence. The Loop and Inverse blocks each show two scene examples. Read each row from left to right and compare the starting and terminal views: Loop returns after an exploratory route, while Inverse returns by reversing the outbound controls. The distinction is the path construction, even when both relations recover a similar starting view. The bottom two rows instead form one Equivalence pair. Their intermediate views differ, but their shared terminal state defines the A/B comparison.

From visual examples to scored pairs. Selected frames make the relations visible, but endpoint identities are assigned from simulator states and recorded poses rather than inferred from image similarity. The evaluator uses the annotated comparison frames, not the spacing of thumbnails in the figure. The complete records contain 40 control steps and 360 frames per trajectory. Construction rules and pose tolerances appear in Appendix D; the exact SC comparison targets are defined in Appendix F. This links the visual examples to the scored relations while preserving the distinction between a plausible-looking return and a pose-validated one.

C Annotations, Modalities, and Release Format

Embodiment and rendering. The agent is a rigid capsule with radius $34\,\text{cm}$ and half-height $88\,\text{cm}$. The camera is positioned $60\,\text{cm}$ above the capsule center and renders at 1280×720 with a 79° horizontal field of view, following the Habitat convention (Savva et al., 2019). We stabilize the reference renderer by freezing directional lighting, using an HDR capture path, clamping auto-exposure, discarding 15 warm-up frames after each teleport, and disabling motion blur, depth of field, lens flares, and ray tracing.

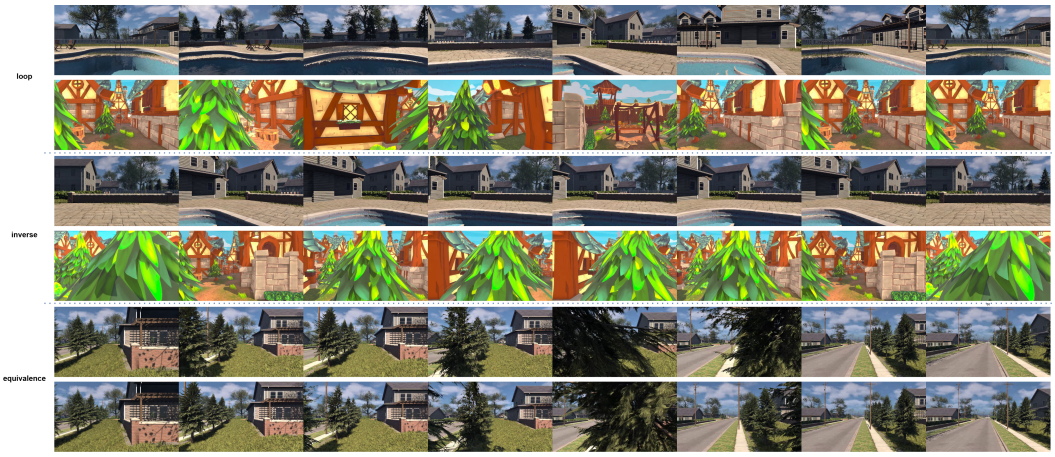

Figure 5: Rendered reference trajectories, with selected frames shown in temporal order. Top: two Loop examples returning after exploration. Middle: two Inverse examples retracing their outbound controls. Bottom: the A/B trajectories of one Equivalence pair, with different intermediate views and a common terminal state. Relation labels are validated using simulator states and camera poses.

Each trajectory graph carries RGB at 1280×720 and 16 fps, `SceneDepth` as 16-bit PNG normalized to 100 m, per-frame action labels and camera poses, graph and branch identifiers, relation labels, and per-step collision metadata. Core tasks require only RGB, one of the recorded control signals, and the relation annotations. Depth is a companion modality withheld from the initial RGB-only release. The evaluation harness wraps a video generator as `rollout(context, signal)`$\rightarrow$`frames` and computes paired SC and GT metrics together with distributional checks; running the evaluator requires no retraining.

The full dataset contains 11,377 videos in three non-overlapping subsets: the 10,000-video **Exploration Corpus**, the 200-video **Relation Evaluation Set**, and the 1,177-video **Relation Development Pool**. The latter two constitute the 1,377-video relation-structured collection, representing 1,135 relation units. This relation-structured collection is publicly available on Hugging Face at the URL in the Reproducibility Statement; evaluation code is provided in the supplementary materials. Users may allocate the development pool to training, validation, or additional testing. Each video contains 360 frames at 16 fps, giving 22.5 s per video, 4,095,720 frames in total, and 71.10625 hours of RGB video (reported as 71.11 h). The Relation Evaluation Set contains 50 graph units per relation, or 150 graphs and 200 rollout branches because Equivalence retains both A and B branches; its graph IDs are held out from training and fine-tuning for the reported evaluation. Table 4 reports subset sizes, durations, and intended uses.

D Trajectory Graph Construction: Details

This appendix supplies the implementation details omitted from §4.2. The full dataset contains 11,377 RGB videos (71.11 h), comprising 10,000 exploration videos and 1,377 relation-structured videos. Table 4 separates the three dataset subsets and their durations; Table 5 breaks down the relation-structured collection by construction type. Of the relation videos, 200 belong to the fixed evaluation set and 1,177 to the development pool. Evaluation graph IDs are held out from training and fine-tuning for the reported results; users may allocate the development pool to training, validation, or additional testing.

Capture controls and root states. The capture setup uses nine discrete actions: four translations, two yaw rotations, two pitch rotations, and `no_op`. Translation, yaw, and pitch increments are 100 cm, 15°, and 7.5°, respectively. Each action is interpolated over nine rendered frames at 16 fps. Roots lie on a 200 cm XY grid over a 4×4 m region with 8 yaw orientations per cell, filtered by capsule-overlap tests against scene geometry.

Table 4: Dataset composition and duration. The three subsets are non-overlapping. Each video contains 360 frames at 16 fps (22.5 s); durations are rounded to two decimal places. The evaluation and development subsets together contain 1,377 relation videos (8.61 h).

Component	Videos	Hours	Primary role
Exploration Corpus	10,000	62.50	Training and broad scene/control coverage
Relation Evaluation Set	200	1.25	Standardized held-out evaluation (150 graph units)
Relation Development Pool	1,177	7.36	User-selected training, validation, or additional testing
Total	**11,377**	**71.11**	4,095,720 frames

Table 5: Video and relation-unit counts in the relation-structured collection (8.61 h in total), including both evaluation and development videos.

Split	Videos	Relation units
Inverse-Easy	246	246 paths
Inverse-Hard	202	202 paths
Loop-Easy	247	247 paths
Loop-Hard	198	198 paths
Equivalence-Easy	484	242 paired graphs (A/B)
Total	1,377	1,135 relation units

Trajectory sampler. For each root, the construction type is sampled uniformly from {Seed, Inverse, Loop, Equivalence}. A constrained weighted sampler with inertia factor $\times 1.4$ (bias toward continuing the current control type), inverse-damping $\times 0.05$ (penalty on immediate reversal), a per-trajectory rotation budget of 3, and a collision-veto weight of $\times 0.01$ with up to 5 resamples per step produces purposeful navigation while avoiding geometry. The same sampler is reused across constructions; exact and tolerance-bounded relation branches are identified in metadata.

Horizon alignment. The choice $T_{\text{steps}} = 40$ reference control steps ($\times 9$ frames/step $= 360$ frames at 16 fps) is set to match the Chunk-AR baseline context length (App. E); it is a free parameter of the benchmark for other models.

Inverse construction. We sample a K-step construction path A with $K \sim \mathcal{U}\{10, 20\}$, interleaving translation blocks of 2–4 steps and rotation blocks of 1–3 steps. We form A^{-1} by reversing the order and inverting every discrete control, then emit $A \,\|\, \texttt{no_op}^{T_{\text{steps}}-2K} \,\|\, A^{-1}$. Because translations and rotations do not commute, this construction tests faithful sequence tracking rather than a pointwise control mirror.

Loop construction. The exploration and return branch samples $K \sim \mathcal{U}\{15, 28\}$ exploratory controls and synthesizes a five-stage discrete return sequence. The stages are yaw, pitch, forward or backward translation, and lateral translation. The exact branch uses geometrically closed rectangles, triangles, and hexagons. The branch identifier and terminal residual are retained in the metadata so that exact and tolerance-bounded loops can be analyzed separately.

Equivalence construction. Stage 1 independently applies Fisher–Yates shuffling within maximal mono-type segments of consecutive translations or rotations, where the actions commute. If fewer than 25% of steps change or the peak intermediate displacement is below 150 cm, Stage 2 emits an L-shape $U^m V^n$ and a zig-zag $(UV)^k U^{m-k} V^{n-k}$ with $m, n \in [3, 6]$, sampled from axis pairs $\{F, B\} \times \{L, R\}$. The paths have the same origin and endpoint but different interiors.

Residual tolerances for construction. Inverse paths are accepted when closure residual is $< 0.5\Delta_t$ horizontal and yaw $< 2^\circ$; the polygon branch of Loop is closed by geometry and accepted unconditionally, while the five-stage return branch is accepted when residuals fall below $0.45\Delta_t$ horizontal / $0.3\Delta_t$ vertical. Successful loops and inverse paths are `no_op`-padded to the fixed horizon.

Equivalent-path slot filling. The remaining slots in the L-shape / zig-zag Stage-2 construction are filled with self-inverse rotation pairs (`look_up`↔`look_down` on A; `turn_left`↔`turn_right` on B), which preserve the endpoint while absorbing frame budget. We restrict $m, n \in [3, 6]$ because the larger-scale configuration $m, n \in [6, 12]$ is geometrically infeasible in human-scale indoor scenes under the joint collision-free constraint.

Reasoning and random-walk assignment. A trajectory whose recorded pose deviates from its expected construction path by more than 1 cm at any step is assigned to the auxiliary `random_walk/` corpus. These trajectories retain frame-level actions, camera poses, and scene metadata for training and control learning, but are excluded from relation-level scoring. The exploration and return Loop branch additionally carries its terminal closure residual, enabling sensitivity analysis restricted to exact polygon loops.

E Chunk-AR Baseline Implementation Details

Chunk-AR is an in-domain reference implementation rather than a modeling contribution. It is a latent chunk-autoregressive transformer trained on the Exploration Corpus with a frozen Wan2.1 VAE (Wan et al., 2025). The reported configuration has approximately 420M parameters, predicts one nine-frame chunk per discrete action, and rolls the generated chunks into a fixed-length context. The evaluation graph IDs are held out from its training data.

The model uses ordinary next-chunk latent MSE and no consistency-specific loss. This makes it a conventional feasibility baseline for the three relation tests rather than a model trained to optimize our metrics. The configuration uses $d_{\text{model}} = 1280$, 20 heads, 8 encoder blocks, 14 decoder blocks, and $d_{\text{ff}} = 5120$.

F Self-Consistency Metric Family: Formal Definitions

This appendix gives the formal definitions sketched in §3.3 and the relation-preservation argument for changing the control signal. Let $\hat{V}_{h_0,A} = (\hat{x}_0, \ldots, \hat{x}_T)$ denote model M's rollout from history h_0 along the recorded trajectory A, supplied either as discrete actions or as camera poses. Let d be a fixed image distance. We report LPIPS as the perceptual measure and PSNR as the pixel-space measure, using the same relation-defined frame pairs for both.

Definitions.

- **Inv-SC**: for $A\|A^{-1}$, the rollout should return to its start if M treats A and A^{-1} as algebraic inverses:
$$\widehat{\Delta}_{\text{Inv-SC}}(M) \;=\; \mathbb{E}_{(h_0,A)\sim\mathcal{P}_{\text{Inv}}}\, d\big(\hat{V}_{h_0,A\|A^{-1}}[T],\; \hat{V}_{h_0,A\|A^{-1}}[0]\big). \tag{1}$$
- **Loop-SC**: for A with $T_A(s_0) = s_0$, a consistent model should close:
$$\widehat{\Delta}_{\text{Loop-SC}}(M) \;=\; \mathbb{E}_{(h_0,A)\sim\mathcal{P}_{\text{Loop}}}\, d\big(\hat{V}_{h_0,A}[T],\; \hat{V}_{h_0,A}[0]\big). \tag{2}$$
- **Equiv-SC**: for an equivalence pair (A, B), the image-distance proxy compares the two generated terminal observations:
$$\widehat{\Delta}_{\text{Equiv-SC}}(M) \;=\; \mathbb{E}_{(h_0,A,B)\sim\mathcal{P}_{\text{Equiv}}}\, d\big(\hat{V}_{h_0,A}[T],\; \hat{V}_{h_0,B}[T]\big). \tag{3}$$

Shared-control robustness. Each SC metric compares endpoints produced by the same model under one recorded trajectory representation. A valid conversion from actions to camera poses, or between pose conventions, must preserve the relation: inverse trajectories must remain inverses, closed paths must remain closed, and equivalent paths must retain a common endpoint. SC is not invariant to an arbitrary biased conversion. For the seven models receiving camera-pose trajectories (§5.2), we supply the same recorded poses and apply only fixed coordinate and unit conversion.

Complementarity with GT-anchored metrics. GT-anchored evaluation asks whether a rollout matches the reference. It is most interpretable when the model's control signal and conditioning match the benchmark. Self-consistency asks whether the model remains internally coherent under

the annotated relation and is silent about absolute correctness. A well-behaved model should score well on both tiers. We use SC as the primary cross-model signal only after checking that the supplied actions or camera poses preserve the relation. Otherwise, we report a model-specific control profile rather than a strict comparison. A static generator could trivially obtain low Inverse and Loop SC, so we also inspect temporal motion and reference fidelity. App. G.3 reports the relation-level motion check.

G Extended Experimental Protocol and Validation

G.1 Conditioning and Scoring Windows

The reference timeline contains frames 0–359. SC uses only the initial visual context and evaluates the relation-defined return or paired endpoints. GT-anchor evaluates a separate reference-conditioned continuation on frames 180–359. The context supplied to each implementation is listed below; HY-WorldPlay retains its initial reference image rather than receiving the midpoint frame. This difference matters when comparing GT-anchor values across models.

Table 6: Reference context for the GT-anchor evaluation. Frame indices refer to the common 360-frame reference timeline.

Configuration	GT context
HY-WorldPlay	frame 0
minWM, Chunk-AR	frames 0–179
MagicWorld-Fast, MiniWorld-0.5B/1B	frame 179
LingBot Light, SANA-WM, Matrix-Game 2.0	frame 179

SC return scores use the saved initial frame, except that the Matrix-Game evaluation uses the supplied root image. Equivalence compares the generated A/B endpoints, with both branches using the same graph-derived seed and control-step horizon. In the paired bootstrap, the two branches remain one graph unit. The same-output GT-terminal comparison in the following subsection is a separate diagnostic: it compares SC outputs with GT without rerunning inference and is not substituted for the GT-anchor rows of the main table.

G.2 Joint Agreement and Reference-Fidelity Analysis

For an Equivalence graph g, let x_g^A, x_g^B be the generated terminal images and y_g^A, y_g^B their matched GT terminals. We keep two distances separate,

$$C_g = d(x_g^A, x_g^B), \qquad F_g = \tfrac{1}{2}\big[d(x_g^A, y_g^A) + d(x_g^B, y_g^B)\big], \tag{4}$$

where C_g is Equiv-SC and F_g is a GT-terminal comparison using the same SC outputs, not another GT-conditioned inference run. A common incorrect output can make C_g small while leaving F_g large; lower values on both axes provide stronger evidence of reference-supported coherence. These are image-space proxies with different references, not an additive error decomposition.

G.3 Metric Calibration and Motion Checks

LPIPS/PSNR are appearance-sensitive, so we calibrate them against controlled GT perturbations. Using released absolute camera poses and only ground-truth videos, we form same-trajectory action-boundary pairs with pure translation, pure rotation, or exactly matched pose. We average within graph/bin over the 60 R20 graph units and use 10,000 graph-bootstrap resamples. As displacement increases from bin 1 to bin 4, both metrics move in the expected direction (Table 7).

Exact-pose pairs define a low floor of 0.021 LPIPS [0.016, 0.029] and 35.08 dB PSNR [33.70, 36.42]. Same-pose JPEG, 5% brightness, and Gaussian-noise controls produce LPIPS in $[0.0025, 0.0285]$, over an order of magnitude below the smallest pose-error bin. These controls establish predictable sensitivity to pose displacement under the tested nuisances; they do not make LPIPS/PSNR pure geometric distances.

Table 7: Controlled pose-error calibration. Spearman intervals are graph-bootstrap 95% CIs.

Perturbation	LPIPS bin 1 → 4	PSNR bin 1 → 4	ρ LPIPS	ρ PSNR
Translation	0.289 → 0.415	19.06 → 16.76	0.348 [0.266, 0.452]	−0.267 [−0.369, −0.185]
Rotation	0.505 → 0.615	15.30 → 13.53	0.448 [0.286, 0.594]	−0.362 [−0.502, −0.204]

We additionally compute temporal RGB-L1 motion relative to matched GT, $M(\hat{V})/(M(V^\star)+10^{-8})$, and predeclare < 0.10 as a static-output flag. For minWM SC, mean relation-level ratios are 0.662 [0.478, 0.911] on Inverse, 0.951 [0.590, 1.420] on Loop, and 0.683 [0.564, 0.817] on Equivalence. Thus no aggregate approaches the static threshold. This motion check rejects relation-wide static copying as an explanation for these means, while individual rollouts require separate inspection.

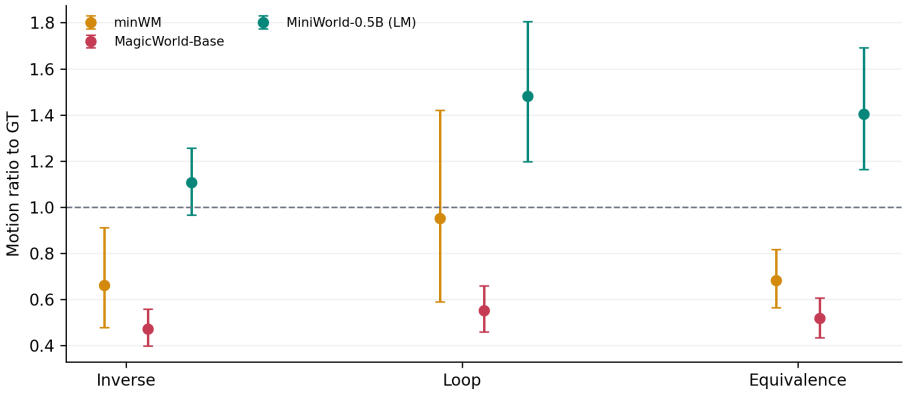

Figure 6: Temporal RGB-L1 motion relative to matched GT on four camera-trajectory R20 configurations. Motion is a supporting check: ratios above the static threshold show that the aggregate outputs are not relation-wide static copies, but motion alone does not establish correct state recovery.

G.4 External-Model Horizon Details

For the eight configurations in Table 2, the revisit analysis uses the same 50 Inverse graphs, LPIPS-Alex, and a 512×288 evaluation size. Symmetric reference-state pairs are used when valid; otherwise a deterministic search selects a pair at the same control-step separation within 1 cm and $0.1°$. At gaps 10, 20, and 30, 28 graphs use the symmetric pair and 22 require this search; all 50 use the symmetric pair at gap 40. The maximum selected pose discrepancy is 0 cm / $0.028°$. Paired bootstrap draws resample graphs while retaining all four gaps jointly (10,000 draws, seed 2357). The GT revisit baseline uses the same selected reference frames. These are within-rollout revisit comparisons, not comparisons of separately generated videos of different lengths. All eight model rows use this pair-selection rule and metric implementation.

On the 20 frozen R20 Inverse graphs, we compare generated frames at GT-pose-certified revisits separated by 10, 20, 30, and 40 reference control steps. The fixed symmetric pairs pass the 1 cm / $0.1°$ pose audit on all 11 Inverse-Easy graphs. For nine Inverse-Hard graphs, the evaluator deterministically selects a revisit with the same step separation that passes the thresholds; every selected pair has zero measured pose error. Table 8 reports paired graph-bootstrap results.

A strict sensitivity analysis restricted to the 11 external graphs where every canonical symmetric pair passes the pose check also preserves the increasing LPIPS trend.

G.5 Endpoint Distribution Details

Endpoint FID uses clean-fid 0.1.35 in clean mode with ImageNet Inception-v3 pool3 features and the logical final frame. Predictions and references share graph IDs, timestamps, and preprocessing; Equivalence contributes two endpoint images while remaining one graph-level cluster. We use graph

Table 8: Inverse SC versus reference control-step separation on the common R20 graphs.

Source	Gap	LPIPS↓ [95% CI]	PSNR↑ [95% CI]
minWM	10	0.4509 [0.3957, 0.5007]	14.12 [12.95, 15.44]
minWM	20	0.6323 [0.6091, 0.6552]	12.36 [11.44, 13.39]
minWM	30	0.6788 [0.6601, 0.6991]	11.10 [10.35, 11.88]
minWM	40	0.7207 [0.7005, 0.7429]	10.47 [9.80, 11.17]
HY-WorldPlay	10	0.3211 [0.2558, 0.3856]	18.00 [15.42, 21.10]
HY-WorldPlay	20	0.3890 [0.3176, 0.4594]	16.70 [14.41, 19.38]
HY-WorldPlay	30	0.4067 [0.3456, 0.4697]	16.39 [14.05, 19.29]
HY-WorldPlay	40	0.4250 [0.3599, 0.4905]	16.07 [13.82, 18.87]
GT floor	10	0.0188 [0.0133, 0.0248]	34.34 [31.32, 37.37]
GT floor	20	0.0177 [0.0134, 0.0241]	33.90 [31.88, 35.98]
GT floor	30	0.0176 [0.0113, 0.0258]	34.73 [32.32, 36.98]
GT floor	40	0.0177 [0.0105, 0.0269]	35.72 [33.49, 37.74]

bootstrap only as a sensitivity check because the fixed set contains 50 graph units per relation and FID is finite-sample sensitive.

The main table reports the primary endpoint-FID values. These small-sample estimates are descriptive checks of reference-distribution proximity, not universal model rankings.

G.6 I3D-FVD Video-Distribution Analysis

We compute FVD (Unterthiner et al., 2018) with the Kinetics-400 I3D checkpoint and estimator used by StyleGAN-V (Skorokhodov et al., 2022). The official TorchScript detector receives RGB frames, rescales them to $[-1, 1]$, and bilinearly resizes the complete frame to 224×224 with `align_corners=False`. We extract its 400-dimensional logits before softmax in FP32. Feature means and covariances use the StyleGAN-V convention, $\mu = \frac{1}{n}\sum_i f_i$ and $\Sigma = \frac{1}{n}\sum_i f_i f_i^\top - \mu\mu^\top$, and the distance is computed with a covariance-matrix square root. Thus, the detector, feature layer, preprocessing and distance estimator follow an established video-generation evaluation implementation; the temporal sampler is specific to our benchmark.

Each video contributes twenty 32-frame windows at 16 fps, uniformly spanning its evaluated generated interval. The fixed schedules are shared by all eight configurations, exclude input contexts and boundary padding, and pair every generated clip with the same time window in its reference video. SC windows span logical frames 9–359; GT-anchor windows span 192–359 to avoid the supplied prefix and boundary padding. This measures the distribution of sampled local clips across the rollout, not dependencies across its entire duration.

Each model contributes 200 videos and 4,000 clips per tier, covering the same 150 Relation Evaluation Set graph units. Equivalence contributes two branches. For the main-table scalar we concatenate the two generated feature sets and recompute the distance against reference features with identical tier and temporal weights; we do not average the two distances or add unrelated GT videos. The resulting 8,000 generated clips remain clustered by 150 graphs and are not independent observations. Conditioning budgets differ by model, so the pooled row describes the released protocol mixture; Table 9 also reports each tier separately.

The computation script and result metadata record the checkpoint hash, upstream implementation and preprocessing. Per-clip features and timeline mappings are cached for reproducibility. The covariance-square-root calculation agrees with an independent nuclear-norm formulation within 2×10^{-8}; GT self-distance and its distance to a permutation are both below 2×10^{-8}. Scores are finite-sample point estimates, not evidence of statistical significance between nearby values. Rank correlations describe these eight configurations, including two MiniWorld sizes, rather than a population of independent model families. Historical R3D-18 and terminal-only I3D results remain separate files and are not mixed into the reported scores.

G.7 Seed Sensitivity

We measure generation-noise variation on the complete Matrix-Game 2.0 Relation Evaluation Set. The original pass and four additional passes use independent generation seeds while keeping the

Table 9: Multi-window I3D-FVD using the StyleGAN-V detector and distance estimator. Twenty 32-frame clips uniformly span each video's evaluated interval, giving 4,000 clips per tier and model. The pooled column combines both tiers against reference clips with matching temporal weights.

Model	SC	GT-anchor	pooled
HY-WorldPlay	181.4	252.3	200.8
minWM	669.7	440.3	471.2
MagicWorld-Fast	418.5	260.6	275.6
MiniWorld-0.5B	1245.2	844.9	969.1
MiniWorld-1B	1379.9	962.5	1097.3
LingBot Light	355.1	233.7	243.8
SANA-WM	849.8	687.2	707.7
Chunk-AR	289.0	239.4	234.2

checkpoint, graph manifest, conditioning, preprocessing, and Equivalence A/B seed coupling fixed. Each pass contains 200 SC and 200 GT-anchor branches, so the analysis covers 2,000 rollouts in total and completes without failure. SC and GT-anchor metrics use the same endpoint and horizon definitions as the main evaluation at 512×288; pass-level results are recorded in JSON files and `seed_variance.csv`.

Table 10 reports the five-pass mean, across-pass standard deviation, and coefficient of variation (CV). The across-pass SD is a sampling-noise estimate, not a graph-bootstrap confidence interval. Relation-level SC means are 0.711 ± 0.009, 0.733 ± 0.008, and 0.558 ± 0.012 LPIPS for Inverse, Loop, and Equivalence; Equivalence remains the lowest-error relation in all five passes, while the nearly tied Inverse and Loop means exchange order once.

Table 10: Five-pass seed-sensitivity analysis on the complete Matrix-Game 2.0 Relation Evaluation Set. Each pass contains 200 SC and 200 GT-anchor branches. Values are the across-pass mean, standard deviation, and coefficient of variation (CV).

		Across-pass variability		
Scope / metric	Relation	Mean	SD	CV
SC endpoint LPIPS	all relations	0.6675	0.0038	0.57%
SC endpoint PSNR (dB)	all relations	10.863	0.237	2.19%
SC motion ratio	all relations	1.190	0.012	1.00%
GT rollout LPIPS	all relations	0.5487	0.0016	0.29%
GT rollout PSNR (dB)	all relations	13.155	0.052	0.40%
GT endpoint LPIPS	all relations	0.6716	0.0028	0.41%
GT H1 LPIPS	nearest bin	0.3602	0.0010	0.29%
GT H4 LPIPS	farthest bin	0.6592	0.0026	0.39%
SC endpoint LPIPS	Inverse	0.711	0.009	1.23%
	Loop	0.733	0.008	1.14%
	Equivalence	0.558	0.012	2.14%

The variability is small relative to graph uncertainty: across the 40 reported metric rows, the ratio of across-pass SD to the corresponding graph-bootstrap interval half-width ranges from 0.01 to 0.46 (median 0.12). The nearest GT horizon bin is more stable than the farthest one (CV 0.29% versus 0.39%), consistent with error accumulation along the rollout. PSNR is more sensitive to the generation seed than LPIPS; SC Equivalence reaches a CV of 5.93% at the relation level, whereas its LPIPS CV is 2.14%. For this Matrix-Game configuration, the five-pass analysis supports the aggregate conclusions from the fixed evaluation set. It does not resolve close model rankings or establish the seed sensitivity of other models.

G.8 Reproducibility Artifact

The supplementary materials include a standalone evaluator for recomputing metrics from saved model rollouts. Applied to 1,371 Matrix-Game SC rollout videos, it audits 448 Inverse graphs, 445

Loop graphs, and 239 complete Equivalence pairs, yielding the results below without rerunning model inference:

Table 11: Independent recomputation of the Matrix-Game SC row (LPIPS / PSNR).

Relation	Recomputed	Paper (rounded)
Inverse	0.7061 / 10.4476	0.71 / 10.45
Loop	0.7173 / 10.6227	0.72 / 10.62
Equivalence	0.5918 / 12.5732	0.59 / 12.57

The supplementary evaluation code covers model-specific control conversion, SC/GT metrics, graph bootstrap, motion checks, and horizon/endpoint evaluation. Metric recomputation requires the corresponding saved model rollouts, which are distinct from the reference videos in the Hugging Face dataset. For minWM/HY-WorldPlay, regenerating rollouts requires their official repositories and evaluated checkpoints.

H QUALITATIVE GALLERIES

The following three relation-first galleries illustrate Inverse, Loop, and Equivalence across nine historical configurations. For Inverse and Loop, compare each row's starting observation with its return; for Equivalence, compare the A/B endpoints with one another and then with GT, since similar generated endpoints may still share an incorrect scene. The supplied control signals and inference budgets differ across configurations, so these galleries illustrate qualitative failure modes rather than establish a controlled ranking. They include MagicWorld-Base, whereas the main table evaluates MagicWorld-Fast.

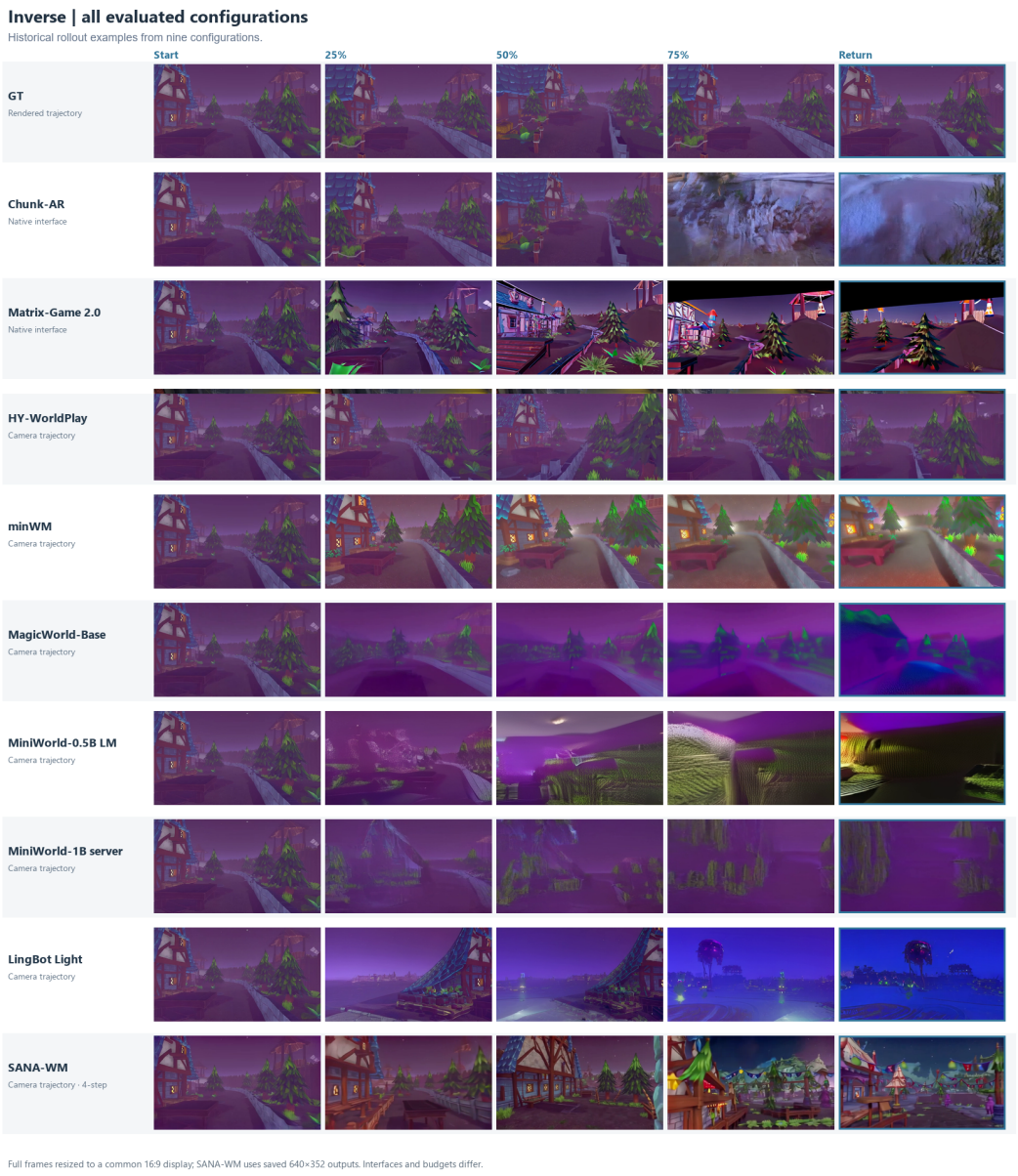

Figure 7: Inverse: GT plus nine historical configurations. Compare the start with the highlighted return.

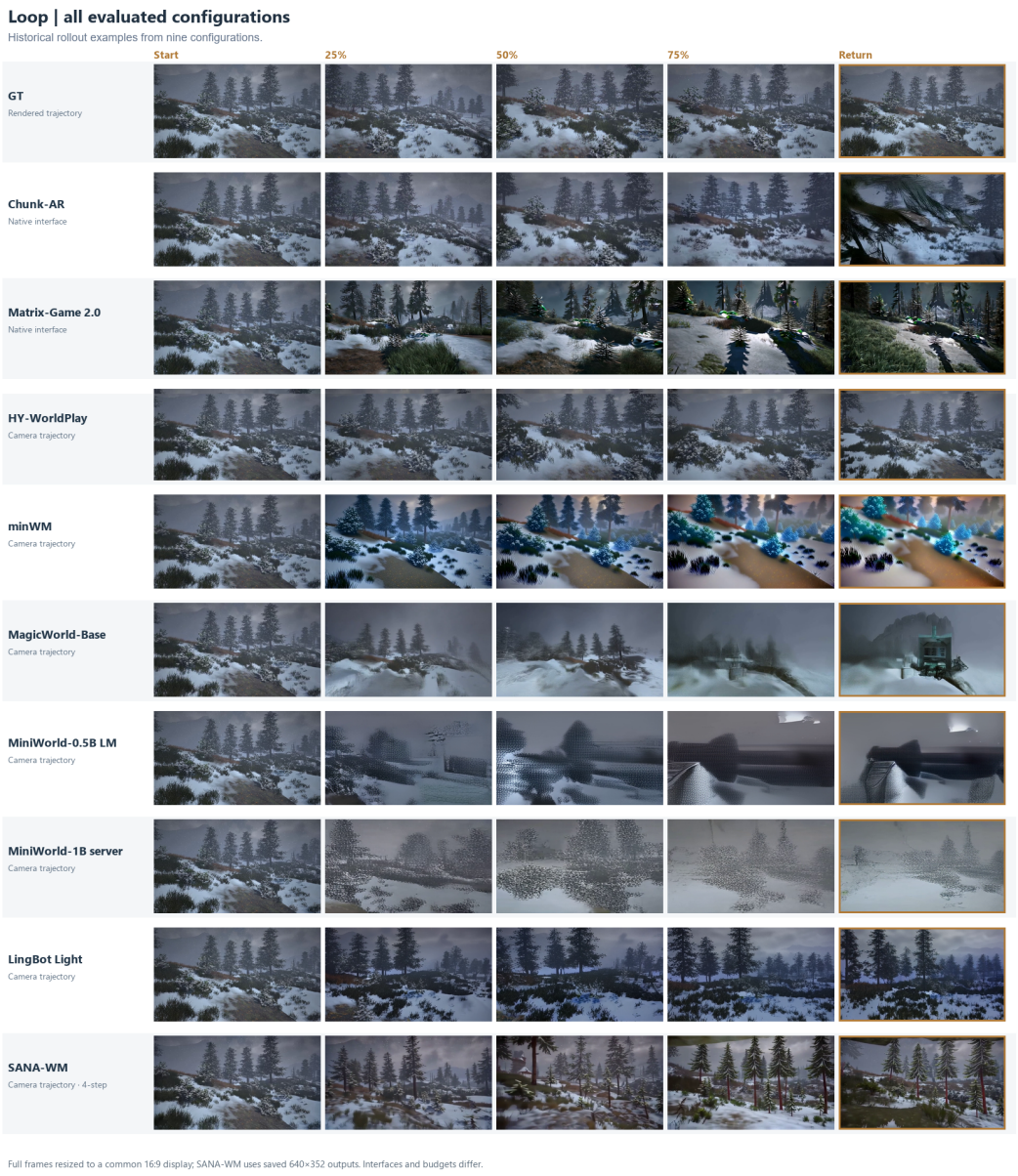

Figure 8: Loop: GT plus nine historical configurations. Start-to-return comparison exposes scene and layout drift.

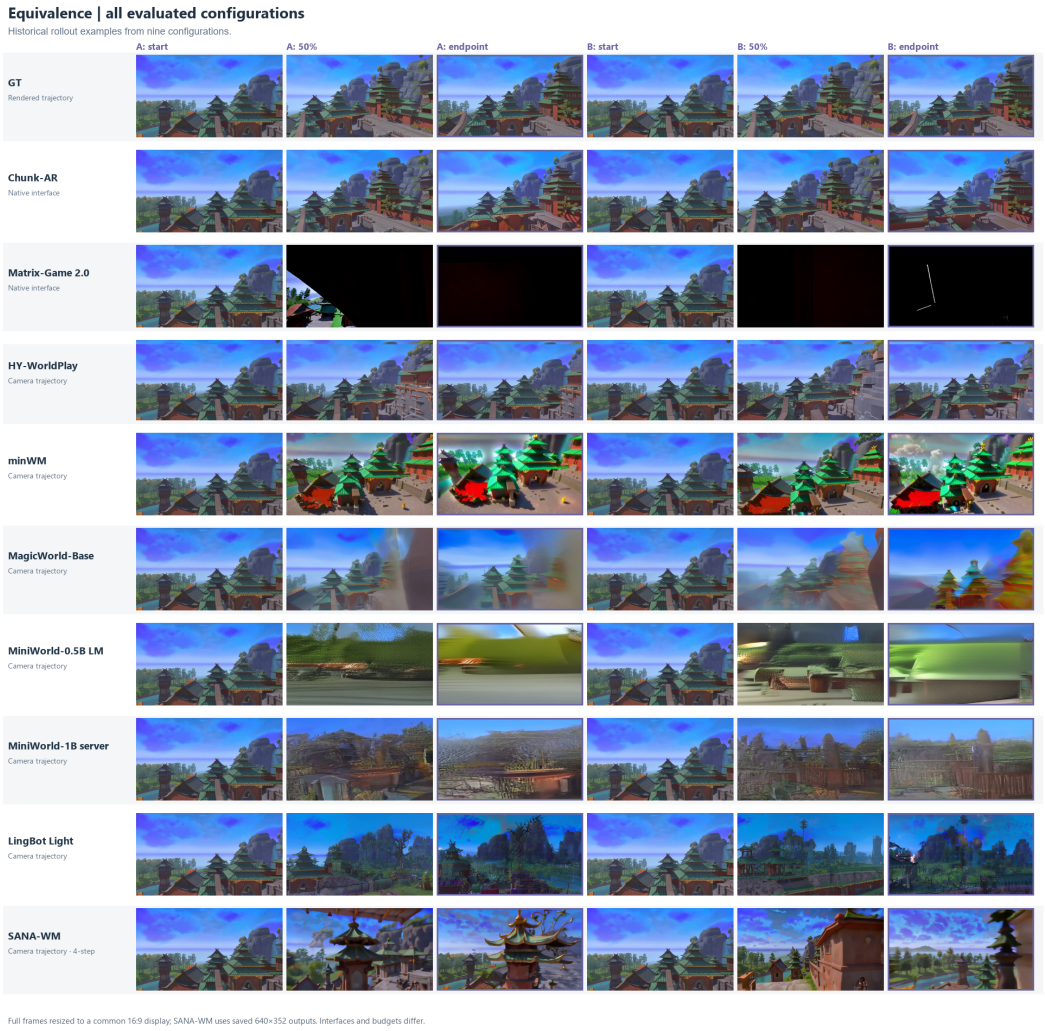

Figure 9: Equivalence: GT plus nine historical configurations, including MagicWorld-Base. Compare the two generated endpoints with one another and then with GT; similar A/B endpoints may still share an incorrect scene.