

PROWBench: Do Video Models Render What the Program Specifies?

Zheng-Hui Huang^{△,*}, Guixu Lin^{△,*}, Yu-Ju Tsai[△], Jian-Kai Zhu, Fengbo Lan[△], Yu-Lun Liu, Yung-Yu Chuang, Kaipeng Zhang^{△,†}, Zhixiang Wang^{△,†}

[△]Alaya Lab

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 assess visual quality, controllability, and instruction or physical adherence, but rarely test fidelity to fine-grained, program-specified world events. We introduce **PROWBench**, comprising 170 programmatically constructed episodes and 600 proxy videos covering diverse scenes and interactions. PROWBench logs entity states and timestamped events, including those outside the camera’s field of view, as replayable world records, from which it renders synchronized views and proxy representations. This enables generated videos to be checked against the observable consequences of program execution. An extensible framework constructs scenes, controls behaviors, and can render each camera view in different representations, such as coarse 3D, and bounding boxes. The benchmark covers first- and third-person perspectives, with synchronized multi-view observations available for a subset of episodes. Grounded in these records, PROWBench evaluates entity control, long-horizon memory, and, with two VLM-based metrics, **Logic-Render Alignment** and **Interaction Success Rate**, adherence to the prescribed timeline and the visual realization of timestamped engine-recorded events.

Project page: <https://alaya-lab.github.io/PROWBench>

Code: <https://github.com/AlayaLab/PROWBench>

Correspondence: Zhixiang Wang (Project Lead), Kaipeng Zhang

Note: * denotes equal contributions

Date: October 2, 2026



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.

1 Introduction

Recent advances in video world models have enabled increasingly realistic, open-ended 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 [18, 9]: coding agents and executable programs maintain states and interaction rules, while video models generate observations conditioned on structured representations of the world. This 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 [17], controllability and dynamics [11], instruction following and physical adherence [24], and interactive control and consistency [57, 44, 55, 56, 58]. However, as summarized in Table A.1, they rarely test fidelity to fine-grained, program-specified world events: without replayable records of entity states and timestamped events, a generated video can be judged for plausibility but not against what the program executed. None of them renders the same recorded scene in paired proxy representations, and synchronized views are rare, although both are needed to isolate the effects of representation and viewpoint while holding scene content and actions fixed.

To address these gaps, we introduce **PROWBench**, a benchmark for evaluating the visual realization of programmable worlds. Figure 1 provides an overview of our extensible framework and benchmark. PROWBench combines a programmatically constructed dataset with complementary metrics for logic adherence and interaction realization. The data engine is available for extension. It logs entity states and timestamped events, including those outside the camera’s field of view, as replayable world records and renders them into synchronized views and proxy representations, so that generated videos can be checked against the observable consequences of program execution.

PROWBench contains 170 dynamic scene episodes covering diverse environments and interactions. Each episode currently 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. Grounded in these records, PROWBench evaluates entity control and the long-horizon memory of entities that leave and re-enter the view, and introduces two VLM-based metrics: **Logic-Render Alignment** assesses adherence to the prescribed timeline, 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 the inputs produced by this data engine: coarse 3D, semantic-proxy, and colored-OBB renders of the recorded states, together with text prompts obtained by converting the engine’s recorded events. Our experiments assess their logic adherence and interaction realization across viewpoints and representation settings. Surprisingly, with appropriate prompts, existing general-purpose video generation models such as MiniMax-H3, without task-specific training, follow the prescribed entity trajectories and event logic well, and perform on par with camera-conditioned world models in camera following without explicit camera injection. Additional task- or domain-specific training can still improve several aspects, such as entity placement and the realization of domain-specific events. The proxy itself also matters: outputs sometimes follow the input geometry too closely and inherit the coarse shapes of low-poly proxies, while box proxies leave open which described appearance belongs to which entity, so designing a good proxy remains an open problem. Long-video generation remains challenging in terms of both camera following and entity persistence, 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 evaluate entity control and long-horizon memory against the recorded world states, and propose Logic-Render Alignment and Interaction Success Rate to assess whether generated videos follow the prescribed timeline 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 [3] learns controllable environments from unlabeled videos, while GameNGen [47] enables real-time action-conditioned game generation. GameFactory [59], Matrix-Game [61], Yume1.5 [36], and LingBot-World [43] advance generalization, interaction, and long-horizon consistency. Memory-enhanced WorldPlay [45], Matrix-Game 3.0 [52], and AlayaWorld [46] 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 [9] uses a coding agent to maintain executable states and interaction logic, while Programmable World Model [18] feeds persistent structured states to a generative renderer. Other approaches similarly separate state or dynamics from appearance [38, 31, 5]. 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 [12, 48, 25, 53, 35, 20, 26]. Geometry-aware methods incorporate 3D boxes, bird’s-eye-view layouts, world volumes, and structured traffic scenes [14, 27, 54, 34, 50, 51, 49]. Generative rendering uses explicit geometry or renderer-derived signals: Coarse-to-Real [15] and VideoFrom3D [22] use coarse 3D scenes, while Diffusion Renderer [29], AlayaRenderer [19], its real-time extension [30], and LiVER [6] 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 [17] and EvalCrafter [33] assess video quality, temporal consistency, motion, and semantic alignment. WorldSimBench [40], WorldModelBench [24], and WorldScore [11] extend evaluation to physical plausibility, controllability, and world dynamics. MIND [57], WorldMark [56], Omni-WorldBench [55], WorldArena [44], and iWorld-Bench [13] further target interactive behavior, memory, and long-horizon consistency.

WBench [58] 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.

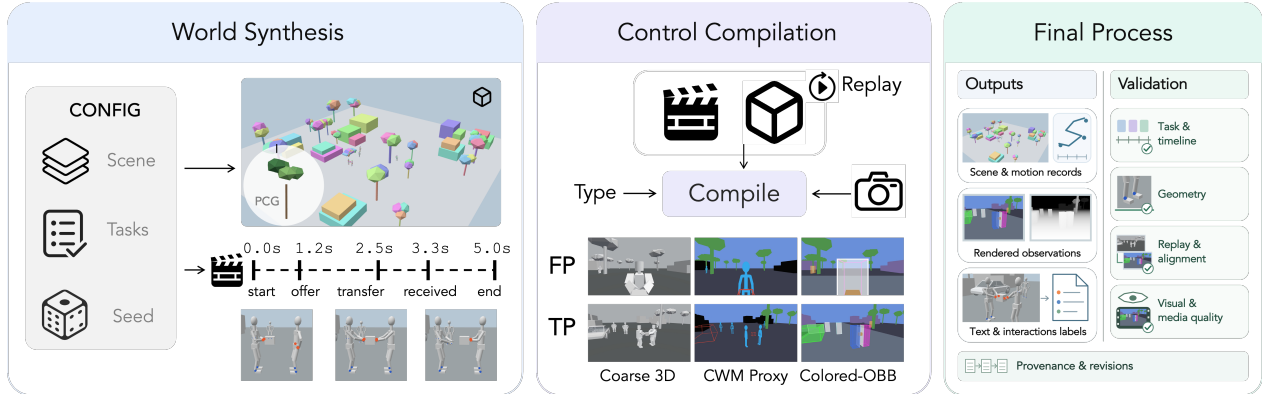


Figure 2 Overview of the world generation pipeline. World synthesis generates replayable trajectories, control compilation produces synchronized conditioning inputs across viewpoints and representations, and the final process validates and exports videos with structured annotations.

3 Benchmark

We construct the PROWBench dataset using a programmable world generation pipeline that produces aligned conditioning inputs and evaluation records. Figure 2 outlines the pipeline, while Figure A.2 illustrates the dataset’s content and diversity. We describe the pipeline and evaluation suite below, with detailed dataset statistics provided in Section A.1.

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.

Specialized 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 to preserve scene shape, object silhouettes, and articulated motion. *Semantic proxy (CWM proxy)* 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, with a semi-transparent white face marking the front of each entity, while retaining geometry for environmental structures and simple objects. OBB bounds are updated over time, while instance colors remain 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 exports three categories of outputs: scene and motion records, rendered observations, and text and interaction labels. Scene and motion records support replay and include scene geometry, entity identities and semantic categories, spatial and articulated states, joint poses, bounding volumes, and entity and camera trajectories with associated camera parameters. Rendered observations comprise synchronized proxy videos across viewpoints and representations, accompanied by metric depth interpreted with respect to its recorded geometry. Text and interaction labels describe action phases, participant relationships, and timestamped events, together with accompanying textual descriptions and contact targets and residuals.

Validation comprises four complementary components. Task and timeline checks assess scenario-specific phases, terminal conditions or state validity, and recording coverage. They also verify task-level interaction labels against the applicable annotation policy and required participant categories. Geometric checks evaluate penetration, support, and contact consistency at saved states using canonical geometry. Replay and alignment checks assess state reconstruction, camera alignment, and consistency with the selected representation. They further establish correspondence among videos, depth, and state labels using source records, frame indices, camera information, and timestamps. Visual and media quality checks combine first-frame visibility measurements based on canonical-geometry pixel counts and view-frustum tests with documented visual inspection of first frames and selected temporal samples. These inspections assess subject visibility and observation readability, while media decoding and hash checks assess output integrity. Visibility measurements and visual reviews are interpreted within the scope of the recorded frames and geometry.

To preserve provenance and revision history, failed attempts and their available diagnostics are retained separately from accepted episodes. For each batch-specific revision, affected outputs are backed up, changes to world records, camera trajectories, or representation policies are documented, and the corresponding media and metadata are updated together. Hash checks verify the integrity of unchanged source records where applicable. Sample metadata and source hashes link released observations to their source records, rendering configurations, and verification reports.

3.2 Evaluation Suite

We evaluate five complementary capabilities: controllability, visual and temporal quality, logic and state 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 evaluate whether entities and the camera follow their prescribed trajectories. For entity control, we consider dynamic and interaction-relevant entities visible in the reference view, excluding terrain, effects, and scene geometry. Generated entities are localized with SAM 3 [7] from semantic-label prompts and matched to engine-projected references by label-compatible Hungarian matching. **BBox IoU** measures spatial alignment over the visible trajectory, with missed detections assigned zero IoU, while **Center Error** measures normalized center displacement conditional on detection. We report clip-level averages over evaluable entities; full visibility, matching, and aggregation details are provided in Appendix B.

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 [41] 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. For a generated scene-camera sample s with event set $\mathcal{E}_s$, we define

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

Table entries average ISR_s equally across samples, preventing samples with more annotated events from receiving greater weight.

We additionally report gISR_s , an IoU-weighted variant of ISR that jointly reflects interaction success and entity placement:

$$\text{gISR}_s = \text{ISR}_s \cdot \text{IoU}_s, \quad (2)$$

where IoU_s is the clip-level BBox IoU of sample s . Table entries average gISR_s equally across samples.

Logic-Render Alignment. Logic-Render Alignment measures whether each segment of the action timeline in the engine-derived prompt is visibly depicted within its time window. 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

$$\text{LRA}_s = \frac{1}{|\mathcal{T}_s|} \sum_{\tau \in \mathcal{T}_s} \mathbf{1}[\text{prescribed action in segment } \tau \text{ is judged present}], \quad (3)$$

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

Similarly, we report gLRA_s , an IoU-weighted variant of LRA that jointly reflects timeline adherence and entity placement:

$$\text{gLRA}_s = \text{LRA}_s \cdot \text{IoU}_s. \quad (4)$$

Table entries average gLRA_s equally across samples.

Figure C.4 illustrates the distinction between the two behavioral metrics. ISR evaluates whether a registered interaction is successfully completed, requiring both the prescribed action and expected end state, whereas LRA evaluates whether the generated video realizes the engine-defined event logic by depicting each prescribed action in its corresponding temporal segment.

Memory over Long Horizons. For the 30-second evaluation, we test whether entities that leave the field of view reappear at the location and in the state prescribed by the engine. We define an *absence event* when

an entity becomes invisible and later re-enters the view. **Reappearance IoU** measures spatial recovery after re-entry by comparing the reference box with label-compatible detections, with missed detections assigned zero IoU. **State Persistence** measures whether a returning person preserves the prescribed locomotion state, classified as *walking* or *standing* and judged from post-return frames by Qwen3.6-27B [41]. Both metrics are averaged first within scenes and then across scenes with valid events.

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

3.3 Benchmark Tracks

PROWBench organizes its evaluation cases into two tracks with four settings. Each setting uses its own evaluation set, and the scenes of the two main-track settings are disjoint.

Main Track. The main track evaluates controlled generation of five-second clips in two settings that differ in the appearance reference and use disjoint scenes. 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. The Verified-FF setting evaluates this case with 40 scene-camera pairs (31 third-person follow, 9 first-person; top of Table 1). Its RGB first frame is part of the benchmark: it is generated with MiniMax-H3 from the colored-OBB proxy and the scene description and checked against the recorded person count, positions, facing and layout (Appendix A.2), and every method that accepts a first frame receives this identical image. In procedural game-world construction, the agent usually has no ready-made RGB appearance reference. The Unverified-FF setting therefore prescribes only the world state, the text, and the control signals, and asks whether models can jointly solve entity control, appearance generation, and identity binding. It contains 90 pairs (78 third-person follow, 12 first-person; bottom of Table 1), and appearance must be determined by each model from the text and the control signal. The curation of this set used MiniMax-H3 outputs, so results involving MiniMax-H3 should be read with this potential bias in mind.

Challenge Track. The challenge track tests whether short-clip control extends to two harder settings. 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. The long-horizon setting (Table 2) contains ten 30-second recordings, one camera each. 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. The multi-view setting (Table 3) contains ten four-player scenes, each with four synchronized third-person follow views, one per player, generated independently.

4 Experiments

Our experiments address a central question: can current video generation and video world models faithfully render a prescribed dynamic world, rather than merely produce visually plausible videos? For game worlds, visual quality alone is insufficient. A successful model must follow prescribed entity trajectories, camera motion, and event timelines, while maintaining consistency over long horizons and across synchronized views. We therefore organize the evaluation into a main track of five-second clips and a challenge track of long-horizon and multi-view settings. Sections 4.3 and 4.4 report the results, and Section 4.5 compares the method classes across both tracks and draws the main takeaways. The experimental setup is detailed in Section 4.2.

4.1 Evaluated Methods

We evaluate 11 models, grouped into three families by the control information they receive. **(a) Camera-conditioned world models.** EchoWM [60], SANA-WM [63], LingBot-World 2.0 [43] and AlayaWorld v1.1 [46] receive an RGB first frame, the text prompt, and the recorded camera trajectory. They obtain explicit camera control but no entity trajectories. **(b) Proxy-video-to-RGB models.** These models translate an engine-rendered control representation into a realistic RGB video. Cosmos-Transfer2.5 [1] receives engine depth and class-level segmentation; C2R [15] receives the coarse 3D proxy and the text; CWM [9] (fine-tuned from MiniMax-H3) receives the CWM proxy, depth, the RGB first frame, and the text; LynnReal-Omni [37],

Table 1 Main track. Results on the Verified-FF setting (top, 40 five-second scene-camera pairs) and the Unverified-FF setting (bottom, 90 pairs); the two settings use different scenes, so best and second-best values are marked within each setting in **bold** and underlined. Input: = first frame checked against the recorded state (Verified-FF), = unchecked first frame, i.e., frame 0 of the MiniMax-H3 output for the same scene and camera (Unverified-FF), = camera pose, = depth, = class-level segmentation, = CWM proxy, = coarse 3D proxy, = colored-OBB proxy. All methods also receive the text prompt. $\uparrow/\downarrow$: higher/lower is better. CE_{err}: normalized box-center error; Rot/Trans: camera rotation (degrees)/translation error; Img/Aes: imaging/aesthetic quality; Warp: warping error ($\times 10^{-3}$); Smooth: motion smoothness; CLIP: CLIP-Score; ISR: Interaction Success Rate; LRA: Logic-Render Alignment; gISR/gLRA: ISR/LRA multiplied by BBox IoU per sample and then averaged, i.e., events realized by correctly placed entities.

Method	Input	Entity control		Logic / State				Camera control			Quality		Temporal		Text
		IoU↑	CErr↓	ISR↑	LRA↑	gISR↑	gLRA↑	Rot↓	Trans↓	CamMC↓	Img↑	Aes↑	Warp↓	Smooth↑	CLIP↑
Verified-FF (40 pairs)															
EchoWM [60]		0.351	0.055	0.506	0.685	0.168	0.241	2.2	0.151	0.166	67.6	0.607	0.619	0.987	25.2
SANA-WM [63]		0.309	0.070	0.590	0.569	0.185	0.180	2.7	0.145	0.168	68.6	0.612	0.871	0.985	25.1
LingBot 2.0 [43]		0.327	0.063	0.733	0.794	0.241	0.256	2.8	0.100	0.133	69.9	0.615	1.026	0.978	25.4
AlayaWorld v1.1 [46]		0.303	0.073	0.583	0.622	0.174	0.185	9.4	0.229	0.356	66.8	0.592	0.530	0.990	25.0
Cosmos 2.5 [1]		0.451	0.061	0.381	0.560	0.165	0.246	1.7	0.166	0.177	67.4	0.552	0.430	0.991	24.3
C2R [15]		0.375	0.064	0.606	0.694	0.218	0.259	7.8	0.352	0.443	70.6	0.600	0.786	0.990	25.7
CWM [9]		0.454	0.040	0.637	0.795	0.290	0.359	2.0	0.130	0.154	66.7	0.629	0.398	0.991	25.3
LynnReal-Omni [37]		0.518	0.036	0.615	0.832	0.324	0.429	2.0	0.074	0.100	73.8	0.625	1.259	0.982	25.2
Seedance 2.5 [4]		0.436	0.043	0.640	0.817	0.277	0.356	3.4	0.098	0.148	66.9	0.628	0.314	0.993	25.3
MiniMax-H3 [39]		0.494	0.038	0.675	0.738	0.334	0.368	3.4	0.080	0.133	70.2	0.634	0.444	0.991	25.2
Unverified-FF (90 pairs)															
EchoWM [60]		0.255	0.099	0.669	0.712	0.172	0.180	3.1	0.165	0.193	70.2	0.629	0.823	0.986	24.8
SANA-WM [63]		0.240	0.075	0.564	0.591	0.145	0.142	4.7	0.228	0.275	69.9	0.636	0.676	0.986	24.5
LingBot 2.0 [43]		0.240	0.083	0.713	0.799	0.180	0.196	4.0	0.104	0.152	71.8	0.634	1.037	0.978	25.0
AlayaWorld v1.1 [46]		0.204	0.101	0.714	0.712	0.159	0.146	10.7	0.324	0.461	69.3	0.602	0.564	0.990	24.8
Cosmos 2.5 [1]		0.376	0.057	0.440	0.457	0.154	0.158	5.3	0.259	0.326	71.9	0.551	0.603	0.990	24.3
C2R [15]		0.315	0.052	0.540	0.625	0.181	0.204	8.8	0.456	0.548	69.9	0.609	0.666	0.990	25.3
CWM [9]		0.356	0.050	0.741	0.784	0.269	0.284	4.7	0.151	0.208	67.5	0.637	0.384	0.990	24.8
LynnReal-Omni [37]		0.457	0.036	0.809	0.852	0.379	0.395	1.4	0.084	0.101	74.8	0.670	1.676	0.976	24.2
Seedance 2.5 [4]		0.336	0.058	0.758	0.829	0.259	0.282	4.5	0.164	0.216	66.4	0.616	0.257	0.993	24.9
MiniMax-H3 [39]		0.409	0.055	0.758	0.806	0.326	0.337	3.6	0.119	0.168	71.6	0.663	0.407	0.990	24.8

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 [18] receives colored-OBB proxy, the RGB first frame, and the text prompt. Because PWM is trained only on gunfight scenes, it is evaluated only on the gunfight set of Section 4.5. **(c) Video foundation models.** MiniMax-H3 [39], Seedance 2.5 [4] 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 Experimental Setup

All evaluation samples come from recorded game-world states and are scored against the engine state; each sample is one scene observed by one camera. Every recorded state is provided in three proxy representations, Coarse 3D, CWM proxy and Colored-OBB, together with engine depth, the camera trajectory, a scene description and the state of every object, and each method receives the control representation of its native interface. 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. The tracks and settings are defined in Section 3.3. 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, and analyses of appearance identity account separately for effects inherited from it.

4.3 Main Track

Table 1 reports ten methods in the two main-track settings.

Table 2 Long-horizon evaluation of controllability and memory. Results on ten 30-second recordings, with one camera per recording and one output per method; best and second-best values per column in **bold** and underlined. Input: = unchecked first frame, i.e., frame 0 of the MiniMax-H3 output for the same scene and camera, = depth, = CWM proxy, = coarse 3D proxy, = colored-OBb. $\uparrow/\downarrow$: higher/lower is better; Rot in degrees. Re-IoU: Reappearance IoU; Re-State: State Persistence of returning entities; other metrics as in Table 1.

Method	Input	Entity control		Logic / State				Memory		Camera control		
		IoU $\uparrow$	CErr $\downarrow$	ISR $\uparrow$	LRA $\uparrow$	gISR $\uparrow$	gLRA $\uparrow$	Re-IoU $\uparrow$	Re-State $\uparrow$	Rot $\downarrow$	Trans $\downarrow$	CamMC $\downarrow$
C2R [15]		0.212	0.066	<u>0.519</u>	0.718	<u>0.127</u>	0.170	<u>0.111</u>	0.405	41.9	0.571	1.119
CWM [9]		0.210	0.070	0.504	<u>0.854</u>	0.111	<u>0.182</u>	0.053	0.886	<u>30.2</u>	0.381	<u>0.793</u>
LynnReal-Omni [37]		0.280	0.054	0.515	0.897	0.149	0.247	0.103	0.528	22.7	<u>0.273</u>	0.610
Seedance 2.5 [4]		0.175	0.077	0.497	0.777	0.087	0.132	0.131	0.650	39.4	0.242	0.819
MiniMax-H3 [39]		0.206	0.088	0.553	0.844	0.121	0.179	0.059	<u>0.833</u>	31.8	0.351	0.814

Entity Control. Proxy-conditioned models lead in both settings. In Verified-FF (Table 1, top), LynnReal-Omni, MiniMax-H3 and CWM reach a BBox IoU of 0.518, 0.494 and 0.454, whereas the four camera-conditioned world models, which receive the camera trajectory but no entity trajectories, reach only 0.303–0.351. In Unverified-FF (Table 1, bottom), LynnReal-Omni and MiniMax-H3 reach 0.457 and 0.409, versus 0.204–0.255 for the camera-conditioned world models.

Logic and State. LingBot 2.0 has the highest Interaction Success (ISR) in Verified-FF (0.733) despite the third-lowest BBox IoU (0.327), whereas Cosmos-Transfer2.5, which receives dense depth and segmentation, ranks fourth and third in BBox IoU but has the lowest Interaction Success (ISR) and Logic-Render Alignment (LRA) in both settings. LynnReal-Omni and Seedance 2.5 reach the highest Logic-Render Alignment (LRA) in both settings (0.832 and 0.817 in Verified-FF; 0.852 and 0.829 in Unverified-FF). Weighting the behavioral scores by entity placement changes the ordering: in Verified-FF, MiniMax-H3 achieves the highest gISR (0.334), followed by LynnReal-Omni (0.324), while LynnReal-Omni leads in gLRA (0.429), followed by MiniMax-H3 (0.368). LingBot 2.0 ranks fifth in gISR (0.241), showing that its leading unweighted interaction score does not translate into the strongest joint score for interaction success and spatial alignment. In Unverified-FF, LynnReal-Omni leads in both gISR (0.379) and gLRA (0.395), followed by MiniMax-H3 (0.326 and 0.337).

Camera Control. The gap in camera control is much smaller: in Verified-FF, 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. LynnReal-Omni has the lowest CamMC in both settings.

4.4 Challenge Track

Tables 2 and 3 report evaluation results in the long-horizon and multi-view settings of the challenge track.

In the long-horizon setting (Table 2), Over the full 30 s, BBox IoU is only 0.175–0.280. The Reappearance IoU, measured on entities that return after leaving the frame, is only 0.053–0.131. 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 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.

In the multi-view setting (Table 3), 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.

4.5 Discussion and Takeaways

We now compare the three method classes of Section 4.1: camera-conditioned world models, task-specific renderers that translate a proxy into RGB video (Cosmos-Transfer2.5, C2R, CWM, LynnReal-Omni and, on

Table 3 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; best and second-best values per column in **bold** and underlined. Input: = unchecked first frame, i.e., frame 0 of the MiniMax-H3 output for the same scene and camera, = depth, = CWM proxy, = coarse 3D proxy, = colored-OBb proxy. All methods also receive the text prompt. $\uparrow/\downarrow$: higher/lower is better; Rot in degrees. Comp./Cons.: appearance compliance/cross-view consistency (%); Miss: miss rate; other metrics as in Table 1.

Method	Input	Entity control			Logic / State				Appearance			Camera control		
		IoU $\uparrow$	CErr $\downarrow$	Miss $\downarrow$	ISR $\uparrow$	LRA $\uparrow$	gISR $\uparrow$	gLRA $\uparrow$	Comp. $\uparrow$	Cons. $\uparrow$	MEt3R $\downarrow$	Rot $\downarrow$	Trans $\downarrow$	CamMC $\downarrow$
C2R [15]		0.414	0.039	0.192	0.637	0.750	0.271	0.318	32.2	25.5	0.397	2.2	0.155	0.173
CWM [9]		0.439	<u>0.037</u>	0.064	<u>0.812</u>	<u>0.950</u>	0.345	0.414	<u>92.9</u>	88.1	0.422	2.0	0.047	0.071
LynnReal-Omni [37]		<u>0.490</u>	0.038	0.047	0.700	0.900	<u>0.347</u>	<u>0.435</u>	87.6	89.7	0.448	0.4	0.023	0.026
Seedance 2.5 [4]		0.341	0.049	0.089	<u>0.812</u>	0.963	0.284	0.331	82.2	81.0	0.499	1.0	0.076	0.083
MiniMax-H3 [39]		0.524	0.029	<u>0.049</u>	0.825	0.938	0.428	0.489	94.0	<u>88.6</u>	<u>0.400</u>	<u>0.9</u>	<u>0.034</u>	<u>0.043</u>

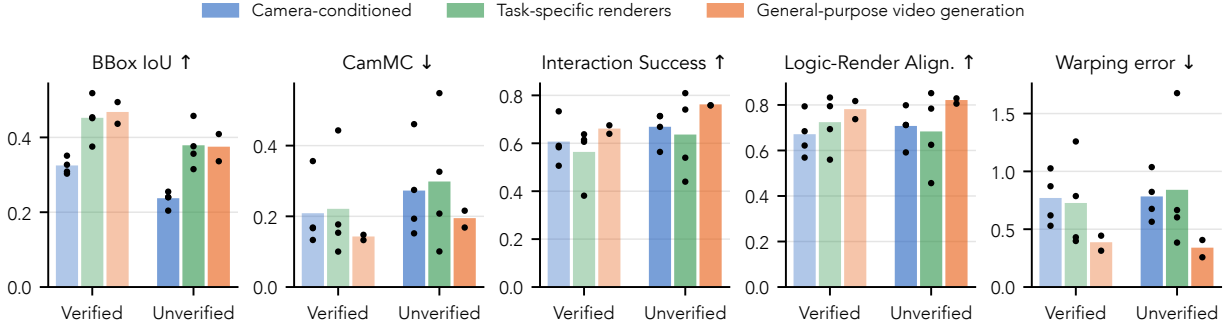


Figure 3 Method classes at a glance. Main-track results per method class in Verified-FF (light) and Unverified-FF (dark). Bars show class means, dots the individual methods. Camera-conditioned: EchoWM, SANA-WM, LingBot 2.0, AlayaWorld v1.1; task-specific renderers: Cosmos-Transfer2.5, C2R, CWM, LynnReal-Omni; general-purpose video generation: MiniMax-H3, Seedance 2.5.

the gunfight set, PWM), and general-purpose video foundation models that receive the proxy video as a reference (MiniMax-H3, Seedance 2.5). Figure 3 summarizes the Main-track results by class; unless stated otherwise, comparisons are paired over scenes with a sign test. Method rankings are closely across the Verified-FF and Unverified-FF settings. The Spearman rank correlations between the per-method means in the two parts of Table 1 are 0.98 for BBox IoU, 0.93 for translation error, 0.95 for CamMC, and 0.91 for Logic-Render Alignment. We organize the following discussion around seven takeaways.

T1 Rendering from a proxy beats camera conditioning in following entity trajectories and event timelines.

Both proxy-conditioned classes place entities far better than the camera-conditioned world model: the class-mean BBox IoU is higher in 38 of 40 Verified-FF scenes and 87 of 89 Unverified-FF scenes (median +0.14, $p < 10^{-8}$), with class means of 0.450 (task-specific) and 0.465 (general-purpose) against 0.322 in Verified-FF, and 0.376 and 0.373 against 0.235 in Unverified-FF. In Figure C.3, the proxy-conditioned models keep both cars in their reference boxes while one of them passes the walking protagonist (per-entity IoU 0.62–0.86); 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). Camera-conditioned world models do control entities implicitly, through the text prompt, the first frame and, for the protagonist, a follow camera that keeps it in view. This implicit control places the protagonist about as well as the proxy (protagonist BBox IoU 0.44 and 0.42 against 0.46–0.54 for the proxy-conditioned classes, over third-person views) but not the other entities (0.26 and 0.20 against 0.38–0.44). Events follow the same pattern when the proxy is paired with a strong video prior. Interaction Success checks whether each engine-recorded event reaches both its prescribed action and end state, and Logic-Render Alignment whether each segment of the prescribed timeline is visibly depicted (Figure C.4). General-purpose models exceed the camera-conditioned world models in Logic-Render Alignment in both settings (26 of 32 and 57 of 77 non-tied scenes, $p < 10^{-3}$) and in Interaction Success in Unverified-FF (42 of 61, $p = 0.004$), and the renderers

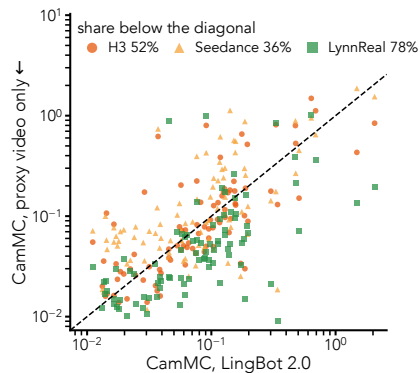


Figure 4 Proxy only vs. camera input. Per-sample CamMC in Unverified-FF of three models without camera trajectory input against LingBot 2.0; points below the diagonal favor the proxy-only model.

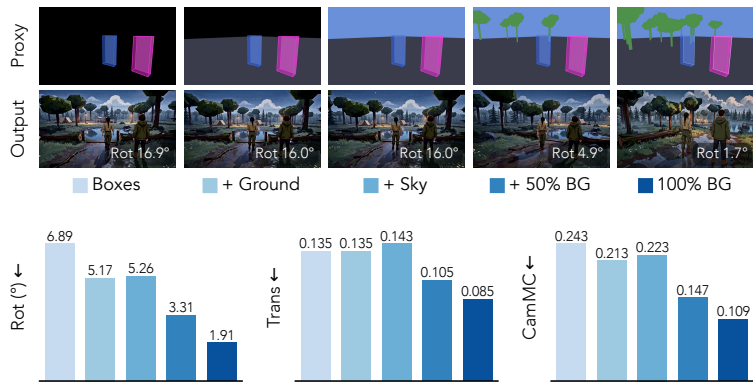


Figure 5 Proxy composition and camera control. LynnReal-Omni on 20 scene-camera pairs, from entity boxes only to adding ground, sky and 50% of the background objects. Top: scene 043 at 4.25s; the overlay gives the mean rotation error of that video. Bottom: means over the 20 pairs (lower is better). The 100% BG reference uses the original prompts and first-frame setting and is not a controlled step.

fine-tuned from MiniMax-H3 are also strong (LynnReal-Omni has the highest Logic-Render Alignment in both settings). The task-specific class as a whole, however, is not better than the camera-conditioned world models (Interaction Success in 15 of 27 and 29 of 66 scenes), because Cosmos-Transfer2.5 has the lowest Interaction Success and Logic-Render Alignment in both settings. LingBot achieves the highest Interaction Success in Verified-FF, but correct actions and end states do not guarantee correct interaction locations, so this result should be interpreted alongside entity control results.

T2 On camera trajectories, the two are on par: the proxy already encodes the camera.

The proxy video is rendered from the prescribed camera, so it carries the camera motion implicitly. General-purpose models, which receive no camera trajectory, tie with EchoWM, SANA-WM and LingBot 2.0 in CamMC: the class mean is lower in 20 of 40 Verified-FF and 45 of 90 Unverified-FF scenes ($p = 1.0$). Per sample, MiniMax-H3 does not differ significantly from EchoWM, SANA-WM and LingBot 2.0 in CamMC in Verified-FF (sign test, $p > 0.2$) and has a lower translation error than EchoWM and SANA-WM in 74% of the Unverified-FF pairs ($p < 10^{-5}$), whereas LingBot 2.0 has lower camera errors than Seedance 2.5 in most pairs of both settings. MiniMax-H3, Seedance 2.5 and LynnReal-Omni all outperform AlayaWorld v1.1 in 87–98% of the pairs in translation error and CamMC. Figure 4 compares each sample with LingBot 2.0, the camera-conditioned model with the lowest CamMC in Unverified-FF: MiniMax-H3 has the lower camera error in 52% of the samples, Seedance 2.5 in 36%, and LynnReal-Omni in 78%.

The composition of the proxy also matters for camera control. Figure 5 examines how proxy components affect camera control across 20 scene-camera pairs. Starting from entity boxes, adding ground reduces CamMC from 0.243 to 0.213, while adding sky provides no further improvement (0.223). Incorporating 50% of the background objects produces larger gains, reducing rotation error from 6.89° to 3.31° , translation error from 0.135 to 0.105, and CamMC from 0.243 to 0.147 relative to boxes-only proxies. The illustrated scene exhibits the same pattern: mean rotation error remains near 16° after adding ground and sky, but falls to 4.89° when background objects are included. These results suggest that explicit background structure supplies useful spatial cues for reproducing prescribed camera motion. The full-background reference achieves still lower errors, with a rotation error of 1.91° and CamMC of 0.109. However, this setting also differs in prompts, first-frame inputs, and some entity-box sets, so its additional gains cannot be attributed solely to increased background coverage.

T3 General-purpose video generation models already render short clips well.

On the short, single-camera clips of the main track, general-purpose models are competitive with task-specific renderers across several metrics. They achieve comparable BBox IoU (23 of 40 and 43 of 89 scenes),

Table 4 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 and second-best values are **bold** and underlined. Input: = first frame, = depth, = CWM proxy, = coarse 3D proxy, = colored-OBB proxy. All methods also receive the text prompt. CWM = Code World Model. Metrics as in Table 1.

Method	Input	Entity control		Logic / State				Camera control			Quality		Temporal		Text
		IoU $\uparrow$	CErr $\downarrow$	ISR $\uparrow$	LRA $\uparrow$	gISR $\uparrow$	gLRA $\uparrow$	Rot $\downarrow$	Trans $\downarrow$	CamMC $\downarrow$	Img $\uparrow$	Aes $\uparrow$	Warp $\downarrow$	Smooth $\uparrow$	CLIP $\uparrow$
C2R [15]		0.437	0.041	0.450	0.725	0.191	0.330	2.7	1.008	1.043	69.6	0.544	0.301	<u>0.994</u>	24.3
CWM [9]		0.413	0.039	<u>0.650</u>	<u>0.900</u>	0.287	0.381	0.7	0.740	0.744	72.6	<u>0.562</u>	<u>0.157</u>	0.992	<u>25.6</u>
PWM [18]		0.602	0.021	0.850	0.875	0.523	<u>0.529</u>	<u>1.1</u>	0.348	0.359	71.0	0.579	0.182	0.993	26.0
LynnReal-Omni [37]		<u>0.552</u>	<u>0.033</u>	0.625	0.975	<u>0.340</u>	0.535	1.3	<u>0.563</u>	<u>0.584</u>	75.0	0.527	0.613	0.988	25.2
MiniMax-H3 [39]		0.526	0.035	0.600	<u>0.900</u>	0.302	0.475	5.1	0.770	0.830	<u>73.9</u>	0.556	0.107	0.995	25.4

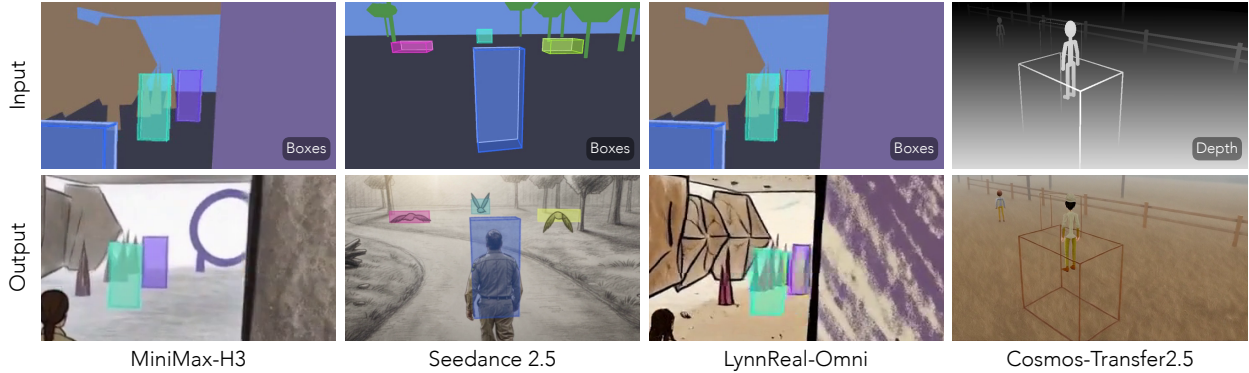


Figure 6 Proxy leakage. The top row shows proxy inputs, and the bottom row shows corresponding generated frames. In these representative failure cases, colored bounding boxes or wireframe outlines from the proxies remain visible in the outputs, revealing unintended leakage of control representations into the rendered appearance.

lower camera errors (CamMC in 30 of 40 and 76 of 90 scenes, $p \leq 0.003$), and higher Interaction Success and Logic-Render Alignment in Unverified-FF (53 of 67 and 61 of 79 scenes, $p \approx 10^{-6}$). They also achieve the lowest warping error, with class means of 0.38 and 0.33, compared with 0.72 and 0.83 for task-specific renderers.

T4 Generic fine-tuning gives mixed gains; domain-specific events still need dedicated training.

Fine-tuning does not, however, yield uniform improvements. CWM and LynnReal-Omni both build on MiniMax-H3, enabling paired comparisons with their general-purpose base. CWM has lower BBox IoU in 71% of the Unverified-FF pairs ($p = 10^{-4}$) and lower imaging quality in approximately 90% of the pairs in both settings, but lower warping error in approximately 80%. LynnReal-Omni achieves higher BBox IoU in 62% of the Unverified-FF pairs ($p = 0.03$) and higher Logic-Render Alignment in 81% of the non-tied Verified-FF pairs ($p = 0.02$), but higher warping error in every pair in both settings. Because lower warping error can also reflect reduced motion, it should not be interpreted as an unqualified improvement in temporal consistency. These comparisons show that the benefits of fine-tuning depend on the model and the control objective.

Despite their strong performance on short clips, general-purpose models can still struggle to execute prescribed events reliably. We illustrate this limitation on a dedicated evaluation set of 20 gunfight scenes in which one person is shot and collapses according to a prescribed schedule (Table 4). We compare the general-purpose pipelines with PWM, a domain-specific model trained on a limited set of scenes. The first frames are generated by MiniMax-H3 [39]. MiniMax-H3 achieves a BBox IoU of 0.526 and an Interaction Success Rate of 0.600, while PWM reaches 0.602 and 0.850, respectively. The gap in Interaction Success Rate highlights a limitation of general-purpose rendering: reasonable spatial alignment does not guarantee reliable execution of a prescribed event. Furthermore, as shown in Figure 6, outputs from general-purpose models sometimes exhibit proxy leakage, whereas such artifacts are rarely observed in PWM’s outputs.

T5 The next step is finding the best proxy.

The proxy decides what the renderer is told, but its form matters little for placement. Table 5 compares three proxies on ten Verified-FF pairs (nine follow, one first-person): the coarse-3D-proxy, CWM-proxy and colored-BBox rows are generated with the same first frames, seed and generation settings. All metrics follow the protocol of Table 1 but cover only this subset. The layout barely changes with the proxy (BBox IoU 0.503–0.514), and camera and quality metrics differ little. The event metrics differ more, but only through a few scenes: the colored-OBB proxy has the highest Interaction Success (0.733 against 0.625–0.683; higher than the coarse 3D proxy in three scenes and lower in none, and higher than the CWM proxy in two scenes and lower in one), and the coarse 3D proxy has the highest Logic-Render Alignment (0.802; higher than the colored-OBB proxy in four scenes and lower in one); with ten pairs, none of these differences is significant (sign test, $p \geq 0.25$). From our visualizations, we observe that, compared with coarse 3D and CWM proxies, colored-OBB proxies generate more natural human motions, with richer joint articulation and more refined strides. Therefore, we use the colored-OBB proxy as the representation input to Minimax-H3 in our evaluation.

However, the proxy leaves a key ambiguity unresolved: which appearance belongs to which entity. A box proxy fixes where each entity is and how it moves, but not what it looks like; which described appearance belongs to which box must be inferred from the text. 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 C.10a (Appendix C.2), 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.

The same ambiguity affects orientation. 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 C.10b, 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. The mixed representation in Figure C.5 (Appendix C.2) likewise omits limbs, whereas the coarse 3D proxy keeps them. A proxy that encodes identity, facing direction and articulation while remaining cheap to render is therefore a direct lever for faithful rendering, and the benchmark can compare candidate proxies under a fixed generator as in Table 5.

Table 5 MiniMax-H3 under different proxy representations on a Verified-FF subset. Input:  = first frame,  = CWM proxy,  = coarse 3D proxy,  = colored-OBB proxy. Metrics as in Table 1.

Method	Input	Entity control		Logic / State				Camera control			Quality		Temporal		Text
		IoU↑	CErr↓	ISR↑	LRA↑	gISR↑	gLRA↑	Rot↓	Trans↓	CamMC↓	Img↑	Aes↑	Warp↓	Smooth↑	CLIP↑
MiniMax-H3 [39]	   	0.514	0.030	0.625	0.802	0.304	0.391	4.8	0.072	0.156	70.1	0.655	0.570	0.987	25.1
MiniMax-H3 [39]	  	0.507	0.031	0.683	0.665	0.338	0.333	4.5	0.079	0.156	70.7	0.650	0.573	0.987	25.1
MiniMax-H3 [39]	  	0.503	0.033	0.733	0.690	0.347	0.339	5.0	0.065	0.155	70.7	0.652	0.574	0.987	25.1

T6 Open problems remain in the challenge track: long horizons, re-entering entities and multi-view identity.

No class solves the challenge track: over 30 s, BBox IoU is 0.175–0.280 and returning entities reach a Reappearance IoU of only 0.053–0.131 (Table 2, Figures 7 and 8). The long-horizon BBox IoU is below the 0.32–0.52 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 7 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. In Figure 8, 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

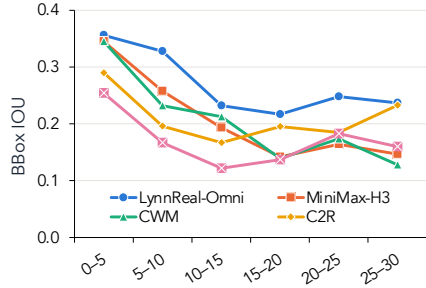


Figure 7 BBox IoU over the 30-s recordings. Averaged in 5-s windows over the ten recordings and all entities with a reference box: every method is best in the first five seconds and loses the prescribed layout as the sequence continues (the visible entities differ between windows, so the curves show a trend).



Figure 8 A returning pedestrian. 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).

does not imply restoring its position after a long absence across windows. Multi-view identity is the second open problem. The four synchronized views of a scene are generated independently, so each player must be re-created from its description in every view. Per-view control remains good (BBox IoU 0.34–0.52, CamMC 0.026–0.173; Table 3), but identity depends on whether the input binds each description to one person. Methods that receive identity-colored boxes (LynnReal-Omni, Seedance 2.5, MiniMax-H3) or a first frame (CWM) reach 82–94% Appearance Compliance and 81–90% Cross-view Consistency, whereas C2R, whose grey coarse 3D proxy draws all players alike and which receives no first frame, reaches 32% and 26%. Even the best methods render a differently dressed player in 10–19% of the judged view pairs, and these scores cover only observations the judge could assess. State Persistence ranks the methods differently from Reappearance IoU (Section 4.4): The two metrics capture different aspects of long-term consistency. Domain-specific events still require dedicated training: on the gunfight set, general models reach an Interaction Success of 0.450–0.650 against 0.850 for PWM. Without a verified first frame, appearance descriptions can be bound to the wrong person (T5). Together, these results suggest developing rendering models on top of strong video foundation models, targeting event fidelity, long-horizon memory and identity binding.

T7 The benchmark reveals what existing metrics miss, but has limitations.

The task metrics capture what placement and perceptual metrics miss. Interaction Success and Logic-Render Alignment are largely independent of entity placement. Within a scene, the Spearman correlation between BBox IoU and Interaction Success across methods averages 0.05 in Verified-FF and 0.06 in Unverified-FF (and 0.14 and 0.12 for Logic-Render Alignment). Interaction Success and Logic-Render Alignment also differ from each other: in Figure C.4, LingBot 2.0 depicts all three prescribed actions (Logic-Render Alignment 1.00) but misses the end state of the first event (Interaction Success 0.67). Logic-Render Alignment responds to the prescription rather than to the video alone: when the judge receives a mismatched timeline, the per-method mean drops from 0.46–0.85 to 0.07–0.20, and the matched timeline scores higher in 68–97% of the samples. Across methods, the mean CLIP-Score varies by only 4.5–5.6% of its average and motion smoothness by 1.5%–1.6%, compared with 53–79% for BBox IoU and 38–59% for Interaction Success and Logic-Render Alignment. In Unverified-FF, LynnReal-Omni has the lowest CLIP-Score (24.2) while leading all entity, camera and logic metrics. The multi-view setting shows the same for geometric metrics: The results reveal a clear gap between geometric agreement and participant consistency across independently generated views. 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. The benchmark also has limitations. Interaction Success, Logic-Render Alignment and the appearance metrics rely on a VLM judge that has not yet been calibrated against human annotation;

camera metrics inherit the error of the pose estimator, which reaches $1.1\text{--}5.4^\circ$ ($1.1\text{--}2.7^\circ$ in Verified-FF) even on the engine renders; the Unverified-FF set was curated with MiniMax-H3 outputs and all first frames are generated by MiniMax-H3, which may favor it; and the challenge-track sets are small (ten recordings or scenes each).

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 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.

References

- [1] A. Ali, J. Bai, M. Bala, Y. Balaji, A. Blakeman, T. Cai, J. Cao, T. Cao, E. Cha, Y.-W. Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.
- [2] M. Asim, C. Wewer, T. Wimmer, B. Schiele, and J. E. Lenssen. Met3r: Measuring multi-view consistency in generated images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6034–6044. IEEE, 2025.
- [3] J. Bruce, M. D. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, Y. Aytar, S. Bechtle, F. Behbahani, S. C. Y. Chan, N. Heess, L. Gonzalez, S. Osindero, S. Ozair, S. Reed, J. Zhang, K. Zolna, J. Clune, N. de Freitas, S. Singh, and T. Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4603–4623. PMLR, 2024.
- [4] ByteDance Seed Team. Seedance 2.5. <https://seed.bytedance.com/en/blog/one-take-creation-flexible-referencing-introducing-seedance-2-5>, July 2026.
- [5] Z. Cai, S. Yang, Y. Wang, Z. Gao, Y. Liu, S. Weng, E. Wu, K. Zhang, and B. Shi. MASS: Multiplayer world models with authoritative shared state, 2026.
- [6] Z. Cai, T. Yang, Z. Chang, S. Li, H. Jiang, S. Weng, and B. Shi. Lighting-grounded video generation with renderer-based agent reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20055–20065, 2026.
- [7] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris Coll-Vinent, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pages 138846–138923, 2026.

- [8] H. Chen, K. Zhou, H. Hua, K. Zhang, J. Qian, W. Ma, H. Chen, C. Liu, Y. Zhao, X. Wang, et al. Memobench: Benchmarking world modeling in dynamically changing environments. *arXiv preprint arXiv:2606.27537*, 2026.
- [9] Y. Chen, G. Lin, and C. Zhang. Code world model: Coding agent as world brain. *arXiv preprint arXiv:2608.25927*, 2026.
- [10] K. Ding, X. Chen, M. Cai, Z. Xu, Y. Wang, Y. Lu, J. Li, S. Chen, Y. Gao, X. Tao, et al. Playworld: Benchmarking world models with agent players over long-horizon objectives. *arXiv preprint arXiv:2608.13552*, 2026.
- [11] H. Duan, H.-X. Yu, S. Chen, L. Fei-Fei, and J. Wu. WorldScore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 27713–27724, Oct. 2025.
- [12] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [13] J. Fang, Y. Lei, Q. Wan, Z. Wang, Y. Huang, Y. Xu, B. Zhao, W. Zhang, C. Gao, X. Chen, and Y. Li. iworld-bench: A benchmark for interactive world models with a unified action generation framework, 2026.
- [14] R. Gao, K. Chen, E. Xie, L. Hong, Z. Li, D.-Y. Yeung, and Q. Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *International Conference on Learning Representations (ICLR)*, 2024.
- [15] G. Gomez-Nogales, Y. Hong, C. Ge, P. Zhuang, M. Comino-Trinidad, D. Casas, and Y. Zhou. Coarse-to-real: Generative rendering for populated dynamic scenes. *arXiv preprint arXiv:2601.22301*, 2026.
- [16] J. Huang, Q. Zhou, H. Rabeti, A. Korovko, H. Ling, X. Ren, T. Shen, J. Gao, D. Slepichev, C.-H. Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.
- [17] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, June 2024.
- [18] Z.-H. Huang, G. Lin, J. Lin, Y.-C. Huang, R. Yu, M. Niu, S. Yang, Y.-L. Liu, Y.-Y. Chuang, K. Zhang, et al. Programmable world model. *arXiv preprint arXiv:2609.10540*, 2026.
- [19] Z.-H. Huang, Z. Wang, J. Tan, R. Yu, Y. Zhang, B. Zheng, Y.-L. Liu, Y.-Y. Chuang, and K. Zhang. Generative world renderer. *arXiv preprint arXiv:2604.02329*, 2026.
- [20] Y. Jain, A. Nasery, V. Vineet, and H. Behl. PEEKABOO: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8079–8088, 2024.
- [21] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [22] G. Kim, J. Han, and S. Cho. Videofrom3d: 3d scene video generation via complementary image and video diffusion models. In *SIGGRAPH Asia 2025 Conference Papers*, pages 110:1–110:11. ACM, 2025.
- [23] W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.-H. Yang. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision*, 2018.
- [24] D. Li, Y. Fang, Y. Chen, S. Yang, S. Cao, J. Wong, M. Luo, X. Wang, H. Yin, J. E. Gonzalez, I. Stoica, S. Han, and Y. 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, editors, *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

- [25] P. Li, K. Chen, Z. Liu, R. Gao, L. Hong, D.-Y. Yeung, H. Lu, and X. Jia. Trackdiffusion: Tracklet-conditioned video generation via diffusion models. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 3539–3548, 2025.
- [26] Q. Li, Z. Xing, R. Wang, H. Zhang, Q. Dai, and Z. Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12112–12123, 2025.
- [27] X. Li, Y. Zhang, and X. Ye. Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In *Computer Vision – ECCV 2024*, pages 469–485, 2024.
- [28] Z. Li, Z.-L. Zhu, L.-H. Han, Q. Hou, C.-L. Guo, and M.-M. Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023.
- [29] R. Liang, Z. Gojcic, H. Ling, J. Munkberg, J. Hasselgren, C.-H. Lin, J. Gao, A. Keller, N. Vijaykumar, S. Fidler, and Z. 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)*, pages 26069–26080, 2025.
- [30] G. Lin, Z.-H. Huang, S. Yang, M.-H. Yang, K. Zhang, and Z. Wang. Generative world renderer at the speed of play. *arXiv preprint arXiv:2607.18703*, 2026.
- [31] Z. Lin, Z. Wang, C. Tan, B. Wen, and Y. Jin. Stateplay: State-aware game world models for mechanics-consistent generation, 2026.
- [32] X. Liu, H. Wang, R. Li, D. Yu, R. Wan, R. Zhang, S. Tao, X. Yang, S. Zhang, Z. Zhang, J. Zhang, and S. Ma. Triworldbench: A tri-view consistency perspective on embodied world models, 2026.
- [33] Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. H. Chan, and Y. Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22139–22149, 2024.
- [34] J. Lu, Z. Huang, Z. Yang, J. Zhang, and L. Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *Computer Vision – ECCV 2024*, pages 329–345, 2024.
- [35] G. Y. Luo, Z. H. Luo, A. Gosselin, A. Jolicœur-Martineau, and C. Pal. Ctrl-v: Higher fidelity autonomous vehicle video generation with bounding-box controlled object motion. *Transactions on Machine Learning Research*, 2025.
- [36] X. Mao, Z. Li, C. Li, X. Xu, K. Ying, and K. Zhang. Yume1.5: A text-controlled interactive world generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7752–7761, 2026.
- [37] X. Mao, P. Lin, S. Rui, Y. Zhang, H. Wan, and W. Ma. Lynnreal-omni: Native multi-modal video generation for agentic visual workflows, 2026.
- [38] Z. Meng, Z. Li, C. Li, Q. Li, and K. Zhang. Marionette: Predicting world states, rendering geometry, painting appearance, 2026.
- [39] MiniMax-AI. Minimax h3. Official GitHub repository, Aug. 2026. Open-sourced on August 3, 2026.
- [40] Y. Qin, Z. Shi, J. Yu, X. Wang, E. Zhou, L. Li, Z. Yin, X. Liu, L. Sheng, J. Shao, L. Bai, and R. 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*, pages 50338–50362. PMLR, 2025.
- [41] Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026.
- [42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language

- supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763, 2021.
- [43] Robbyant Team, Z. Gao, Q. Wang, Y. Zeng, J. Zhu, K. L. Cheng, Y. Li, H. Wang, Y. Xu, S. Ma, Y. Chen, J. Liu, Y. Cheng, Y. Yao, J. Zhu, Y. Meng, K. Zheng, Q. Bai, J. Chen, Z. Shen, Y. Yu, X. Zhu, Y. Shen, and H. Ouyang. Advancing open-source world models. *CoRR*, abs/2601.20540, 2026.
 - [44] Y. Shang, Z. Li, Y. Ma, W. Su, X. Jin, Z. Wang, L. Jin, X. Zhang, Y. Tang, H. Su, C. Gao, W. Wu, X. Liu, D. Shah, Z. Zhang, Z. Chen, J. Zhu, Y. Tian, T.-S. Chua, X. Wang, W. Zhu, and Y. Li. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *AI Open*, 7:208–226, 2026.
 - [45] W. Sun, H. Zhang, H. Wang, J. Wu, Z. Wang, Z. Wang, Y. Wang, J. Zhang, T. Wang, and C. Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *CoRR*, abs/2512.14614, 2025.
 - [46] A. Team, K. Zhang, C. Li, Y. Zhan, Y. Ge, Y. Yin, J. Tan, K. He, L. Fan, M. Zhai, et al. Alayaworld: Interactive long-horizon world modeling—full technical report. *arXiv preprint arXiv:2607.18367*, 2026.
 - [47] D. Valevski, Y. Leviathan, M. Arar, and S. Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, 2025.
 - [48] J. Wang, Y. Zhang, J. Zou, Y. Zeng, G. Wei, L. Yuan, and H. 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*, pages 52274–52289. PMLR, 2024.
 - [49] Q. Wang, Y. Luo, X. Shi, X. Jia, H. Lu, T. Xue, X. Wang, P. Wan, D. Zhang, and K. Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *ACM SIGGRAPH 2025 Conference Papers*, pages 60:1–60:10. ACM, 2025.
 - [50] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *Computer Vision – ECCV 2024*, volume 15106, pages 55–72. Springer, 2024.
 - [51] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. 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)*, pages 14749–14759, 2024.
 - [52] Z. Wang, Z. Liu, J. Li, K. Huang, B. Xu, F. Kang, M. An, P. Wang, B. Jiang, Y. Wei, Y. Xietian, J. Pei, L. Hu, B. Jiang, H. Xue, Z. Wang, H. Sun, W. Li, W. Ouyang, X. He, Y. Liu, Y. Li, and Y. Zhou. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *CoRR*, abs/2604.08995, 2026.
 - [53] Z. Wang, Z. Yuan, X. Wang, Y. Li, T. Chen, M. Xia, P. Luo, and Y. Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11. ACM, 2024.
 - [54] Y. Wen, Y. Zhao, Y. Liu, F. Jia, Y. Wang, C. Luo, C. Zhang, T. Wang, X. Sun, and X. Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6902–6912, 2024.
 - [55] M. Wu, Z. Cai, F. Zhao, X. Feng, R. Dang, B. Song, R. Tian, J. Zhu, J. Lei, H. Dou, J. Tang, L. Sun, J. Wu, X. Chu, Z. Liu, and K. Huang. Omni-WorldBench: Towards a comprehensive interaction-centric evaluation for world models. *arXiv preprint arXiv:2603.22212*, 2026.
 - [56] X. Xu, Z. Lin, K. He, Y. Feng, X. Mao, Y. Yin, K. Zhang, and Y. Ge. WorldMark: A unified benchmark suite for interactive video world models. *arXiv preprint arXiv:2604.21686*, 2026. Version 1.
 - [57] Y. Ye, X. Lu, Y. Jiang, Y. Gu, R. Zhao, Q. Liang, J. Pan, F. Zhang, W. Wu, and A. J. Wang. MIND: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026.

- [58] K. Ying, H. Hu, S. Ren, J. Li, F. Chen, Z. Wang, X. Cao, X. Cai, and H. Ding. WBench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026.
- [59] J. Yu, Y. Qin, X. Wang, P. Wan, D. Zhang, and X. Liu. Gamefactory: Creating new games with generative interactive videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11590–11599, 2025.
- [60] S. Zhang, Y. Li, J. Zhuang, W. Jin, H. Wang, X. Lu, Y. Sun, S. Zhang, H. Li, X. Ma, Y. Li, Y. Liu, Y. Su, Y. Ma, H. Wu, Z. Su, Y. Ma, L. Zhang, H. Huang, Z. Xue, A. Rao, and N. Duan. Echowm: Open and enterable omnimodal world models. *arXiv preprint arXiv:2608.23189*, 2026.
- [61] Y. Zhang, C. Peng, B. Wang, P. Wang, Q. Zhu, F. Kang, B. Jiang, Z. Gao, E. Li, Y. Liu, and Y. Zhou. Matrix-game: Interactive world foundation model. *CoRR*, abs/2506.18701, 2025.
- [62] G. Zheng, T. Li, R. Jiang, Y. Lu, T. Wu, and X. Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.
- [63] H. Zhu, H. Liu, Y. Zhao, T. Ye, J. Chen, J. Yu, T. He, S. Han, and E. Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026.

Appendix

A Benchmark Data

This section reports dataset statistics and describes how the first frames and text prompts of the benchmark are constructed.

A.1 Data Statistics

Figure A.1 shows all 170 cases. 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: coarse 3D, CWM proxy, and Colored-OB. 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 Verified-FF and 90-case Unverified-FF settings of the main track. 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.

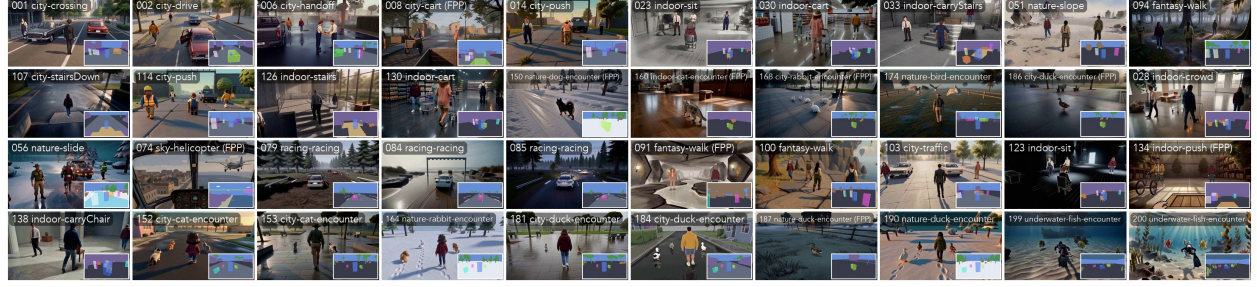
The event-driven collection, Gunfight, contains 20 cases derived from ten source recordings. Each case provides one third-person view and three proxy videos, contributing 20 views and 60 videos. These cases vary camera and actor motion around a recorded actor-death event. Episode IDs therefore identify 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 4.4 and are included in the dataset totals above.

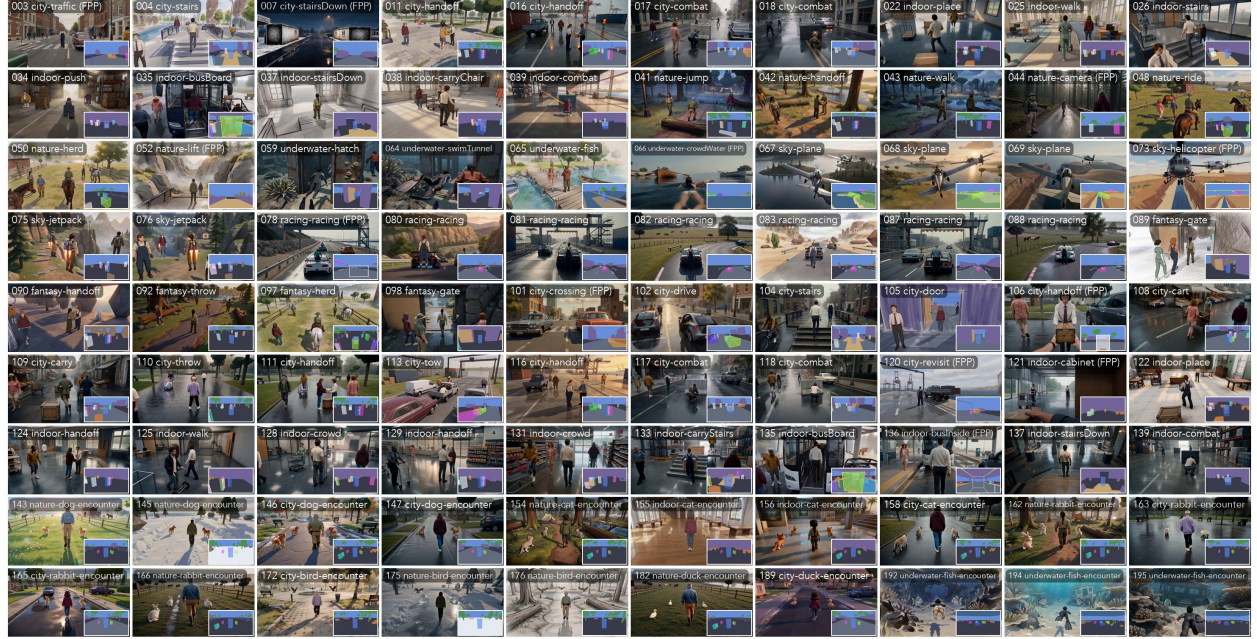
Scene-domain labels combine 130 native annotations with 40 source-reviewed annotations. The resulting counts are city (64), indoor (38), nature (31), racing (12), underwater (10), fantasy (8), and sky (7); city includes airport aprons and industrial outdoor settings. We review interaction labels across all 170 episodes 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 Gunfight, 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 Gunfight 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

Main track, Verified-FF (40 pairs)



Main track, Unverified-FF (90 pairs)



Challenge track, long-horizon (10 recordings, 30 s; frame at 15 s)



Challenge track, multi-view (10 four-player scenes; view A shown)



Analysis, domain-specific gunfight set (20 scenes)



Figure A.1 Gallery of all 170 PROWBench cases. Each tile shows the MiniMax-H3 output at the middle of the clip (at 15 s for the 30-second long-horizon recordings); the inset shows the colored-OBB proxy input at the same instant. MiniMax-H3 is used only because it has outputs for every collection; its frames are generated results, not ground truth. Cases are grouped by collection: the Verified-FF (40) and Unverified-FF (90) settings of the main track, the long-horizon (10) and multi-view (10, view A shown) settings of the challenge track, and the domain-specific gunfight set (20). FPP marks first-person views.

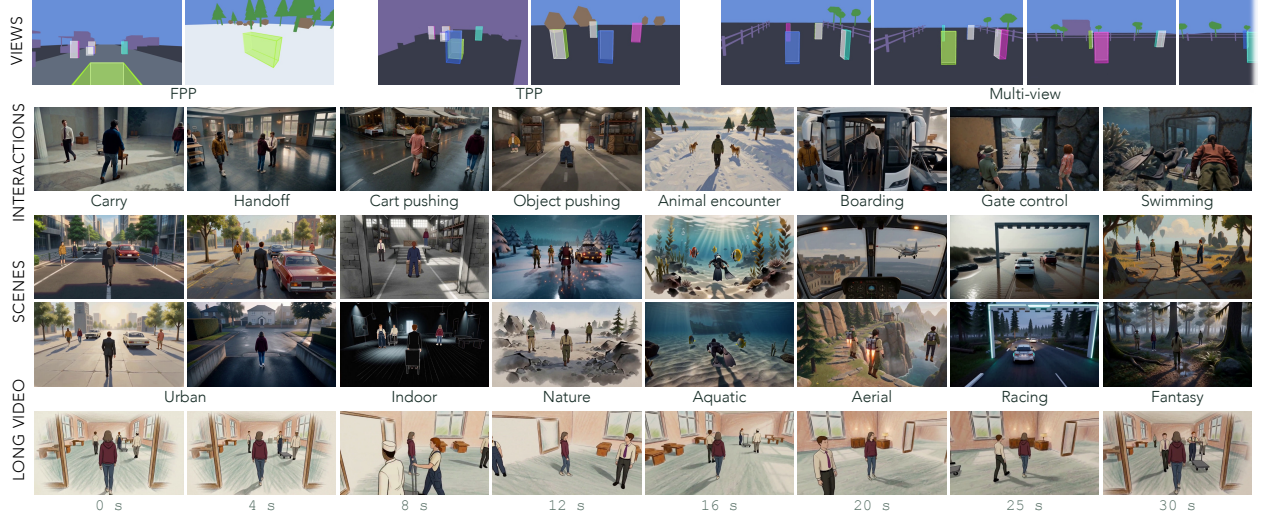


Figure A.2 Diversity of PROWBench. PROWBench provides diverse camera perspectives, interactions, scene domains, and long-horizon videos.

orientation is 6.60° . Figure A.3 summarizes the interaction, style, scene-domain, camera-motion, and perspective distributions.

A.2 First-Frame Construction

The Verified-FF setting (Section 4.3) models the case in which an agent already has access to a concept image or character design. To provide all first-frame-conditioned methods with a common appearance reference while remaining consistent with the prescribed world state, we construct one RGB first frame for each selected view from its colored-OBV proxy and scene description. Because this image becomes part of the benchmark input, it must satisfy both appearance and geometric constraints, including the recorded person count, spatial layout, subject positions, and facing directions. We evaluated several image-generation models and selected MiniMax-H3 because it showed the strongest alignment with the proxy geometry, particularly in person count and layout. The final Verified-FF set contains 40 first frames (17 realistic and 23 stylized), all of which pass the automatic verification described below. Every method that accepts a first frame receives the same RGB image.

Generation. Before generation, the appearance of every entity is fixed in text: each person, vehicle, and animal is assigned a concrete appearance from a reference library matched to the scene and target style, preventing the generator from implicitly choosing identities or appearances. The generation prompt combines three components: (i) a frame-0 layout contract specifying the screen position, extent, and facing direction of every visible entity according to the engine record; (ii) the proxy-control block used by the text prompts (Appendix A.3), which binds geometry to the proxy and appearance to text; and (iii) the scene appearance and style description. MiniMax-H3 receives the first 22 frames of the colored-OBV proxy together with this prompt, without an RGB image reference, and generates a short video. We use frame 0 of the generated clip, at 1280×720 , as the candidate first frame. For each view-specific instance, we generate three candidates using different random seeds.

Automatic Verification. Each candidate is subjected to hard checks against the frame-0 engine record:

- *Person count:* the number of persons detected by SAM 3 [7], using the same detector as in evaluation, must equal the number of visible persons recorded by the engine.
- *Box alignment:* the mean IoU between detected person boxes and their corresponding engine boxes must be at least 0.30; each visible animal must individually achieve an IoU of at least 0.30.
- *Image defects:* Qwen3.6-27B [41] is given the candidate image together with the scene’s style, world, and character descriptions, and must verify that the target style is respected and that the image contains none of the following artifacts: copied proxy geometry (e.g., wireframes, flat-colored mannequins, or

Table A.1 Benchmark protocols and their reference information. Behavior specification summarizes motion and interaction conditions; initial images and video context are omitted for compactness. State/trajectory references and task/event references are distinguished from conditions supplied to a model. **Replay** requires shared geometry and saved dynamic states sufficient to re-render an episode without rerunning its behavior controllers. **Views** denotes synchronized videos of the same world evolution; paired first frames or sequential perspective switches do not qualify. **Proxies** denotes alternative visual representations derived from the same states and cameras. ✓: documented support; −: not documented in the compared protocol or not applicable; Sim.: simulator-based evaluation, without confirmed recorded-state re-rendering; Subset: selected episodes. PROWBench combines multi-entity state and event records with replay and paired observations to evaluate visual realization under fixed world evolution.

Benchmark	Behavior specification	State / trajectory reference	Task / event reference	Replay	Views	Proxies
VBench [17]	Text prompts	−	Prompt semantics	−	−	−
WorldScore [11]	Text; camera paths	Target camera paths	Text-specified dynamics	−	−	−
WorldModelBench [24]	Text instructions	−	Instructions; physics criteria	−	−	−
MIND [57]	Agent navigation; camera rotation	Actor and camera poses	−	−	−	−
WorldArena [44]	Text; robot actions	Robot states and actions	Task goals; simulator outcomes	Sim.	−	−
Omni-WorldBench [55]	Text interactions; optional camera paths	Camera targets where provided	Specified object effects and event order	−	−	−
MemoBench [8]	Text; optional camera paths	Camera poses; synthetic depth	Reference-video object states	−	−	−
WorldMark [56]	Navigation via native adapters	Specified navigation controls	−	−	−	−
WBench [58]	Text interactions; navigation controls	Synthetic navigation paths	Per-turn action and event checks	−	−	−
PlayWorld [10]	Text goals; adaptive navigation	−	Goal-specific VQA rubrics	−	−	−
TriWorldBench [32]	Text instructions; robot actions	Reference robot trajectories	Derived action phases; task checks	−	✓	−
PROWBench	Recorded evolution via proxies and text	Multi-entity states; camera poses	Timestamped engine events; timeline	✓	Subset	✓

Scope notes. MIND records navigation inputs and actor/camera poses, without structured interaction-event records in its published schema. Omni-WorldBench, WBench, and PlayWorld do evaluate interactions or events; the distinction here is the provenance of their references, not the presence of interaction evaluation. TriWorldBench derives coarse action phases from reference robot trajectories; these are not verified task-success labels. Engine event records specify the authored execution, not independently verified physical causality.

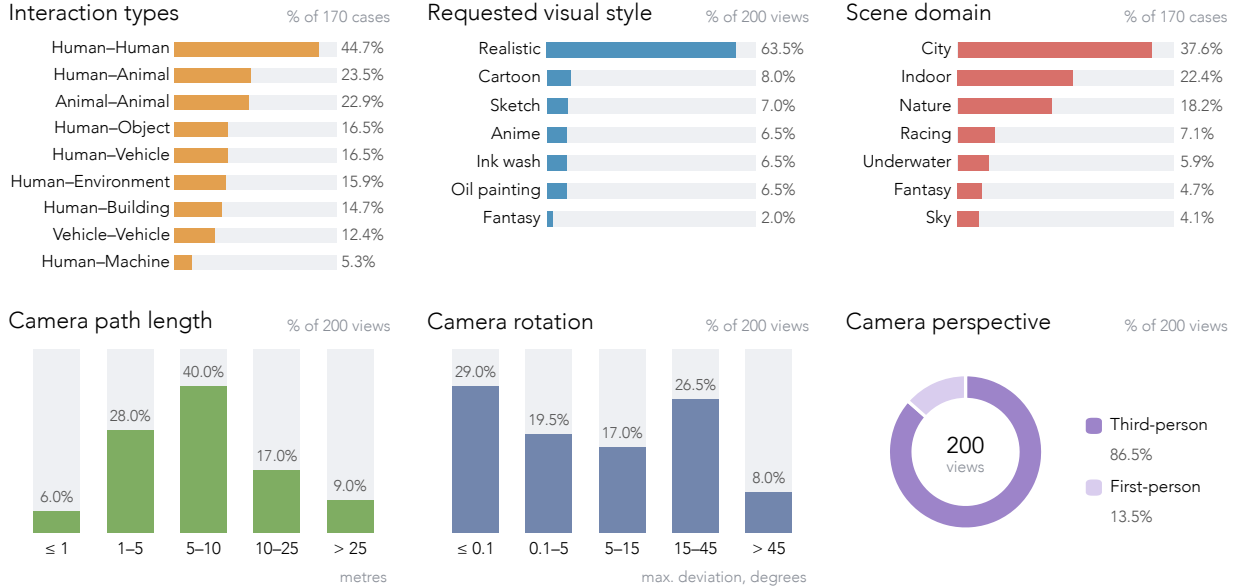


Figure A.3 Dataset composition of PROWBench. 170 evaluation cases, 200 selected camera views, and 600 proxy videos. The panels show interaction categories, requested visual styles, scene domains, camera path length, maximum angular deviation from the initial camera orientation, and camera perspectives; each panel states its denominator (170 episodes or 200 selected views). Interaction labels follow a unified task-level annotation policy and may overlap within a case. 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.

boxes), large unlit black surfaces inherited from the proxy, structural defects such as detached doors, or placeholder geometry such as disc-shaped tree crowns or box-shaped props.

- *Heading*: for vehicles and mounts with an unambiguous front direction, the visible heading (toward, away from, left of, or right of the camera) must agree with the orientation of the corresponding engine box.

Candidates that fail any check are regenerated with a new seed and additional constraint sentences targeted to the failed criteria, e.g., explicitly requiring every proxy primitive to be replaced by a fully rendered object. Across regeneration rounds, we retain the best candidate that satisfies all hard checks.

A.3 Text Prompts

Each method receives a text prompt for every rendered view of a scene. The underlying engine provides structured metadata for each view, including scene attributes, entity annotations, temporal events, and view-dependent observations. We first construct a unified base prompt from this metadata. Except for the appearance description, all prompt content is generated deterministically from the engine records; the appearance block is the only component produced by a vision-language model. The base prompt is then converted, again deterministically, into the input format required by each method. As a result, prompt construction is reproducible for a fixed set of engine records, and differences across methods arise from their input interfaces and proxy representations. We describe below the source fields, base-prompt construction, the system prompt, representation-specific wording, and the differences between the two main-track settings and across methods. Prompts C.1 and C.2 at the end of this section show the base prompt and a converted prompt for one view.

Source Fields. For each rendered view, the base prompt is constructed from the following records:

- *Scene description*: an authored textual description consisting of a target rendering-style line, an environment paragraph, and a protagonist paragraph. The style line is copied verbatim into every prompt. When no first frame is provided, the environment and protagonist descriptions also serve as the source of appearance information.

- *Entity annotations*: all dynamic entities and interaction objects, together with their engine identifiers, semantic classes, and role descriptions (e.g., “the receiving companion”). For each entity, we record whether it is visible at frame 0, based on the engine visibility flags and its projected bounding box. If an entity enters the view later, its first entry time is also recorded. In the colored-OBB representation, each entity is additionally assigned a unique identity color.
- *Action timeline*: phase boundaries are recomputed from the recorded frames, while retaining the engine-authored sentence associated with each phase. We augment these descriptions with measured facts derived from the recorded trajectories, including the path length of each moving entity (e.g., “moves about 10 metres along its route”), inter-entity relations, final resting locations of moving objects, and occlusion events. These descriptions do not use camera-relative direction terms.
- *Appearance source*: in Verified-FF, appearance is derived from the verified first frame; in Unverified-FF, it is taken from the authored scene description.

Camera motion is deliberately excluded from the text prompt and is provided only through the method’s control input, i.e., the proxy video or camera trajectory.

Base-Prompt Construction. Except for the appearance block, the base prompt is generated entirely by a deterministic script, so identical engine records yield identical text. The prompt contains five components. (1) A fixed *control block* specifies the roles of the two references: the proxy video provides the exact position, extent, orientation, and timing of every subject, but never its target appearance; the RGB first frame, when available, determines subject identity, appearance, and rendering style, but does not prescribe subsequent motion. The control block also defines the precedence of information sources in case of disagreement: geometry, camera, and motion follow the proxy; style and appearance follow the first frame; articulated actions follow the timestamped text. (2) *Subject definitions* assign each subject a unique name and explicitly bind it to its proxy representation. (3) A one-sentence *task summary* states the generation objective and shot duration. (4) A *retention analysis* specifies, for each subject, which properties must be preserved from each reference. (5) A *detailed description* specifies the rendering style, environment, appearance of every subject, and a contiguous action timeline covering the full five-second clip. The timeline is expressed as consecutive intervals from 00:00.000 to 00:05.000, for example: “From 00:00.000 to 00:04.667, the protagonist walks across the road while the approaching vehicles brake and yield; the protagonist moves about 10 metres along its route; ...”.

The appearance block is produced by Qwen3.6-27B [41] with greedy decoding. Its input consists of two images: the RGB first frame and frame 0 of the CWM proxy render. The latter is used solely as a layout reference for matching people and objects in the first frame to engine entities; it is not treated as an appearance reference. The model outputs a concise description of the opening frame and environment, the appearance and initial pose of each entity, and any discrepancy between the first frame and the recorded layout. To prevent the proxy representation from being mistaken for target appearance, proxy-specific terms such as mannequin, wireframe, and bounding box are disallowed.

Every composed prompt is checked by an automatic validator. The validator requires that (i) every subject is defined, preserved, and described; (ii) only the allowed references are mentioned; (iii) the timeline is contiguous and covers exactly 0–5 s; (iv) prompts without a first frame contain no first-frame references; and (v) the required representation-specific wording is present. If validation fails, the validation errors are returned to the VLM for one regeneration attempt.

System Prompt. The control block plays the role of a system prompt, but how it reaches each model depends on the model’s text interface. CWM is the only method whose interface has a separate system role. Its text encoder applies the Qwen chat template with two turns, a fixed system prompt followed by a user turn containing the <Picture 1> and <Video 1> references and the caption, and encodes the formatted sequence in a single pass. The release ships two such system prompts, one for the first generation window and one for continuation windows. Our control block is adapted from this system prompt, with three changes for the box proxy: the proxy gives each subject’s exact extent and must be followed precisely (CWM’s prompt treats its proxy as a coarse reference), the white face of a box marks the entity’s front, and the hexadecimal colour tags are identity codes. MiniMax-H3 and LynnReal-Omni use no chat template and no system role: their text encoder receives one token sequence consisting of a label and the visual tokens of each reference

(<Picture 1> for the first frame, <Video 1> for the proxy video), followed by the prompt verbatim, so the control block enters as the first part of the prompt (Prompt C.1). The Seedance 2.5 API likewise accepts a single prompt. In Unverified-FF, Seedance 2.5 receives the base prompt with the control block at its start, as MiniMax-H3 does. In Verified-FF, it receives a deterministic conversion of the same base prompt into its reference syntax, as C2R and CWM receive conversions into theirs: the conversion keeps the style line, the appearance of every subject, the binding of each subject to its box by colour name and legend colour, the proxy semantics, and the timestamped timeline, but states the proxy semantics as ordinary sentences instead of a control block (Table A.3). C2R, Cosmos-Transfer2.5, and the camera-conditioned world models receive no system prompt, and their converted prompts contain no proxy terminology (Prompt C.2).

Representation-Specific Wording. Only two parts of the base prompt depend on the proxy representation: the sentence defining the semantics of the proxy video, and the rule used to bind each subject to its proxy.

- *Coarse 3D proxy:* the full scene is represented as an untextured light-grey low-poly proxy. People are articulated grey block figures whose recorded limb motion must be preserved. All grey surfaces encode geometry only and must not be interpreted as target colors or materials.
- *CWM proxy:* people are represented as flat blue upright mannequins with a fixed pose; they translate and rotate but carry no limb-motion information. Interaction objects are shown as red boxes. The environment is rendered using flat semantic colors that act only as proxy codes and do not correspond to target materials or colors.
- *Colored-OBb:* people, vehicles, animals, and interaction objects are represented as semi-transparent solid boxes, with one unique identity color per entity. The color serves only to identify which box corresponds to which subject and must not be interpreted as a target color or material. The white face marks the front of the entity; hence, if the white face is not visible, the entity is facing away from the camera. The white face itself is likewise only a proxy marker and must not appear in the generated result.

For colored-OBbs, each subject definition explicitly identifies its proxy using the hexadecimal legend color. For example: “<Subject 3> is the passenger car: in <Picture 1> a classic red sedan with a white license plate, parked on the right side of the street facing away from the camera; in <Video 1> the object proxy tagged #f79e36, whose white face is its front.” Subjects that are not visible at frame 0 are additionally annotated with their first entry time. In the main track, the methods that receive the box proxy (MiniMax-H3, LynnReal-Omni and Seedance 2.5) use the colored-OBb prompt; C2R, which receives the coarse 3D proxy, and CWM, which receives the CWM proxy, use converted prompts without proxy wording (Table A.3). The coarse 3D and CWM proxy base prompts are used for MiniMax-H3 only in the representation study (Table 5).

Verified-FF versus Unverified-FF. Table A.2 summarizes the two main-track settings. They share the same structural skeleton, including subject definitions, proxy bindings, the action timeline, and all measured facts. They differ only in the source of appearance information and in the role of the first frame.

Table A.2 Base prompt in the two main-track settings.

	Verified-FF	Unverified-FF
References	<Picture 1> (first frame) and <Video 1> (proxy)	<Video 1> only
Task	keyframe completion and reference generation; the generated clip must start exactly from the provided first frame	reference generation
Appearance source	Qwen3.6-27B description of the verified first frame	authored scene description
Style source	authored style line; the first frame also visually exhibits this style	authored style line
Control rule	for overall style and the appearance of subjects visible in the first frame, the first frame takes precedence over text	the appearance of every subject is specified by text
Structural skeleton	identical: subject definitions, proxy bindings, action timeline, and measured facts	

Prompt Formats across Methods. The base prompt is designed for models that can consume explicitly tagged image and video references. For methods with different interfaces, we apply deterministic transformations that modify the format while preserving the underlying semantic content to the extent permitted by the target interface (Table A.3). No transformation relies on a language model. In Unverified-FF, the appearance

description used by CWM is derived from frame 0 of the MiniMax-H3 output, which also serves as CWM’s RGB anchor frame.

Table A.3 Prompt format by method.

Methods	Prompt format
MiniMax-H3, LynnReal-Omni	base prompt verbatim, including the control block, tagged references, subject bindings with legend colors, and the complete timestamped timeline
Seedance 2.5	Verified-FF: converted to Seedance’s reference syntax (@Image1 , @Video1); retains the style line, the appearance of every subject, a “who is who” list identifying each box by color name and legend color, proxy semantics, the timestamped timeline, and field-of-view information; removes the control block, subject tags, retention analysis, and soundscape. Unverified-FF: base prompt with the control block, as for MiniMax-H3 (see <i>System prompt</i>)
C2R, Cosmos-Transfer2.5, EchoWM, SANA-WM, LingBot 2.0, AlayaWorld v1.1	converted to the single-paragraph caption style used to train C2R: an opening description of the style, setting, number of people, and appearance of every subject, followed by the action progression in prose without timestamps and a final quality sentence; removes tags, proxy terminology, and camera-specific text, and is shortened to satisfy C2R’s 512-token text limit
CWM	converted to CWM’s caption format, consisting of <i>appearance and scene</i> , a timestamped <i>protagonist action timeline</i> , and <i>other visible actors</i> ; removes subject tags and proxy-specific wording, while CWM adds its own system prompt for the mixed proxy render and RGB anchor frame

Prompt C.1 Base prompt, Verified-FF, [...] marks omitted text

Control block (placed at the start of the prompt; plays the role of a system prompt)

AWM_PROXY_CONTROL. Two visual references accompany this text. <Video 1> is the conditioning proxy video: treat it as the exact spatiotemporal reference for where every subject is, how large it is and how it moves, but never as an appearance template or literal target geometry. <Picture 1> is the global RGB anchor image: it defines identity and appearance, never motion. <Picture 1> is the ABSOLUTE standard for the worldview, art direction and overall style of the entire take, and on worldview, art direction and style it outranks even the description below. It provides NOTHING about any later frame’s camera, composition, motion or interactions. Each subject’s position and motion come from <Video 1>. A subject’s appearance follows <Picture 1> when that subject is visible in it; subjects not visible in <Picture 1> take their appearance from the text, rendered consistently in <Picture 1>’s worldview and art direction.

Human proxies in <Video 1> provide only framewise root position, occupied scale (in this representation the box is the entity’s exact extent), and whole-body root orientation, which are exact and must be followed precisely. [...] On every human, vehicle and animal proxy in <Video 1> the white face is the front of that entity’s body: white there is a direction code, never an object, a material or a light, and it must not be painted. The hex colour tags used below only say which proxy stands for which subject: they are identity codes, never colours or materials of the result.

[...] (further proxy rules; vehicle proxies; environment cues)

When sources conflict, resolve in this order: (1) geometry, camera, position and motion follow <Video 1>; (2) the worldview, art direction and overall style follow <Picture 1> absolutely - above the description below; identity, appearance and materials follow <Picture 1> for subjects visible in it, and follow the text for subjects not visible in it; (3) the timestamped text defines the concrete events, actions and rewrite targets within those constraints. Clip contract: this clip is the very beginning of the take - nothing precedes it. <Picture 1> is the global appearance anchor for the whole take, and the first frame of the target is locked to this exact image: the clip begins at precisely this visual state - identities, appearance, materials, art direction, lighting, camera framing and layout - and continues directly from it while performing the timestamped actions. Do not restart, redesign or re-interpret any identity.

subject_definitions:

<Picture 1> is the first frame of [Shot 1]: the opening frame of the target video, already rendered in the target art style of this take, showing A man in a dark suit walks down the center of a city street, flanked by two parked cars and three pedestrians standing on the sidewalks, exactly as they must appear at 00:00.000.

<Video 1> is the conditioning proxy video of the same 5-second shot (its role is fixed by the control block above). The clip contains exactly 4 human proxies standing for <Subject 1> (the protagonist), <Subject 5> (independently moving adult participant 1), <Subject 6> (independently moving adult participant 2), <Subject 7> (independently moving adult participant 3) and 2 object proxies standing for <Subject 3> (the passenger car), <Subject 4> (the passenger car); nothing else in <Video 1> is a person or an interaction object.

<Subject 1> is the protagonist: in <Picture 1> A man with short brown hair, wearing a dark charcoal suit and black shoes, walking away from the camera down the center of the road; in <Video 1> the human proxy tagged #366ef7, whose white face is its front.

<Subject 3> is the passenger car: in <Picture 1> A classic red sedan with a white license plate, parked on the right side of the street facing away from the camera; in <Video 1> the object proxy tagged #f79e36, whose white face is its front.

<Subject 4> is the passenger car: in <Picture 1> A classic black sedan with chrome bumpers and round headlights, parked on the left side of the street facing the camera; in <Video 1> the object proxy tagged #6636f7, whose white face is its front.

<Subject 5> is independently moving adult participant 1: in <Picture 1> A young man with dark hair, wearing a mustard hooded jacket, dark shorts, and white sneakers, standing on the left sidewalk facing the street; in <Video 1> the human proxy tagged #a7f736 that starts on the left of the frame, whose white face is its front.

<Subject 6> is independently moving adult participant 2: in <Picture 1> A person with short blonde hair, wearing a maroon hoodie and blue jeans, standing on the right sidewalk facing the street; in <Video 1> the human proxy tagged #f736df that starts on the right of the frame, whose white face is its front.

<Subject 7> is independently moving adult participant 3: in <Picture 1> A man with dark hair, wearing a white t-shirt under dark blue overalls and brown shoes, standing on the left sidewalk near the black car; in <Video 1> the human proxy tagged #36fd6d that starts on the left of the frame, whose white face is its front.

[...] (Subject 2: environment)

summary:

[keyframe completion + reference generation] Starting exactly from <Picture 1> as the first frame of [Shot 1], generate a 5.00-second continuous shot in <Subject 2> of <Subject 1> crossing a road as approaching vehicles brake and yield, involving <Subject 3> and <Subject 4> and <Subject 5> and <Subject 6> and <Subject 7>. Use <Video 1> for layout, root motion, object trajectories, action timing and the camera path, and keep the appearance,

materials and lighting established in <Picture 1>.

retention_analysis:

<Picture 1> ([Shot 1] first frame): *fully_preserved* - the first frame of the target video reproduces this image exactly, including every person's appearance, the objects, the scene layout, materials, lighting and framing; later frames evolve continuously from it.

<Subject 1> (appears in [Shot 1]): *partially_preserved* - retain the proxy's framewise root position, occupied scale, root orientation (the white face of the box is the front of the body), identity continuity and interaction timing from <Video 1>; reconstruct its appearance from <Picture 1> and its articulated motion from the timeline below.

[...]

detailed_description:

Anime-style rendering: a 2D anime look with hand-drawn character design and painted backgrounds over the full 3D world — modern digital anime: gradient shading, glossy highlights, clean thin contour lines, vivid high-contrast colours; [...]

[Shot 1] The shot opens exactly on <Picture 1> and evolves continuously from that frame. [...] (*setting and appearance of every subject, as in subject_definitions*) From 00:00.000 to 00:04.667, the protagonist walks across the road while the approaching vehicles brake and yield; the protagonist moves about 10 metres along its route; independently moving adult participant 1 moves about 5 metres along its route; independently moving adult participant 2 moves about 4 metres along its route; independently moving adult participant 3 moves about 4 metres along its route; <Subject 3> moves about 9 metres along its route; <Subject 4> moves about 6 metres along its route. From 00:04.667 to 00:05.000, the protagonist reaches the far side of the road and keeps walking; independently moving adult participant 1 moves a short distance along its route; independently moving adult participant 2 moves a short distance along its route; independently moving adult participant 3 moves a short distance along its route; <Subject 3> moves about half a metre along its route; <Subject 4> moves a short distance along its route. <Subject 3> is one continuous object throughout and ends the clip about 7 metres from the protagonist. <Subject 4> is one continuous object throughout and comes to rest about 5 metres from the protagonist. Keep the reference's fixed 53-degree vertical field of view. [...]

overall_soundscape:

The ambient hum of a quiet city street with distant traffic noise and the rhythmic sound of footsteps on asphalt.

non_diegetic_music:

N/A

Prompt C.2 Converted prompt for C2R, Cosmos-Transfer2.5 and the camera-conditioned world models

A high-quality video of a person crossing a road as approaching vehicles brake and yield. Anime-style rendering: a 2D anime look with hand-drawn character design and painted backgrounds over the full 3D world — modern digital anime: gradient shading, glossy highlights, clean thin contour lines, vivid high-contrast colours; no proxy colours, wireframes, block figures or game UI; the opening frame establishes this look and every later frame keeps it. The setting is a wide urban street paved with asphalt and marked with white lane lines, flanked by sidewalks with moss-toned metal railings. There are 4 persons in this video. The protagonist is a man with short brown hair, wearing a dark charcoal suit and black shoes, walking away from the camera down the center of the road. The passenger car 1 is a classic red sedan with a white license plate, parked on the right side of the street facing away from the camera. The passenger car 2 is a classic black sedan with chrome bumpers and round headlights, parked on the left side of the street facing the camera. Independently moving adult participant 1 is a young man with dark hair, wearing a mustard hooded jacket, dark shorts, and white sneakers, standing on the left sidewalk facing the street. Independently moving adult participant 2 is a person with short blonde hair, wearing a maroon hoodie and blue jeans, standing on the right sidewalk facing the street. Independently moving adult participant 3 is a man with dark hair, wearing a white t-shirt under dark blue overalls and brown shoes, standing on the left sidewalk near the black car. Over the five seconds, the protagonist walks across the road while the approaching vehicles brake and yield, and then the protagonist reaches the far side of the road and keeps walking. Human motion is natural and continuous, presented in a cinematic style with sharp details, clean textures, and no blur or visual artifacts.

B 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 [7] from text prompts derived from their semantic labels. Generated tracks are matched to reference entities by label-compatible Hungarian matching using track-level mean IoU.

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

$$\text{IoU}_j = \frac{1}{|\mathcal{V}_j|} \sum_{t \in \mathcal{V}_j} \text{IoU}(\hat{b}_j(t), b_j(t)), \quad \text{CErr}_j = \frac{1}{|\mathcal{M}_j|} \sum_{t \in \mathcal{M}_j} \frac{\|\hat{c}_j(t) - c_j(t)\|_2}{\sqrt{2}}, \quad (5)$$

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 [16] 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 [62], with R_t, T_t the reference and $\hat{R}_t, \hat{T}_t$ the estimated relative poses on the grid $\mathcal{G}$,

$$\begin{aligned} \text{RotErr} &= \frac{1}{|\mathcal{G}|} \sum_{t \in \mathcal{G}} \arccos \frac{\text{tr}(\hat{R}_t R_t^\top) - 1}{2}, & \text{TransErr} &= \frac{1}{|\mathcal{G}|} \sum_{t \in \mathcal{G}} \|\bar{T}_t - \hat{T}_t\|_2, \\ \text{CamMC} &= \frac{1}{|\mathcal{G}|} \sum_{t \in \mathcal{G}} \|[R_t | \bar{T}_t] - [\hat{R}_t | \hat{T}_t]\|_F, \end{aligned} \quad (6)$$

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 in all tables. 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.

Visual and temporal quality. **Imaging Quality** (MUSIQ [21]) and **Aesthetic Quality** (the LAION aesthetic predictor) follow VBench [17]. **Warp Error** [23, 33] 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 [28]. **CLIP-Score** [42] 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 and Consistency. For each intended participant in the multi-view setting, *Appearance Compliance (Comp.)* asks whether the person at the engine-projected location matches the authored appearance description shared by all methods, judged per view at 0.5, 2.5, and 4.5 s from the full frame and a target crop. *Cross-view Appearance Consistency (Cons.)* asks whether the participant looks the same across views: the judge compares every pair of synchronized views without the description, ignoring changes in pose, viewpoint, lighting, and front/back visibility. A participant can thus be consistent yet non-compliant. Both metrics use Qwen3.6-27B with greedy decoding on five attributes (clothing color, clothing style, hair, accessories, and build) and skip attributes that are undescribed or not visible. An observation o (person-view-time for Comp., person-view-pair-time for Cons.) with judgeable attributes $\mathcal{A}_o$

scores

$$q_o = \prod_{a \in \mathcal{A}_o} \mathbf{1}[\text{attribute } a \text{ matches or agrees}], \quad (7)$$

i.e., every judged attribute must pass. Observations with an absent target or no judgeable attribute are unscored. Scores are averaged over views or view pairs, then over participant–time groups, then equally over scenes, and reported in percent; because unscored observations are excluded, both metrics should be read together with the BBox miss rate.

Geometric multi-view consistency. We also report **MEt3R** [2], 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 pairs of the four views at nine synchronized times per scene, average both comparison directions, and average scenes equally; lower is better.

C Experiments

This appendix gives implementation details of the evaluated models and qualitative examples for Section 4.

C.1 Model Configurations and Implementation Details

Table C.1 summarizes the conditioning inputs, spatial resolutions, frame rates, and delivered frame counts of the evaluated methods. The specifications apply to the five-second main-track experiments; PWM is evaluated only on the gunfight set. Spatial dimensions are reported as width $\times$ height. Input sizes describe the supplied conditioning files, with internal resizing stated where applicable; output sizes and frame counts describe the delivered videos. EchoWM [60] uses the Base checkpoint, while SANA-WM [63] and LingBot-World 2.0 [43] use their bidirectional checkpoints; all three generate the full clip without autoregressive rollout. For AlayaWorld v1.1 [46], we use the camera-controlled Stage2b AR teacher with the Stage3 DMD adapter, using four denoising steps per rollout round, because the released Stage1 bidirectional checkpoint supports first-frame and text conditioning but does not provide camera-trajectory control.

Table C.1 Evaluated Model Configurations and Input-Output Specifications. Figure C.1 shows the full test condition of one Unverified-FF sample, and Figure C.2 the conditions of the other settings.

Method	Conditioning	Supplied spatial input	Output	FPS	Frames
EchoWM [60]	RGB first frame, camera trajectory	1280 $\times$ 720 image	1280 $\times$ 704	24	121
SANA-WM [63]	RGB first frame, camera trajectory	1280 $\times$ 720 image	1280 $\times$ 704	16	80
LingBot-World 2.0 [43]	RGB first frame, camera trajectory	1280 $\times$ 720 image	1280 $\times$ 720	16	81
AlayaWorld v1.1 [46]	RGB first frame, camera trajectory; chunk-autoregressive generation, 4 rounds, 4 denoising steps per round	1280 $\times$ 720 image, resized and cropped to 960 $\times$ 544	960 $\times$ 544	24	120
Cosmos-Transfer2.5 [1]	Depth, class-level segmentation	1280 $\times$ 720, 24 fps, 121 frames	1280 $\times$ 720	24	121
C2R [15]	Coarse 3D proxy video	832 $\times$ 480, 16 fps, 81 frames	832 $\times$ 480	16	81
CWM [9]	RGB first frame, CWM proxy, depth	First frame 1280 $\times$ 720; CWM proxy and depth 336 $\times$ 192	1344 $\times$ 768	24	120
PWM [18]	RGB first frame, instance-ID maps from 3D boxes	832 $\times$ 480	832 $\times$ 480	16	81
LynnReal-Omni [37]	Colored-OBb proxy video; first frame as reference	1280 $\times$ 720, encoded at 1344 $\times$ 768	1344 $\times$ 768	24	120
MiniMax-H3 [39]	Colored-OBb proxy video; first frame as reference	1344 $\times$ 768, 24 fps, 124 frames	1344 $\times$ 768	24	120
Seedance 2.5 [4]	Colored-OBb proxy video; first frame as reference	1280 $\times$ 720, 30 fps, 150 frames	1280 $\times$ 720	24	121

C.2 Qualitative Examples

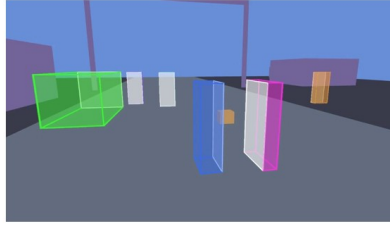
Figure C.3 illustrates entity control (T1) of Section 4.5, and Figure C.4 illustrates Interaction Success Rate and Logic-Render Alignment on one scene. Figures C.5–C.8 show the inputs and outputs of nine methods on four scenes (AlayaWorld v1.1 omitted). Figures C.9 and C.10 show qualitative examples from the Unverified-FF setting discussed in Section 4.5.

Without a verified 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, most of them synthesize the scene in the prescribed style and keep it aligned with the proxy: in Figure C.9, LynnReal-Omni, Seedance 2.5 and MiniMax-H3 reproduce the requested marker-sketch style with varying fidelity, CWM inherits it from MiniMax-H3’s frame 0, and C2R renders the scene photorealistically although its prompt requests the same style, while in every method the protagonist 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.



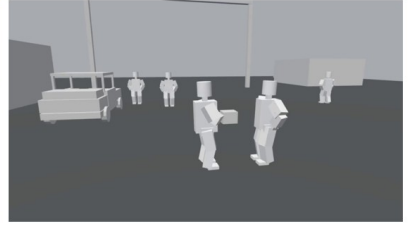
First frame (unchecked)

used by: camera-cond. models, CWM



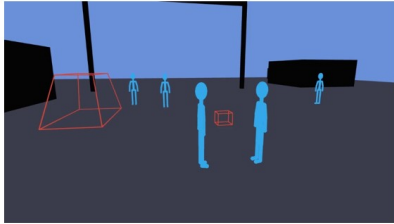
Colored-OBb proxy

used by: LynnReal-Omni, Seedance 2.5, MiniMax-H3



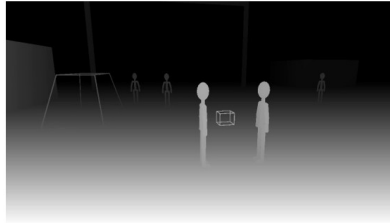
Coarse 3D proxy

used by: C2R



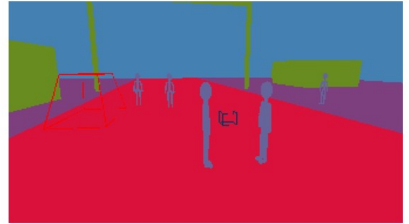
Mixed render

used by: CWM



Depth (from mixed)

used by: Cosmos-Transfer2.5, CWM



Semantic segmentation

used by: Cosmos-Transfer2.5

Text prompt (excerpts; every method's prompt is generated from the same engine records)

Scene and appearance (all methods):

A realistic, high-quality video of a person offering a parcel to another person and releasing it to the receiver. [...] The setting is a working port with stacked shipping containers, a quay, cranes and cargo vehicles. There are 7 persons in this video. A mechanic with auburn hair tied back, wearing navy work overalls over a cream tee and steel-toe boots, standing in the foreground. The person receiving the object is a security guard with a shaved head [...]

Proxy binding (proxy-video models, e.g. MiniMax-H3):

<Video 1> is the conditioning proxy video: treat it as the exact spatiotemporal reference for where every subject is, how large it is and how it moves, but never as an appearance template [...] Human proxies provide only framewise root position, occupied scale and whole-body root orientation [...] reconstruct articulated motion from the timestamped text.

Timeline (from the engine action log):

From 00:00.000 to 00:01.167, the protagonist carries the parcel toward the waiting companion [...]. From 00:01.167 to 00:02.467, the protagonist extends the parcel toward the companion, who reaches out for it [...]

Camera-conditioned world models additionally receive the recorded camera trajectory (150 poses at 30 fps).

Figure C.1 Test condition of one sample. Unverified-FF, handoff scene (follow camera): every conditioning input provided for this sample, the methods that receive it (Table C.1), and excerpts of the text prompt. Proxy frames are shown at 0.5s; the first frame is frame 0 of the MiniMax-H3 output, which is not checked against the recorded state.

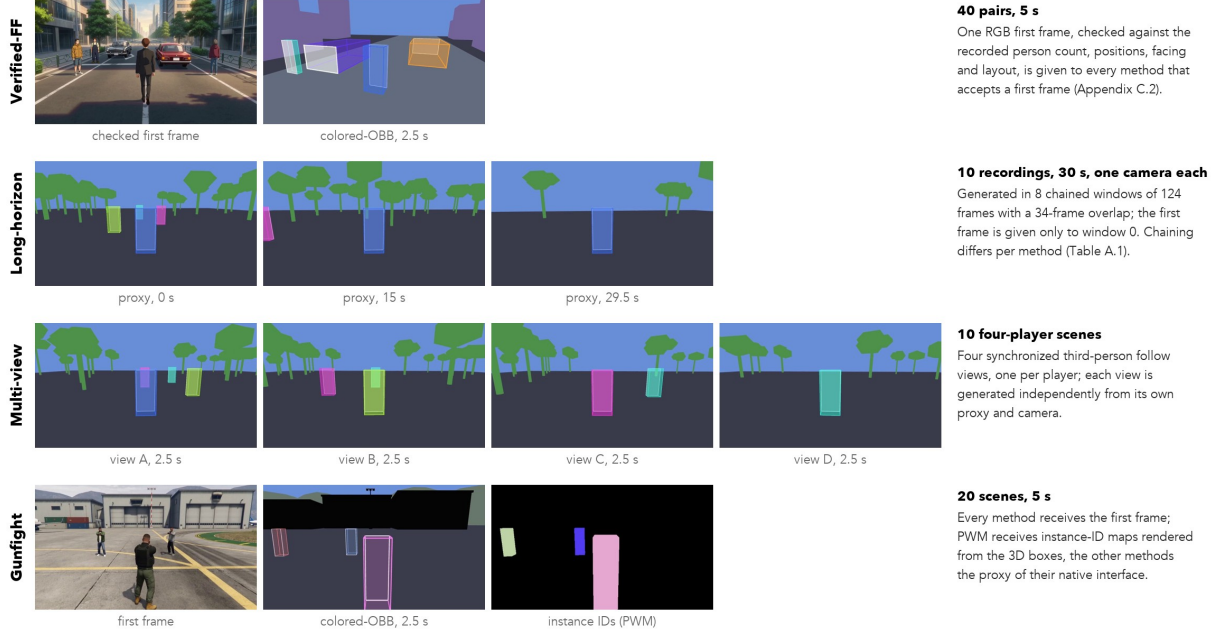


Figure C.2 Test conditions of the other settings. One example per setting: Verified-FF (main track), long-horizon and multi-view (challenge track), and the gunfight set used for the PWM comparison. Compared with the Unverified-FF example in Figure C.1, the settings differ in the first frame, the clip length and generation protocol, and the number of synchronized views. All methods also receive a text prompt generated from the engine records.

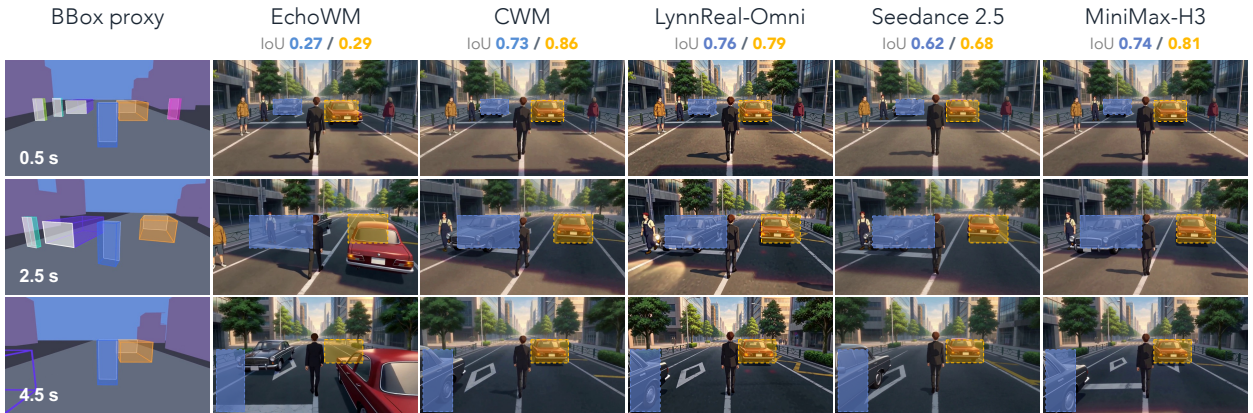


Figure C.3 Entity-level control. A pedestrian walking down a city street, third-person follow camera, same first frame for all methods. The semi-transparent white face on each proxy box marks the entity's front. Dashed: engine reference boxes of the oncoming car (cyan) and the car ahead (orange). The 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).

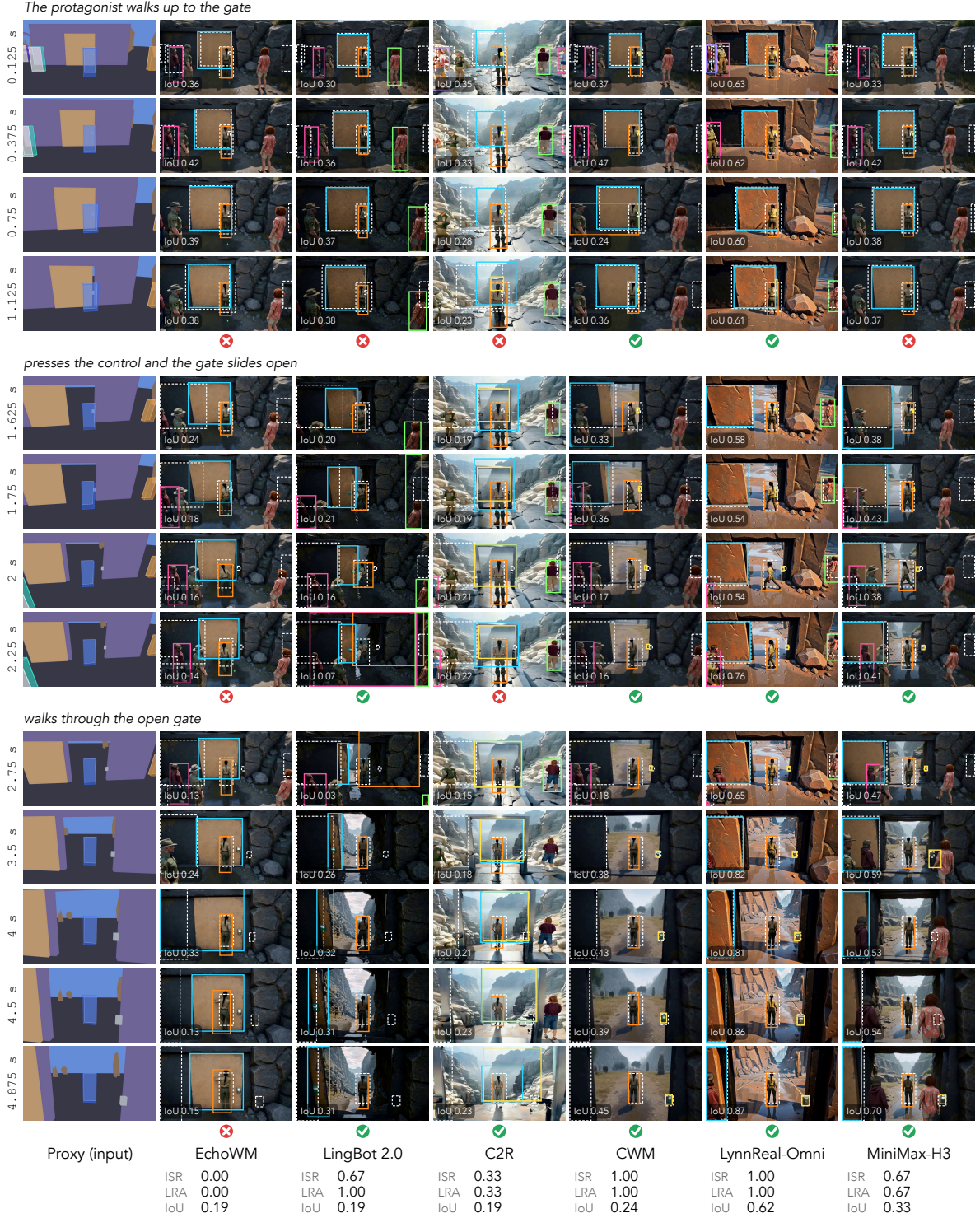


Figure C.4 Interaction Success Rate (ISR) and Logic-Render Alignment (LRA) on one scene. Unverified-FF, gate scene. Columns: the colored-OBb proxy and six methods; rows: the three recorded events (approach, opening, passed-through), four to five frames each. Badges below each event: whether the judge, which sees frames from the whole event window, finds the event realized, i.e., both its action and its end state are judged present. Dashed white boxes are the engine reference boxes and solid boxes the boxes detected in the generated video (one color per entity); the overlay gives the mean BBox IoU of each frame. Bottom: ISR, the fraction of realized events, LRA, the fraction of timeline segments whose action is depicted, and the BBox IoU of the whole video. The two can differ, e.g., LingBot 2.0 depicts all three actions (LRA 1.00) but misses the end state of the first event (ISR 0.67).



Figure C.5 All methods on one scene. Unverified-FF, handoff scene: a courier hands a parcel to a receiver between 1.0 and 2.6 s. For each method, the conditioning input (left pair) and the generated output (right pair) at 0.5 and 2.6 s. Camera-conditioned world models receive frame 0 of the MiniMax-H3 output and the camera trajectory but no proxy; Cosmos-Transfer2.5 receives semantic segmentation (shown) and depth from the CWM proxy representation; CWM additionally receives depth and frame 0 of the MiniMax-H3 output.



Figure C.6 All methods on one scene: entering a car. Verified-FF, city scene, follow camera: the protagonist walks to a parked car, gets in (1.1–2.2s) and drives away (from 2.6s); frames at 1.6 and 4.0s. Methods that accept a first frame also receive the verified first frame. EchoWM and SANA-WM leave the protagonist outside the car, and LingBot 2.0 opens the door but the car does not leave, whereas the proxy-conditioned methods drive the car away as recorded.

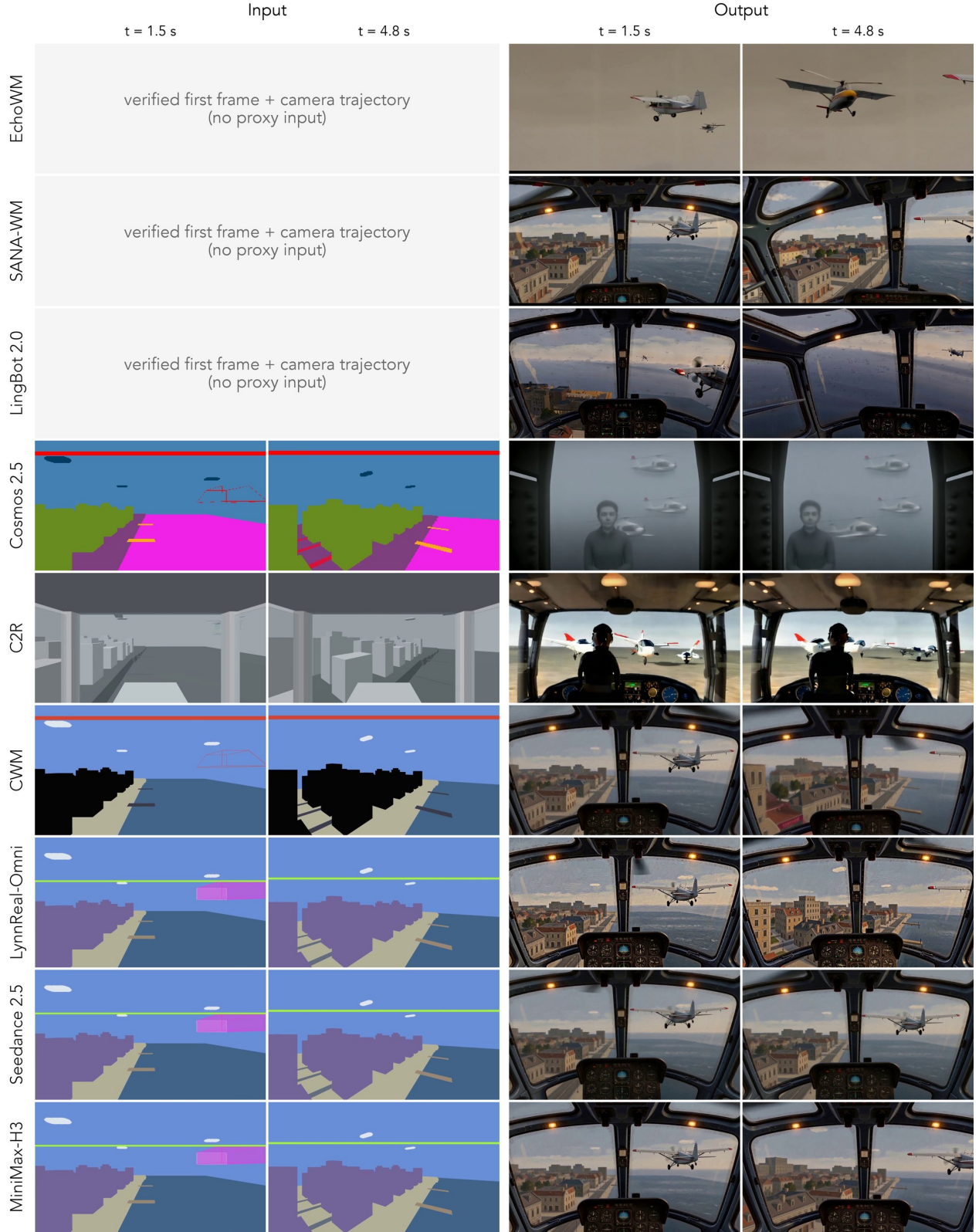


Figure C.7 All methods on one scene: flying a helicopter. Verified-FF, first-person view from a helicopter cockpit that ascends (0–2.3 s), flies forward (2.3–4.1 s) and descends over a harbour town, with an aircraft ahead; frames at 1.5 and 4.8 s. Methods that accept a first frame also receive the verified first frame. EchoWM drops the cockpit and shows the aircraft from outside; Cosmos-Transfer2.5 renders a person facing the camera and C2R a pilot seen from behind, so neither keeps the prescribed first-person view.

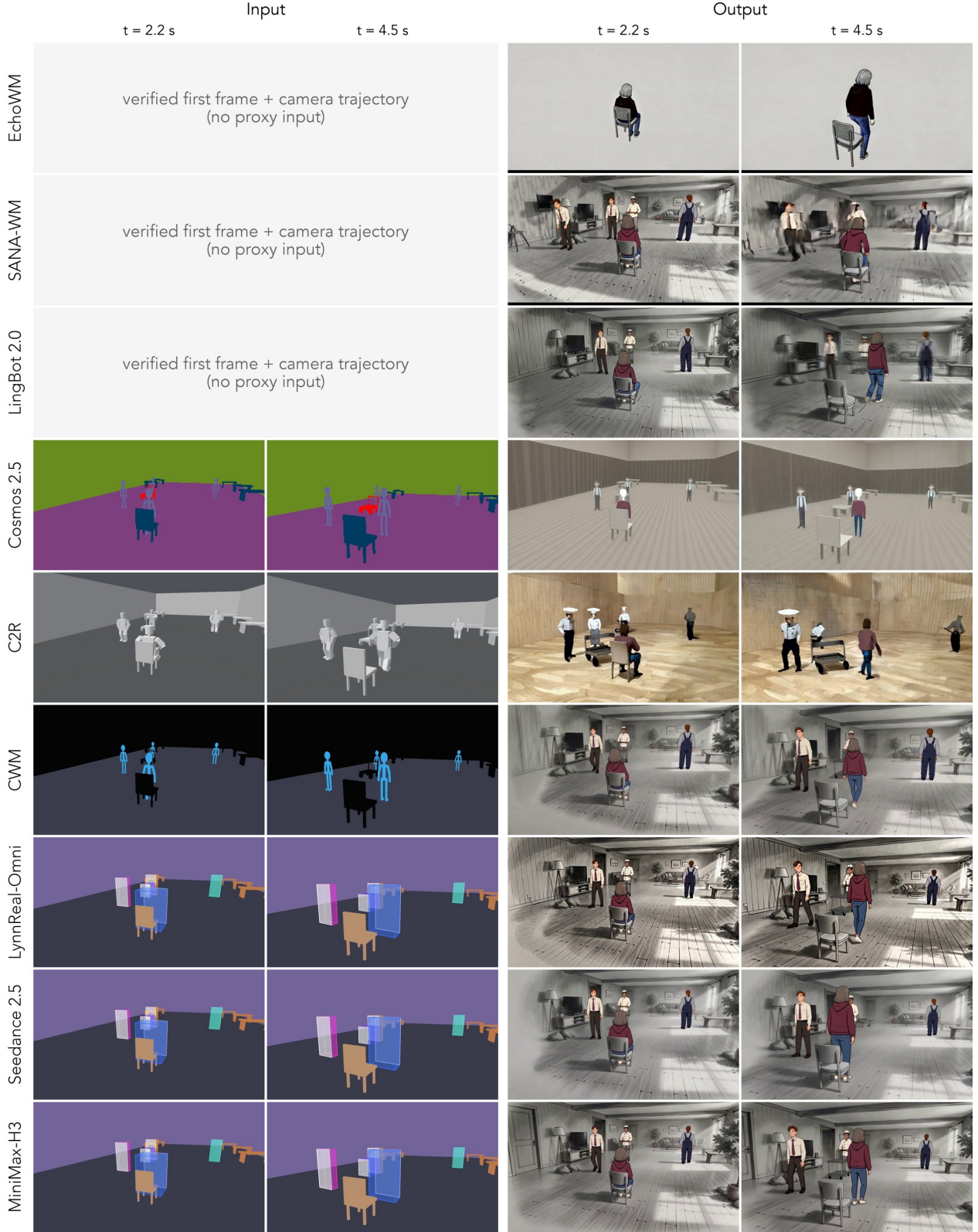


Figure C.8 All methods on one scene: sitting down and standing up. Verified-FF, indoor scene, follow camera: the protagonist sits down on a chair (0.6–1.5 s), stays seated, and stands up again (3.0–3.9 s); frames at 2.2 s (seated) and 4.5 s (standing). Methods that accept a first frame also receive the verified first frame; EchoWM loses the room and renders the protagonist on an empty background.



Figure C.9 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 although its prompt requests the marker-sketch style. All methods follow the proxy as the protagonist climbs toward the upper landing (pair IoU 0.52–0.61).

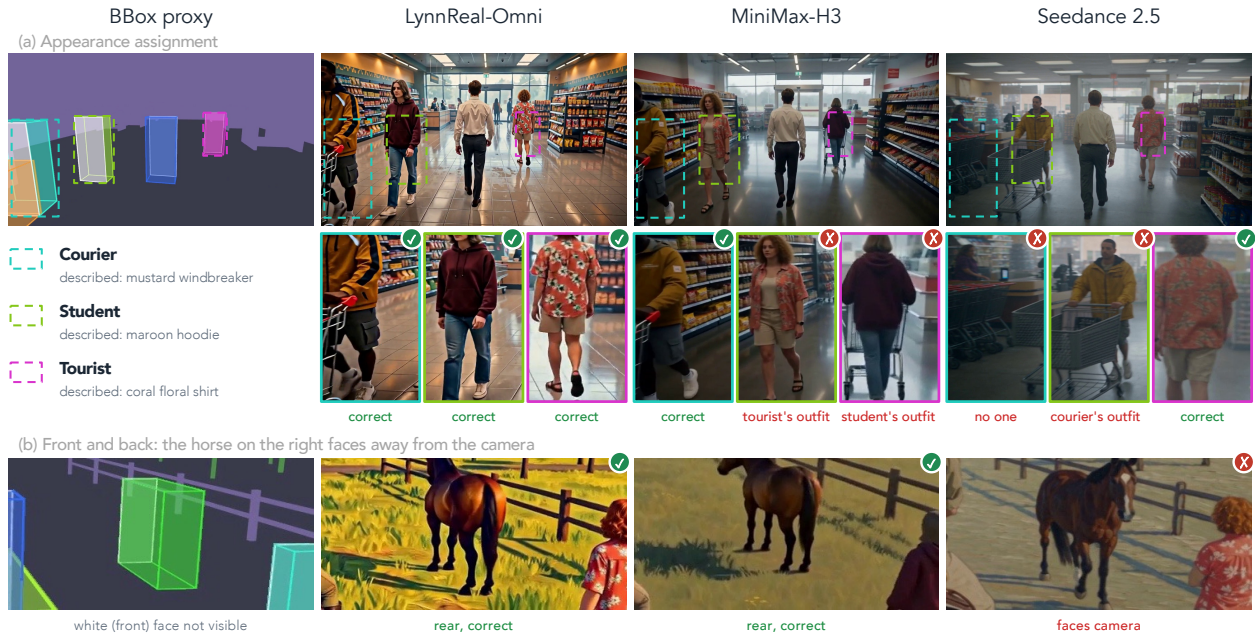


Figure C.10 Appearance and facing errors without a first frame. (a) Supermarket scene at 3.0s: dashed boxes are the engine reference boxes of a courier, a student and a tourist, each with a described outfit; the zoomed crops show what each method renders inside each box. MiniMax-H3 exchanges the student's and the tourist's outfits; Seedance 2.5 puts the courier's outfit at the student's position. (b) Horseback scene at 0.5s: the horse on the right faces away from the camera; Seedance 2.5 renders it facing the camera.