PROWBENCH: AN EXTENSIBLE BENCHMARK FOR PROGRAMMABLE WORLD MODELS

Anonymous authors
Paper under double-blind review



Figure 1: **Overview of our extensible framework and benchmark.** Our extensible framework pairs visual proxies with ground-truth world records, enabling reliable evaluation of how faithfully generated videos realize prescribed states, rules, and interactions. This assessment would otherwise be difficult. Here, we compare the CLIP score with our proposed interaction metric. CLIP assigns a higher score to the clearly inferior case, whereas one of our metric correctly reflects whether the prescribed interaction is faithfully executed.

ABSTRACT

Programmable world models separate executable dynamics from visual generation, offering a promising foundation for next-generation game engines. However, their visual adherence to explicit rules and interactions remains insufficiently evaluated. Existing benchmarks emphasize video quality and camera controllability, with limited support for evaluation grounded in replayable world records. We introduce **PROWBench**, comprising 170 programmatically constructed episodes and 600 proxy videos covering diverse scenes and interactions. An extensible framework constructs scenes, controls behaviors, and renders three synchronized representations per camera view, including coarse 3D, semantic proxy, and colored oriented bounding boxes, from shared state and interaction records. The benchmark covers first- and third-person perspectives, with synchronized multi-view observations available for a subset of episodes. Two VLM-based metrics, **Logic–Render Alignment** and **Interaction Success Rate**, assess prompt-rule adherence and the visual realization of timestamped engine-recorded events, respectively.

1 INTRODUCTION

Recent advances in video world models have enabled increasingly realistic and controllable interactive environments. However, maintaining persistent world states and enforcing user-defined rules remain challenging when entities, attributes, and interaction outcomes are represented implicitly through visual histories or latent context. Programmable World Model and Code World Model address these limitations by separating world-state evolution from visual generation (Huang et al., 2026a; Chen et al., 2026b): coding agents and executable programs maintain states and interaction rules, while video models generate observations conditioned on structured representations of the world. This

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.

separation combines explicit state control with visual generation, but correct program execution does not guarantee faithful visual realization. A generated video may appear plausible while violating a rule or failing to depict a specified interaction. This raises a central evaluation question: *can video models faithfully render the scene structure, rules, and interactions of a programmable world?*

Existing benchmarks evaluate video quality (Huang et al., 2024), controllability and dynamics (Duan et al., 2025), instruction following and physical adherence (Li et al., 2025a), and interactive control and consistency (Ye et al., 2026; Shang et al., 2026; Wu et al., 2026; Xu et al., 2026; Ying et al., 2026). However, as summarized in Table 3, they offer limited support for replayable scene records and paired renderings across viewpoints and proxy representations. Shared geometry, state trajectories, and camera records are needed to isolate the effects of viewpoint and representation while holding scene content and actions fixed. For programmable worlds, these controlled comparisons should be complemented by evaluation against explicit rules and timestamped interaction records, assessing whether generated videos faithfully realize the specified behavior.

To address these gaps, we introduce **PROWBench**, a benchmark for evaluating the visual realization of programmable worlds. PROWBench combines a programmatically constructed dataset, complementary metrics for logic adherence and interaction realization, and an extensible generation framework. It grounds evaluation in shared world-state trajectories and recorded interactions, enabling controlled comparisons across perspectives and visual representations.

PROWBench contains 170 dynamic scene episodes covering diverse environments and interactions. Each episode provides three synchronized representations—coarse 3D, semantic proxy, and colored oriented bounding boxes—from shared state, action, and interaction records, enabling controlled comparisons across representations. Some episodes additionally provide synchronized multi-view observations for cross-view comparisons. In addition, we introduce two VLM-based metrics: **Logic-Render Alignment** assesses adherence to prompt-specified rules extracted by an agent, while **Interaction Success Rate** checks whether recorded events are visually realized at the corresponding times using engine logs as ground truth. An extensible programmatic framework separates scene construction, behavior control, state recording, and rendering, allowing researchers to expand scene layouts, actions, and representations while generating aligned observations from shared state records.

We evaluate state-of-the-art models with different input representations, including coarse 3D representation, semantic proxy, colored oriented bounding boxes, and text-only prompts. Our experiments assess their logic adherence and interaction realization across viewpoints and representation settings. The results show that, with appropriate prompts, existing training-free methods can perform well, including in camera following. However, specialized scenarios such as gunfights benefit from additional training. Generated outputs sometimes adhere too closely to the input geometry, inheriting the coarse shapes of low-poly proxies. Long-video generation remains challenging in terms of both camera following and visual quality, while direct multi-view generation still yields unsatisfactory results.

Our contributions are threefold:

- **Benchmark.** We introduce PROWBench, comprising 170 episodes and 600 synchronized proxy videos covering multiple perspectives, visual representations, and diverse interactions, with shared world-state and action records.
- **Metrics.** We propose Logic-Render Alignment and Interaction Success Rate to assess whether generated videos follow specified rules and visually realize engine-recorded interaction events.
- **Extensible framework.** We provide a reusable programmatic generation framework that supports extensions to scene layouts, actions, and visual representations, enabling aligned multi-view and multi-representation data generation from shared state records.

2 Related Work

2.1 Video World Models and Programmable World States

Video world models predict future observations from visual context and actions. Genie (Bruce et al., 2024) learns controllable environments from unlabeled videos, while GameNGen (Valevski

et al., 2025) enables real-time action-conditioned game generation. GameFactory (Yu et al., 2025), Matrix-Game (Zhang et al., 2025), Yume1.5 (Mao et al., 2026a), and LingBot-World (Robbyant Team et al., 2026) advance generalization, interaction, and long-horizon consistency. Memory-enhanced WorldPlay (Sun et al., 2025), Matrix-Game 3.0 (Wang et al., 2026), and AlayaWorld (Team et al., 2026) improve persistence through spatial-context retrieval or reconstruction.

While these models infer states from visual history and memory, programmable world models externalize state evolution. Code World Model (Chen et al., 2026b) uses a coding agent to maintain executable states and interaction logic, while Programmable World Model (Huang et al., 2026a) feeds persistent structured states to a generative renderer. Other approaches similarly separate state or dynamics from appearance (Meng et al., 2026; Lin et al., 2026b; Cai et al., 2026a). PROWBench evaluates whether generated observations faithfully realize these explicit states and interactions.

2.2 Structured Representations for Visual Generation

Structured conditions offer direct spatial control through depth, trajectories, bounding boxes, masks, and camera motion (Esser et al., 2023; Wang et al., 2024a; Li et al., 2025b; Wang et al., 2024d; Luo et al., 2025; Jain et al., 2024; Li et al., 2025c). Geometry-aware methods incorporate 3D boxes, bird's-eye-view layouts, world volumes, and structured traffic scenes (Gao et al., 2024; Li et al., 2024; Wen et al., 2024; Lu et al., 2024; Wang et al., 2024b;c; 2025). Generative rendering uses explicit geometry or renderer-derived signals: Coarse-to-Real (Gomez-Nogales et al., 2026) and VideoFrom3D (Kim et al., 2025) use coarse 3D scenes, while Diffusion Renderer (Liang et al., 2025), AlayaRenderer (Huang et al., 2026b), its real-time extension (Lin et al., 2026a), and LiVER (Cai et al., 2026b) use denser rendering cues. These approaches span sparse object controls to dense geometric representations. Comparisons across representations are often confounded by differences in models, datasets, and tasks. PROWBench derives paired proxy representations from the same recorded world-state trajectory and camera configuration, isolating representation effects on visual realization under fixed scene evolution.

2.3 World Model Benchmarks and Evaluation

VBench (Huang et al., 2024) and EvalCrafter (Liu et al., 2024) assess video quality, temporal consistency, motion, and semantic alignment. WorldSimBench (Qin et al., 2025), WorldModelBench (Li et al., 2025a), and WorldScore (Duan et al., 2025) extend evaluation to physical plausibility, controllability, and world dynamics. MIND (Ye et al., 2026), WorldMark (Xu et al., 2026), OmniWorldBench (Wu et al., 2026), WorldArena (Shang et al., 2026), and iWorld-Bench (Fang et al., 2026) further target interactive behavior, memory, and long-horizon consistency.

WBench (Ying et al., 2026) evaluates multi-turn navigation, subject actions, event editing, and perspective switching for instruction adherence, consistency, and physical plausibility. These benchmarks generally assess complete pipelines, coupling errors in action understanding, state prediction, memory, dynamics, and rendering.

PROWBench isolates rendering fidelity using engine-recorded frame-level states, rules, and timestamped interactions, enabling comparisons across synchronized perspectives and proxy representations under the same trajectory. Logic–Renderer Alignment measures whether specified rules are visibly realized, while Interaction measures whether recorded events appear at the corresponding times.

3 Benchmark

We construct the PROWBench dataset through a programmable world generation pipeline that produces aligned conditioning inputs and evaluation records. Figure 3 provides an overview of the dataset's content and diversity. Below, we describe the pipeline and evaluation suite. Detailed dataset statistics are provided in Section A.

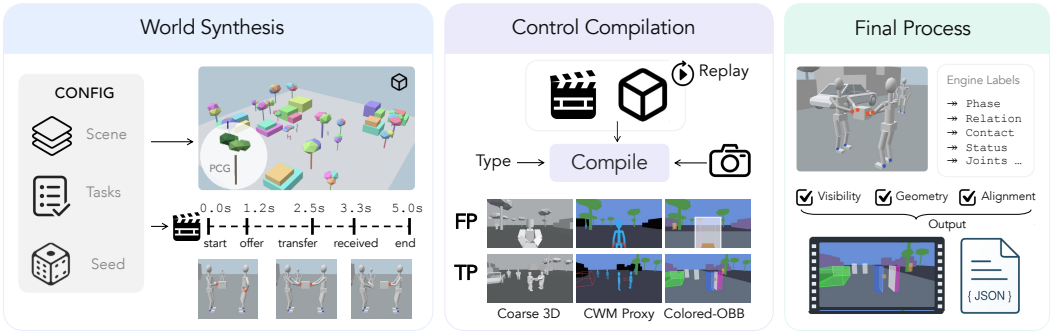


Figure 2: **Overview of the world generation pipeline.**World synthesis generates replayable trajectories, control compilation produces synchronized conditioning inputs across viewpoints and representations, and the final process validates and exports videos with structured annotations.

3.1 World Generation Pipeline

We introduce a pipeline for constructing programmable worlds and producing structured conditioning inputs and evaluation annotations for generative video models. The pipeline provides three complementary capabilities: *world synthesis*, *control compilation*, and *final process*. Together, these capabilities produce replayable world trajectories, synchronized conditioning inputs, and structured evaluation records.

World Synthesis. The pipeline constructs dynamic worlds from configurations specifying the environment, entity layout, participating actors, task controllers, random seeds, and execution duration. Procedural routines generate terrain, buildings, roads, interiors, and interaction objects, together with collision surfaces, collision constraints, and interaction anchors. Entities have persistent episode-level identifiers, semantic categories, component hierarchies, and task-relevant attributes. Random seeds control variations in scene layouts, motion paths, speeds, and animation phases.

Task-specific controllers coordinate articulated poses, object trajectories, and participant interactions through procedural animation and motion models. They support three broad categories of behavior. Interaction controllers organize tasks into explicit phases: for example, carrying proceeds through approaching, grasping, lifting, transporting, and releasing an object, with corresponding updates to hand targets and object ownership. Activity and locomotion controllers support animal behaviors and simplified flight and swimming dynamics. Driving controllers provide closed-loop control of vehicle motion. During execution, these controllers report action phases, participating entities, and relationship changes for recording and validation. External agents can query entities, submit supported actions or plans, advance the world by fixed ticks, and inspect execution feedback.

Each execution produces a structured episode record containing shared scene geometry, entity definitions and hierarchies, initial conditions, and temporal records of entity transforms, articulated poses, visibility, and appearance changes. Controller events and available input traces accompany these records, with explicit temporal mappings linking world execution, saved states, and video frames. Each saved state can be reconstructed independently from the shared scene description and its corresponding dynamic record. Configurations, random seeds, and source information are retained for provenance tracking.

Control Compilation. The pipeline compiles recorded world states and camera trajectories into synchronized visual conditioning inputs for generative video models.

Third-person cameras follow task-dependent trajectories, while first-person cameras are derived from recorded head positions, character or vehicle orientations, and task-specific gaze information. The observing character's display geometry is hidden in its first-person view, while its entity and articulated state remain in the shared world record. Final camera trajectories include intrinsics, camera-to-world and world-to-camera matrices, projection parameters, and clipping planes. Additional cameras can be associated with different participants to provide multiple perspectives of the same episode.

Rendering adapters map shared world records to conditioning representations. Our current implementation provides three representations. *Coarse 3D representation (Coarse 3D)* uses neutral materials

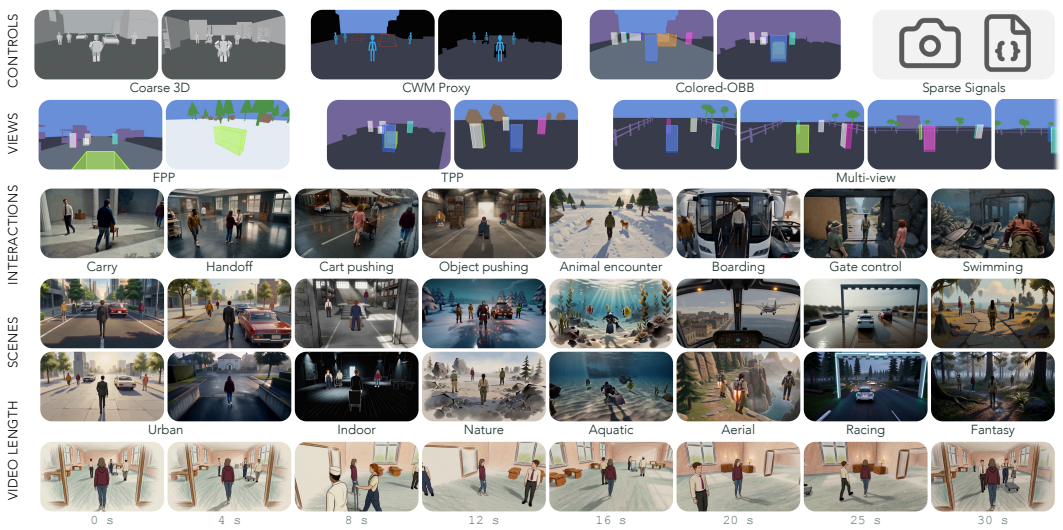


Figure 3: **Diversity of PROWBench.** PROWBench provides diverse conditioning inputs, camera perspectives, interactions, scene domains, and long-horizon videos.

to preserve scene shape, object silhouettes, and articulated motion. *Semantic proxy (Mixed)* combines semantically colored environments with simplified entity proxies, retaining global trajectories, orientations, and task-relevant posture while reducing local articulation and shape detail. *Colored oriented bounding boxes (Colored-OBB)* represent selected entities through spatial envelopes, stable instance colors, and directional cues, while retaining geometry for environmental structures and simple objects. OBBs are computed in selected frames, with bounds updated over time and instance colors consistent across frames and views.

Rendering adapters specify entity selection, proxy templates, transparency, occlusion, and clipping behavior. Given shared scene geometry G, recorded world state S_i, camera parameters C_i^v for view v, and representation r, a conditioning frame is generated as $I_i^{v,r} = \mathcal{R}_r(G, S_i, C_i^v)$, where $\mathcal{R}_r$ is the representation-specific rendering adapter. Different views observe the same saved state, and representations within each view share camera parameters. Additional camera policies and rendering adapters can reuse recorded episodes without rerunning behavior controllers, provided the required information is available in the world records, whereas new behaviors produce new recorded executions. This separation preserves temporal alignment and enables controlled comparisons across representations and viewpoints while keeping the underlying world evolution fixed.

Final Process. The pipeline provides structured annotations that connect each observation to its underlying scene state and interaction history. These include entity identities, semantic categories, spatial and articulated states, camera parameters, action phases, participant relationships, and timestamped events. Bounding volumes, visibility, and depth provide additional geometric information, interpreted according to the associated scene or proxy representation. Interaction annotations retain their provenance, allowing controller-specified targets to be interpreted alongside independent geometric checks.

Validation covers task execution, geometric consistency, observation quality, and annotation alignment. Task-level checks verify expected action phases and temporal coverage. Geometric checks examine penetration, support, and contact consistency at recorded states, while replay and rendering checks assess state reconstruction, camera alignment, and representation consistency. Visual review evaluates subject visibility and the readability of selected observations. Final checks assess media integrity and verify that videos and annotations correspond to matching states, cameras, and time intervals.

Failed candidates are retained with provenance and diagnostic reports, while accepted episodes are managed as consistent data versions. Revisions identify changes to world records, camera trajectories, or representation policies. Camera and representation repairs update affected observations and annotations together while retaining unchanged world trajectories. Each released episode remains linked to its source records, rendering configuration, and verification reports, preserving traceability throughout subsequent revisions.

3.2 Evaluation Suite

We evaluate five complementary capabilities: controllability, visual and temporal quality, behavioral alignment, multi-view consistency, and long-horizon memory. Engine-recorded trajectories, events, and states provide references for evaluating adherence to the prescribed world evolution, while established perceptual metrics assess video quality.

Controllability. We measure whether entities and the camera follow their prescribed trajectories. **BBox IoU** and **Center Error** compare tracked entity boxes with engine-projected references; missing detections receive zero IoU, while Center Error measures localization conditional on detection.

Interaction Success Rate. Global similarity alone does not determine whether a prescribed interaction is successfully executed. We therefore evaluate registered engine events, each defined by its participants, temporal window, action, and expected end state. For each event, Qwen3.6-27B is provided with eight frames sampled around the event window, including both full-frame context and an engine-localized crop, and independently judges the prescribed action, expected end state, and participant presence. An event is considered successful only when the response is valid and both the action and end state are judged present. Participant presence is reported separately as a diagnostic and is not included in the success criterion. For a generated scene–camera sample s with event set $\mathcal{E}_s$, we define

$$\mathrm{ISR}_s = \frac{1}{|\mathcal{E}_s|} \sum_{e \in \mathcal{E}_s} \mathbf{1}[\text{valid response, action=yes, end state=yes}]. \tag{1}$$

Unknown judgments for either the action or end state, as well as invalid responses, remain in the denominator and do not count as successes; their frequencies are reported separately. Table entries average ISR_s equally across samples, preventing samples with more annotated events from receiving greater weight. In the tables, this metric is abbreviated as *Interaction Success*.

Logic–Render Alignment. Logic–Render Alignment measures whether the generated video visually follows the engine-defined event logic over time. For each annotated timeline segment, the judge receives twelve frames sampled from the corresponding temporal window, together with the expected event description and neighboring descriptions as context. The metric evaluates whether the action prescribed by the engine logic is visibly realized within the corresponding segment. For a sample s with annotated timeline segments $\mathcal{T}_s$, we define

$$\mathrm{LRA}_s = \frac{1}{|\mathcal{T}_s|} \sum_{\tau \in \mathcal{T}_s} \mathbf{1}[\text{prescribed action in segment } \tau \text{ is judged present}], \tag{2}$$

so that LRA_s is the fraction of timeline segments whose prescribed actions are visually consistent with the engine-defined logic.

Additional metric definitions, implementation details, and evaluation protocols are provided in Appendix C.1.

4 Experiments

Our experiments address one question: can current video generation and world models faithfully render a prescribed dynamic world, rather than merely produce a visually plausible video? For a game world built and controlled by an agent, visual quality is not enough: the model must follow the prescribed entity trajectories, camera motion and event timeline, and keep the world state consistent over long horizons and across synchronized views. We therefore evaluate four complementary scenarios: controlled generation with a prescribed RGB first-frame appearance reference (Section 4.2), controlled generation without one (Section 4.3), world-state preservation and entity re-entry over long horizons (Section D.1), and consistency across independently generated views of a multi-player scene (Section D.2). The experimental setup is detailed in Section B.

Across these scenarios we study three questions. First, how accurately do generated videos follow the prescribed entity trajectories, camera trajectories and event timelines? Second, how does adherence depend on the conditioning information, namely the form of the proxy representation, the presence of a shared RGB appearance reference, and the identity and position descriptions in the text? Third, does short-clip control extend to harder settings: long sequences composed of several generation

windows, entities that leave the view and re-enter, and synchronized views generated independently of each other? Finally, on a dedicated gunfight set we compare the methods with PWM, which is trained only on gunfight scenes, as a case study of a domain-specific model (Section D.4).

4.1 Evaluated Methods

We evaluate eleven models, grouped into three families by the control information they receive (the Input column of the tables).

General-purpose world models. EchoWM (Zhang et al., 2026), SANA-WM (Zhu et al., 2026), LingBot-World 2.0 (Robbyant Team et al., 2026) and AlayaWorld v1.1 (Team et al., 2026) receive an RGB first frame, the text prompt and the recorded camera trajectory. They obtain explicit camera control but no entity trajectories; we refer to them as *camera-conditioned* models.

Proxy-video-to-RGB models. These models translate an engine-rendered control representation into a realistic RGB video. Cosmos-Transfer2.5 (Ali et al., 2025) receives engine depth and class-level segmentation; C2R (Gomez-Nogales et al., 2026) the White proxy and the text; CWM (Chen et al., 2026b) the Mixed render, depth, the RGB first frame and the text; LynnReal-Omni (Mao et al., 2026b), which builds directly on MiniMax-H3 and extends it with native support for heterogeneous visual controls, including editable 3D renders and game recordings; and PWM (Huang et al., 2026a) 3D bounding boxes and the RGB first frame. Because PWM is trained only on gunfight scenes, it is evaluated only on the gunfight set of Section D.4.

Video foundation models. MiniMax-H3 (MiniMax-AI, 2026), Seedance 2.5 (ByteDance Seed Team, 2026) are used without any adaptation to our proxy representations. They take the colored-OBB proxy video as a reference video together with the text prompt, and RGB references as specified in each table.

4.2 Evaluation with First-Frame Conditioning

When the agent already has an appearance reference, such as a concept image or a prescribed character design, it can supply it as the RGB first frame. We first ask whether, under this comparatively favourable condition, models render the given world state; Table 1 reports the 40 pairs.

Models with entity-level proxy control keep the prescribed trajectories. LynnReal-Omni, MiniMax-H3 and CWM reach a BBox IoU of 0.518, 0.494 and 0.454. They preserve the scene and character appearance given by the first frame and also reproduce the motion of the other entities: in Figure 4, they keep both cars in their reference boxes while one of them passes the walking protagonist (per-entity IoU 0.62–0.86). The four camera-conditioned world models, which receive the camera trajectory but no entity trajectories, reach only 0.302–0.351; in the same scene EchoWM keeps the protagonist (0.63) but does not bring the oncoming car past the protagonist and lets the car ahead fall back toward the camera instead of keeping its distance (0.27 and 0.29). The gap in camera control is much smaller: the rotation error of most methods (1.7–3.4°) is comparable to that of the pose estimator on the engine renders (1.1–2.7°), and only AlayaWorld v1.1 and C2R are clearly worse. With current models and inputs, explicit camera information constrains the camera motion but does not substitute for entity-level motion control.

4.3 Evaluation without First-Frame Conditioning

In procedural game-world construction, the agent usually has no ready-made RGB appearance reference. We therefore ask whether models can jointly solve entity control, appearance generation and identity binding when the benchmark prescribes only the world state, the text and the control signals; Table 2 reports the 90 pairs. The control gap of Section 4.2 carries over to this set (LynnReal-Omni 0.457 and MiniMax-H3 0.409 versus 0.204–0.255 for the camera-conditioned world models) and again lies in the entities other than the protagonist. We first show what current models achieve without a reference and then the difficulties specific to its absence.

Without a first frame, models still render the described world in alignment with the proxy. Current models do not need an appearance reference to render a coherent scene. From the proxy and the text alone, they synthesize the scene in the prescribed style and keep it aligned with the proxy: in Figure 7, every method renders the photorealistic street described by the text, and the protagonist

Table 1: **Evaluation with first-frame conditioning.** Results on 40 five-second scene–camera pairs. Best value per column in **bold**.

Input: FF = first frame; P = 3D proxy; B = 3D bounding boxes; M = mixed render; D = depth; C = camera pose; CS = class-level segmentation. All methods use text prompts (omitted from Input).

Method	Trained	Input	Control					Quality		Temporal		Text Align.	Interaction Success ↑	Logic Align. ↑
			Entity		Camera									
			IoU ↑	Center Err. ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓	Imaging ↑	Aesthetic ↑	Warp Err. ($\times 10^{-3}$) ↓	Motion Smooth. ↑	CLIP-Score ↑		
EchoWM (Zhang et al., 2026)	✓	FF + C	0.351	0.055	2.2	0.151	0.166	67.6	0.607	0.619	0.987	25.2	0.506	0.685
SANA-WM (Zhu et al., 2026)	✓	FF + C	0.309	0.070	2.7	0.145	0.168	68.6	0.612	0.871	0.985	25.1	0.590	0.569
LingBot 2.0 (Robbyant Team et al., 2026)	✓	FF + C	0.302	0.064	3.3	**0.067**	0.116	70.1	0.614	1.360	0.974	25.1	**0.690**	0.780
AlayaWorld v1.1 (Team et al., 2026)	✓	FF + C	0.303	0.073	9.4	0.229	0.356	66.8	0.592	0.530	0.990	25.0	0.583	0.622
Cosmos 2.5 (Ali et al., 2025)	✓	D + CS	0.451	0.061	**1.7**	0.166	0.177	67.4	0.552	0.430	0.991	24.3	0.381	0.560
C2R (Gomez-Nogales et al., 2026)	✓	P	0.375	0.064	7.8	0.352	0.443	70.6	0.600	0.786	0.990	**25.7**	0.606	0.694
CWM (Chen et al., 2026b)	✓	M + FF + D	0.454	0.040	2.0	0.130	0.154	66.7	0.629	0.398	0.991	25.3	0.637	0.795
LynnReal-Omni (Mao et al., 2026b)	✓	FF + B	**0.518**	**0.036**	2.0	0.074	**0.100**	**73.8**	0.625	1.259	0.982	25.2	0.615	**0.832**
Seedance 2.5 (ByteDance Seed Team, 2026)	✗	FF + B	0.436	0.043	3.4	0.098	0.148	66.9	0.628	**0.314**	**0.993**	25.3	0.640	0.817
MiniMax-H3 (MiniMax-AI, 2026)	✗	FF + B	0.494	0.038	3.4	0.080	0.133	70.2	**0.634**	0.444	0.991	25.2	0.675	0.738



Figure 4: **Entity-level control versus camera-only conditioning.** A pedestrian walking down a city street, third-person follow camera, same first frame for all methods. Dashed: engine reference boxes of the oncoming car (cyan) and the car ahead (orange). The proxy-conditioned models keep both cars in their boxes (IoU 0.62–0.86); EchoWM, conditioned on the camera trajectory only, does not (0.27 and 0.29).

climbs toward the upper landing at the prescribed times (pair IoU 0.52–0.61). What the first frame would otherwise fix, however, is left to the model: which way each entity faces and which described appearance belongs to which person.

Appearance descriptions can be assigned to the wrong person. Without a visual reference, the model must infer from text which appearance description belongs to which person. As a result, even when the overall scene and actions appear plausible, identity–appearance associations can be incorrect. In the supermarket example in Figure 5a, which contains a courier, a student, and a tourist, MiniMax-H3 swaps the student's and tourist's outfits throughout the clip. Seedance 2.5 assigns the courier's outfit to the student's position while leaving the courier's box empty, whereas LynnReal-Omni preserves the intended appearance assignment for all three characters. In a manual audit of 62 multi-person scenes, MiniMax-H3 swaps the appearances of two people in 12 cases. These examples suggest that visually plausible multi-person generation can still exhibit errors in maintaining entity identity and associated appearance attributes.

Front and back are not always controlled. Without an image that fixes how each entity looks from the camera, models sometimes render an entity facing the opposite way. In the horseback scene of Figure 5b, the horse on the right faces away from the camera: LynnReal-Omni and MiniMax-H3 show its rear, whereas Seedance 2.5 renders it facing the camera. The entity is placed in its box, but its front–back direction is not controlled reliably.

5 Conclusion

We introduced **PROWBench**, a benchmark and extensible framework that grounds the evaluation of programmable world models in replayable world states and interaction records. Synchronized visual proxies enable controlled comparisons across representations, viewpoints, and time, while

Table 2: Evaluation without first-frame conditioning. Results on 90 scene–camera pairs. Best value per column in **bold**.

Input: P = 3D proxy; B = 3D bounding boxes; M = mixed render; D = depth; C = camera pose; FF = frame 0 of the MiniMax-H3 output for the same scene and camera; CS = class-level segmentation. All methods use text prompts (omitted from Input).

Method	Trained	Input	Control					Quality		Temporal		Text Align.	Interaction Success ↑	Logic Align. ↑
			Entity		Camera									
			IoU ↑	Center Err. ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓	Imaging ↑	Aesthetic ↑	Warp Err. ($\times 10^{-3}$) ↓	Motion Smooth. ↑	CLIP-Score ↑		
EchoWM (Zhang et al., 2026)	✓	FF + C	0.255	0.099	3.1	0.165	0.193	70.2	0.629	0.823	0.986	24.8	0.669	0.712
SANA-WM (Zhu et al., 2026)	✓	FF + C	0.240	0.075	4.7	0.228	0.275	69.9	0.636	0.676	0.986	24.5	0.564	0.591
LingBot 2.0 (Robbyant Team et al., 2026)	✓	FF + C	0.205	0.110	3.8	0.111	0.158	72.8	0.633	1.203	0.973	25.1	0.739	0.811
AlayaWorld v1.1 (Team et al., 2026)	✓	FF + C	0.204	0.101	10.7	0.324	0.461	69.3	0.602	0.564	0.990	24.8	0.714	0.712
Cosmos 2.5 (Ali et al., 2025)	✓	D + CS	0.376	0.057	5.3	0.259	0.326	71.9	0.551	0.603	0.990	24.3	0.440	0.457
C2R (Gomez-Nogales et al., 2026)	✓	P	0.315	0.052	8.8	0.456	0.548	69.9	0.609	0.666	0.990	**25.3**	0.540	0.625
CWM (Chen et al., 2026b)	✓	M + FF† + D	0.356	0.050	4.7	0.151	0.208	67.5	0.637	0.384	0.990	24.8	0.741	0.784
LynnReal-Omni (Mao et al., 2026b)	✓	B	**0.457**	**0.036**	**1.4**	**0.084**	**0.101**	**74.8**	**0.670**	1.676	0.976	24.2	**0.809**	**0.852**
Seedance 2.5 (ByteDance Seed Team, 2026)	✗	B	0.336	0.058	4.5	0.164	0.216	66.4	0.616	**0.257**	**0.993**	24.9	0.758	0.829
MiniMax-H3 (MiniMax-AI, 2026)	✗	B	0.409	0.055	3.6	0.119	0.168	71.6	0.663	0.407	0.990	24.8	0.758	0.806

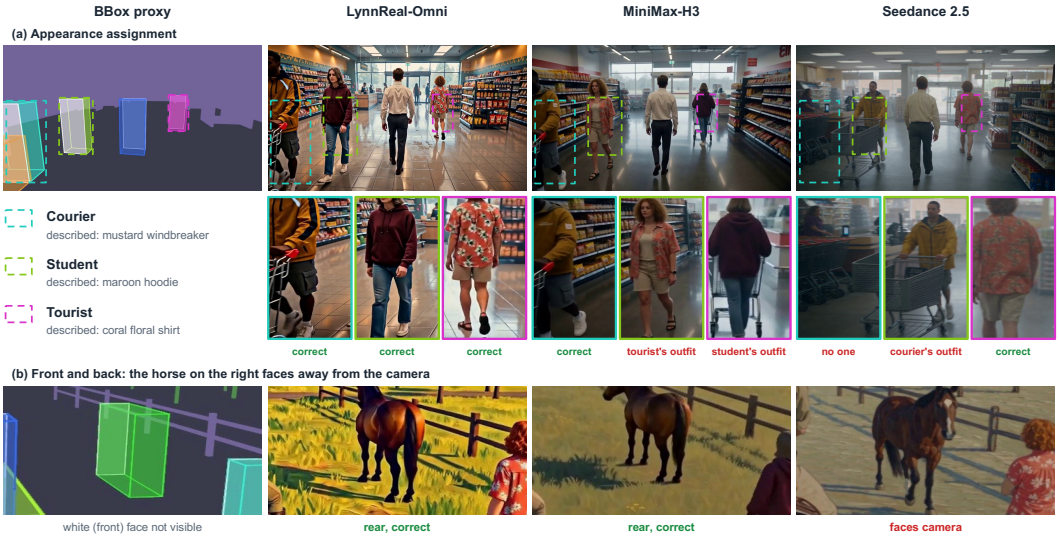


Figure 5: **Appearance and facing errors without a first frame.** (a) Supermarket scene at 3.0 s: dashed boxes are the engine reference boxes of a courier, a student and a tourist, each with a described outfit; the zoomed crops show what each method renders inside each box. MiniMax-H3 exchanges the student's and the tourist's outfits; Seedance 2.5 puts the courier's outfit at the student's position. (b) Horseback scene at 0.5 s: the horse on the right faces away from the camera; Seedance 2.5 renders it facing the camera.

Logic–Render Alignment and Interaction Success Rate complement geometric and perceptual metrics by measuring whether engine-defined rules and events are visibly realized.

Our experiments demonstrate the value of explicit entity-level control: proxy-conditioned models adhere to prescribed trajectories more closely than models receiving camera control alone, and can generate coherent scenes even without an RGB first-frame reference. At the same time, visual plausibility does not guarantee faithful world realization. Appearance assignment, identity binding, and facing direction remain unreliable without a visual reference; entity control and reappearance degrade over longer horizons; and geometric agreement across independently generated views does not ensure consistent participant identity or appearance.

These findings position programmable-world rendering as a problem of preserving executable state across entities, events, time, and viewpoints, rather than optimizing visual quality alone. PROWBench provides a reproducible foundation for measuring progress toward this goal. Reliable generative world renderers will require stronger entity–appearance binding, event-aware generation, long-horizon state memory, and joint multi-view consistency.

AI USE STATEMENT

We used MiniMax-H3 to generate RGB first-frame conditioning images for the evaluation in Section 4.2 and used its outputs to assist in curating the evaluation set without a prescribed first frame in Section 4.3. Vision-language models, including Qwen3.6-27B, supported automatic evaluation of interaction success, logic–render alignment, appearance compliance, cross-view appearance consistency, and state persistence (Section 3.2 and Appendix C.1). AI also assisted in reviewing interaction labels across 170 benchmark cases using source descriptions, entity roles, recorded events and trajectories, and sampled proxy frames. Original labels and supporting evidence were retained (Appendix A).

For the 40-case first-frame evaluation set, generated reference images underwent automatic checks of person count, bounding-box alignment, visual style, and facing direction. We also manually audited appearance assignments in 62 multi-person scenes (Section 4.3). The VLM-based appearance judgments have not been systematically calibrated against human annotations, as discussed in Appendix C.1.

OpenAI Codex (GPT-6) assisted with adapting the manuscript to the ICLR 2027 LaTeX template, formatting tables, editing language, adjusting citations and cross-references, and drafting this disclosure. Formatting changes were checked through source comparisons, LaTeX compilation, and visual inspection of the rendered PDF. The authors take responsibility for the final manuscript and all accompanying artifacts, including AI-assisted content.

REFERENCES

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 2024.

ByteDance Seed Team. Seedance 2.5. https://seed.bytedance.com/en/blog/one-take-creation-flexible-referencing-introducing-seedance-2-5, July 2026.

Ziqi Cai, Siqi Yang, Yimu Wang, Zixian Gao, Yunheng Liu, Shuchen Weng, Erwin Wu, Kaipeng Zhang, and Boxin Shi. MASS: Multiplayer world models with authoritative shared state, 2026a.

Ziqi Cai, Taoyu Yang, Zheng Chang, Si Li, Han Jiang, Shuchen Weng, and Boxin Shi. Lighting-grounded video generation with renderer-based agent reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20055–20065, 2026b.

Haoyu Chen, Kaichen Zhou, Hang Hua, Kaile Zhang, Jingwen Qian, Wufei Ma, Haonan Chen, Chunjiang Liu, Yizhou Zhao, Xiaoyuan Wang, et al. Memobench: Benchmarking world modeling in dynamically changing environments. *arXiv preprint arXiv:2606.27537*, 2026a.

Yiwen Chen, Guosheng Lin, and Chi Zhang. Code world model: Coding agent as world brain. *arXiv preprint arXiv:2608.25927*, 2026b.

Kaixin Ding, Xi Chen, Minghong Cai, Zhiyuan Xu, Yiyang Wang, Yuxiang Lu, Junyi Li, Shuyang Chen, Yuan Gao, Xin Tao, et al. Playworld: Benchmarking world models with agent players over long-horizon objectives. *arXiv preprint arXiv:2608.13552*, 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 (ICCV)*, pp. 27713–27724, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Duan_WorldScore_A_Unified_Evaluation_Benchmark_for_World_Generation_ICCV_2025_paper.html.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *International Conference on Learning Representations (ICLR)*, 2024.

Gonzalo Gomez-Nogales, Yicong Hong, Chongjian Ge, Peiye Zhuang, Marc Comino-Trinidad, Dan Casas, and Yi Zhou. Coarse-to-real: Generative rendering for populated dynamic scenes. *arXiv preprint arXiv:2601.22301*, 2026.

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

Zheng-Hui Huang, Guixu Lin, Jiacheng Lin, Yi-Chuan Huang, Ruihan Yu, Muyao Niu, Siqi Yang, Yu-Lun Liu, Yung-Yu Chuang, Kaipeng Zhang, et al. Programmable world model. *arXiv preprint arXiv:2609.10540*, 2026a.

Zheng-Hui Huang, Zhixiang Wang, Jiaming Tan, Ruihan Yu, Yidan Zhang, Bo Zheng, Yu-Lun Liu, Yung-Yu Chuang, and Kaipeng Zhang. Generative world renderer. *arXiv preprint arXiv:2604.02329*, 2026b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. PEEKABOO: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8079–8088, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Geonung Kim, Janghyeok Han, and Sunghyun Cho. Videofrom3d: 3d scene video generation via complementary image and video diffusion models. In *SIGGRAPH Asia 2025 Conference Papers*, pp. 110:1–110:11. ACM, 2025. doi: 10.1145/3757377.3763871.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision*, 2018.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025a. doi: 10.52202/085713-1834. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4ec03ed08a3fcb59e1c815b5598beff1-Abstract-Datasets_and_Benchmarks_Track.html.

Pengxiang Li, Kai Chen, Zhili Liu, Ruiyuan Gao, Lanqing Hong, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Tracklet-conditioned video generation via diffusion models. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 3539–3548, 2025b.

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12112–12123, 2025c. doi: 10.1109/ICCV51701.2025.01126.

Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In *Computer Vision – ECCV 2024*, pp. 469–485, 2024. doi: 10.1007/978-3-031-73229-4_27.

Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9801–9810, 2023.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26069–26080, 2025. doi: 10.1109/CVPR52734.2025.02428.

Guixu Lin, Zheng-Hui Huang, Siqi Yang, Ming-Hsuan Yang, Kaipeng Zhang, and Zhixiang Wang. Generative world renderer at the speed of play. *arXiv preprint arXiv:2607.18703*, 2026a.

Zijun Lin, Zeqing Wang, Cheston Tan, Bihan Wen, and Yeying Jin. Stateplay: State-aware game world models for mechanics-consistent generation, 2026b.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond H. Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149, 2024. doi: 10.1109/CVPR52733.2024.02090.

Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *Computer Vision – ECCV 2024*, pp. 329–345, 2024. doi: 10.1007/978-3-031-72989-8_19.

Ge Ya Luo, Zhi Hao Luo, Anthony Gosselin, Alexia Jolicoeur-Martineau, and Christopher Pal. Ctrl-v: Higher fidelity autonomous vehicle video generation with bounding-box controlled object motion. *Transactions on Machine Learning Research*, 2025.

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

Xiaofeng Mao, Peijia Lin, Shaohao Rui, Yibo Zhang, Haibin Wan, and Weijie Ma. Lynnreal-omni: Native multi-modal video generation for agentic visual workflows, 2026b. URL https://arxiv.org/abs/2609.15863.

Zian Meng, Zhen Li, Chuanhao Li, Qiang Li, and Kaipeng Zhang. Marionette: Predicting world states, rendering geometry, painting appearance, 2026.

MiniMax-AI. Minimax h3. Official GitHub repository, August 2026. URL https://github.com/MiniMax-AI/MiniMax-H3. Open-sourced on August 3, 2026.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, and Ruimao Zhang. Worldsimbench: Towards video generation models as world simulators. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50338–50362. PMLR, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *CoRR*, abs/2601.20540, 2026. doi: 10.48550/arXiv.2601.20540.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, Chen Gao, Wei Wu, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Xin Wang, Wenwu Zhu, and Yong Li. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *AI Open*, 7:208–226, 2026. doi: 10.1016/j.aiopen.2026.07.002.

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. *CoRR*, abs/2512.14614, 2025. doi: 10.48550/arXiv.2512.14614.

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

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, 2025.

Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52274–52289. PMLR, 2024a.

Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *ACM SIGGRAPH 2025 Conference Papers*, pp. 60:1–60:10. ACM, 2025. doi: 10.1145/3721238.3730755.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *Computer Vision – ECCV 2024*, volume 15106, pp. 55–72. Springer, 2024b. doi: 10.1007/978-3-031-73195-2_4.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14749–14759, 2024c. doi: 10.1109/CVPR52733.2024.01397.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11. ACM, 2024d. doi: 10.1145/3641519.3657518.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *CoRR*, abs/2604.08995, 2026. doi: 10.48550/arXiv.2604.08995.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6902–6912, 2024. doi: 10.1109/CVPR52733.2024.00659.

Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, Jing Tang, Lei Sun, Jiahong Wu, Xiangxiang Chu, Zeming Liu, and Kaiqi Huang. Omni-WorldBench: Towards a comprehensive interaction-centric evaluation for world models. *arXiv preprint arXiv:2603.22212*, 2026. doi: 10.48550/arXiv.2603.22212. URL https://arxiv.org/abs/2603.22212v1.

Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Kaipeng Zhang, and Yongtao Ge. WorldMark: A unified benchmark suite for interactive video world models. *arXiv preprint arXiv:2604.21686*, 2026. doi: 10.48550/arXiv.2604.21686. URL https://arxiv.org/abs/2604.21686v1. Version 1.

Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. MIND: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026. doi: 10.48550/arXiv.2602.08025. URL https://arxiv.org/abs/2602.08025v1.

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. doi: 10.48550/arXiv.2605.25874. URL https://arxiv.org/abs/2605.25874v1.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11590–11599, 2025.

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

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, and Yahui Zhou. Matrix-game: Interactive world foundation model. *CoRR*, abs/2506.18701, 2025. doi: 10.48550/arXiv.2506.18701.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

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

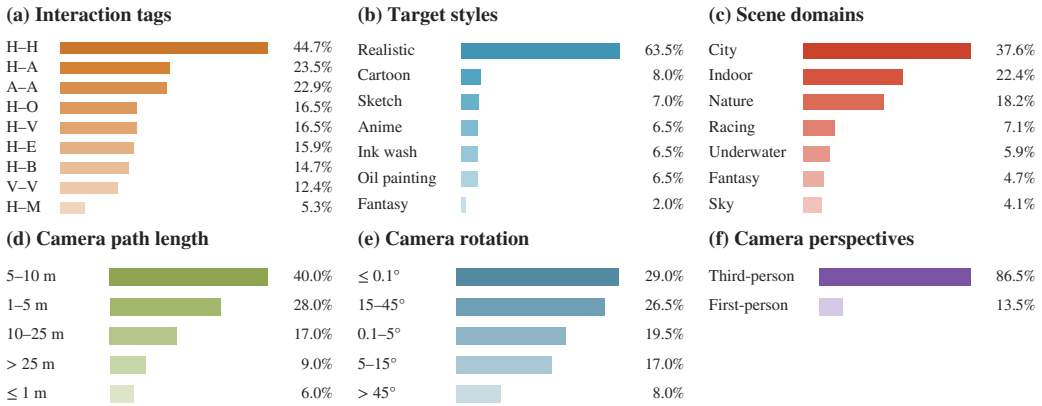


Figure 6: Dataset composition of PROWBench: 170 evaluation cases, 200 selected camera views, and 600 proxy videos. The six panels show (a) interaction categories, (b) requested visual styles, (c) scene domains, (d) camera path length, (e) maximum angular deviation from the initial camera orientation, and (f) camera perspectives. Percentages in (a,c) use all 170 cases as the denominator; those in (b,d,e,f) use all 200 selected views. Interaction labels follow a unified task-level annotation policy and may overlap within a case. H/V/B/O/A/M/E denote human/vehicle/building/object/animal/machine/environment. Camera statistics cover each full source clip, including both five-second and 30- second sequences. Each selected view provides three synchronized proxy representations derived from the same source world-state records.

A Data Statistics

PROWBench comprises 170 evaluation cases identified by unique episode IDs, 200 selected camera views, and 600 proxy input videos across four collections: diverse scenarios, multi-participant trajectories, event-driven state transitions, and long-horizon trajectories. Each selected view provides three synchronized representations: White, Mixed, and Colored-OBB. All source proxy inputs are rendered at 30 FPS: 570 videos are five seconds long with 150 frames each, and 30 videos are 30 seconds long with 900 frames each. There are 540 videos at 1280×720 and 60 event-driven videos at 832×480. This yields 37,500 view-specific camera-frame observations and 112,500 proxy video frames, totaling 62.5 minutes of aggregate proxy-video duration across the three representations.

The diverse-scenario collection contains 130 episodes, divided between the 40-case first-frame and 90-case no-first-frame comparison groups. Each episode contributes one selected view and three proxy videos, yielding 390 videos. This collection spans seven scene domains and, under the common interaction policy below, nine non-exclusive interaction categories with 227 episode–tag assignments.

The multi-participant collection, Multi4, contains ten episodes with four participants following distinct scripted trajectories each. It covers six trajectory patterns: staggered following, opposing movement, crossing, convergence with turning, overtaking, and parallel turns. Four synchronized third-person following views are selected per episode, yielding 40 views and 120 proxy videos.

The event-driven collection, Neon, contains 20 cases derived from ten source recordings. Each case provides one third-person view and three proxy videos, contributing 20 views and 60 videos. These cases vary camera and actor motion around a recorded actor-death event. Episode IDs therefore identify benchmark cases and need not correspond to independent source recordings.

The long-horizon collection contains ten continuous 30-second episodes covering traversal, landmark revisitation, camera look-around, combat, and racing. One view is selected per episode: six first-person and four third-person views, yielding ten views and 30 proxy videos. These cases support the long-horizon evaluation in Section D.1 and are included in the dataset totals above.

Scene-domain labels combine 130 native annotations with 40 source-reviewed annotations. The resulting counts are city (64), indoor (38), nature (31), racing (12), underwater (10), fantasy (8),

Table 3: Comparison of video and world-model benchmarks. Inputs (union): T=text, I=image, V=video context, C=camera trajectory, and P=proxy video. FPP/TPP denote first-/third-person perspectives. Focus summarizes the primary capability emphasized by each benchmark. **Definitions.** States: per-frame multi-entity 3D transforms and joint states; Replay: geometry and recorded scene states sufficient for re-rendering; Multi-view: synchronized videos from different viewpoints sharing the same underlying scene states; Proxies: alternative visual representations sharing the same scene states and cameras. **Symbols.** ✓: documented support; –: undocumented or not applicable under these definitions; Pos.: partial camera/character position records; Sim.: simulator-based evaluation without documented recorded-state replay; Init.: paired initial views only.

Benchmark	Inputs	Perspective		Scene records		Paired renderings		Focus
		FPP	TPP	States	Replay	Multi-view	Proxies	
VBench (Huang et al., 2024)	T	–	–	–	–	–	–	Video quality
WorldScore (Duan et al., 2025)	T,I,C	✓	–	–	–	–	–	Camera & dynamics
WorldModelBench (Li et al., 2025a)	T,I	✓	–	–	–	–	–	Physical reasoning
MIND (Ye et al., 2026)	V,C	✓	✓	Pos.	–	–	–	Visual memory
WorldArena (Shang et al., 2026)	T,I,C	✓	–	–	Sim.	–	–	Embodied utility
Omni-WorldBench (Wu et al., 2026)	T,I,C	✓	–	–	–	–	–	Interaction
MemoBench (Chen et al., 2026a)	T,I,C	–	–	–	–	–	–	Memory
WorldMark (Xu et al., 2026)	I,C	✓	✓	–	–	Init.	–	Control
WBench (Ying et al., 2026)	T,I,C	✓	✓	–	–	–	–	Multi-turn
PlayWorld (Ding et al., 2026)	T,I,C	✓	✓	–	–	–	–	Long-horizon
PROWBench (Ours)	T,I,C,P	✓	✓	✓	✓	✓	✓	**State GT**

and sky (7); city includes airport aprons and industrial outdoor settings. We review interaction labels across all 170 cases under a common task-level policy, using source descriptions, entity roles, recorded events and trajectories, and sampled proxy frames. Labels cover the authored focal task and its participants, including explicit relative group motion; incidental background co-occurrence is excluded. Weapons, handcarts and furniture count as objects; fixed architectural access counts as human–building interaction; landmark observation and revisitation count as environment exploration. Original labels and per-tag evidence are retained. This AI-assisted review yields 293 episode–tag assignments across nine non-exclusive categories. These are source-defined task labels, including description-supported weapon use in Neon, and do not imply independently verified physical contact or hit causality.

Across the four collections, the annotations cover 30 environment labels and 47 semantic action labels for all 170 episodes, including source-supported combat labels for the 20 Neon cases. Seven requested visual styles cover all 200 views; the 40 Multi4 views are labeled realistic from explicit conditioning prompts. These labels describe requested styles rather than measured output appearance. The selected cameras comprise 173 third-person views (86.5%) and 27 first-person views (13.5%). Measured over each full recorded clip, the median camera path length is 6.11 m, and the median maximum angular deviation from the initial orientation is 6.60°. Figure 6 summarizes the interaction, style, scene-domain, camera-motion, and perspective distributions.

B Experimental Setup

Evaluation data. We build one evaluation set per scenario. All samples come from recorded game-world states and are scored against the engine state (Section 3.2); each sample is one scene observed by one camera. Every recorded state is provided in three proxy representations, White, Mixed and colored OBB (Section 3.1), together with engine depth, the camera trajectory, a scene description and the state of every object. Each method receives the control representation of its native interface.

With a prescribed RGB first frame (Table 1). This set contains 40 scene–camera pairs (31 third-person follow, 9 first-person), and models an agent that already has a concept image or character design. To give every method that accepts a first frame the same appearance reference, consistent with the prescribed world state, we generate one RGB first frame per pair from the colored-OBB proxy of that camera and the scene description. Because this image is part of the benchmark, it must also match the recorded person count, positions, facing and layout; after comparing several image

generators, we use MiniMax-H3, whose first frames aligned best with the proxy geometry in person count and layout. The set contains 17 realistic and 23 stylized first frames, each admitted only after automatic checks of person count, box alignment, style and facing. All methods that accept a first frame receive the identical image.

Without a prescribed RGB first frame (Table 2). This set contains 90 pairs from disjoint scenes (78 follow, 12 first-person), with the same three proxy representations but no appearance reference shared by all methods: appearance must be determined by each model from the text and the control signal. Methods whose native interface requires an RGB image (the world models and CWM) receive frame 0 of the MiniMax-H3 output for the same scene and camera; this frame is not a prescribed reference, but it can transmit MiniMax-H3's appearance assignments. Analyses of appearance identity account separately for effects inherited from this frame. The curation of this set used MiniMax-H3 outputs, so results involving MiniMax-H3 should be read with this potential bias in mind.

Long-horizon (Table 4). Ten 30-second recordings, one camera each, evaluate long-horizon temporal continuity, particularly whether entities remain consistent when they leave the view, re-enter later, and preserve their state over time.

Multi-view (Table 5). Ten four-player scenes are each observed by four synchronized third-person follow cameras, one per player. The four views depict the same instants and the same world state, which tests whether positions, identities and appearances of the players, and the scene structure, agree across views.

Domain-specific comparison (Table 7). Twenty gunfight scenes with a follow camera serve the comparison with PWM (Huang et al., 2026a). Because PWM (Huang et al., 2026a) is trained only on this kind of scene, the set compares PWM (Huang et al., 2026a) with the general pipelines within its training domain;

Conditioning and protocol. Text prompts are generated automatically from the recorded scene, entities and action timeline. All models are evaluated in their native output formats against the same engine records; the engine renders of the proxies are scored as diagnostic references.

C Evaluation Details

C.1 Metric Details

Entity control. We evaluate entities marked as dynamic or interaction candidates, excluding terrain, effects, and scene geometry. Entities that are not visible in the reference view are excluded from evaluation, with additional filtering for first-person views to remove the camera wearer and other entities that dominate the field of view.

Generated entities are localized using SAM 3 (`facebook/sam3`) from text prompts derived from their semantic labels. If SAM 3 fails to detect a label throughout a clip, we fall back to SAM 2 (Ravi et al., 2024), initialized from the reference box. Generated tracks are matched to reference entities by label-compatible Hungarian matching using track-level mean IoU.

For entity j with visible instants $\mathcal{V}_j$,

$$\mathrm{IoU}_j = \frac{1}{|\mathcal{V}_j|} \sum_{t \in \mathcal{V}_j} \mathrm{IoU}\big(\hat{b}_j(t), b_j(t)\big), \qquad \mathrm{CErr}_j = \frac{1}{|\mathcal{M}_j|} \sum_{t \in \mathcal{M}_j} \frac{\|\hat{c}_j(t) - c_j(t)\|_2}{\sqrt{2}}, \tag{3}$$

where $b_j(t)$ and $\hat{b}_j(t)$ denote the reference and generated bounding boxes in normalized image coordinates, and $\mathcal{M}_j \subseteq \mathcal{V}_j$ contains the instants at which the generated track is detected. Missing boxes contribute zero IoU, so **BBox IoU** captures both localization accuracy and persistence, while **Center Error** measures localization error conditional on detection and is normalized by the image diagonal. Clip-level BBox IoU is averaged over all evaluated entities, whereas Center Error is averaged over entities with at least one matched instant. Clips with no evaluable entities are excluded.

Camera control. Camera trajectories of generated videos are recovered with ViPE (Huang et al., 2025) in pose-only mode with intrinsics fixed to the recorded camera (after the method's crop and resize) and with dynamic entities masked using their engine labels. Poses are converted to the engine

convention and expressed relative to the first instant. Following CamI2V (Zheng et al., 2024), with R_t, T_t the reference and $\hat{R}_t$, $\hat{T}_t$ the estimated relative poses on the grid $\mathcal{G}$,

$$\begin{aligned} \text{RotErr} &= \frac{1}{|\mathcal{G}|}\sum_{t\in\mathcal{G}} \arccos\frac{\operatorname{tr}(\hat{R}_t R_t^\top)-1}{2}, \qquad \text{TransErr} = \frac{1}{|\mathcal{G}|}\sum_{t\in\mathcal{G}} \big\|\bar{T}_t - \bar{\hat{T}}_t\big\|_2, \\ \text{CamMC} &= \frac{1}{|\mathcal{G}|}\sum_{t\in\mathcal{G}} \big\|[R_t \,|\, \bar{T}_t] - [\hat{R}_t \,|\, \bar{\hat{T}}_t]\big\|_F, \end{aligned} \tag{4}$$

where each translation sequence is divided by its own maximum norm, $\bar{T}_t = T_t/(\max_s\|T_s\| + 10^{-9})$, to remove the scale ambiguity of monocular estimation. RotErr is reported in degrees. The normalization also removes differences in the overall magnitude of camera motion, so TransErr and CamMC measure the shape of the path rather than its length. Because the estimator is itself imperfect, the engine renders are scored in the same way; their errors indicate the resolution of the camera columns but are not strict error bounds.

Visual and temporal quality. **Imaging Quality** (MUSIQ (Ke et al., 2021)) and **Aesthetic Quality** (the LAION aesthetic predictor) follow VBench (Huang et al., 2024). **Warp Error** (Lai et al., 2018; Liu et al., 2024) estimates RAFT optical flow between consecutive frames at 672×384, warps the next frame back to the current one, masks occlusions by forward–backward consistency and reports the masked mean squared RGB error ($\times 10^{-3}$). **Motion Smoothness** is the VBench score that reconstructs frames with a video frame-interpolation model (Li et al., 2023). **CLIP-Score** (Radford et al., 2021) is the mean cosine similarity ($\times 100$) between CLIP ViT-B/32 embeddings of the 40 grid frames and the episode's action title. These metrics are computed without reference to the engine state and therefore measure plausibility, not adherence.

Appearance Compliance. For multi-view generation, we first ask whether each intended participant looks as described. The VLM compares the person at the engine-projected location with the canonical authored appearance description, shared across methods rather than replaced by a method-specific shortened caption. At 0.5, 2.5, and 4.5 s, it inspects full-frame context and a target crop, checking clothing color, clothing style, hair, accessories, and build. Attributes absent from the description or not visually observable are not guessed.

Cross-view Appearance Consistency. We separately ask whether the same intended participant has a consistent appearance across synchronized views. The VLM sees two to four views together *without* the authored appearance description and compares the same five attributes for every unordered view pair. Ordinary changes in pose, perspective, lighting, or front/back visibility are not inconsistencies. A participant can therefore be consistently depicted across views while failing to match the authored description; the two appearance metrics answer different questions. The table abbreviates this metric as *Consistency*.

Appearance scoring and coverage. Both appearance metrics use Qwen3.6-27B with greedy decoding. For an observation o (one person–view–time for compliance, or one person–view-pair–time for consistency), let $\mathcal{A}_o$ contain its judgeable attributes. If the target is present and $\mathcal{A}_o$ is nonempty, its score is

$$q_o = \prod_{a\in\mathcal{A}_o} \mathbf{1}[\text{attribute } a \text{ matches or agrees}]. \tag{5}$$

Thus every judged attribute must pass; this is not an average of attribute scores. An absent or ambiguous target, or an observation with no judgeable attribute, is unscored; both targets must be present for a view pair. We first average scored views or view pairs per participant and time, then average available participant–time scores within each scene, and finally average scenes equally and multiply by 100. Empty participant–time groups are omitted; a scene with no score makes the overall result undefined. *Judgeable coverage* uses the same hierarchy with a 1 for a scored observation and 0 otherwise. Unlike the event scores, unjudgeable appearance observations reduce coverage rather than count as failures. These conditional percentages should therefore be read with coverage and BBox miss rate. The judgments have not been systematically calibrated against human annotations; localization drift can also affect the appearance assessment.

Geometric multi-view consistency. We also report **MEt3R**, which compares image features after aligning views through geometry predicted from the generated frames. It measures mutual consistency of the generated views, not agreement with the engine geometry, and complements the person-level appearance judgments rather than measuring adherence to an appearance description. We evaluate all six view pairs at nine synchronized times per scene, average both comparison directions, and average scenes equally; lower is better.

Memory over long horizons. For the 30-second evaluation, we test whether a dynamic entity that leaves the view returns where the engine places it and in the state it should have. An *absence event* is read from engine visibility on the 8 Hz grid: an entity is visible, then invisible, and visible again at a return time t_r at least $1.0\,\mathrm{s}$ after its last visible instant. **Reappearance IoU** is the best IoU between the reference box and any label-compatible SAM 3 detection at each visible instant in $[t_r, t_r + 1\,\mathrm{s})$, with a missing detection counted as 0; it is averaged per event, then per scene, then over scenes with at least one event. It deliberately uses per-instant best matches rather than persistent track identities, because trackers routinely assign a new identity after a long disappearance. **State Persistence** concerns the locomotion state of a returning person, labelled *walking* or *standing* from the mean engine speed over the second before leaving and over $[t_r, t_r + 1.5\,\mathrm{s}]$ after returning (walking above $0.2\,\mathrm{m/s}$). An event is judgeable if some post-return detection overlaps the reference box or, failing that, the nearest detection centre lies within the reference box's largest side. Qwen3.6-27B sees six frames of the post-return window with a zoom on the matched detection and names the state; *unknown* counts as incorrect and invalid responses are excluded. The score is the accuracy over judgeable events whose state is unchanged by the absence, averaged per scene and then over scenes. This judge has not been calibrated on engine renders of this set.

D Additional Results

D.1 Long-Horizon Evaluation

Control in short clips does not guarantee that a model maintains one world state over longer horizons, in which entities leave the view, re-enter later and keep their state and motion across generation windows. Table 4 evaluates ten 30-second sequences.

Alignment degrades over long horizons. Over the full 30 s, BBox IoU is only 0.175–0.280, below the 0.31–0.52 of the same five methods on the five-second sets (different scenes), and 42 of the 50 method–recording results stay below 0.3. Figure 8a shows how this arises: every method is best in the first five seconds (0.25–0.36) and falls to 0.12–0.25 after ten seconds, so the prescribed layout is followed at the start of a sequence but increasingly lost as it continues.

Returning entities are rarely restored to the prescribed position. The Reappearance IoU, measured on entities that return after leaving the frame, is only 0.053–0.131. In Figure 8b, every method localizes a pedestrian before it leaves (IoU about 0.6); after it re-enters some 13 s later, the post-return IoU is 0.34 (CWM), 0.30 (LynnReal-Omni), 0.17 (Seedance 2.5) and 0.01 (MiniMax-H3). Following an entity within a local window does not imply restoring its position after a long absence across windows.

Spatial recovery and state preservation are different dimensions. State Persistence, which checks whether a returning entity shows its prescribed walking or standing state, ranks the methods differently from Reappearance IoU: CWM has the lowest Reappearance IoU but a State Persistence of 0.886. The two metrics capture different aspects of long-term consistency. The state scores cover only 17–21 events per method from an uncalibrated judge, so these numbers indicate tendencies rather than a stable ranking. The window-chaining protocols also differ (for example, Seedance 2.5 additionally receives the last frame of the previous window), so this section measures end-to-end long-horizon preservation under each method's actual protocol, not an isolated internal memory.

D.2 Multi-Player Multi-View Evaluation

In a multi-player game, each player observes the same world through their own camera. Even when each view is generated independently, the views must agree on where each participant is, who they

are, what they look like, and the scene structure. Table 5 evaluates ten four-player scenes, each with four synchronized third-person follow views, one per player, generated independently.

Geometric consistency does not imply identity consistency. The results reveal a clear gap between geometric agreement and participant consistency across independently generated views. C2R obtains the lowest MEt3R point estimate (0.397), yet achieves only 32.2% Appearance Compliance and 25.5% Cross-view Consistency. In contrast, methods whose inputs explicitly distinguish the participants reach 82–94% Appearance Compliance and 81–90% Cross-view Consistency on judgeable observations. Thus, a model can generate views that are geometrically coherent while failing to preserve the same participant appearance across viewpoints. Multi-view evaluation should therefore measure both geometric and identity/appearance consistency, since geometric agreement alone is not a reliable proxy for participant consistency.

Table 4: **Long-horizon evaluation of controllability and memory.** Results on ten 30-second recordings, with one camera per recording and one output per method. Best value per column in **bold**.

Input: FF = first frame; P = 3D proxy; B = 3D bounding boxes; M = mixed render; D = depth.

Method	Trained	Input	Control					Memory	
			Entity		Camera				
			IoU ↑	Center Err. ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓	Reappear. IoU ↑	State Persist. ↑
C2R (Gomez-Nogales et al., 2026)	✓	P	0.212	0.066	0.7	0.571	1.119	0.111	0.405
CWM (Chen et al., 2026b)	✓	M + FF + D	0.210	0.070	0.5	0.381	0.793	0.053	**0.886**
LynnReal-Omni (Mao et al., 2026b)	✓	B	**0.280**	**0.054**	**0.4**	0.273	**0.610**	0.103	0.528
Seedance 2.5 (ByteDance Seed Team, 2026)	✗	B	0.175	0.077	0.7	**0.242**	0.819	**0.131**	0.650
MiniMax-H3 (MiniMax-AI, 2026)	✗	B	0.206	0.088	0.6	0.351	0.814	0.059	0.833

Table 5: **Multi-view evaluation of controllability and consistency.** Five methods are evaluated on 10 scenes with four synchronized views each (40 clips per method). Views are generated independently.

Input: P = 3D proxy; B = 3D bounding boxes; M = mixed render; D = depth; FF = frame 0 of the MiniMax-H3 output for the same scene and camera. All methods use text prompts (omitted from Input).

Method	Trained	Input	Appearance			BBox control			Camera control†			Interaction Success ↑	Logic Align. ↑
			Compliance (%) ↑	Consistency (%) ↑	MEt3R ↓	IoU ↑	Center err. ↓	Miss rate ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓		
C2R (Gomez-Nogales et al., 2026)	✓	P	32.2	25.5	**0.397**	0.414	0.039	0.192	2.2	0.155	0.173	0.637	0.750
CWM (Chen et al., 2026b)	✓	M + FF + D	92.9	88.1	0.422	0.439	0.037	0.064	2.0	0.047	0.071	0.812	0.950
LynnReal-Omni (Mao et al., 2026b)	✓	B	87.6	**89.7**	0.448	0.490	0.038	**0.047**	**0.4**	**0.023**	**0.026**	0.700	0.900
Seedance 2.5 (ByteDance Seed Team, 2026)	✗	B	82.2	81.0	0.499	0.341	0.049	0.089	1.0	0.076	0.083	0.812	0.963
MiniMax-H3 (MiniMax-AI, 2026)	✗	B	**94.0**	88.6	0.400	**0.524**	**0.029**	0.049	0.9	0.034	0.043	0.825	0.938

D.3 PROXY REPRESENTATION.

As shown in Table 6, for a fixed generator, first frame and seed, the choice of proxy changes layout little 10 (BBox IoU 0.506–0.550), while in our examples box proxies yield more articulated strides than 3D proxy and mixed proxies. We therefore use the bounding-box proxy as the representation input to Minimax-H3 in the evaluation.

Table 6: MiniMax-H3 under different proxy representations of the released benchmark.

Input: FF = first frame; P = 3D proxy; B = 3D bounding boxes 1–5); M = mixed render; AB = all-entity boxes.

Method	Input	Control					Quality		Temporal		Text Align.	Interaction Success ↑	Logic Align. ↑
		Entity		Camera									
		IoU ↑	Center Err. ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓	Imaging ↑	Aesthetic ↑	Warp Err. ($\times 10^{-3}$) ↓	Motion Smooth. ↑	CLIP-Score ↑		
MiniMax-H3 (MiniMax-AI, 2026)	FF + P	0.550	0.019	5.94	0.159	0.257	76.7	0.541	0.256	0.990	26.6	0.948	0.924
MiniMax-H3 (MiniMax-AI, 2026)	FF + M	0.506	0.024	5.08	0.126	0.213	76.8	0.541	0.261	0.990	26.6	0.984	0.889
MiniMax-H3 (MiniMax-AI, 2026)	FF + B	0.537	0.023	6.84	0.155	0.263	76.7	0.539	0.263	0.990	26.7	0.969	0.864
MiniMax-H3 (MiniMax-AI, 2026)	FF + AB	0.536	0.022	6.21	0.138	0.247	76.8	0.542	0.260	0.990	26.8	0.938	0.888



Figure 7: **Style and trajectory without a first frame.** Stair-climbing scene, follow camera, sampled at 0.5, 2.5 and 4.9 s; the text prescribes a graphic, non-photorealistic marker-sketch style. LynnReal-Omni, Seedance 2.5 and MiniMax-H3 render this style with varying fidelity, and CWM inherits it from MiniMax-H3's frame 0; C2R renders the scene photorealistically. All methods follow the proxy as the protagonist climbs toward the upper landing (pair IoU 0.52–0.61).



Figure 8: **Alignment over the 30-s recordings.** (a) BBox IoU averaged in 5-s windows over the ten recordings and all entities with a reference box: every method is best in the first five seconds and loses the prescribed layout as the sequence continues (the visible entities differ between windows, so the curves show a trend). (b) Park scene, first-person camera, the recording with the largest drop: the dashed box is the engine reference of a pedestrian who leaves the view at 3.9 s and returns at 17.25 s. All methods localize it before it leaves (IoU about 0.6); after the return, the IoU is 0.34 (CWM), 0.30 (LynnReal-Omni), 0.17 (Seedance 2.5) and 0.01 (MiniMax-H3).

D.4 Case Study with a Domain-Specific Model

Finally, we compare the general pipelines with a model trained for a specific domain, on events of that domain. PWM is trained only on gunfight scenes, so we built a dedicated set of 20 gunfight scenes in which one person is shot and collapses on a prescribed schedule (Table 7). The first frame given which is generated by Minimax-H3 (MiniMax-AI, 2026). PWM obtains a BBox IoU of 0.602 and an Interaction Success Rate of 0.850. MiniMax-H3 reaches a BBox IoU of 0.526 but an Interaction Success Rate of only 0.600: models that place the person roughly correctly can still differ clearly in whether the event is executed.

Table 7: **Comparison with a domain-specific model on armed-confrontation scenes.** The 20-scene set includes first frame conditioning and is designed for comparison with PWM, which is trained only on gunfight scenes (third-person; an armed confrontation in which one person is shot and collapses). Metrics use native model outputs; best values are **bold**.

Input: FF = first frame; P = 3D proxy; B = 3D bounding boxes; M = mixed render; D = depth. All methods use text prompts (omitted from Input). CWM = Code World Model.

Method	Trained	Input	Control					Quality		Temporal		Text Align.	Interaction Success ↑	Logic Align. ↑
			Entity		Camera									
			IoU ↑	Center Err. ↓	RotErr (°) ↓	TransErr ↓	CamMC ↓	Imaging ↑	Aesthetic ↑	Warp Err. ($\times 10^{-3}$) ↓	Motion Smooth. ↑	CLIP-Score ↑		
C2R (Gomez-Nogales et al., 2026)	✓	P	0.437	0.041	2.7	1.008	1.043	69.6	0.544	0.301	0.994	24.3	0.450	0.725
CWM (Chen et al., 2026b)	✓	M + FF + D	0.413	0.039	**0.7**	0.740	0.744	72.6	0.562	0.157	0.992	25.6	0.650	0.900
PWM Huang et al. (2026a)	✓	FF + B	**0.602**	**0.021**	1.1	**0.348**	**0.359**	71.0	**0.579**	0.182	0.993	**26.0**	**0.850**	0.875
LynnReal-Omni (Mao et al., 2026b)	✓	FF + B	0.552	0.033	1.3	0.563	0.584	**75.0**	0.527	0.613	0.988	25.2	0.625	**0.975**
MiniMax-H3 (MiniMax-AI, 2026)	✗	FF + B	0.526	0.035	5.1	0.770	0.830	73.9	0.556	0.107	0.995	25.4	0.600	0.900

Table 8: **Evaluated method specifications.**

Method	Conditioning	Supplied spatial input	Output	FPS	Frames
General-purpose world models					
EchoWM	RGB first frame, camera trajectory	1280 × 720 image	1280 × 704	24	121
SANA-WM	RGB first frame, camera trajectory	1280 × 720 image	1280 × 704	16	80
LingBot-World 2.0	RGB first frame, camera trajectory	1280 × 720 image	1280 × 720	16	81
AlayaWorld v1.1	RGB first frame, camera trajectory	1280 × 720 image, resized and cropped to 960 × 544	960 × 544	24	120
Proxy-video-to-RGB models					
Cosmos-Transfer2.5	Engine depth, class-level segmentation	1280 × 720, 24 fps, 121 frames	1280 × 720	24	121
C2R	White proxy video	832 × 480, 16 fps, 81 frames	832 × 480[a]	16	81
CWM	RGB first frame, depth, semantic-ID maps	First frame 1280 × 720; maps 336 × 192	1344 × 768	24	120
PWM	RGB first frame, instance-ID maps from 3D boxes	832 × 480	832 × 480	16	81
LynnReal-Omni	Colored-OBB proxy video; first frame as reference	1280 × 720, encoded at 1344 × 768	1344 × 768	24	120
Video foundation models					
MiniMax-H3	Colored-OBB proxy video; first frame as reference	1344 × 768, 24 fps, 124 frames	1344 × 768	24	120
Seedance 2.5	Colored-OBB proxy video; first frame as reference	1280 × 720, 30 fps, 150 frames	1280 × 720	24	121