CONSISWORLD: BENCHMARKING SPATIAL AND CONTROL CONSISTENCY IN INTERACTIVE VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive video world models allow users to explore generated environments through movement and turning commands. Reliable interaction requires both consistent scene structure across viewpoints and the ability to anticipate subsequent action effects from prior experience. We introduce ConsisWorld to jointly evaluate spatial and action consistency through three tests: scene geometry preservation, single-action execution, and multi-action composition. The benchmark contains 140 initial views from 30 synthetic and real-world scenes, covering 36,400 videos generated by ten models. The spatial test constructs a fixed 3D reference from the initial image, depth, and camera parameters, and checks whether generated views remain consistent with the original scene geometry as the camera moves. The action tests execute each of six basic actions repeatedly, estimate its single-step displacement and rotation from the first four steps, and use these estimates to predict subsequent motion and the endpoint relations of independently generated composite routes. Each model is thus evaluated against its own measured action effects. We find that some models preserve scene geometry well yet exhibit substantial motion deviations during repeated execution of the same action. Composing different actions exacerbates these consistency challenges, causing actual endpoint relations to deviate significantly from the predictions of isolated actions. Moreover, some models are stable across repeated generations from the same initial view, but the displacement induced by an action changes substantially across initial views. These results show that good scene preservation does not imply stable continuous control. There remains substantial room for improvement in reducing motion deviations during sustained execution and action transitions, and improving control consistency across initial views. ConsisWorld evaluates scene preservation and action execution together, providing a benchmark for comparing model capabilities and assessing future improvements.

1 INTRODUCTION

Interactive video world models update scene views in response to movement and turning commands, allowing users to continuously explore generated environments (He et al., 2025; Sun et al., 2025; Shen et al., 2026; Zhu et al., 2026; Team et al., 2026b; Gao et al., 2026). Reliable interaction requires high consistency in scenes and actions: spatial layouts should remain coherent as the camera moves, and users who have learned the visual effects of moving forward or turning left should be able to anticipate the outcomes of sustained forward motion or moving left after moving forward. Therefore, interactive evaluation must examine not only whether generated views preserve the original scene structure, but also whether actions produce predictable results when sustained or combined.

Existing benchmarks primarily evaluate visual quality (Huang et al., 2024; Liu et al., 2024), camera control, trajectory following, and memory (Duan et al., 2025; Fang et al., 2026; Ying et al., 2026). For instance, WorldMark assesses command response latency and stability (Xu et al., 2026), while related work explores the identity and composition relations of actions (Wang et al., 2026a). However, we focus on a more fundamental predictive question: can separately measured short-term action effects accurately predict subsequent repetitions of that action and the outcomes of combining different actions? Action consistency determines whether users can familiarize themselves

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.

with a model through a few interactions and reliably anticipate subsequent motion. Because existing models employ diverse control interfaces such as keyboard events, camera parameters, or text prompts (He et al., 2025; Wang et al., 2026b; Shen et al., 2026; Chen et al., 2026), the displacement and rotation induced by identical commands vary significantly across models. Furthermore, a model's inherent motion dynamics are often asymmetrical; for example, a single model's forward and backward motion magnitudes may not be identical. Therefore, rather than penalizing models with idealized, rigid geometric rules (e.g., assuming opposite actions must strictly cancel out), a more reasonable evaluation should test whether a model continues to follow its own empirical patterns when executing complex actions.

Based on this philosophy, we introduce the **ConsisWorld** benchmark to comprehensively evaluate interactive consistency across three dimensions: spatial geometry preservation, single-action execution, and multi-action composition. To answer the aforementioned predictive question, ConsisWorld first estimates the native motion magnitude induced by a single command for a given model and initial view, and then holds these initial estimates fixed as a baseline. In the spatial test, we construct a fixed 3D reference from the initial image, depth, and camera parameters, evaluating geometry preservation by comparing the reprojection of reference points with their tracked locations in generated images. In the action tests, to extract the model's empirical patterns, we repeatedly execute six basic actions and extract the actual step lengths, turning angles, and any accompanying lateral drift from the first four steps. Keeping these measured baselines fixed, the single-action test examines whether subsequent sustained execution deviates from this baseline. The composition test superimposes these single-step baselines in command order to predict the final pose of composite routes, comparing it against observations from independently generated multi-action videos. Finally, to enable fair comparisons across models with vastly different step sizes, we divide the calculated prediction errors by the model's own single-step magnitude, converting them into a relative error ratio for quantitative analysis.

ConsisWorld contains 24 synthetic scenes from Unreal Engine 5 (UE5) (Epic Games, 2022) and six real-world scenes from ETH3D (Schops et al., 2017), spanning natural terrain, built environments, and indoor and outdoor spaces, providing a total of 140 initial views. Each view includes an image, depth, and camera parameters, accompanied by a complete set of single-action and composite-route tests. Using this benchmark, we systematically evaluate ten open-source interactive video world models, generating and analyzing 36,400 videos to deeply compare their action behaviors across repeated generations and different initial views.

Our experiments reveal distinctly different performance characteristics in scene preservation and continuous control among existing models. While some models maintain spatial geometry reasonably well, they exhibit substantial motion deviations when repeating the same action. When combining different actions, the actual route endpoints often deviate significantly from predictions based on single-step baselines. Furthermore, even if a model performs stably from one initial view, the displacement induced by the identical action command can shift drastically when switching to other views. These findings demonstrate that **good scene preservation" does not necessarily translate into stable continuous control"**. There remains substantial room for improvement in maintaining sustained motion, handling action transitions, and achieving cross-view control consistency.

In summary, our contributions are as follows:

1. We construct the ConsisWorld benchmark with 140 carefully curated initial views from 30 diverse synthetic and real-world scenes, combining fixed 3D geometric references, six basic actions, and four composite route families to jointly evaluate both scene preservation and action consistency.

2. We introduce a customized motion reference mechanism based on a model's own short-sequence execution. By holding this baseline fixed, we innovatively quantify the accuracy of a model's empirical motion patterns in predicting its subsequent repeated actions and complex action compositions.

3. We systematically evaluate 36,400 videos generated by ten open-source models, revealing that even models with strong geometry preservation widely suffer from consistency flaws, including sustained motion deviations, composite-route endpoint prediction errors, and substantial shifts in action magnitude across different views.

2 Related Work

2.1 Video generation evaluation

VBench (Huang et al., 2024), VBench++ (Huang et al., 2025), and EvalCrafter (Liu et al., 2024) assess perceptual quality, temporal consistency, and semantic alignment in generated videos. VideoPhy and VideoPhy-2 evaluate physical commonsense in object interactions (Bansal et al., 2025; 2026), while WorldModelBench examines instruction following and physical adherence (Li et al., 2025). While these metrics target passive video synthesis, they complement our specific focus on scene geometry and camera motion predictability during active, command-driven interaction.

2.2 Interactive world model evaluation

WorldScore evaluates controllability and dynamics along prescribed camera trajectories (Duan et al., 2025). WorldMark standardizes heterogeneous control interfaces and measures direction, purity, response latency, and motion stability (Xu et al., 2026). iWorld-Bench and WBench extend evaluation to trajectory following, memory, and multi-turn interaction (Fang et al., 2026; Ying et al., 2026). Because evaluating absolute trajectories can penalize models with inherently asymmetric dynamics, ConsisWorld uses empirical motion measured in separate single-action executions as a fixed reference, testing whether it predicts subsequent repetitions and new action combinations.

2.3 Spatial and action consistency

MEt3R measures multi-view consistency through feature comparisons using reconstructed geometry (Asim et al., 2025), while LoopNav evaluates spatial consistency through loop navigation (Lian et al., 2025). WorldMark assesses global consistency using geometry reconstructed post-hoc from generated videos (Xu et al., 2026). Our spatial reference instead derives directly from the initial depth and calibration, remaining fixed to prevent compounding generation errors. Regarding action dynamics, World Models as Group Actions evaluates theoretical identity, inverse, and local composition relations (Wang et al., 2026a). For action composition, ConsisWorld predicts endpoint relations from independently measured empirical motion rather than assuming opposite commands perfectly cancel out. Each model's unique step lengths and accompanying rotational motion are faithfully retained in the prediction.

3 Data Construction and Benchmark Design

3.1 Scenes and geometric references

ConsisWorld comprises 24 synthetic scenes from Unreal Engine 5 (UE5) and 6 real-world ETH3D scenes, spanning diverse natural and built environments (e.g., mountains, deserts, grasslands, offices, and courtyards), as shown in Figure 1. These two sources provide 104 and 36 initial views, respectively, totaling 140 views across the 30 scenes. Each scene contains at least three views with varying camera poses. An initial view—comprising an RGB image, aligned depth, camera intrinsics, and initial pose—serves as a distinct evaluation case from which models generate videos.

UE5 provides rendered images, depth, and camera records, while ETH3D supplies calibrated images and laser-scanned real-world geometry (Schops et al., 2017). For ETH3D's sparse depth, as illustrated in the bottom row of Figure 1, we establish 3D references exclusively at pixels with valid depth. All depth is standardized to the camera axis. To preserve strict alignment between generation inputs and reference geometry, camera intrinsics and reference pixel coordinates are jointly updated during any image cropping or resizing. Appendix A specifies the data and coordinate conventions.

3.2 Evaluation tasks and test design

ConsisWorld jointly evaluates interactive consistency through three specific tests: spatial preservation, single-action execution, and multi-action composition.

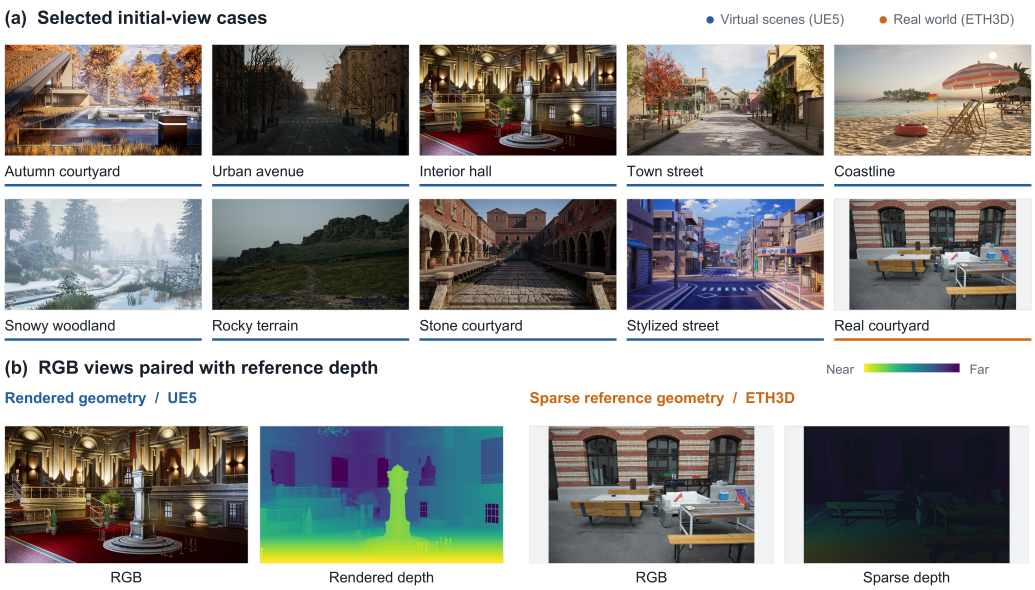

Figure 1: Initial-view cases and reference depth. Top: selected synthetic UE5 scenes and a real-world ETH3D scene. Bottom: RGB views paired with rendered UE5 depth and sparse ETH3D reference depth.

Spatial preservation tests. To assess whether generated views maintain consistent scene structure, we project fixed 3D reference points into the generated video frames using recovered camera poses. Geometric consistency is evaluated by measuring the spatial discrepancy between these reprojected points and their tracked locations in the generated images. This design yields the Ray score (detailed in Section 4.1).

Single-action tests. For each initial view, we generate videos of six basic actions: forward (W), backward (S), left translation (A), right translation (D), left turn (YL), and right turn (YR). Each action is executed for 16 native control units (steps). We estimate the empirical single-step motion from the cumulative displacement and rotation after the first four steps. Keeping this baseline fixed, we extrapolate the expected cumulative motion at steps 12 and 16, comparing it against the actual camera motion recovered from the video. This tests whether short-term action effects remain consistent during sustained execution, yielding the Atomic and directional accuracy (Dir.) scores (Section 4.2).

Table 1: Shortened schematic route pairs. Bold actions mark differences.

Route family	Route 1	Route 2
G1: Inverse composition	W $\rightarrow$ **A** $\rightarrow$ **D** $\rightarrow$ S	W $\rightarrow$ **W** $\rightarrow$ **S** $\rightarrow$ S
G2: Translation order	**W** $\rightarrow$ **D** $\rightarrow$ S $\rightarrow$ A	**D** $\rightarrow$ **W** $\rightarrow$ S $\rightarrow$ A
G3: Order after turning	YR $\rightarrow$ **W** $\rightarrow$ **D** $\rightarrow$ S $\rightarrow$ A	YR $\rightarrow$ **D** $\rightarrow$ **W** $\rightarrow$ S $\rightarrow$ A
G4: Order across two turns	YR $\rightarrow$ **W** $\rightarrow$ **D** $\rightarrow$ YL $\rightarrow$ **S** $\rightarrow$ **A**	YR $\rightarrow$ **D** $\rightarrow$ **W** $\rightarrow$ YL $\rightarrow$ **A** $\rightarrow$ **S**

Composite-route tests. Each case executes eight routes (at least eight steps each), organized into four pairs based on action content and execution order. The two routes in a pair are generated separately from the same initial image by replacing selected actions or altering their order. We compose the fixed single-action measurements to predict the relative pose between the two endpoints, then compare it with the observed relative pose to obtain Trans. and Yaw (Section 4.3). Shared pose biases can cancel in this relative comparison, so individual-route endpoint scores are also reported in Appendix H.2.

G1 to G4 denote the four route-pair constructions. G1 replaces a middle segment of opposing actions, while G2 to G4 alter the translation order at the initial orientation, after a shared turn, and

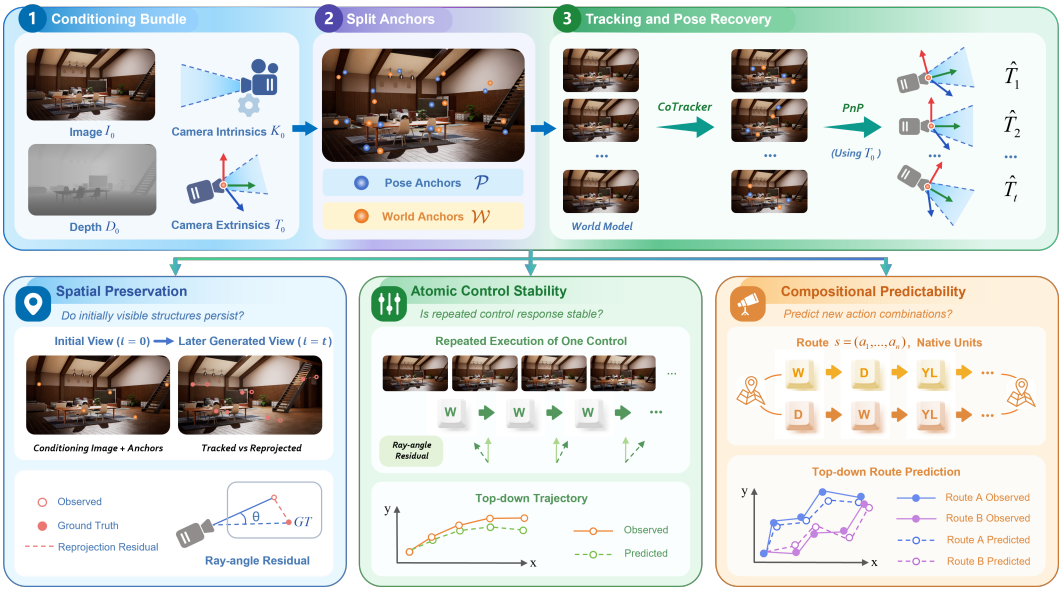

Figure 2: **Overview of ConsisWorld.** Two disjoint anchor sets from the initial 3D reference support camera recovery and geometry evaluation. Single-action measurements from the first four steps predict later repetitions and the relative endpoint poses of independently generated route pairs.

across two turning stages, respectively. Table 1 provides a shortened schematic pair for each family, where each action label denotes one step and arrows indicate execution order. These sequences illustrate construction rules rather than specific evaluated cases.

We assign 120 evaluation route pairs to the 30 scenes (one pair from each family per scene). All initial views and models within a scene share the same eight routes. Consequently, each model generates $1,120$ composite videos across the 140 cases, yielding 560 evaluated route pairs. Appendix B.2 details route construction and assignment.

3.3 Control Interfaces and Execution Settings

Adapters map the six basic actions to each model's native interfaces (e.g., keyboard and mouse inputs, camera-motion parameters, or text prompts) and define each step's input magnitude, duration, and video-frame interval. Single-action and composition tests maintain consistent control and image-preprocessing settings within each model (Appendix B.1).

Because identical commands can induce varying absolute displacements and rotations across models, we retain each model's native action magnitudes and measure its single-step motion at each initial view. Subsequent predictions use these measurements, and prediction errors are normalized by the corresponding motion magnitude. This relative error metric effectively evaluates how consistently models maintain their own action effects during sustained execution and composition, independent of intrinsic step lengths or turning angles.

4 Consistency Evaluation

ConsisWorld jointly evaluates spatial and action consistency using fixed initial 3D geometry and camera motion recovered from generated videos. Figure 2 summarizes the pipeline.

4.1 Spatial Geometry Preservation

We back-project valid-depth pixels from the initial image into a fixed 3D reference and track them with CoTracker3 (Karaev et al., 2024). Disjoint Pose Anchors $\mathcal{P}$ and World Anchors $\mathcal{W}$ support camera recovery and geometry evaluation, respectively. Camera poses are initialized with

EPnP/RANSAC and refined through nonlinear optimization and ray-space refinement (Lepetit et al., 2009; Fischler & Bolles, 1981); World Anchors are excluded from pose fitting. We average the angular discrepancies between tracked World Anchors and their projections under the recovered poses, normalize by 1°, and convert the result to Ray. Only valid point–frame observations, excluding the input frame, enter this average. Appendix C gives the projection equations, solver, and observation checks.

4.2 Single-step motion estimation and sustained execution

For each model, initial view, seed, and action a, let $\mathbf{H}_{a,k} \in SE(2)$ represent horizontal displacement and yaw after step k relative to the action's start. Specifically, $\mathbf{H}_{a,k} = \Pi(\widehat{\mathbf{T}}_{c_0}^{-1}\widehat{\mathbf{T}}_{c_k})$, where $\widehat{\mathbf{T}}_t$ is the recovered camera-to-world pose, c_0, c_k denote the start and step-k boundaries, and Π retains planar motion. We estimate a single-step transform from the first four steps of each single-action video:

$$\bar{\boldsymbol{\xi}}_a = \frac{1}{K}\sum_{k=1}^{K}\frac{\mathrm{Log}(\mathbf{H}_{a,k})}{k}, \qquad \mathbf{U}_a = \mathrm{Exp}(\bar{\boldsymbol{\xi}}_a), \qquad K = 4. \tag{1}$$

Here, Log maps each cumulative transform to the Lie algebra; division by k expresses it per step before averaging and mapping back with Exp. The six actions are estimated separately, retaining their individual step lengths, lateral motion, and accompanying turns.

With $\mathbf{U}_a$ fixed, $\widehat{\mathbf{H}}_{a,k} = \mathbf{U}_a^k$ predicts motion at steps 12 and 16 of the same video. Matrix multiplication updates both position and orientation, preserving the effect of turning on later displacement. Atomic scores translation errors for movement commands and yaw errors for turns, normalized by k times the corresponding single-step magnitude. Dir. independently checks whether cumulative motion follows the requested direction. Appendix D.1 specifies the transforms, error formulas, and direction tests.

4.3 Action composition prediction

Composite routes are generated independently from the same initial image and predicted using fixed single-action estimates for the same model, initial view, and seed. These estimates retain measured differences between opposing actions and are not fitted to the composite videos. For a route $q = (a_1, \ldots, a_n)$, we multiply the transforms in command order. Given observed endpoints $\mathbf{H}_A, \mathbf{H}_B$ relative to their respective starts, the predicted endpoints and relative-pose error are

$$\begin{gathered}\widehat{\mathbf{H}}_q = \mathbf{U}_{a_1}\mathbf{U}_{a_2}\cdots\mathbf{U}_{a_n}, \qquad \widehat{\boldsymbol{\Delta}} = \widehat{\mathbf{H}}_A^{-1}\widehat{\mathbf{H}}_B, \\ \boldsymbol{\Delta} = \mathbf{H}_A^{-1}\mathbf{H}_B, \qquad \mathbf{E} = \widehat{\boldsymbol{\Delta}}^{-1}\boldsymbol{\Delta}.\end{gathered} \tag{2}$$

Here, $\widehat{\boldsymbol{\Delta}}$ and $\boldsymbol{\Delta}$ are the predicted and observed endpoint relative poses. The translation and yaw residuals of $\mathbf{E}$ are normalized by the median single-step displacement of the four translation actions and the median absolute angle of the two turns, respectively. These errors yield Trans. and Yaw and are not further divided by route length. Shared pose biases may cancel in relative comparisons, so these scores do not replace single-route endpoint evaluation. Exact definitions are in Appendix D.3, with single-route results in Appendix H.2.

4.4 Scoring and validity

Ray, Atomic, Trans., and Yaw use $s(r) = 100/(1 + r)$ to convert normalized errors into scores. Pose checks require tracking coverage and inlier ratios of at least 0.1, both relative to the initial Pose Anchor count. Spatial scoring additionally requires at least 12 valid World Anchors per frame. Composition additionally requires all six single-step estimates, valid poses at the starts and every action boundary, input-frame matching, and per-step motion of at least 10% of the corresponding estimated axial magnitude. Appendices C and D specify all checks and thresholds.

For the valid pair set $\mathcal{V}$ among $N = 560$ planned route pairs, support and overall scores are

$$Q = \frac{|\mathcal{V}|}{N}, \qquad S_d = \frac{1}{N}\sum_{i\in\mathcal{V}} s(r_{d,i}), \qquad d \in \{t, \psi\}. \tag{3}$$

Support is $100Q$, and failed pairs contribute zero. When $\mathcal{V}$ is nonempty, $S_d = Q\, S_{d,\mathrm{valid}}$, where $S_{d,\mathrm{valid}}$ is the valid-pair mean. Ray assigns zero to videos without any valid observations; Atomic assigns zero to unscorable checkpoints.

5 Experimental Setup

5.1 Models and video generation

The main evaluation includes ten models: Matrix-Game 2.0 (He et al., 2025), Matrix-Game 3.0 (Wang et al., 2026b), SANA-WM (Zhu et al., 2026), DreamX-World (Team et al., 2026b), Hunyuan-World 1.5 (WorldPlay) (Sun et al., 2025), HY-World 2.0 (Team HY-World, 2026), LingBot-World v2 (Gao et al., 2026), AlayaWorld (Team et al., 2026a), H3-World (Chen et al., 2026), and Lyra 2.0 (Shen et al., 2026). For the 140 initial views (Section 3), each single action uses three random seeds, with motion estimated and scored per seed. Composite routes use seed 0 and are evaluated against the corresponding seed-0 single-action estimates. This yields 18 single-action and eight composite videos per view, totaling $36,400$ videos across ten models.

5.2 Aggregation and visual metrics

Errors are converted to scores before aggregation. **Ray** averages the eight composite video scores per view, then across all 140 views. For each action, **Atomic** and **Dir.** first average steps 12 and 16 and the three seeds, then average views within each scene and across scenes; the six action scores are averaged with equal weight. Composition scores (**Trans.** and **Yaw**) aggregate all 560 route pairs (Equation (3)), assigning a score of zero to invalid pairs. **Support** reports the valid pair fraction (valid-only means are in Appendix H.2).

Action stability (**RSD**) is the relative standard deviation of the signed requested-axis motion (Equation (1)). Repeatability computes RSD across the three seeds, while cross-view variation computes it across initial views within the same scene. Appendix F details the exact groupings.

We use $2,000$ scene-level bootstrap replicates for 95% intervals on paired Full-minus-Ideal differences (Appendix H.3). Selected main-table model differences use paired resampling stratified by UE5 and ETH3D (Appendix H.5). **MUSIQ** (Ke et al., 2021) and **Aesthetic** quality are evaluated using VBench (Huang et al., 2024) implementations. We score the midpoint frame of each of 16 equal temporal intervals, average within each video and then over all $3,640$ videos per model, and report VBench-normalized values scaled by 100.

6 Results and Analysis

6.1 Similar geometry scores do not imply consistent sustained motion

Under the evaluated controls, models with similar Ray scores can differ substantially in sustained-action consistency. In Table 2, AlayaWorld and Lyra 2.0 score 88.9 and 89.4 on Ray, but 28.8 and 88.4 on Atomic. DreamX-World achieves 95.2% directional accuracy yet an Atomic score of 46.9, showing that correct direction does not ensure motion remains consistent with the first four steps.

Lyra 2.0 leads both Ray and Atomic, while HY-World 2.0 achieves the highest composition translation score. These models use 3D geometry or retrieved-view memory during generation (Shen et al., 2026; Team HY-World, 2026). Ray must nevertheless be interpreted alongside actual motion: smaller movements can reduce the difficulty of preserving visible geometry. Paired confidence intervals and observation support are reported in Appendix H.5; the available records do not support a whole-video motion-matched comparison between AlayaWorld and Lyra 2.0.

6.2 Action magnitudes vary across initial views

All ten models have higher translation RSD across initial views than across repeated generations from the same view (Table 3). HY-World 2.0 and Lyra 2.0 have repeat-generation RSDs of 2.5% and 3.6%, but cross-view RSDs of 38.2% and 29.2%. Stable repetition therefore does not imply that the same action magnitude carries over to other initial views. AlayaWorld and H3-World also vary

Table 2: Ten-model evaluation (↑). Best in bold; second best underlined.

Model	Geometry	Single action		Composition			Visual quality	
	Ray	Dir. (%)	Atomic	Support (%)	Trans.	Yaw	MUSIQ	Aesthetic
Matrix-Game 2.0	67.8	62.7	58.4	85.2	36.3	71.0	67.1	56.1
Matrix-Game 3.0	68.0	71.6	63.2	93.8	57.1	80.3	72.1	56.3
SANA-WM	79.4	93.3	76.5	45.7	16.6	33.7	69.6	55.4
DreamX-World	81.3	95.2	46.9	78.6	16.4	34.2	<u>74.3</u>	60.7
Hunyuan-World 1.5	84.5	**99.5**	77.8	31.2	12.1	25.8	71.6	58.6
HY-World 2.0	79.3	95.7	<u>87.1</u>	**98.6**	**80.4**	**92.0**	72.9	**62.7**
LingBot-World v2	76.6	94.0	83.0	90.7	45.4	73.9	72.2	<u>61.5</u>
AlayaWorld	<u>88.9</u>	91.4	28.8	36.4	3.0	17.7	71.5	60.9
H3-World	75.2	41.1	26.5	19.8	5.2	10.4	63.3	55.5
Lyra 2.0	**89.4**	<u>98.3</u>	**88.4**	<u>98.2</u>	<u>68.0</u>	<u>91.4</u>	**75.5**	59.1

substantially at fixed views, with translation RSDs of 53.0% and 77.0%. Complete-group counts are reported in Appendix F.

Table 3: Action RSD (%, ↓). Best in bold; second best underlined.

Model	Repeated generation		Across initial views	
	Trans.	Turn	Trans.	Turn
Matrix-Game 2.0	13.6	1.1	40.8	6.1
Matrix-Game 3.0	4.8	3.2	**24.9**	6.5
SANA-WM	14.4	3.1	36.8	5.3
DreamX-World	23.8	33.0	38.6	19.3
Hunyuan-World 1.5	46.2	6.6	47.6	9.0
HY-World 2.0	**2.5**	<u>0.7</u>	38.2	<u>4.8</u>
LingBot-World v2	19.2	4.0	32.7	10.7
AlayaWorld	53.0	58.7	98.1	106.9
H3-World	77.0	52.0	84.0	51.9
Lyra 2.0	<u>3.6</u>	**0.4**	<u>29.2</u>	**0.9**

6.3 Insufficient motion and endpoint deviations in composition

Composition scores reflect both the fraction of routes meeting the evaluation conditions and prediction accuracy on those routes. Figure 3 separates these two aspects.

Weak motion reduces support. Removing the motion-magnitude check raises Hunyuan-World 1.5's support from 31.2% to 100.0%, and SANA-WM's from 45.7% to 98.4% (Figure 3a). The additional pairs pass the other checks but contain at least one step below the motion threshold. This loss of support should be distinguished from endpoint prediction error; it does not by itself establish that the camera stopped completely.

Supported routes can still depart from predictions. DreamX-World's conditional translation scores are only 19.6 for individual endpoints and 20.9 for endpoint relations (Figure 3b). LingBot-World v2 scores 32.9 and 50.0, respectively. Shared pose biases can cancel in a relative comparison, so agreement between endpoint relations does not guarantee accurate individual endpoints. Support and conditional scores are therefore reported together.

Longer single-action measurements do not generally resolve these combination errors. On common valid samples with fixed K4 normalization, extending estimation from four to eight steps raises DreamX-World's step-16 Atomic score from 47.4 to 54.4, while its composition translation score remains 21.3. Predictor differences are concentrated in the inverse-route family rather than the order families. Appendix H.5 gives the route audit, family-specific comparisons, and window analysis.

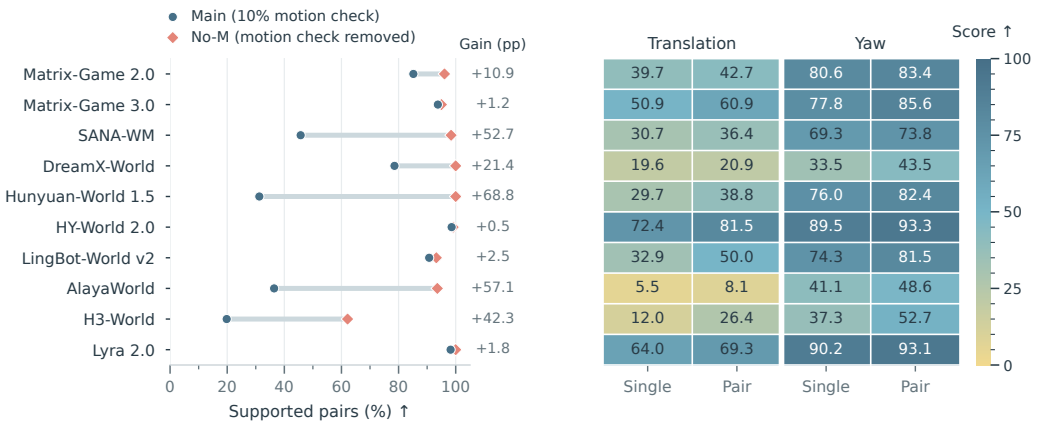

Figure 3: **Composition diagnostics. Left**: support under Main and No-M, which removes only the per-step 10% relative motion-magnitude check, with 560 planned route pairs per model. Gain reports the No-M-minus-Main support increase in percentage points. **Right**: single-route endpoint and route-pair relative-pose scores, averaged over the same Main-valid pairs within each model. Higher support and scores are better. Data are from Tables H.5 and H.4.

6.4 Measurement accuracy validation

On 30 known-motion UE5 segments, median tracking error is 0.117% of the image diagonal; median translation and rotation errors are 1.480% and 0.306% of ground-truth motion. Across five UE5 control groups, each with six eight-step single-action videos and two composite routes, we estimate motion from steps 1–4 and evaluate steps 5–8. Scores are computed within each group and then averaged across the five groups. Ground-truth and recovered single-action scores average 100.0 and 98.8; final route-pair translation scores average 99.8 and 99.1, and yaw scores 100.0 and 98.2. Intermediate-boundary differences are larger, reaching 11.0 translation and 11.1 yaw points in the five-group means. Full results and anchor-distribution tests are in Appendices E and H.1.

6.5 Pitch control

While eight models achieve at least 99.1% pitch-direction accuracy at step 8, proper spatial disentanglement remains challenging. AlayaWorld achieves only 78.1% accuracy and suffers from severe axis coupling, exhibiting a median additional-yaw-and-roll to pitch ratio of 42.6%. This indicates that vertical commands heavily induce unintended planar rotations. In contrast, the top eight models effectively isolate pitch motion, maintaining minimal coupling ratios of 2.4%–6.9%. The 16-step videos, eight-step evaluation window, and complete results are described in Appendix G.2.

7 Conclusion

We introduce ConsisWorld, a benchmark that tests spatial scene preservation against fixed 3D geometric references. Crucially, it evaluates sustained execution and multi-action composition using a self-referential empirical baseline derived from each model's measured single-step motion, explicitly accommodating inherent asymmetric step dynamics. Our systematic ten-model evaluation reveals a fundamental decoupling: strong geometry preservation frequently coexists with substantial trajectory deviations during continuous control. Action composition exposes critical failure modes, notably premature motion decay and endpoint relations that severely depart from baseline predictions. Furthermore, the displacement induced by identical commands varies considerably across different initial views, indicating models fail to internalize globally consistent physical scales. By jointly evaluating these entangled capabilities, ConsisWorld provides a rigorous diagnostic framework and quantitative evidence necessary for reducing continuous-control drift, improving action transitions, and ensuring reliable interaction across diverse environments.

AI USE STATEMENT

Generative AI tools were used for translation and language editing, feedback on methodology, assistance with evaluation and reporting code, data format conversion and the collection and organization of experimental files, figure preparation, and discussions of result interpretation. The authors have reviewed the AI-assisted text, code, figures, and analytical suggestions incorporated into this work. All code used in this work has been tested, and the data have been manually verified by the authors. The authors take responsibility for the final manuscript, analyses, and scientific conclusions.

REPRODUCIBILITY STATEMENT

The benchmark uses synthetic environments rendered in UE5 and public ETH3D reference scenes. Data processing and control settings are described in Appendices A and B. Measurement and scoring procedures are specified in Appendices C and D. We plan to release evaluation and reporting code, model adapters, input command manifests, and the associated parameter and environment configurations to support reproduction of the evaluation protocol.

REFERENCES

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6034–6044. IEEE, 2025.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, volume 2025, pp. 102075–102121, 2025.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, volume 2026, pp. 118456–118470, 2026.

Danze Chen, Zeqing Wang, Ziyue Lin, Xingyi Yang, and Yeying Jin. H3-world: Turning language understanding into world control. *arXiv preprint arXiv:2609.01560*, 2026.

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.

Epic Games. Unreal engine 5. `https://www.unrealengine.com`, 2022. Computer software. Available: https://www.unrealengine.com.

Jianjie Fang, Yingshan Lei, Qin Wan, Ziyou Wang, Yuchao Huang, Yongyan Xu, Baining Zhao, Weichen Zhang, Chen Gao, Xinlei Chen, et al. iworld-bench: A benchmark for interactive world models with a unified action generation framework. *arXiv preprint arXiv:2605.03941*, 2026.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

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.

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.

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

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proc. arXiv:2410.11831*, 2024.

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.

Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81(2):155–166, 2009.

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. 2025.

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

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149. IEEE, 2024.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

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.

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Ruicong Liu, et al. Alayaworld: Long-horizon and playable video world generation. *arXiv preprint arXiv:2607.06291*, 2026a.

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

Team HY-World. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv preprint arXiv:2604.14268*, 2026. doi: 10.48550/arXiv.2604.14268. URL https://arxiv.org/abs/2604.14268.

Zijie Wang, Wei Zhang, Weiming Zhang, Fanqi Zhang, Xiao Tan, Yipeng Qin, and Guanbin Li. World models as group actions. *arXiv preprint arXiv:2605.24578*, 2026a.

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*, 2026b.

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.

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.

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.

Supplementary Material and Appendices

A Data Specifications and Execution Counts

A.1 Coordinates and depth processing

Section 3.1 describes the contents of each case. Here we specify distance units and preprocessing conventions. Camera coordinates follow the OpenCV right-handed convention: x points right, y down, and z forward. Depth is standardized to camera-axis depth and stored in centimeters. UE5 depth is obtained by multiplying the EXR red channel by the conversion factor s_c recorded for each case; ETH3D depth is converted from meters to centimeters, with 3D references established only at pixels with valid depth.

Image cropping or resizing is represented by a pixel transform $\mathbf{A}$. Intrinsics are updated as $\mathbf{K}' = \mathbf{AK}$, and reference pixel coordinates are transformed consistently, preserving pixel-center conventions. Both sources are screened for completeness of reference information, and initial views are grouped by their source scenes. Figure 1 illustrates the scene images and reference depth.

A.2 Execution counts

Each model generates 2,520 single-action videos and 1,120 composite videos, totaling 3,640. The single-action set covers 140 initial views, six actions, and three random seeds. The composition set executes eight routes per view with seed 0, yielding 560 evaluated pairs. Composite videos are shared by the spatial and action-composition tests.

For each model, the composition set contains 280 eight-step, 280 eleven-step, and 560 fourteen-step videos. Seeds 0, 1, and 2 account for 1,960, 840, and 840 videos, respectively. The main evaluation contains 36,400 videos across ten models.

B Model Adaptation and Route Design

B.1 Control settings

Table B.1 lists the control and output configurations for the six basic actions in each model. Control magnitudes use each model's native input units. Single-action and composition tests use consistent settings within each model.

Table B.1: Model control and output configurations.

Model	Frames/ unit	FPS	Resolution	Control and preprocessing
Matrix-Game 2.0	4	16	640×352	Keyboard 1; mouse 0.1; center crop
Matrix-Game 3.0	4	17	1280×704	Keyboard 1; mouse 0.1; center crop
SANA-WM	4	16	1280×704	Native action sequence; center crop
DreamX-World	4	16	1280×704	Translation 0.075 and yaw 1.5° per frame interval; stretch resize
Hunyuan-World 1.5	4	24	832×480	Native w/s/a/d/left/right control tokens; native crop
HY-World 2.0	2	16	832×480	Translation 0.025 and yaw 3° per action unit; stretch resize
LingBot-World v2	4	16	832×464 or 768×512	Translation 0.025 and yaw 0.75° per frame interval; stretch resize
AlayaWorld	8	24	960×544	Forward/backward 1.27, lateral 0.0846, yaw 2° per unit; center crop
H3-World	17	24	832×480	Fixed camera-motion text prompts; model-specific crop
Lyra 2.0	10	16	832×480	Translation 0.10 and yaw 2° per action unit; stretch resize

B.2 Route construction and assignment

Table 1 illustrates the pairing rules for G1 to G4 with shortened schematic examples. By varying translation directions, action-block lengths, turning directions, and replacement positions, the route library constructs 96 pairs per family, totaling 384 pairs and 768 distinct action sequences. Evaluation selects 120 pairs for the 30 scenes, assigning one pair from each family to each scene.

All initial views and models within a scene share these eight routes, while different scenes receive different route combinations. Construction checks action-label validity, equal lengths within each pair, uniqueness of expanded sequences, and consistency of scene assignments. G1 is labeled the inverse group, and G2 to G4 the order group; these names identify route construction.

C Pose Recovery and Geometric Scoring Details

C.1 Depth back-projection and coordinate conversion

For a reference pixel $\mathbf{p} = [u, v, 1]^\top$ with camera-axis depth $z(\mathbf{p})$, its 3D location in the initial camera coordinates is $\mathbf{P}_{\text{cam}} = z(\mathbf{p})\mathbf{K}^{-1}\mathbf{p}$. Let $\mathbf{T}_0^{\text{c2w}}$ be the initial camera-to-world transform and $\hat{\mathbf{T}}_t^{\text{w2c}}$ the recovered world-to-camera transform at frame t. The current camera pose in the initial camera coordinate system is

$$\mathbf{G}_t = \left(\mathbf{T}_0^{\text{c2w}}\right)^{-1} \left(\hat{\mathbf{T}}_t^{\text{w2c}}\right)^{-1}. \tag{C.1}$$

Geometric scoring uses the full $SE(3)$ pose; single-action and main composition scoring use the planar transform defined in Appendix D.1.

C.2 Anchors and solver

Reference pixels with valid depth are sampled on a grid and split by checkerboard parity into disjoint Pose Anchors $\mathcal{P}$ and World Anchors $\mathcal{W}$. CoTracker3 extracts 2D trajectories for both sets. A valid track must have finite in-frame coordinates, visibility of at least 0.9, and confidence of at least 0.5.

We initialize poses with EPnP/RANSAC (Lepetit et al., 2009; Fischler & Bolles, 1981) and refine them with LM (Levenberg–Marquardt) and ray-space least squares. RANSAC uses a pixel threshold of $\sqrt{f_x f_y}\tan(1^\circ)$. Initialization requires at least 12 valid Pose Anchors, and the final solution must retain at least six inliers with angular errors no greater than 1°.

Let v_{it} indicate a valid observation of Pose Anchor i at frame t, and let ϕ_{it}^{Pose} be its angular error. Tracking coverage and the inlier ratio are

$$\begin{aligned} c_t &= \frac{\sum_{i\in\mathcal{P}} v_{it}}{|\mathcal{P}|}, \\ \rho_t &= \frac{\#\{i \in \mathcal{P} : v_{it} = 1,\ \phi_{it}^{\text{Pose}} \le 1^\circ\}}{|\mathcal{P}|}. \end{aligned} \tag{C.2}$$

Both ratios use the initial number of Pose Anchors as their denominator. The main configuration requires a valid, finite rigid transform with $c_t \ge 0.1$ and $\rho_t \ge 0.1$. An empty initial Pose Anchor set is unmeasurable.

C.3 Geometric observation filtering and aggregation

Let $\widehat{\mathbf{T}}_t = (\hat{\mathbf{T}}_t^{\text{w2c}})^{-1} \in SE(3)$ be the recovered camera-to-world transform at frame t. For a fixed World Anchor $\mathbf{X}_i \in \mathcal{W}$, let $\widetilde{\mathbf{x}}_{i,t}$ and $\widehat{\mathbf{x}}_{i,t}$ be its tracked and projected image locations. The projected

location, angular residual, and mean video residual are

$$\begin{aligned} \widehat{\mathbf{x}}_{i,t} &= \pi_{\mathbf{K}}\left(\widehat{\mathbf{T}}_t^{-1}\mathbf{X}_i^h\right), \\ e_{i,t}^{\mathrm{ray}} &= \angle\left(\mathbf{K}^{-1}\widetilde{\mathbf{x}}_{i,t}^h, \mathbf{K}^{-1}\widehat{\mathbf{x}}_{i,t}^h\right), \\ \bar{e}_{\mathrm{ray}}(v) &= \frac{1}{|\mathcal{E}_v|}\sum_{(i,t)\in\mathcal{E}_v} e_{i,t}^{\mathrm{ray}}. \end{aligned} \tag{C.3}$$

Here, $\mathbf{K}$ is the image's intrinsic matrix, $\pi_{\mathbf{K}}$ is perspective projection, and superscript h denotes homogeneous coordinates. $\mathcal{E}_v$ contains the valid point–frame observations in video v, excluding the input frame. Angular residuals express projection discrepancies in the same angular units across image resolutions; the reference 3D points remain fixed.

A scored frame must pass the pose checks and contain at least 12 valid World Anchors. Each validation point must have a valid track, finite projected coordinates, and positive camera depth. All observations meeting these conditions enter the residual calculation, without further filtering by residual magnitude.

The video's mean angular error is normalized by 1° and converted to Ray with the common scoring map used in Section 4.4:

$$s(r) = \frac{100}{1+r}. \tag{C.4}$$

If only some frames are measurable, we average the valid point-frame observations; a video with none receives 0. The Pose-Anchor ratios in Appendix C.2 determine pose validity; $\mathcal{E}_v$ records the World-Anchor observations contributing to Ray. The model score averages the eight videos within each case, then the 140 cases.

Actual Ray observation support. For each video, let T be the number of planned noninitial frames, W the initial World-Anchor set, and E the point–frame observations actually admitted to Ray scoring under the existing pose and minimum-point checks. Frame support is the fraction of these T frames with a nonempty scored set; observation support is $|E|/(|W|T)$. Ratios are computed per video and aggregated with the original case weights. Empty W is reported separately as undefined, without changing Ray scoring. Temporal thirds use planned frame indices, including unsupported frames. Appendix H.5 reports the results.

D Single-Step Motion and Composition Scoring Details

D.1 Planar transforms and single-step estimation

Let c_0 denote the action's start and c_k the end of step k. The relative pose is $\widehat{\mathbf{T}}_{c_0}^{-1}\widehat{\mathbf{T}}_{c_k} = \mathbf{G}_{c_0}^{-1}\mathbf{G}_{c_k}$; let $\mathbf{R}_{a,k}$ and $(t_{x,a,k}, t_{y,a,k}, t_{z,a,k})^\top$ denote its rotation and translation. The mapping Π retains horizontal displacement and yaw to form the cumulative $SE(2)$ transform used in Equation (1):

$$\begin{aligned} \mathbf{H}_{a,k} &= \Pi\left(\mathbf{G}_{c_0}^{-1}\mathbf{G}_{c_k}\right) \\ &= \begin{bmatrix} \cos\theta_{a,k} & \sin\theta_{a,k} & t_{x,a,k} \\ -\sin\theta_{a,k} & \cos\theta_{a,k} & t_{z,a,k} \\ 0 & 0 & 1 \end{bmatrix}, \\ \theta_{a,k} &= \mathrm{atan2}((\mathbf{R}_{a,k})_{02}, (\mathbf{R}_{a,k})_{00}). \end{aligned} \tag{D.1}$$

Matrix indices start at zero, transforms act on column vectors, and $\mathbf{H}_{a,0} = \mathbf{I}$. Positive yaw turns the camera toward $+x$, and action transforms are successively right-multiplied in execution order. For a planar transform, $\mathbf{t}(\mathbf{H})$ extracts the translation column, and $\theta(\mathbf{H}) = \mathrm{atan2}(H_{01}, H_{00})$.

For $\boldsymbol{\xi} = (\rho_x, \rho_z, \theta)^\top$, the Lie algebra matrix and exponential and logarithmic maps are

$$\boldsymbol{\xi}^\wedge = \begin{bmatrix} 0 & \theta & \rho_x \\ -\theta & 0 & \rho_z \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathrm{Exp}(\boldsymbol{\xi}) = \exp(\boldsymbol{\xi}^\wedge), \qquad \big(\mathrm{Log}(\mathbf{H})\big)^\wedge = \log\mathbf{H}. \tag{D.2}$$

Equation (1) maps each cumulative transform from the first $K = 4$ steps to the Lie algebra, divides by its step count, averages these estimates, and maps back to a single-step transform. The six actions are estimated separately, retaining their individual step lengths, lateral motion, and accompanying turns. We denote the resulting transform by $\mathbf{U}_a^{[4]}$ in this appendix. The start and first four action boundaries must pass the pose checks, and the absolute cumulative yaw used in estimation must be below 170°. If any required boundary is unavailable, the action's single-step estimate is unavailable.

To quantify action variation and check observed motion magnitudes, let $\bar{\boldsymbol{\xi}}_a^{[4]} = (\rho_{x,a}, \rho_{z,a}, \omega_a)^\top$ and extract an axial component that is positive in the requested direction:

$$u_a = \begin{cases} \rho_{z,a}, & a = \mathrm{W}, \\ -\rho_{z,a}, & a = \mathrm{S}, \\ -\rho_{x,a}, & a = \mathrm{A}, \\ \rho_{x,a}, & a = \mathrm{D}, \\ -\omega_a, & a = \mathrm{YL}, \\ \omega_a, & a = \mathrm{YR}. \end{cases} \tag{D.3}$$

u_a is used for RSD in Appendix F, motion checks in Appendix D.3, and predictor definitions in Appendix H.3. Atomic uses the full single-step transform's displacement length or yaw magnitude for normalization in Equation (D.4). Angular computations use radians, and wrap returns the shortest signed angular difference.

Sustained-action prediction. With $\mathbf{U}_a$ fixed, $\widehat{\mathbf{H}}_{a,k} = \mathbf{U}_a^k$ predicts motion at steps 12 and 16 of the same video. Matrix multiplication updates both position and orientation. Here, $\mathcal{A}_t$ denotes the four translation actions (W, S, A, D), and $\mathcal{A}_r$ the two turns (YL, YR). The normalized prediction error is

$$r_{a,k} = \begin{cases} \dfrac{\|\mathbf{t}(\mathbf{H}_{a,k}) - \mathbf{t}(\widehat{\mathbf{H}}_{a,k})\|_2}{k\|\mathbf{t}(\mathbf{U}_a)\|_2}, & a \in \mathcal{A}_t, \\ \dfrac{\left|\mathrm{wrap}\Big(\theta(\mathbf{H}_{a,k}) - \theta(\widehat{\mathbf{H}}_{a,k})\Big)\right|}{k|\theta(\mathbf{U}_a)|}, & a \in \mathcal{A}_r, \end{cases} \qquad k \in \{12, 16\}. \tag{D.4}$$

The denominator is the step count times the action's single-step displacement length or absolute yaw, not the predicted endpoint displacement. Each error is converted to an Atomic score using Equation (C.4). Atomic is computed when the single-step estimate is available, its normalization magnitude is finite and positive, and the start and target poses are valid. Unavailable checkpoints receive 0.

Directional accuracy. Dir. independently evaluates the cumulative motion at steps 12 and 16. For W, S, A, D, YL, and YR, respectively, we check $t_z, -t_z, -t_x, t_x, -\theta, \theta$ relative to the start. An observation receives 1 if the starting and target poses are valid and the corresponding component is positive, and 0 otherwise. Zero response, reversed motion, and unmeasurable observations retain their distinct statuses. Atomic and Dir. follow the aggregation order in Section 5.2.

Figure D.1 summarizes response profiles across all six basic actions over execution steps. Its power-law fits describe the observed response shape; Atomic uses the fixed first-four-step estimates in Equations (1) and (D.4).

D.2 Vertical displacement and pitch during planar actions

We separately record vertical displacement and pitch during planar action execution. At step 12, we remove yaw from the relative rotation to obtain $\mathbf{R}_{\mathrm{res}} = \mathbf{R}_y(\theta)^\top \mathbf{R}$, then extract the rotation vector $\boldsymbol{\omega} = \mathrm{Log}(\mathbf{R}_{\mathrm{res}})$. Relative vertical displacement and pitch are defined as

$$\lambda_y = \frac{|t_y(12)|}{12\|\mathbf{t}(\mathbf{U}_a^{[4]})\|_2}, \qquad \lambda_{\mathrm{pitch}} = \frac{|\omega_x(12)|}{12|\theta(\mathbf{U}_a^{[4]})|}. \tag{D.5}$$

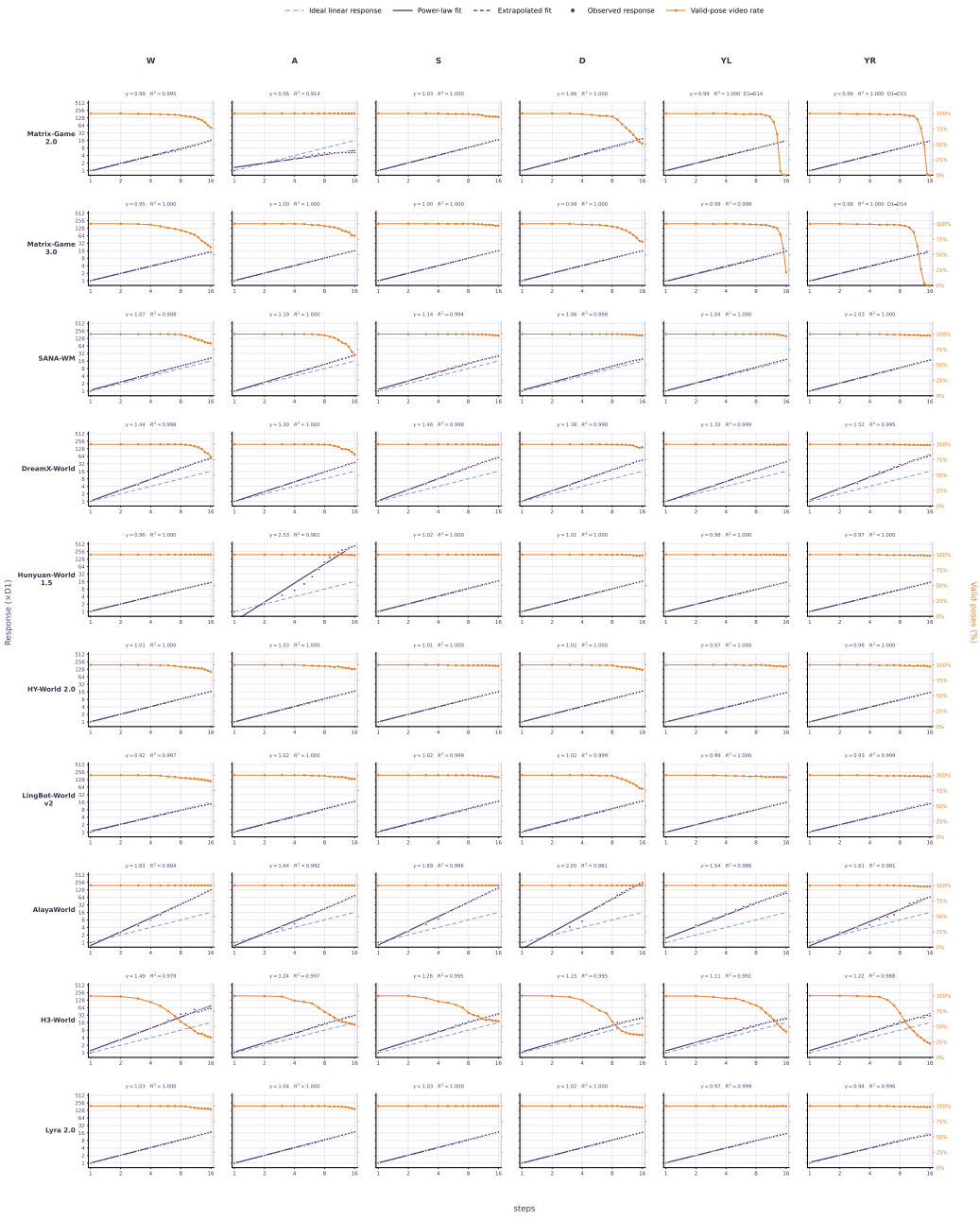

Figure D.1: Single-action response profiles for ten models across six basic actions, normalized by the corresponding first-step response. Annotations report the exponent γ obtained by fitting a power-law response $r(k) \propto k^{\gamma}$, where k is the execution step, and the goodness of fit R^2. $\gamma = 1$ indicates linear response growth. Purple curves show the power-law fits and their dashed extrapolations; the blue dashed lines indicate ideal linear responses. Orange curves show the fraction of videos with valid poses (right axis).

The two quantities are normalized by the corresponding translation and yaw magnitudes, computed when the denominators are finite and positive, and multiplied by 100 for reporting. Appendix G.2 evaluates explicit upward and downward pitch commands.

D.3 Composition scoring conditions and aggregation

Composition prediction follows Equation (2), using the six fixed single-step transforms for the same model, case, and seed. These transforms retain measured differences between opposing actions and are not updated using the composite routes. For the relative-pose error $\mathbf{E}$, translation and yaw residuals are normalized as

$$\begin{aligned} b_t &= \mathrm{median}_{a\in\mathcal{A}_t}\,\|\mathbf{t}(\mathbf{U}_a)\|_2, \quad & r_t &= \frac{\|\mathbf{t}(\mathbf{E})\|_2}{b_t}, \\ b_r &= \mathrm{median}_{a\in\mathcal{A}_r}\,|\theta(\mathbf{U}_a)|, \quad & r_\psi &= \frac{|\,\mathrm{wrap}(\theta(\mathbf{E}))|}{b_r}. \end{aligned} \tag{D.6}$$

Here, b_t is the median displacement length across the four translation actions, and b_r is the median absolute turning angle across the two turning actions. The functions $\mathbf{t}(\cdot)$, $\theta(\cdot)$, and wrap are defined in Appendix D.1. We convert r_t and r_ψ to Trans. and Yaw using Equation (C.4). Composition errors are expressed in single-step motion units and are not further divided by route length.

All six estimates must be available, u_a must be finite and nonzero, and b_t, b_r must be finite and positive.

The starts and all action boundaries of both routes must pass the pose checks in Appendix C.2. At each step, the observed planar displacement length or absolute yaw increment must reach the corresponding $0.1|u_a|$. The mean L1 difference between the first frame and the adapter-processed input image must not exceed 32 on the 0–255 RGB scale.

Pairs satisfying these conditions enter $\mathcal{V}$. Overall scores aggregate all 560 pairs using Equation (3); failed pairs receive 0 and retain their failure reasons. The appendix also reports $S_{d,\mathrm{valid}}$, the average over valid pairs. If that set is empty, the overall score is 0 and the conditional average is unavailable.

E Error Calibration for Pose Recovery

We calibrate pose recovery on 30 segments from ten known-motion UE5 sequences, using ground truth recorded by the rendering engine. Anchors are sampled at valid-depth pixels at each segment's start, and all valid sampled points are used to recover camera motion.

Tracking errors are normalized by the image diagonal. Across 30 segments, the median is 0.117% and the 90th percentile (P90) is 0.175%. Pose errors are normalized by each segment's ground-truth motion magnitude, with 20 segments contributing to translation statistics and ten to rotation statistics.

Table E.1: PnP component ablation with identical tracking inputs. Errors are percentages (↓).

Solver	Solved ↑	Trans. Median ↓	Trans. P90 ↓	Rot. Median ↓	Rot. P90 ↓
EPnP + RANSAC	30/30	2.277	5.160	0.457	0.756
+ LM	30/30	**1.125**	**3.728**	<u>0.310</u>	**0.671**
+ LM + ray refinement	30/30	<u>1.480</u>	<u>3.980</u>	**0.306**	<u>0.675</u>

As shown in Table E.1, all three configurations recover all 30 segments. All three configurations recover all 30 segments. LM yields the lowest translation error. Adding ray refinement reduces median rotation error from 0.310% to 0.306%. Compared with EPnP/RANSAC, the full pipeline reduces median translation and rotation errors from 2.277% and 0.457% to 1.480% and 0.306%, respectively. The best error values are bold and the second-best underlined.

F Action Variation: Statistical Rules and Complete Groups

Table 3 summarizes RSD across repeated generations and initial views. Here we define the statistic and provide complete group counts. Let u_i be the signed axial component in Appendix D.1, and $\bar{u}$ the group mean. For groups containing all planned members, we compute

$$\mathrm{RSD}(u) = 100 \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})^2}}{\frac{1}{n}\sum_{i=1}^{n}|u_i|}. \tag{F.1}$$

The standard deviation retains response signs, while the denominator uses mean absolute response. Valid zero responses participate; groups with missing members or all-zero responses are unavailable.

Cross-seed comparisons fix the scene, case, and action and compare the three seeds. Case RSDs are averaged within each scene, then across scenes. Cross-view comparisons fix the scene, action, and seed and compare all initial views of that scene. Seed-specific RSDs are averaged within each scene, then across scenes. Both settings finally aggregate the four translation actions and two turning actions separately. Tables F.1 and F.2 mark the best and second-best RSD values and provide complete group counts to document statistical coverage.

Table F.1: Cross-seed RSD at fixed initial views and complete group counts.

Model	Trans. RSD ↓	Turn RSD ↓	Trans. groups	Turn groups
Matrix-Game 2.0	13.6	1.1	557/560	278/280
Matrix-Game 3.0	4.8	3.2	553/560	278/280
SANA-WM	14.4	3.1	560/560	280/280
DreamX-World	23.8	33.0	560/560	280/280
Hunyuan-World 1.5	46.2	6.6	560/560	280/280
HY-World 2.0	**2.5**	<u>0.7</u>	559/560	280/280
LingBot-World v2	19.2	4.0	559/560	276/280
AlayaWorld	53.0	58.7	560/560	280/280
H3-World	77.0	52.0	423/560	260/280
Lyra 2.0	<u>3.6</u>	**0.4**	560/560	280/280

Table F.2: Cross-view RSD within the same scene and complete group counts.

Model	Trans. RSD ↓	Turn RSD ↓	Trans. groups	Turn groups
Matrix-Game 2.0	40.8	6.1	353/360	176/180
Matrix-Game 3.0	**24.9**	6.5	350/360	176/180
SANA-WM	36.8	5.3	360/360	180/180
DreamX-World	38.6	19.3	360/360	180/180
Hunyuan-World 1.5	47.6	9.0	360/360	180/180
HY-World 2.0	38.2	<u>4.8</u>	359/360	180/180
LingBot-World v2	32.7	10.7	359/360	176/180
AlayaWorld	98.1	106.9	360/360	180/180
H3-World	84.0	51.9	269/360	162/180
Lyra 2.0	<u>29.2</u>	**0.9**	360/360	180/180

G Supplementary Models and Pitch Tests

G.1 Directional tests for Genie 3 and GameCraft

Genie 3. We collect 672 videos from 48 cases in ten UE5 scenes, numbered 01 to 10, through the interactive interface: 288 single-action videos and 384 composite videos. The platform manages native random seeds and internal states. This test assesses direction from net motion between video endpoints; the timing map from input events to native frames has not yet been calibrated.

Starting and ending poses must be valid and finite, with anchor inlier ratios and coverage of at least 0.3. Of the 672 videos, 436 have valid endpoint poses, 179 fail endpoint quality checks, and 57 have

no valid pose result. Table G.1 measures direction on videos with valid endpoints. Axial alignment is the inner product of the requested-direction unit vector and horizontal net displacement, divided by displacement length and multiplied by 100.

Table G.1: Genie 3 translation direction and axial alignment.

Action	Endpoint support ↑	Correct/supported ↑	Median axial alignment (%) ↑
W	11/48	11/11	100.0
S	42/48	42/42	100.0
A	33/48	33/33	99.3
D	29/48	29/29	98.9

All 115 valid translation videos have net displacement in the requested direction, with action-specific median axial alignment between 98.9% and 100.0%. The other 77 translation videos fail endpoint pose checks. YL and YR have valid endpoints in 2/48 and 3/48 videos, respectively. The four composite-route families have valid endpoints for 39/48, 40/48, 38/48, and 30/48 pairs.

GameCraft. We conduct a supplementary test using 180 BF16 videos from five UE5 scenes. Single-action videos execute two native control intervals; requested labels and adapter settings remain unchanged. With inlier-ratio and coverage thresholds of 0.1, 172 videos have valid poses at all required boundaries, six fail quality checks, and two have missing or nonfinite boundary poses. Each action in Table G.2 contains 15 videos, and directional accuracy is conditioned on valid poses.

Table G.2: GameCraft direction results for short actions.

Action	Support ↑	Correct/supported ↑	Conditional direction accuracy (%) ↑
W	12/15	12/12	100.0
S	15/15	14/15	93.3
A	15/15	4/15	26.7
D	12/15	7/12	58.3
YL	15/15	3/15	20.0
YR	15/15	12/15	80.0

GameCraft's conditional directional accuracies are 100.0% and 93.3% for forward and backward motion, and 20.0% and 80.0% for left and right turns. Both turn requests produce 12 positive-yaw and three negative-yaw outcomes, indicating similar turning distributions for left and right commands under this test configuration.

G.2 Pitch control

The pitch test covers 19 initial views, three random seeds, and upward/downward pitch commands (PU/PD), with 114 videos of 16 steps per model. Sustained pitch in wide-aspect-ratio images, such as 16:9, can move initial reference geometry out of view. We therefore use the first eight steps consistently, checking nine action boundaries including the start. Both pose-threshold settings use identical videos, evaluation windows, and recovered poses.

Directional accuracy is determined by the sign of pitch at step 8, with all 114 videos per model as the denominator. To quantify yaw and roll accompanying pitch, let ϑ_8, ψ_8, and φ_8 be pitch, yaw, and roll relative to the initial camera. We compute the additional rotation ratio:

$$C_{\mathrm{rot}}(8) = 100\frac{\sqrt{\psi_8^2 + \varphi_8^2}}{|\vartheta_8|}. \tag{G.1}$$

All three angles use the same units. The ratio is computed for valid poses with $|\vartheta_8| > 10^{-9}$; the tables report its median and valid sample count. The best and second-best directional accuracies and additional rotation ratios are marked in both tables.

Table G.3: Pitch results over the first eight steps of 16-step videos, with pose-ratio thresholds of 0.1.

Model	Support ↑	Direction (%) ↑	Extra rotation median (%) ↓	Valid ratios
Matrix-Game 2.0	114/114	**100.0**	3.9	114
Matrix-Game 3.0	114/114	**100.0**	<u>2.9</u>	114
SANA-WM	114/114	**100.0**	3.7	114
DreamX-World	114/114	**100.0**	6.9	114
Hunyuan-World 1.5	114/114	**100.0**	5.1	114
HY-World 2.0	114/114	<u>99.1</u>	**2.4**	114
LingBot-World v2	114/114	**100.0**	6.5	114
AlayaWorld	114/114	78.1	42.6	114
H3-World	87/114	57.9	7.2	87
Lyra 2.0	114/114	**100.0**	3.9	114

Table G.4: The same pitch videos and evaluation window, with pose-ratio thresholds of 0.0.

Model	Support ↑	Direction (%) ↑	Extra rotation median (%) ↓	Valid ratios
Matrix-Game 2.0	114/114	**100.0**	3.9	114
Matrix-Game 3.0	114/114	**100.0**	<u>2.9</u>	114
SANA-WM	114/114	**100.0**	3.7	114
DreamX-World	114/114	**100.0**	6.9	114
Hunyuan-World 1.5	114/114	**100.0**	5.1	114
HY-World 2.0	114/114	<u>99.1</u>	**2.4**	114
LingBot-World v2	114/114	**100.0**	6.5	114
AlayaWorld	114/114	78.1	42.6	114
H3-World	90/114	60.5	7.2	90
Lyra 2.0	114/114	**100.0**	3.9	114

At threshold 0.1, 1,113/1,140 videos across ten models have valid poses at the required boundaries, and 1,066/1,140 have the correct endpoint direction. Reducing the ratio thresholds to 0.0 increases H3-World's valid count from 87 to 90 and directional accuracy from 57.9% to 60.5%; the other models are unchanged.

The main composition test uses horizontal translation and yaw. Pitch changes camera orientation and the reference direction for subsequent movement, so it is evaluated separately. Camera poses are always recovered in full $SE(3)$, and vertical displacement and pitch during planar commands are recorded as specified in Appendix D.2.

H Evaluation-System Analysis

H.1 Anchor distribution and full-pipeline scoring controls

On the 30 known-motion segments, we retain 5%, 10%, and 25% of anchors and compare uniform with spatially concentrated sampling. At 25% retention, concentrated sampling yields median and P90 translation errors 2.17 and 2.93 times those of uniform sampling.

The full-pipeline controls use five groups of UE5 rendered cases. Each group contains six eight-step single-action videos, two composite routes, and 60 independently recorded camera transforms at action endpoints. Camera ground truth and CoTracker/PnP poses are passed separately through the same scoring pipeline, using the same coordinate convention and Equation (1) to estimate single-step motion. All results in Tables H.1 and H.2 and the boundary-wise comparisons are computed within each group and then averaged across the five groups.

Single-action controls estimate motion from the first four steps and evaluate steps 5 to 8. Relative residuals in Table H.1 are normalized by step count times single-step magnitude. Within each group,

the mean residual, maximum residual, and mean score are computed over the 24 held-out action-boundary observations. Each statistic and the support count are then averaged across the five groups. Each composite route has six recorded action boundaries.

Table H.1: Held-out single-action prediction over steps 5–8. All columns report means across five UE5 groups.

Trajectory source	Mean residual (%) ↓	Mean group max. residual (%) ↓	Score ↑	Mean support per group
Independent camera GT	0.05	0.08	100.0	24/24
CoTracker/PnP	1.22	2.99	98.8	24/24

Table H.2: Final scores averaged across five groups of UE5 rendered cases.

Object	GT Trans. ↑	Recovered Trans. ↑	GT Yaw ↑	Recovered Yaw ↑
Route pair	99.8	99.1	100.0	98.2
Route 1	95.0	95.9	100.0	98.4
Route 2	95.0	95.3	100.0	96.7

Routes 1 and 2 denote the corresponding composite routes in each group. Their ground-truth translation scores both average 95.0, indicating nonzero residuals between actual execution and fixed single-step predictions. Final route-pair ground-truth and recovered mean scores differ by 0.7 points in translation and 1.8 in yaw. The second of the six corresponding action boundaries has the largest differences between five-group mean scores, with recovered scores 11.0 and 11.1 points below ground truth.

H.2 Results by reference domain and individual route

Domain-specific statistics use the main configuration and the same four-step motion estimates. UE5 contains 416 route pairs and ETH3D 144. We compute support fractions and mean translation scores over valid pairs separately for each domain.

Table H.3: Composition results on UE5 and ETH3D. Scores are averaged over valid pairs within each domain.

Model	UE5 Support (%) ↑	UE5 Cond. trans. ↑	ETH3D Support (%) ↑	ETH3D Cond. trans. ↑
Matrix-Game 2.0	88.5	43.0	75.7	41.4
Matrix-Game 3.0	<u>98.3</u>	59.9	80.6	64.3
SANA-WM	45.2	34.5	47.2	41.5
DreamX-World	78.6	20.4	78.5	22.3
Hunyuan-World 1.5	24.5	33.6	50.7	46.1
HY-World 2.0	98.1	**78.7**	**100.0**	**89.6**
LingBot-World v2	88.7	48.1	<u>96.5</u>	55.1
AlayaWorld	40.1	8.3	25.7	7.6
H3-World	18.0	26.6	25.0	25.9
Lyra 2.0	**98.8**	<u>66.3</u>	<u>96.5</u>	<u>78.1</u>

Table H.4 compares single-route endpoint prediction with route-pair relative-pose prediction on the same valid route pairs. Single-route scores evaluate each observed endpoint against its prediction from calibrated single-action transforms. Within each pair, we average the two individual-route scores, then average these pair-level scores over valid pairs. Route-pair scores are averaged over the same pairs. These are conditional averages over valid pairs, rather than the overall scores in Table 2,

which assign zero to invalid pairs. Tables H.3 and H.4 mark the best and second-best values within each score or percentage column.

Table H.4: Single-route and route-pair scores on the same valid route set.

Model	Supported pairs	Single-route Trans. ↑	Route-pair Trans. ↑	Single-route Yaw ↑	Route-pair Yaw ↑
Matrix-Game 2.0	477	39.7	42.7	80.6	83.4
Matrix-Game 3.0	525	50.9	60.9	77.8	85.6
SANA-WM	256	30.7	36.4	69.3	73.8
DreamX-World	440	19.6	20.9	33.5	43.5
Hunyuan-World 1.5	175	29.7	38.8	76.0	82.4
HY-World 2.0	552	**72.4**	**81.5**	<u>89.5</u>	**93.3**
LingBot-World v2	508	32.9	50.0	74.3	81.5
AlayaWorld	204	5.5	8.1	41.1	48.6
H3-World	111	12.0	26.4	37.3	52.7
Lyra 2.0	550	<u>64.0</u>	<u>69.3</u>	**90.2**	<u>93.1</u>

The route-pair advantage ranges from 1.3 to 17.1 points for translation and from 2.8 to 15.4 points for yaw across the ten models. The two evaluations can also yield different model rankings: Lyra 2.0 has a slightly higher single-route yaw score than HY-World 2.0 (90.2 versus 89.5), while HY-World 2.0 has a slightly higher route-pair yaw score (93.3 versus 93.1).

Figure 3b visualizes these four conditional scores in the same model order as the main table. Its sample set is the Main-valid set for each model, not the larger No-M set shown in the left panel. DreamX-World scores 19.6 and 20.9 for individual and relative translation predictions, while LingBot-World v2 scores 32.9 and 50.0. Pose biases shared by two routes can cancel in their relative relation; this does not establish that either individual endpoint is accurate or that all execution errors have been removed.

H.3 Motion-magnitude checks and predictor comparisons

Motion-magnitude checks. No-M removes only the per-step relative motion threshold of 10%, retaining single-step estimate validity, pose quality, and input-frame matching. Valid-route counts and overall scores are recomputed on the same videos, with all scores still averaged over the 560 planned pairs.

Figure 3a converts the support counts in Table H.5 to percentages using the same denominator of 560. For Hunyuan-World 1.5, removing the motion check adds 385 pairs and raises the overall translation score from 12.1 to 37.6, while the valid-pair mean changes from 38.8 to 37.6. SANA-WM gains 295 pairs, from 256 to 551. These additional pairs meet the remaining requirements but have at least one step below the motion threshold. The comparison identifies the effect of this check, not a separate estimate of complete freezing, transition latency, or endpoint error on every excluded pair.

Table H.5: Support and overall scores before and after removing the motion-magnitude check.

Model	Main Support	No-M Support	Main Trans. ↑	No-M Trans. ↑	Main Yaw ↑	No-M Yaw ↑
Matrix-Game 2.0	477/560	538/560	36.3	40.9	71.0	80.1
Matrix-Game 3.0	525/560	532/560	57.1	57.8	80.3	81.3
SANA-WM	256/560	551/560	16.6	36.9	33.7	71.8
DreamX-World	440/560	560/560	16.4	21.3	34.2	42.5
Hunyuan-World 1.5	175/560	560/560	12.1	37.6	25.8	83.4
HY-World 2.0	552/560	555/560	80.4	80.8	92.0	92.5
LingBot-World v2	508/560	522/560	45.4	46.7	73.9	75.9
AlayaWorld	204/560	524/560	3.0	13.4	17.7	51.3
H3-World	111/560	348/560	5.2	19.0	10.4	33.9
Lyra 2.0	550/560	560/560	68.0	68.7	91.4	92.5

Predictor definitions. Ideal, Axial, and Full use the first four steps of each action for the same model, case, and seed. They retain shared symmetric magnitudes, individual axial magnitudes, and full planar motion, respectively. Using the translation and turning sets $\mathcal{A}_t$ and $\mathcal{A}_r$, define command directions in Lie algebra coordinates (ρ_x, ρ_z, ω) as

$$\begin{aligned} \mathbf{e}_{\mathrm{W}} &= (0,1,0)^\top, & \mathbf{e}_{\mathrm{S}} &= -\mathbf{e}_{\mathrm{W}}, \\ \mathbf{e}_{\mathrm{D}} &= (1,0,0)^\top, & \mathbf{e}_{\mathrm{A}} &= -\mathbf{e}_{\mathrm{D}}, \\ \mathbf{e}_{\mathrm{YR}} &= (0,0,1)^\top, & \mathbf{e}_{\mathrm{YL}} &= -\mathbf{e}_{\mathrm{YR}}. \end{aligned}$$

The axial component from Appendix D.1 is $u_a = \mathbf{e}_a^\top \bar{\boldsymbol{\xi}}_a^{[4]}$. Ideal's shared translation and turning magnitudes are

$$m_t = \mathrm{median}_{a\in\mathcal{A}_t} |u_a|, \qquad m_r = \mathrm{median}_{a\in\mathcal{A}_r} |u_a|.$$

Let $m_a = m_t$ for $a \in \mathcal{A}_t$ and $m_a = m_r$ for $a \in \mathcal{A}_r$. The three single-step transforms are

$$\begin{aligned} \mathbf{U}_a^{\mathrm{ideal}} &= \mathrm{Exp}(m_a \mathbf{e}_a), \\ \mathbf{U}_a^{\mathrm{axial}} &= \mathrm{Exp}(u_a \mathbf{e}_a), \\ \mathbf{U}_a^{\mathrm{full}} &= \mathrm{Exp}\Big(\bar{\boldsymbol{\xi}}_a^{[4]}\Big) = \mathbf{U}_a. \end{aligned} \tag{H.1}$$

Ideal acts in the standard command direction, with one shared step length for the four translations and one shared angle for both turns. Axial retains each action's signed axial component and sets other components to zero, allowing unequal opposing steps and reversed responses. Full retains lateral displacement, accompanying yaw, and translation during turning.

For route $q = (a_1, \ldots, a_n)$, every predictor composes transforms in the original command order:

$$\widehat{\mathbf{H}}_q^{(p)} = \mathbf{U}_{a_1}^{(p)} \cdots \mathbf{U}_{a_n}^{(p)}, \qquad p \in \{\mathrm{ideal}, \mathrm{axial}, \mathrm{full}\}. \tag{H.2}$$

Predictors are fixed after single-action measurement. Comparisons use the same videos, observed poses, and valid route pairs under No-M, normalizing all errors with b_t, b_r from the original Full definition. Whereas m_t, m_r determine Ideal's action magnitudes, b_t, b_r determine scoring for all three predictors.

References and route structure. With exact observations of fixed action transforms $\mathbf{U}_a^\star$, cumulative motion satisfies $\mathbf{H}_{a,k} = (\mathbf{U}_a^\star)^k$. Within the logarithm branch specified in Appendix D.1, Equation (1) recovers these transforms. Full therefore represents stable unequal step lengths and accompanying motion in its predictions.

For the shortened schematic pairs in Table 1, Ideal predicts coincident endpoints, $\widehat{\mathbf{\Delta}}^{\mathrm{ideal}} = \mathbf{I}$: opposing translations cancel, and translations within each fixed-heading block commute. The G2–G4 examples also predict identity under Axial. Full instead composes the measured transforms, which can predict a nonidentity endpoint relation. The predictor comparison tests agreement with these different references.

Table H.6: Predictor scores on the same valid routes under No-M. Best and second-best scores are marked within each model and motion component.

	Translation ↑			Yaw ↑		
Model	Ideal	Axial	Full	Ideal	Axial	Full
Matrix-Game 2.0	**43.9**	43.7	42.5	83.4	**83.5**	83.4
Matrix-Game 3.0	**64.5**	63.8	60.8	85.6	**85.8**	85.6
SANA-WM	**38.6**	37.8	37.5	**73.4**	73.2	73.0
DreamX-World	21.0	21.0	**21.3**	42.3	**42.6**	42.5
Hunyuan-World 1.5	**40.1**	38.0	37.6	**84.8**	83.7	83.4
HY-World 2.0	**85.0**	83.7	81.5	**94.3**	93.8	93.4
LingBot-World v2	**55.9**	50.1	50.1	**82.2**	82.0	81.4
AlayaWorld	**14.7**	14.5	14.3	**56.7**	54.9	54.8
H3-World	**33.3**	30.7	30.6	**57.1**	55.6	54.5
Lyra 2.0	**69.4**	68.8	68.7	**93.1**	**93.1**	92.5

Paired intervals. We resample scenes with replacement 2,000 times using random seed 240924 and compute the conditional mean score difference, Full minus Ideal, together with its 95% percentile bootstrap confidence interval.

Table H.7: Paired Full-minus-Ideal score differences and 95% confidence intervals.

Model	Matched pairs	Translation difference	Yaw difference
Matrix-Game 2.0	538	-1.4 [-3.3, 0.4]	0.0 [-0.8, 0.7]
Matrix-Game 3.0	532	-3.7 [-5.1, -2.3]	0.0 [-0.8, 0.8]
SANA-WM	551	-1.1 [-1.8, -0.4]	-0.3 [-1.0, 0.5]
DreamX-World	560	0.3 [0.1, 0.5]	0.2 [-0.1, 0.5]
Hunyuan-World 1.5	560	-2.6 [-3.6, -1.6]	-1.4 [-2.5, -0.3]
HY-World 2.0	555	-3.5 [-4.5, -2.5]	-0.9 [-1.4, -0.6]
LingBot-World v2	522	-5.8 [-6.9, -4.5]	-0.9 [-1.6, -0.2]
AlayaWorld	524	-0.4 [-0.8, -0.1]	-1.9 [-3.1, -0.8]
H3-World	348	-2.7 [-4.1, -1.3]	-2.6 [-4.4, -1.2]
Lyra 2.0	560	-0.7 [-1.2, -0.2]	-0.6 [-0.9, -0.3]

Under the current route pairs and scoring configuration, nine models have lower average translation scores with Full than Ideal, and eight paired intervals lie below zero. DreamX-World's Full score is 0.3 points higher, with an interval of [0.1, 0.5]. LingBot-World v2 decreases from 55.9 with Ideal to 50.1 with Axial, remaining at 50.1 with Full. Matrix-Game 3.0 decreases from 64.5 to 63.8 and then 60.8. These changes correspond to introducing action-specific magnitudes and then additionally retaining lateral motion and accompanying turns.

H.4 Sensitivity to normalization scales

Keeping the valid route set fixed, we multiply both the translation and yaw reference scales by $\lambda \in \{0.5, 1, 2\}$ and recompute overall scores and model orderings. The main configuration uses $\lambda = 1$.

Table H.8: Overall scores under different reference scales.

	Translation ↑			Yaw ↑		
Model	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$
Matrix-Game 2.0	24.7	36.3	49.1	63.0	71.0	76.7
Matrix-Game 3.0	43.6	57.1	69.1	72.0	80.3	85.9
SANA-WM	11.0	16.6	23.2	28.4	33.7	38.1
DreamX-World	9.8	16.4	25.6	24.5	34.2	44.7
Hunyuan-World 1.5	8.1	12.1	16.7	23.0	25.8	27.9
HY-World 2.0	71.0	80.4	87.3	87.6	92.0	94.8
LingBot-World v2	32.9	45.4	57.9	64.9	73.9	80.6
AlayaWorld	1.8	3.0	4.5	14.5	17.7	21.0
H3-World	3.6	5.2	7.3	8.2	10.4	12.7
Lyra 2.0	54.9	68.0	78.9	86.5	91.4	94.5

Halving or doubling the reference scales produces one translation and one yaw ordering reversal among the 45 model pairs. Relative orderings among the main models remain stable over this scale range.

H.5 Additional route and sensitivity analyses

Reference identities on the actual routes. We audit the 120 designed route pairs (30 scenes, four pairs per scene), separately from the 560 case-level executions per model. Exact coefficient cancellation along the actual action sequences establishes that Ideal predicts coincident endpoints for all 120 pairs. Axial also predicts coincident endpoints for all 90 order pairs (G2–G4), but not identically for the 30 inverse pairs (G1): independently calibrated opposite actions need not cancel. These are properties of the evaluated sequences, not arbitrary action routes. Numerical composition on all ten models confirms these identities wherever calibration is valid, using a normalized tolerance of 10^{-9}. The following comparisons retain identical No-M support and Full-K4 normalization within each model. They separate the inverse and order families, because an identity endpoint reference and a calibrated nonidentity prediction answer different questions.

Differences between the references are concentrated in G1. For LingBot-World v2, Ideal, Axial, and Full have conditional translation scores of 60.0, 37.5, and 37.4 on inverse routes, but 54.4, 54.4, and 54.5 on order routes. The empirical reference preserves unequal opposing actions when forming a prediction; the resulting score measures agreement with that prediction. These comparisons do not separately estimate errors caused by calibration noise and changes in action effects during composition.

Table H.9: Predictor scores on G1 routes. Cells give conditional translation / yaw scores (higher is better); all three predictors share support and normalization.

Model	Support	Ideal ↑	Axial ↑	Full ↑
Matrix-Game 2.0	133/140	40.7 / 79.2	39.8 / 79.5	34.8 / 79.2
Matrix-Game 3.0	133/140	64.8 / 80.0	61.9 / 80.9	50.2 / 80.0
SANA-WM	138/140	36.4 / 68.5	33.4 / 67.7	32.3 / 67.1
DreamX-World	140/140	24.3 / 32.5	24.1 / 33.8	25.5 / 33.1
Hunyuan-World 1.5	140/140	41.4 / 75.7	33.1 / 71.3	31.1 / 70.2
HY-World 2.0	139/140	86.0 / 94.8	81.0 / 92.8	72.6 / 91.0
LingBot-World v2	134/140	60.0 / 78.2	37.5 / 77.3	37.4 / 74.8
AlayaWorld	133/140	7.4 / 40.6	6.6 / 33.6	5.8 / 33.2
H3-World	88/140	30.5 / 49.1	20.2 / 43.3	20.0 / 38.7
Lyra 2.0	140/140	64.3 / 89.8	61.8 / 89.7	61.6 / 87.3

Table H.10: Predictor scores on G2–G4 routes. Cells give conditional translation / yaw scores (higher is better); all three predictors share support and normalization.

Model	Support	Ideal ↑	Axial ↑	Full ↑
Matrix-Game 2.0	405/420	45.0 / 84.8	45.0 / 84.8	45.1 / 84.8
Matrix-Game 3.0	399/420	64.4 / 87.4	64.4 / 87.4	64.3 / 87.4
SANA-WM	413/420	39.3 / 75.0	39.3 / 75.0	39.2 / 75.0
DreamX-World	420/420	19.9 / 45.6	19.9 / 45.6	19.9 / 45.6
Hunyuan-World 1.5	420/420	39.7 / 87.8	39.7 / 87.8	39.7 / 87.8
HY-World 2.0	416/420	84.6 / 94.1	84.6 / 94.1	84.4 / 94.1
LingBot-World v2	388/420	54.4 / 83.6	54.4 / 83.6	54.5 / 83.6
AlayaWorld	391/420	17.2 / 62.1	17.2 / 62.1	17.2 / 62.1
H3-World	260/420	34.2 / 59.8	34.2 / 59.8	34.2 / 59.8
Lyra 2.0	420/420	71.1 / 94.2	71.1 / 94.2	71.1 / 94.2

Estimation-window sensitivity. We recompute $K \in \{2, 4, 6, 8\}$ from the same cumulative observations. Atomic comparisons at steps 12 and 16 use their respective common samples across all four windows and the original K4 amplitude times dose as denominator. Original aggregation weights are renormalized over these common observations. Composition retains the original K4 No-M quality conditions and normalization, and uses pairs whose six atomic operators exist at every window. Thus the following are common-support conditional diagnostics, not replacements for the main fixed-denominator scores. Larger windows observe more of the execution, so improved extrapolation alone does not identify measurement noise. For H3-World, the common composition subset contains only 55 of 560 pairs.

Table H.11: Atomic scores for different calibration windows. Each score cell is step 12 / step 16; n gives the two common-support counts (out of 2,520 each). Higher scores are better.

Model	n_{12}/n_{16}	$K = 2$	$K = 4$	$K = 6$	$K = 8$
Matrix-Game 2.0	2245/1356	81.1 / 73.2	83.6 / 75.8	85.4 / 77.8	86.9 / 79.4
Matrix-Game 3.0	2135/1393	87.7 / 84.7	89.5 / 86.2	90.7 / 87.4	91.7 / 88.4
SANA-WM	2410/2272	78.6 / 79.0	81.8 / 81.8	84.0 / 83.6	85.6 / 84.8
DreamX-World	2453/2337	47.2 / 44.3	50.8 / 47.4	54.9 / 50.9	59.2 / 54.4
Hunyuan-World 1.5	2509/2496	77.0 / 76.2	78.7 / 77.7	80.3 / 79.2	81.8 / 80.5
HY-World 2.0	2441/2378	89.3 / 89.1	91.0 / 90.5	92.2 / 91.5	93.2 / 92.2
LingBot-World v2	2400/2316	87.3 / 86.1	88.8 / 87.4	90.1 / 88.5	91.2 / 89.5
AlayaWorld	2515/2514	28.2 / 24.5	31.0 / 26.8	33.8 / 29.0	36.4 / 30.8
H3-World	1185/948	57.0 / 57.4	61.2 / 60.9	64.3 / 63.8	67.2 / 66.2
Lyra 2.0	2489/2460	89.7 / 87.6	90.9 / 88.6	91.7 / 89.3	92.3 / 89.8

Table H.12: Composition on the same pairs across all four windows. Cells give translation / yaw scores (higher is better); normalization is fixed at K4.

Model	$n/560$	$K = 2$	$K = 4$	$K = 6$	$K = 8$
Matrix-Game 2.0	499	44.1 / 84.1	43.2 / 84.2	42.9 / 84.4	43.0 / 84.4
Matrix-Game 3.0	477	61.5 / 85.9	60.9 / 86.1	60.6 / 86.0	60.6 / 86.0
SANA-WM	535	37.5 / 73.2	37.4 / 73.4	37.4 / 73.4	37.3 / 73.4
DreamX-World	560	21.2 / 42.5	21.3 / 42.5	21.3 / 42.5	21.3 / 42.4
Hunyuan-World 1.5	548	37.4 / 83.1	37.4 / 83.4	37.4 / 83.4	37.4 / 83.5
HY-World 2.0	524	82.8 / 94.0	81.8 / 94.0	81.4 / 93.9	81.0 / 93.8
LingBot-World v2	465	49.9 / 81.7	49.8 / 81.9	49.7 / 82.0	49.4 / 82.1
AlayaWorld	524	14.5 / 55.2	14.3 / 54.8	14.1 / 54.3	14.1 / 54.2
H3-World	55	25.1 / 47.2	25.1 / 47.3	25.0 / 47.3	24.6 / 47.2
Lyra 2.0	556	68.6 / 92.5	68.7 / 92.5	68.6 / 92.5	68.5 / 92.5

Increasing the calibration window from four to eight steps improves sustained-action prediction on the common samples, but does not generally improve composition prediction. For DreamX-World, the step-16 Atomic score rises from 47.4 to 54.4 while the composition translation score stays at 21.3. For AlayaWorld, the corresponding scores change from 26.8 to 30.8 and from 14.3 to 14.1. A longer estimation window supplies additional single-action observations; it does not condition the generator on a longer history or refit the reference to the composite video.

Table H.13: Availability before intersection across windows. Each cell gives identified atomic units (out of 2,520) / available composition pairs (out of 560) under the fixed K4 No-M quality conditions. Counts describe support, not prediction accuracy.

Model	$K = 2$	$K = 4$	$K = 6$	$K = 8$
Matrix-Game 2.0	2519 / 538	2509 / 538	2480 / 509	2467 / 499
Matrix-Game 3.0	2520 / 532	2504 / 532	2465 / 511	2424 / 477
SANA-WM	2520 / 551	2520 / 551	2517 / 551	2506 / 535
DreamX-World	2520 / 560	2520 / 560	2520 / 560	2515 / 560
Hunyuan-World 1.5	2520 / 560	2520 / 560	2516 / 556	2512 / 548
HY-World 2.0	2520 / 555	2519 / 555	2502 / 544	2484 / 524
LingBot-World v2	2520 / 522	2515 / 522	2495 / 501	2474 / 465
AlayaWorld	2520 / 524	2520 / 524	2520 / 524	2520 / 524
H3-World	2508 / 348	2353 / 348	2117 / 202	1740 / 55
Lyra 2.0	2520 / 560	2520 / 560	2520 / 560	2516 / 556

Geometric observation support. Table H.14 reports actual Ray support as defined in Appendix C.3, including planned frames with no scored observations. No model has an empty World-Anchor set or a whole video without scored observations among its 1,120 videos. Temporal thirds follow planned frame indices. Observation loss can include ordinary reference points leaving the field of view; it is not itself evidence of geometric damage. For AlayaWorld versus Lyra 2.0, we additionally predefine matching within scene, case, route, and seed: at every corresponding action boundary, cumulative 3D camera-center path length and cumulative SO(3) rotation must each differ by at most a ratio of $[0.8, 1.25]$. All prefix poses must be finite, rigid, and pass the 0.1 quality ratios. Both numerical zeros are accepted, but a zero/nonzero pair is rejected. None of the 1,120 paired videos matches at every stage; only 28 match at the final accumulated motion, and 197 of 13,160 stages match individually. Consequently, these records do not support a whole-video motion-matched Ray comparison. We do not relax the rule after observing the outcome.

Table H.14: Ray score and actual observation support. Frame and Obs. are percentages of planned noninitial frames and initial-anchor point–frame opportunities. Early/Middle/Late give Obs. in temporal thirds. Ray is higher-better; support is descriptive.

Model	Ray ↑	Frame	Obs.	Early	Middle	Late
Matrix-Game 2.0	67.8	99.5	73.9	82.8	72.1	66.3
Matrix-Game 3.0	68.0	99.2	73.7	81.2	71.3	68.2
SANA-WM	79.4	99.8	79.4	87.2	77.9	72.8
DreamX-World	81.3	100.0	92.1	97.4	91.4	87.3
Hunyuan-World 1.5	84.5	100.0	88.7	92.1	87.7	86.1
HY-World 2.0	79.3	99.9	86.2	91.1	84.9	82.0
LingBot-World v2	76.6	99.1	79.6	87.5	77.0	73.9
AlayaWorld	88.9	98.1	89.0	94.5	88.3	84.0
H3-World	75.2	93.9	64.8	83.8	61.0	49.4
Lyra 2.0	89.4	100.0	92.1	95.4	91.4	89.6

Uncertainty in the main comparisons. Table H.15 reports paired scene-bootstrap intervals for four selected comparisons. Differences use unrounded scores. Ray and yaw intervals include zero; this is insufficient evidence for a directional difference and does not establish equivalence. These are descriptive pointwise intervals without multiplicity correction.

Table H.15: Selected main-table score differences (first model minus second), with 95% paired scene-bootstrap intervals.

Comparison	Metric	Difference	95% interval
AlayaWorld − Lyra 2.0	Ray	-0.5	[-2.0, 1.1]
AlayaWorld − Lyra 2.0	Atomic	-59.5	[-62.1, -57.0]
HY-World 2.0 − Lyra 2.0	Trans.	12.3	[8.6, 16.1]
HY-World 2.0 − Lyra 2.0	Yaw	0.6	[-1.4, 2.9]

Sensitivity near the motion threshold. Only the per-action motion ratio threshold varies over $\{0.05, 0.10, 0.15\}$; K4 estimates, pose quality, input checks, and the 560-pair denominator remain fixed. Overall scores include zeros for unsupported pairs, while conditional scores use supported pairs only. The 0.10 setting reproduces the main configuration. Support is relatively stable for HY-World 2.0 but changes substantially for several other models. Rankings are not universally preserved: SANA-WM exceeds DreamX-World in overall translation at 0.05 and 0.10, but falls below it at 0.15.

Table H.16: Motion-threshold sensitivity. Overall / conditional translation and yaw scores are shown separately; higher scores are better.

Model	τ	Support	Trans.	Cond. T	Yaw	Cond. Y
Matrix-Game 2.0	0.05	514/560	39.0	42.4	76.6	83.5
Matrix-Game 2.0	0.10	477/560	36.3	42.7	71.0	83.4
Matrix-Game 2.0	0.15	429/560	32.6	42.6	63.9	83.4
Matrix-Game 3.0	0.05	529/560	57.5	60.9	80.9	85.6
Matrix-Game 3.0	0.10	525/560	57.1	60.9	80.3	85.6
Matrix-Game 3.0	0.15	505/560	55.3	61.3	77.3	85.7
SANA-WM	0.05	362/560	23.8	36.8	47.3	73.1
SANA-WM	0.10	256/560	16.6	36.4	33.7	73.8
SANA-WM	0.15	198/560	13.0	36.6	26.3	74.5
DreamX-World	0.05	501/560	18.8	21.0	38.3	42.8
DreamX-World	0.10	440/560	16.4	20.9	34.2	43.5
DreamX-World	0.15	388/560	14.5	20.9	30.4	43.8
Hunyuan-World 1.5	0.05	226/560	15.2	37.8	33.5	82.9
Hunyuan-World 1.5	0.10	175/560	12.1	38.8	25.8	82.4
Hunyuan-World 1.5	0.15	136/560	9.5	39.2	20.3	83.7

Table H.17: Motion-threshold sensitivity (continued). Overall / conditional translation and yaw scores are shown separately; higher scores are better.

Model	τ	Support	Trans.	Cond. T	Yaw	Cond. Y
HY-World 2.0	0.05	554/560	80.6	81.5	92.3	93.3
HY-World 2.0	0.10	552/560	80.4	81.5	92.0	93.3
HY-World 2.0	0.15	548/560	79.9	81.6	91.3	93.3
LingBot-World v2	0.05	519/560	46.4	50.0	75.4	81.4
LingBot-World v2	0.10	508/560	45.4	50.0	73.9	81.5
LingBot-World v2	0.15	498/560	44.4	49.9	72.3	81.4
AlayaWorld	0.05	280/560	5.3	10.6	26.4	52.9
AlayaWorld	0.10	204/560	3.0	8.1	17.7	48.6
AlayaWorld	0.15	161/560	2.0	7.1	13.8	48.2
H3-World	0.05	167/560	8.5	28.6	15.8	53.1
H3-World	0.10	111/560	5.2	26.4	10.4	52.7
H3-World	0.15	83/560	3.5	23.3	7.8	52.8
Lyra 2.0	0.05	556/560	68.5	69.0	92.1	92.8
Lyra 2.0	0.10	550/560	68.0	69.3	91.4	93.1
Lyra 2.0	0.15	546/560	67.7	69.4	91.0	93.3