OpenWorldAct: Open-Vocabulary Event Control for Real-Time Interactive World Models

Anonymous authors
Paper under double-blind review

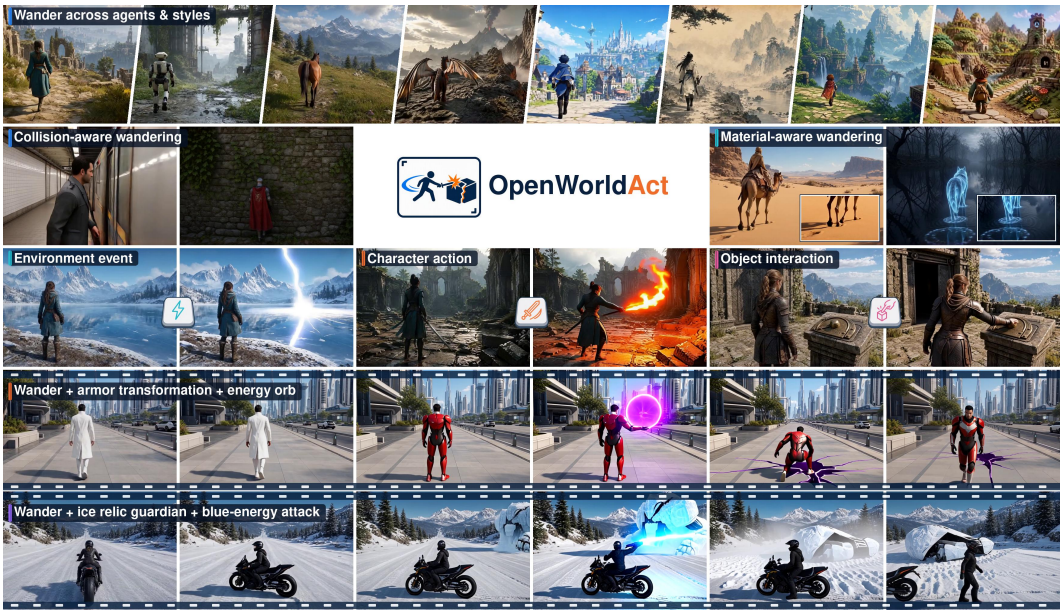

Figure 1: **OpenWorldAct composes free-form wandering with open-vocabulary world interaction.** From top to bottom, the figure shows eight character and visual domains; collision-aware movement and material responses in water and snow; environment, character-action, and character–object events; and two long-horizon compositions of wandering with multiple events.

Abstract

We introduce **OpenWorldAct**, a real-time interactive world model that executes temporally grounded open-vocabulary events while achieving accurate navigation control from a third-person view. Most existing world models support only exploratory action spaces and do not allow users to control open-vocabulary events. The central obstacle is data: gameplay videos provide frame-synchronous controls but cover only a narrow range of semantics, while open or generated videos contain diverse events but can provide only inaccurate control signals obtained through estimation. Data that satisfy both requirements are difficult to obtain. To overcome this bottleneck, we build **ActMosaic**, a complementary data engine that divides its data into two streams. The gameplay stream retains frame-synchronous keyboard and mouse inputs to represent local navigation intent, while the generated event stream is organized around controllable characters, environments, and character–world interactions. At inference time, OpenWorldAct uses a unified time-aligned prompt interface to specify navigation controls and open-vocabulary events during the same rollout. We further develop a dual-resolution streaming few-step autoregressive model, where a low-resolution causal rollout is refined online by a high-resolution branch, enabling long-horizon real-time 720p generation. As an application showcase, we build a Gameplay Harness that maps a user's world setting to state-conditioned skills and tool-assisted observations. Comprehensive experiments show that OpenWorldAct substantially outperforms the evaluated systems in event control while achieving comparable wandering control and visual quality. **We will open source the model weights, code, data and benchmark**.

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

1 INTRODUCTION

Interactive video world models should let a user continuously explore a world in a navigation manner, and trigger diverse semantic events while the world keeps evolving (Bruce et al., 2024; Valevski et al., 2025; Hong et al., 2025). This requires two forms of interaction to coexist in one rollout. Locomotion and camera motion must remain responsive over time, while transient events must begin, develop, and finish at the intended moments. Some methods only inject camera poses or discrete action labels to control movement and viewpoint changes, but these interfaces are difficult to extend to more complex behaviors (He et al., 2025; Li et al., 2025a; Sun et al., 2025; Hong et al., 2025; Wu et al., 2026). Recent prompt-driven world models have achieved promptable chunk-level control, using the text generalization ability of video generation models to support interactions beyond simple spatial exploration (Mao et al., 2025b; DreamX Team et al., 2026; Gao et al., 2026; AlayaWorld Team et al., 2026). However, their semantic interaction remains limited in event breadth or temporal grounding. We target the missing capability: **temporally grounded open-vocabulary event control during continuous interaction**.

The core bottleneck is not simply the amount of video, but the absence of scalable *joint* supervision. Current world-model data typically comes from two regimes. **(a) Gameplay and game-engine synthetic data** provide accurate action and camera trajectories, making them ideal for learning synchronized controls, but their event space is severely limited (Bruce et al., 2024; Decart & Etched, 2024; Che et al., 2025). **(b) Open-domain real-world videos** cover far richer events and visual situations, but their camera motion must be estimated rather than directly logged, and the desired event is often entangled with unrelated character displacement, camera motion, or scene changes (Ho et al., 2022; Villegas et al., 2023; Jiang et al., 2025; Mao et al., 2025a). A direct solution would be to collect large-scale data that simultaneously covers diverse events and provides precise action and camera-trajectory annotations. This Cartesian product over camera motion, displacement, environment changes, character actions, and human–object interactions is impractical to collect and label. Existing work therefore has not solved the problem of scalable joint supervision for precise controls and broad events.

OpenWorldAct addresses this bottleneck with the proposed **ActMosaic**, a data engine that provides complementary supervision instead of joint navigation-event annotation. It combines two primary data sources: **(a) synchronized gameplay** for frame-level locomotion and camera response, and **(b) motion-suppressed event synthesis** generated by video generation models and filtered to remove cases with character displacement or camera motion. For navigation, the control stream keeps frame-synchronous keyboard and mouse inputs as local user intent, avoiding a brittle conversion from navigation commands to fixed global camera-pose trajectories. To preserve the pretrained generator's generalization over free-form prompts, the event stream requires broad event coverage. We decompose events into atomic components and organize them into three categories: character events, environment events, and character–environment interactions. We also filter out excessive character displacement and camera motion in the event data to avoid conflicts with basic control supervision. The model is trained on the supervision each source can reliably provide, and exposes capabilities learned from these disjoint supervision streams through a unified control-injection interface at inference time.

For joint execution, OpenWorldAct uses a unified time-aligned prompt interface. At every causal latent position, the active state contains the persistent scene, the currently held controls, and all temporally overlapping events. All active controls and events are represented as prompts and routed through the same video–text cross-attention path, allowing navigation, camera motion, and free-form events to overlap without task-specific conditioning heads.

Moreover, we develop a dual-resolution streaming few-step autoregressive model that couples low-resolution causal rollout with online high-resolution refinement, enabling long-horizon real-time 720p generation. Building on the model's support for open-vocabulary events, we further implement a Gameplay Harness that converts a user's world setting into a programmatic state machine and a finite action space, and dynamically invokes visual tools to identify the current world state. Figure 1 shows that our model maintains stable wandering control while performing long-horizon open-vocabulary event control.

In evaluation, OpenWorldAct achieves the strongest event control and instruction-following ability compared with evaluated methods across both quantitative and qualitative terms, while retaining competitive navigation control and visual quality. Our contributions can be summarized as follows:

- **Complementary interaction data.** We introduce ActMosaic, a data engine that learns from synchronized gameplay and motion-suppressed event synthesis, avoiding the need to collect paired demonstrations for every control–event combination.
- **Open-vocabulary event control.** We retain the pretrained model's free-form event conditioning while learning navigation from synchronized gameplay, and use time-aligned prompts to specify both during a single autoregressive rollout.
- **Real-time interactive execution.** We propose a dual-resolution streaming strategy for long-horizon real-time 720p generation. The Gameplay Harness is further developed that maps world settings to state-conditioned skills and tool-assisted observations.

2 Related Work

Action-conditioned and camera-controlled worlds. Several world models obtain navigation control by injecting discrete action labels or continuous camera poses into video generators, sometimes through an additional trained control module (Valevski et al., 2025; Decart & Etched, 2024; Che et al., 2025; He et al., 2025; Li et al., 2025a; Sun et al., 2025). This can give users precise control over movement and viewpoint. Yet a navigation interface built from fixed action labels or camera parameters mainly supports exploration; it does not naturally express the much larger space of character actions, object interactions, and environment events. Extending such an interface requires new control definitions and corresponding training data. OpenWorldAct addresses this limitation by unifying all interactions through native prompt interface of the video model, providing a flexible and scalable form of control.

Promptable events and long-horizon interaction. Text-to-video and video-editing models can generate a wide range of actions, effects, and scene changes from text (Ho et al., 2022; Villegas et al., 2023; Kondratyuk et al., 2024; Yang et al., 2025b; Kong et al., 2024; Team Wan et al., 2025; Jiang et al., 2025). Recent promptable world models use text to steer what happens during an ongoing rollout, extending interaction beyond fixed navigation commands (Mao et al., 2025b; DreamX Team et al., 2026; Gao et al., 2026; Wang et al., 2026; AlayaWorld Team et al., 2026). Other work improves the stability and latency of long-running generation, making long-horizon playable world models with open-vocabulary actions feasible (Huang et al., 2025; Li et al., 2025b; Yang et al., 2025a; Zhu et al., 2026; Hong et al., 2025; Sun et al., 2025). However, current interactive models have not fully preserved the prompt interface's generalization to diverse actions and events. The proposed ActMosaic provides complementary supervision, with one stream providing precise navigation signals and the other providing broad and rich events. Benefiting from this design, OpenWorldAct achieves significantly stronger open-vocabulary event control than existing methods.

3 ActMosaic: A Data Engine for Complementary Supervision

Figure 2 gives the complete construction process. The central obstacle is not a lack of videos, but a lack of scalable videos that jointly contain aligned user controls and open-vocabulary event annotations. ActMosaic is built around this mismatch. It constructs two primary supervision streams: a control stream from frame-synchronous gameplay and an event stream from motion-suppressed generation. The event stream filters out samples with character displacement or camera motion so that event supervision does not interfere with the model's learning of basic controls.

We construct complementary data distributions that provide different forms of interaction supervision. Let $x_{1:T}$ denote a video, b its persistent scene description, $u_{1:T}$ synchronized low-level controls, and

$$\mathcal{E} = \{(c_i, d_i, s_i, e_i)\}_{i=1}^{M} \tag{1}$$

a set of semantic events, where c_i is a coarse event type, d_i is a free-form description, and $[s_i, e_i)$ is its temporal support. Rather than assuming access to paired joint demonstrations, ActMosaic constructs

$$\mathcal{D}_{\text{ctrl}} = \{(x_{1:T}, u_{1:T}, b)\}, \qquad \mathcal{D}_{\text{evt}} = \{(x_{1:T}, \mathcal{E}, b)\}. \tag{2}$$

Task-specific training uses these complementary sources separately, while inference explicitly takes joint conditions as input:

$$p_\theta(x_{1:T} \mid x_0, u_{1:T}, \mathcal{E}, b), \tag{3}$$

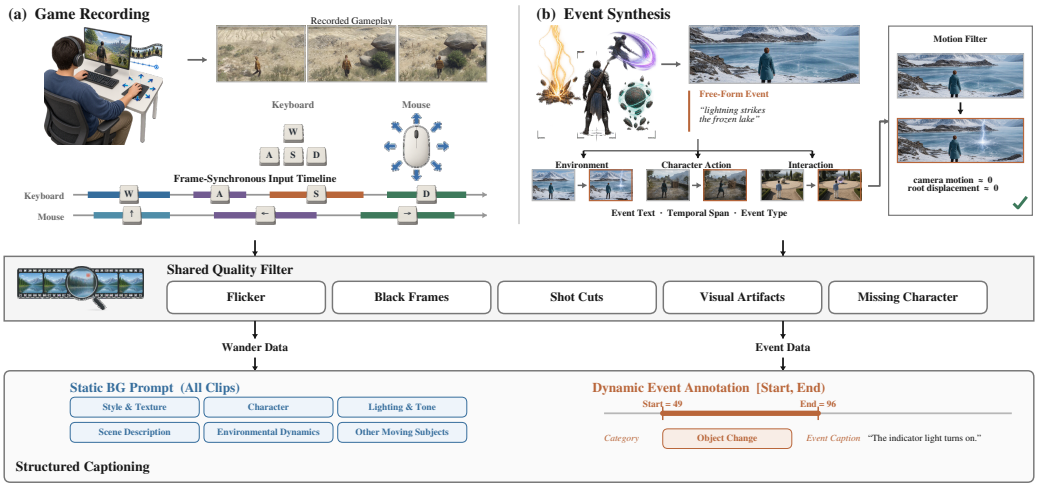

Figure 2: **Illustration of ActMosaic data construction.** **(A)** The control stream synchronizes every gameplay frame with keyboard and mouse state. **(B)** The event stream covers environment, character-motion, and character–world interaction prompts with temporal-span annotations while suppressing character root displacement and camera motion. Both streams pass a shared quality filter before structured captioning.

which avoids collecting the Cartesian product of control trajectories and semantic events. It allows the model to learn precise controls from gameplay data and broad event semantics from generated event data, then compose them at inference without requiring paired control–event clips.

3.1 Frame-Synchronous Gameplay Controls

Gameplay recordings provide dense, accurately timed supervision for persistent control. We record the keyboard WASD state synchronously with captured frames and quantize mouse motion into eight movement directions plus a stationary state. These signals describe local interaction intent at each timestep, such as holding a movement key or applying a mouse-motion increment relative to the current view, rather than prescribing absolute 3D positions or global camera trajectories. Because occlusion, collision, and local scene geometry can make the same navigation input produce different physical and visual outcomes, low-level controls do not correspond one-to-one to fixed 3D displacements or camera-pose labels. Learning camera-pose-conditioned control and then converting navigation commands into a predetermined camera-pose sequence at inference time can therefore miss the state-dependent nature of interactive response. In contrast, frame-synchronous keyboard and mouse signals directly record the user's local control intent at each timestep, allowing the model to learn the correspondence between input actions, the current visual state, and the resulting visual changes for more accurate persistent control.

3.2 Motion-Suppressed Event Generation

In the event stream of ActMosaic, we use GPT-Image-2 (OpenAI, 2026b) generated images or game screenshots as the first frame, use Gemini-3.5-Flash (Google, 2026) to construct event prompts, and use HappyHorse1.1 (Alibaba Cloud, 2026) to generate the final video samples. To preserve the pretrained video generator's generalization over free-form prompts and support open-vocabulary event control, the event corpus must provide broad coverage over possible world changes. However, directly enumerating complete event descriptions is difficult. We therefore perform an atomic decomposition: a world is abstracted into the controllable character and the environment, and event prompts are constructed from elementary changes of these components and their interactions:

- **character events**, including basic actions, equipment use, and character transformations;
- **environment events**, including appearance, disappearance, deformation, destruction, weather, illumination, and scene-state changes;

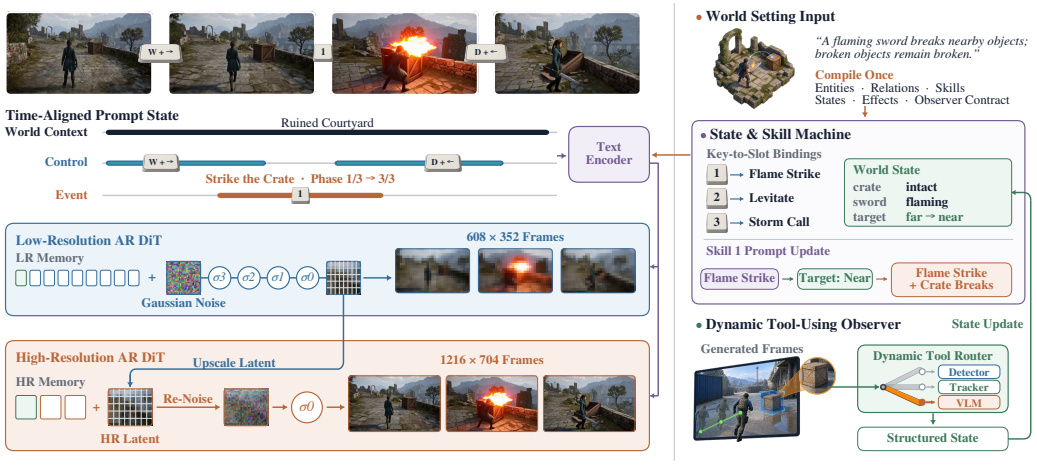

Figure 3: **OpenWorldAct architecture.** *Left:* a time-aligned prompt jointly controls movement, camera, and events through dual-resolution autoregressive rollout. *Right:* Adaptive Gameplay Harness compiles world settings into stateful skills and updates them online with diverse visual tools.

- **character–world interactions**, in which an action is bound to a target and produces an observable consequence.

The generation contract asks the controlled character to avoid sustained root displacement and the camera to stay static. Non-target regions should remain stable, leaving the requested event as the principal visible change. Constrained generation suppresses most non-target motion at the source, while residual filtering removes candidates that still exhibit excessive character displacement or camera motion. Together, these stages reduce motion confounds in the event stream.

All constructed clips pass shared quality filters for flicker, black frames, shot changes, visual artifacts, and missing characters. Event clips additionally require event adherence together with bounded character root displacement and camera motion. Detailed category counts and complementary control-data statistics are reported in Table 3 of Appendix A.

4 Open-Vocabulary Interaction World Model

Figure 3 provides an overview of the inference-time workflow of OpenWorldAct and how the Adaptive Gameplay Harness updates the world state and action space. In this section, we first present a unified, time-aligned prompt stream that enables simultaneous character movement, camera control, and free-form events. We then describe how we train a few-step, dual-resolution autoregressive model to support long-horizon interaction in real time. Finally, we detail the Adaptive Gameplay Harness, which turns the model's open-vocabulary interface into stateful, game-like interactions.

4.1 Unified Prompt-Based Interaction

We retain the text-injection pathway of video diffusion models and use it as the common interface for both dense controls and semantic events. Let I_l be the source-frame interval covered by causal latent position l. The prompt that cross-attends with the video tokens corresponding to I_l is

$$q_l = (U_l, \{(c_i, d_i) : [s_i, e_i) \cap I_l \neq \emptyset\}, b)\,. \tag{4}$$

The prompt q_l is constructed by concatenating the keyboard and mouse state U_l, the set of currently active events $\{(c_i, d_i) : [s_i, e_i) \cap I_l \neq \emptyset\}$, and the time-invariant static scene description b. Each event is represented by its category c_i and free-form caption d_i, while its temporal span $[s_i, e_i)$ determines the latent positions at which it is active.

During training, gameplay samples activate only control prompts and motion-suppressed event samples activate only event prompts. At inference time, the video model's generalization over textual prompts allows the two types of phrases to be concatenated and injected into the model. For complex inputs, a prompt enhancer decomposes the instruction into a composition of atomic event prompts before injection, matching the event decomposition used in ActMosaic.

4.2 From Full-Step AR to Dual-Resolution Generation

We start from the pretrained LTX-2.3 (HaCohen et al., 2024; Lightricks, 2026) and convert it into a block-autoregressive generator using teacher forcing (Williams & Zipser, 1989). Each target latent block attends to the initial frame and a bounded history of clean preceding blocks, under the prompt state aligned to that block.

To reduce inference cost, we distill the autoregressive model to four denoising steps through consistency distillation (Song et al., 2023). We then apply distribution matching distillation (DMD) (Yin et al., 2024) at 480p (832×480) on 721-frame videos to establish long-horizon generation. The student rolls out on its own causal history, while a bidirectional teacher trained at the same resolution and length supplies the target score and a learned fake-score model tracks the student's current distribution. This training follows prior work on distilling autoregressive or streaming video generators for low-latency inference (Huang et al., 2025; Zhu et al., 2026; Zheng et al., 2026).

In order to improve visual quality, we further increase the generation resolution to 720p (1216×704). However, running four denoising steps at 720p remains too slow for real-time generation at 24 FPS. We therefore use a dual-resolution streaming architecture: a four-step causal rollout at 360p (608×352) followed by one refinement step at 720p (1216×704). For each block, a learned $2\times$ latent upscaler maps the clean low-resolution prediction to the high-resolution latent. Then the upscaled latent is renoised at σ, which is the same noise level used by the final step of the low-resolution schedule:

$$\tilde{z}_\sigma^{\mathrm{HR}} = (1-\sigma)G_\phi(z_0^{\mathrm{LR}}) + \sigma\epsilon, \qquad \hat{z}_0^{\mathrm{HR}} = \tilde{z}_\sigma^{\mathrm{HR}} - \sigma v_\theta(\tilde{z}_\sigma^{\mathrm{HR}}, q_l). \tag{5}$$

We train the coupled branches in a second, joint DMD stage on 241-frame videos. A separate bidirectional teacher trained at 720p on 241-frame videos supervises the high-resolution output. Because the low-resolution branch generates most of the content and the refinement branch adds details, we shorten the high-resolution memory window to improve inference speed with almost no loss in quality. The two branches run as a streaming $4{+}1$ pipeline on two separate GPUs, producing each high-resolution block online rather than applying offline super-resolution to a completed video. Full training and optimization settings are provided in Appendix C.1.

4.3 Application: Adaptive Gameplay Harness

Open-vocabulary events make the model substantially more expressive, but a game-like experience also requires persistent rules, reusable skills, and reactions to the evolving scene. Our Gameplay Harness illustrates such an execution layer with three components, shown on the right of Figure 3.

First, a compiler converts a free-form world setting into a structured script containing entities, relations, state variables, callable skills, state-dependent prompt variants, and their effects. The compiler also decomposes complex skill descriptions into a composition of atomic event prompts that match the training prompt format. Second, a state machine binds player inputs to skill slots and selects a prompt variant from the current state. For example in Figure 3, after the target relation changes from *far* to *near*, the prompt bound to Slot 1 can change from *Flame Strike* to *Flame Strike + Crate Breaks*.

Moreover, a dynamic observer updates the state machine from generated video. According to the current script, a router invokes the appropriate detector, tracker, or vision-language model and normalizes its output into read-only entity, relation, and state observations. Only the state machine commits state changes and skill effects, preventing uncertain perception from directly firing an action. Skill prompts are pre-encoded at episode initialization, and state and prompt updates occur at autoregressive block boundaries. This design adapts the available interactions to both the user's world setting and the observed environment while retaining block-level user control. The script schema, observer contract, and qualitative application examples are detailed in Appendix B.

5 Experiments

In this section, we build a benchmark with 120 cases to evaluate OpenWorldAct against current state-of-the-art interactive world models from both quantitative and qualitative perspectives. The evaluation includes diverse automatic metrics, visual comparisons, and a blinded user study. We

Table 1: **Comparison with interactive world models.** We evaluate event execution, wandering control, perceptual quality, and human preference on our benchmark.

Method	Open-vocabulary Event			Wander		Quality ↑	Human Evaluation	
	Avg. (0–3) ↑	Response Rate ↑	IoU_{all} ↑	Camera ↑	Character ↑		Win Rate ↑	Elo ↑
Yume-1.5	0.9467	39.00	25.87	63.67	85.63	78.11	27.7	777
DreamX-World	0.4200	16.00	13.19	73.94	76.51	77.56	23.1	817
AlayaWorld	0.3067	12.67	6.98	59.00	71.07	77.16	25.0	819
LingBot-World-V2	0.9333	41.00	28.34	**88.29**	90.91	78.71	78.6	1253
Alaya-Evoke	0.6733	31.33	20.31	69.47	62.57	**82.11**	53.8	1031
OpenWorldAct	**1.9567**	**72.00**	**51.13**	74.18	**96.77**	79.78	**90.0**	**1303**

further use ablation studies to analyze ActMosaic, examine open-vocabulary generalization, and characterize the quality–speed trade-off of real-time inference.

5.1 Experimental Setup

Benchmark. We build a benchmark containing 120 initial scenes. It contains 30 *Wander* cases with 30-second keyboard and camera trajectories, 60 *Event+Wander* cases with a 30-second navigation signal and three random events, and 30 *Event-only* cases, each with four random events. The camera and character trajectories are procedurally generated random results and include complex combined keyboard actions. All events are randomly generated by an LLM but are not selected to fit the event distribution of the training set. In the ablation study, we specifically build the training-set events as a retrieval bank to analyze the overlap between the benchmark and the training set. We prioritize open, traversable scenes so that the random wander routes have sufficient free space. This can prevent physical collisions from making the accuracy of control difficult to evaluate.

Metrics. We evaluate each method from event control ability, wander control ability, and visual quality. For the event dimension, we use a VLM to grade event completion from 0 (absent) to 3 (all required actions, objects, effects, and causal order are correct). We report the average score and *Response Rate*, which is the fraction of events with a score greater than zero. We also compute mIoU between the event occurrence time and the GT time interval, where events with an event completion score of 0 also receive an IoU of 0. For wander control ability, we use LingBot-Map (Chen et al., 2026) to estimate the camera trajectory and compute its correlation with the requested input direction. We additionally use a VLM to judge the character orientation and pose to evaluate the character control ability. For visual quality, we directly use the computation and weighting method from WBench (Ying et al., 2026).

Compared systems and protocol. We compare with AlayaWorld (AlayaWorld Team et al., 2026), DreamX-World (DreamX Team et al., 2026), LingBot-World-V2 (Gao et al., 2026), Alaya-Evoke (Yin et al., 2026), and Yume-1.5 (Mao et al., 2025b). Outputs are evaluated at their native resolution and frame rate. To ensure a fair comparison, we add corresponding interfaces for baselines whose official code does not support chunk-level prompts. All methods use identical evaluator implementations, rubrics, and judge settings; further metric and aggregation details are provided in Appendix C.2.

5.2 Comparison with Interactive World Models

Quantitative comparison. Table 1 reports all metrics together with the user study results. Quality is evaluated on all 120 videos, control on the 90 videos with wander trajectories, and event metrics on all 300 events. OpenWorldAct improves the strongest baseline from 0.9467 to 1.9567 in average event completion and from 28.34% to 51.13% in IoU_{all}; its response rate is 72.00%, compared with 41.00% for the strongest baseline. For wander control, OpenWorldAct achieves the second-highest camera accuracy at 74.18 and the highest character accuracy at 96.77, improving the strongest baseline on the latter by 5.86 percentage points. At 79.78, OpenWorldAct retains the second-highest aggregate quality score, behind Alaya-Evoke at 82.11, while substantially outperforming it on event execution and character control. Detailed quality results, evaluation on 67 third-person-view cases from the public WBench benchmark, and event-family and compositional breakdowns are provided in Appendix D.

Qualitative comparison. Figure 4 provides visual comparisons on two representative prompts. For the particle-to-mist prompt, several baselines produce only a brief glow or isolated smoke, and some fail to coordinate the effect with the requested two-hand action. In contrast, OpenWorldAct synchronizes the character's raised hands with the release of visible particles and their subsequent



Figure 4: **Qualitative comparison.** All six systems are shown on the same two prompts, with three sampled frames per prompt.

formation into flowing mist. For the rope-and-awning prompt, the baselines largely leave the awning unchanged even when the character approaches the rope or performs a related gesture. OpenWorldAct instead generates the rope-pulling action together with the corresponding fold of the awning. Across both examples, OpenWorldAct realizes the requested action, effect evolution, and object-state change while preserving the surrounding scene and continuous character control.

Blinded human evaluation. Automatic metrics do not fully capture instruction following or perceptual quality for open-ended events. We therefore conduct a blinded user study for human preference assessment. Each task presents an anonymous pair of complete rollouts under matched initial conditions; raters choose which video better realizes the requested interaction while retaining visual coherence and quality, or select a tie. The resulting Elo (Elo, 1978) and tie-aware win rate are included in Table 1; detailed pairwise outcomes and independent ratings of event execution, wandering control, and visual quality are provided in Appendix E.

5.3 Ablation Studies

We organize the ablations around the three contributions introduced in Section 1. (a) Analysis of the complementary interaction data in ActMosaic, (b) examines whether open-vocabulary event control generalizes beyond training-set event captions, and (c) studies the quality–speed trade-off.

Analysis of ActMosaic. Table 2(a) first examines the data design of ActMosaic. Starting from a Wander-only baseline, we add unsuppressed event data to introduce semantic event supervision, then suppress unintended character displacement and camera motion in the event stream. Because the full training pipeline contains multiple stages, including few-step and high-resolution distillation, training every variant end-to-end would be costly and would introduce additional effects from later stages. We therefore use a full-step autoregressive model as a proxy setting to compare the data configurations. This ablation is intended to analyze the effect of the event data and motion-suppression strategy, rather than reproduce the final performance of each variant after the complete training pipeline.

The results show that a Wander-only model retains strong navigation control but has difficulty executing diverse semantic events. Adding event data substantially strengthens event execution, but unintended character and camera motion in the event stream can interfere with navigation control. Motion suppression mitigates this interference while retaining event-control ability, allowing the two capabilities to coexist more effectively.

Table 2: **Ablations of the three core contributions.** We evaluate ActMosaic design, open-vocabulary generalization, and the quality–speed trade-off. Control accuracy, response rate, and IoU_{all} are percentages; ✓ denotes real-time speed above 24 FPS.

(a) Analysis of ActMosaic

Variant	Open-vocabulary Event			Wander		Quality ↑
	Avg. (0–3) ↑	Response Rate ↑	IoU_{all} ↑	Camera ↑	Character ↑	
w/o event data	0.811	32.78	18.83	88.35	99.17	78.36
w/o motion filter	1.944	71.67	48.33	82.92	95.55	77.09
Full	1.911	70.00	49.21	87.16	97.38	77.42

(b) Seen vs. Unseen Event Generalization

Subset	Events Count	Avg. ↑	Response ↑	IoU_{all} ↑
Seen	110	2.082	73.64	52.22
Unseen	190	1.884	71.05	50.51
All	**300**	**1.957**	**72.00**	**51.13**

(c) Quality–Speed Frontier

Variant	FPS ↑		Quality ↑
	1 GPU	2 GPUs	
4-step 360p	25.18 ✓	35.48 ✓	78.10
4-step 480p	14.72	21.44	78.69
4-step 720p	8.080	11.58	**80.35**
Dual-resolution (Full)	13.06	25.06 ✓	79.78

Demonstration of Open-Vocabulary. Panel (b) tests whether the model can execute event descriptions that are absent from task-specific training. To construct the split in panel (b), we compare every benchmark event against the event captions used for task-specific training using both BGE semantic similarity (Xiao et al., 2024) and TF–IDF lexical similarity (Salton & Buckley, 1988), and apply fixed thresholds to partition the benchmark into Seen and Unseen subsets. We then use GPT-5.6-Sol (OpenAI, 2026a) to review the assignments twice to ensure the accuracy of the split. This yields the final partition of 110 Seen and 190 Unseen events. Compared with Seen events, OpenWorldAct has a completion-score gap of only 0.1976 on Unseen events, while the gaps in response rate and IoU_{all} are only 2.584 and 1.707 percentage points, respectively. These small gaps indicate that our model retains the pretrained backbone's responsiveness to general prompts rather than overfitting to the task-specific training data, further supporting its open-vocabulary event-control capability.

Trade-off Between Quality and Speed. Visual quality generally improves as the generation resolution increases, but higher resolution also slows down inference. We therefore conduct the ablation in Table 2(c) to analyze how to balance visual quality and inference speed. Specifically, we compare four-step generation at 360p, 480p, and 720p with the full dual-resolution system. For each variant, we report the DiT inference FPS on one and two GPUs, together with visual quality. For standard four-step generation, we use tensor parallelism on two GPUs for acceleration. Configurations that reach real-time inference, i.e., FPS > 24, are marked with ✓ in the table.

As shown in Table 2(c), our method achieves real-time inference while obtaining the second-best visual quality. Although 720p four-step generation gives the best visual quality, its 11.58 FPS is far from real time. This is because our dual-resolution design assigns the burden of global content generation to the low-resolution branch, while the high-resolution branch only enriches details. In addition, the two resolution branches can run independently on different GPUs, avoiding the communication overhead of tensor parallelism.

6 Limitations and Conclusion

We introduce OpenWorldAct for temporally grounded open-vocabulary event control in interactive world models. ActMosaic combines complementary control and event supervision without requiring paired joint demonstrations, and the unified prompt interface, temporal training objectives, and dual-resolution streaming design enable stronger event execution while retaining competitive control and visual quality. Two limitations remain. First, the model lacks explicit 3D geometry or persistent spatial memory. Second, inference speed still has substantial room for improvement, for example through attention acceleration and lightweight pruning of the LTX decoder. Despite these limitations, OpenWorldAct moves interactive world models beyond fixed action spaces toward real-time, semantically expressive interaction.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code development and data processing. Specifically, GPT-Image-2 generated initial images for the event-data stream, Gemini-3.5-Flash helped construct event prompts, and HappyHorse1.1 generated event videos (Section 3). An LLM generated benchmark event descriptions (Section 5), and GPT-5.6-Sol assisted the semantic-equivalence review of the Seen/Unseen assignments using a predefined rubric (Appendix D). Additionally, we used generative AI tools to polish the language of the manuscript. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including code, data, text, results, and claims.

ETHICS STATEMENT

This work trains on gameplay recordings and model-generated images and videos, and evaluates generated interactive scenes in a human study (Appendix E). All participants provided informed consent, and their ratings were collected anonymously and reported in aggregate. Synthetic video generation can be misused to produce misleading visual content, while the training sources may carry representational biases or third-party rights. We describe the data sources and filtering procedures in Appendix A; the public release will document data provenance, applicable licenses, and any redistribution conditions. These considerations should inform the interpretation and downstream use of the system.

REPRODUCIBILITY STATEMENT

Section 4 describes the model and training objectives, while Appendix A specifies the two data streams and filtering protocol. Appendix C reports training, prompt construction, and evaluation settings; Appendix D provides additional benchmark analyses; and Appendix E gives the human-evaluation protocol and detailed results. As stated in the abstract, we will release the model weights, code, data, and benchmark to support replication and further study.

REFERENCES

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Mingliang Zhai, Ruicong Liu, Xiaojie Xu, Xuangeng Chu, Zhen Li, Zhengyuan Lin, Zhixiang Wang, Zian Meng, and Zihui Gao. AlayaWorld: Interactive long-horizon world modeling – full technical report. *arXiv preprint arXiv:2607.18367*, 2026.

Alibaba Cloud. Alibaba Cloud Model Studio: happyhorse-1.1-i2v. Alibaba Cloud Documentation Center, 2026.

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

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao CHEN. GameGen-X: Interactive Open-world Game Video Generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 37546–37593, 2025.

Lin-Zhuo Chen, Jian Gao, Shangzhan Zhang, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. LingBot-Map: Geometric context transformer for streaming 3d reconstruction. *arXiv preprint arXiv:2604.14141*, 2026.

Decart and Etched. Oasis: A universe in a transformer. Project page, 2024.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, Geng Li, Jifan Li, Ruimin Lin, Qingfeng Shi, Bingze Song, Lei Sun, Jing Tang, Ruitian Tian, Jun Wang, Jiahong Wu, Pengfei Zhang, Shen Zhang, and Jiashu Zhu. DreamX-World 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Batsford, 1978. ISBN 0713418605.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

Google. Gemini 3.5 Flash. Google AI for Developers documentation, 2026.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo DAI, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling Camera Control for Video Diffusion Models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 100433–100464, 2025.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646. Curran Associates, Inc., 2022. doi: 10.52202/068431-0628.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, Kalyan Sunkavalli, Feng Liu, Zhengqi Li, and Hao Tan. RELIC: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A Ross, Bryan Seybold, and Lu Jiang. VideoPoet: A Large Language Model for Zero-Shot Video Generation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25105–25124. PMLR, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 15984–16009. Curran Associates, Inc., 2025a. doi: 10.52202/085713-0540.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212*, 2025b.

Lightricks. LTX-2.3 and LTX Desktop: Production-ready engine. designed to be built on. LTX Blog, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095, 2025.

Fangyuan Mao, Aiming Hao, Jintao Chen, Dongxia Liu, Xiaokun Feng, Jiashu Zhu, Meiqi Wu, Chubin Chen, Jiahong Wu, and Xiangxiang Chu. Omni-Effects: Unified and spatially-controllable visual effects generation. *arXiv preprint arXiv:2508.07981*, 2025a.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025b.

OpenAI. GPT-5.6-Sol. OpenAI API changelog, 2026a.

OpenAI. GPT-Image-2 model. OpenAI API documentation, 2026b.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988. doi: 10.1016/0306-4573(88)90021-0.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 2023.

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

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

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 73754–73776, 2025.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2023.

Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Qingyan Bai, Ka Leong Cheng, Yue Yu, Yixuan Li, Yihao Meng, Zichen Liu, et al. WorldDirector: Building controllable world simulators with persistent dynamic memory. *arXiv preprint arXiv:2607.02517*, 2026.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989. doi: 10.1162/neco.1989.1.2.270.

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

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024. doi: 10.1145/3626772.3657878.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. LongLive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Xu Bin, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025b.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 47455–47487. Curran Associates, Inc., 2024. doi: 10.52202/079017-1505.

Yuanyang Yin, Gongxuan Wang, Yifan Zhan, Chuanhao Li, Kaipeng Zhang, and Feng Zhao. Alaya-EVOKE: From linear-scaling supervision to endless world. *arXiv preprint arXiv:2608.13546*, 2026.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. WBench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rCM: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. *arXiv preprint arXiv:2606.25473*, 2026.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *International Conference on Machine Learning*, 2026.

APPENDIX OVERVIEW

This appendix is organized into five sections. Section A documents the dataset and its filtering protocol. Section B specifies the Gameplay Harness and presents qualitative application examples. Section C gives the training, prompt-construction, and evaluation details needed to reproduce the system and its measurements. Section D reports additional benchmark results and analyses of event breadth and control–event composition. Section E describes the human evaluation protocol and its detailed results.

A ACTMOSAIC DATASET DETAILS

Table 3 summarizes the number of clips in the gameplay and event streams and the number of annotated events in each event category. Table 4 reports the fraction of held-out generated candidates retained by the motion filter and the character-root and camera drift measured on the retained clips.

Table 3: ActMosaic data statistics.

Source	Clips	Events	Primary Supervision
Gameplay	118,898	–	locomotion and camera
Environment	1,608	4,827	world-state change
Character	1,783	3,777	character motion
Interaction	1,299	4,969	character–world relation
Event synthesis total	4,690	13,573	free-form event semantics

Filter audit. We audit the motion filter separately for the environment, motion, and interaction event streams on held-out generated candidates. For each stream, we report the fraction kept and the root and camera drift of the kept clips. Since we emphasized in the prompt to the video generation model that "the camera and character should not move," most of the generated results already have no camera motion or character displacement, so the filtering rate is not very high.

Table 4: Data-filter audit under the motion filter, per event stream. Keep% is the fraction of generated candidates that pass; root and camera drift are normalized displacement scores over the kept set (0 = static/in-place, 1 = obvious displacement; lower is better).

Stream	Keep %	Root Drift ↓	Camera Drift ↓
Environment	83.6	0.12	0.02
Character	93.8	0.05	0.02
Interaction	77.6	0.15	0.08

B GAMEPLAY HARNESS DETAILS

B.1 GAMEPLAY HARNESS SCHEMA

The game-script schema contains entities, relations, persistent state, key-to-skill bindings, variants, preconditions, atomic events, effects, prompt projections, and an observer contract. The observer registry records each detector, tracker, and VLM tool's supported predicates and returns normalized read-only facts with confidence and frame indices. Separate task definitions and prompts specify the compiler, observer tools, and state-and-skill machine; qualitative application examples are provided in Appendix B.2.

Case 1 Cyber Samurai vs. Periodic Combat Drone

User request

"A black cyber samurai carries an energy katana in a future city, viewed from a stable third-person camera. Start without enemies. Every 12 seconds, spawn one hovering combat drone in front of the character. Skill 1 is a blue energy slash; against a drone, it locks on and destroys it. Skill 2 is an EMP that disables and drops the drone. Each defeat increases the kill count. After two defeats, permanently add purple lightning and unlock Skill 3, Lightning Flash Slash."

Parsed setting

Identity Black cyber samurai; stable third-person camera; futuristic-city visual style.

Entities The controlled samurai and at most one hovering combat drone.

Persistent state `drone_present`, `kill_count`, and `lightning_empowered`; initialized as no drone, zero kills, and no lightning.

Prompt projections Drone presence inserts/removes the drone description; the kill count records progression; lightning empowerment adds persistent purple electric arcs to the character.

Compiled state machine

Current state	Trigger and guard	Next state / visible result
No drone; 0 kills; no lightning	12-second timer	Drone becomes visible in front of the character.
Drone visible; 0 kills	User invokes Skill 1 or Skill 2, and the observer reports a visible drone with confidence ≥ 0.75	Drone is removed; kill count becomes 1.
No drone; 1 kill; no lightning	12-second timer	Drone becomes visible in front of the character.
Drone visible; 1 kill	User invokes Skill 1 or Skill 2, and the observer reports a visible drone with confidence ≥ 0.75	Drone is removed; kill count becomes 2; permanent purple lightning is enabled and Skill 3 unlocks.

Skill slots selected from the current state

1. Blue Energy Slash
No drone: a forward swing. Drone visible: lock on, strike, and destroy the drone in sparks.

2. Electromagnetic Pulse
No drone: emit a purple pulse. Drone visible: disable the drone and make it fall.

3. Lightning Flash Slash
Locked until the second defeat. Once unlocked: dash forward with a purple lightning slash.

Interactive rollout

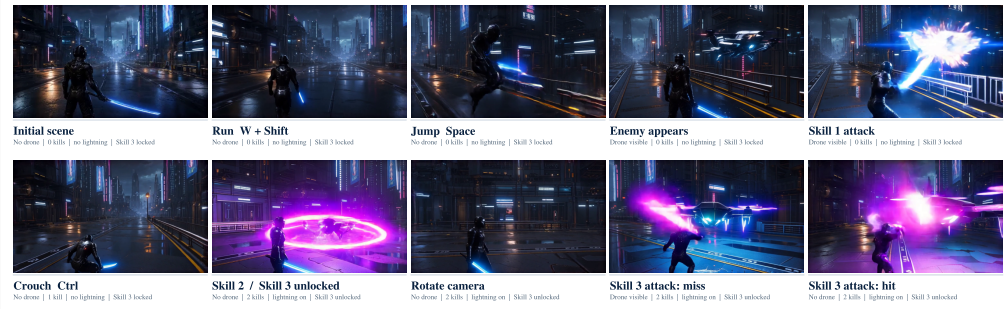

Figure 5: **Harness showcase: one complete cyber-samurai case.**

Case 2 Winter Survival: Building a Campfire

User request

"In a snowy scene, place a stone pile on the left and a tree on the right. The character holds an axe. Start a backpack with only the axe and offer chop-tree and collect-stone skills. Chopping a tree adds wood; collecting a stone adds flint; together they activate a campfire skill."

Parsed setting

Identity A dark-clothed adult holding an axe; stable third-person camera; realistic cold, snowy scene.

Entities The controlled character, one standing conifer tree, and one available stone pile.

Persistent state `backpack.{axe,wood,flint}`, `world.tree_felled`, and `campfire.stage`; initially only the axe is present, the tree stands, and the campfire is `none`.

Prompt projections Inventory changes alter the carried-item description; tree state switches the scene from standing to felled; campfire stage changes from clear snow to a woodpile and then a lit fire.

Compiled state machine

Current state	Trigger and guard	Next state / visible result
Axe only; tree standing; no flint	User invokes Slot 2 and the VLM observes *near* stone pile (confidence ≥ 0.75)	Flint enters the backpack; the stone pile visibly loses one stone.
Axe + flint; tree standing; no wood	User invokes Slot 1 and the VLM observes a standing tree and *near* tree (each confidence ≥ 0.75)	Tree falls in the snow; wood enters the backpack.
Axe + wood + flint; tree felled; campfire `none`	User invokes Slot 3	A woodpile appears beside the character; `campfire.stage` becomes `woodpile`.
Same inventory; campfire `woodpile`	The final atom of the same Slot 3 skill commits	The woodpile ignites; `campfire.stage` becomes `lit`.

Skill slots selected from the current state

1. Chop Tree
With recent VLM evidence of a nearby standing tree, swing the axe, fell it, and add wood; otherwise, swing briefly with no state change.

2. Collect Flint
With recent VLM evidence of a nearby stone pile, collect flint; otherwise, look around with the axe and keep inventory unchanged.

3. Light Campfire
Requires wood, flint, and `campfire.stage=none`. First create a woodpile, then light that same pile.

Interactive rollout with visual resume

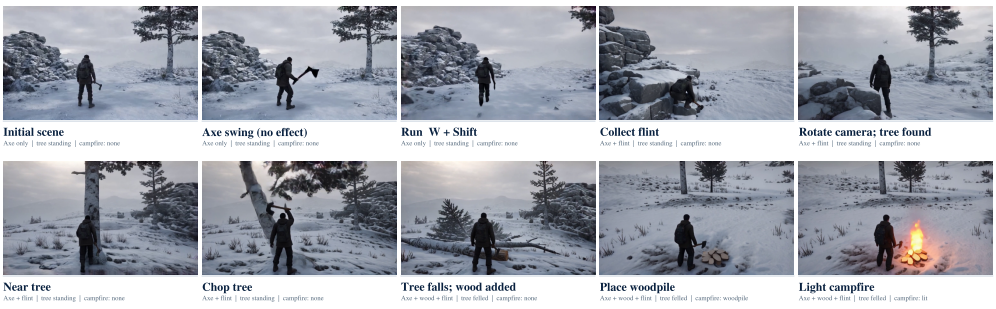

Figure 6: **Harness showcase. Winter Survival: Building a Campfire.**

B.2 Gameplay Harness Application Showcase

We use two saved worlds as concrete end-to-end examples. Figures 5 and 6 present each case in one card: the user request, its compiled representation, the state-controlled skill behavior, and the corresponding interactive trajectory.

C Implementation Details

C.1 Training Details and Hyperparameters

Training begins with autoregressive adaptation and four-step consistency distillation, followed by two sequential DMD stages: 480p DMD and joint 360p+720p DMD. First, we initialize from a pretrained LTX video diffusion transformer and adapt it into a block-autoregressive model with teacher forcing. Each target block conditions on the initial frame, a bounded sequence of preceding latent blocks, and the time-aligned interaction prompt. Gameplay and event samples are drawn according to a fixed data mixture. We then distill the autoregressive model to four denoising steps through consistency distillation.

In the first DMD stage, we train at 480p (832×480) over 721-frame videos. A bidirectional teacher separately trained at this resolution and length provides the target score; the four-step student is trained on its own causal histories, with a fake-score model tracking the student distribution. This stage establishes long-horizon generation at 480p. In the second DMD stage, we jointly distill the 360p low-resolution rollout, latent upscaler, and 720p one-step refinement branch on spatially aligned 360p (608×352) and 720p (1216×704) views of 241-frame videos. A second bidirectional teacher, separately trained at 720p (1216×704) and 241-frame length, supervises the high-resolution output. This stage improves visual quality at 720p while building on the long-horizon capability acquired in the 480p stage. At inference, the resulting schedule uses four low-resolution denoising evaluations followed by one high-resolution refinement evaluation per autoregressive block.

We use AdamW (Loshchilov & Hutter, 2019) for the optimization stages. Table 5 summarizes the implementation settings.

C.2 Evaluation Protocol

Control. We manually inspect the benchmark trajectories and favor scenes in which the requested motion has a clear traversable corridor, avoiding initial views dominated by walls, large foreground obstacles, or constrained terrain. Camera motion is estimated over the complete video and compared with each requested translation or rotation direction. Direction similarity is clipped to $[0, 1]$; translation below 0.02 and rotation below 3° are treated as unobserved. Character control is judged independently on frames sampled at 8 FPS against the time-aligned interaction request. Because engine geometry and free-space masks are not available consistently across the compared systems, these metrics do not evaluate obstacle avoidance; the collision-aware examples in Figure 1 are qualitative demonstrations.

Events. Each authored event interval is evaluated with two seconds of context on either side, sampled at 4 FPS. The judge receives the event description but not the hidden inner boundary, predicts a completion score and the observed start/end time, and is held fixed across methods. We map the prediction back to the original 24-FPS timeline before computing temporal IoU against the authored interval. The completion rubric is: 0 for absent or unrelated behavior; 1 for a localizable but incorrect or severely malformed response; 2 for a recognizable event missing a required component or result; and 3 when the required action, entities, effects, result, and causal order are all present. When the completion score is zero, we assign temporal IoU zero even if the judge returns a spurious interval. Accordingly, the nonzero Response Rate reported in the paper captures response/localization frequency rather than successful completion; Table 6 exposes the full ordinal distribution.

Quality. Quality is the unweighted mean of aesthetic quality, imaging quality, temporal flicker, dynamic degree, motion smoothness, and HPSv3 (Ma et al., 2025).

Aggregation. Quality and control metrics are first computed at the case level and then macro-averaged so that long videos do not receive disproportionate weight. The three groups in each

Table 5: **Training and model hyperparameters.**

Setting	Value
Architecture and inference	
Pretrained backbone	LTX 2.3 w/o audio branch
First DMD training size	480p (832×480)
Dual-resolution LR size	360p (608×352)
Dual-resolution HR size	720p (1216×704)
480p teacher / DMD clip	721 frames
720p teacher / joint DMD clip	241 frames
Inference schedule	4 LR + 1 HR
Autoregressive block length	2
LR / HR context length	1+8 / 1+2
Full-Step AR Optimization	
Optimizer	AdamW
Base learning rate	3×10^{-5}
Weight decay	0.01
Global batch size	64
Warm-up and LR schedule	5-step warm-up (factor $1/3$), then constant
Gradient clipping	None
Numerical precision	BF16
Training steps	15000
Bidirectional Teacher Optimization	
Optimizer	AdamW
Base learning rate	3×10^{-5}
Weight decay	0.01
Global batch size	64
Warm-up and LR schedule	5-step warm-up (factor $1/3$), then constant
Gradient clipping	None
Numerical precision	BF16
Training steps	15000
DMD Optimization	
Optimizer	AdamW (student and critic)
Base learning rate	2×10^{-6} / 4×10^{-7} (student / critic)
Weight decay	0.01
Global batch size	64
Warm-up and LR schedule	5-step warm-up (factor $1/3$), then constant
Gradient clipping	10.0 / 10.0 (student / critic)
Numerical precision	BF16
Training steps (480P / Dual-resolution)	2,000 / 1,000
Data and training budget	
Gameplay / event mixture	55 / 45
Training hardware	80 GB GPUs (989 TFLOPS in FP16 / BF16)

Event+Wander case and the four groups in each Event-only case are the elementary event observations, giving 300 groups in total. Average completion, Response Rate, and IoU_{all} weight these groups equally; IoU_{all} includes score-zero events through the zero-IoU convention above.

Automatic Seen/Unseen partition. We build the retrieval index exclusively from event captions that are present in the task-specific training mixture. Before retrieval, structured category and duration tags are removed so that they cannot create trivial matches. For benchmark event i and training event j, we compute BGE cosine similarity $s_{\text{sem}}(i, j)$ and TF–IDF cosine similarity $s_{\text{lex}}(i, j)$. We label i

Table 6: **Distribution of event-completion scores over all 300 event groups.** Each entry gives count (percentage). Score 0 denotes absent or unrelated behavior, score 1 a localizable but incorrect or severely malformed response, score 2 a recognizable but incomplete event, and score 3 a complete event. Response Rate in the main paper equals the combined percentage of scores 1–3.

Method	Score 0	Score 1	Score 2	Score 3
Yume-1.5	183 (61.00%)	15 (5.00%)	37 (12.33%)	65 (21.67%)
DreamX-World	252 (84.00%)	4 (1.33%)	10 (3.33%)	34 (11.33%)
AlayaWorld	262 (87.33%)	5 (1.67%)	12 (4.00%)	21 (7.00%)
LingBot-World-V2	177 (59.00%)	27 (9.00%)	35 (11.67%)	61 (20.33%)
Alaya-Evoke	206 (68.67%)	23 (7.67%)	34 (11.33%)	37 (12.33%)
OpenWorldAct	84 (28.00%)	14 (4.67%)	33 (11.00%)	**169 (56.33%)**

as Seen if there exists a single training event j for which both $s_{\text{sem}}(i,j) \geq \tau_{\text{sem}}$ and $s_{\text{lex}}(i,j) \geq \tau_{\text{lex}}$; otherwise it is Unseen. The two thresholds are fixed using training-only similarity distributions, before inspecting any benchmark output or evaluation score, and the same rule is applied to all 300 events. This conservative conjunction prevents a high score caused only by generic lexical overlap or broad semantic relatedness from marking an event as Seen. GPT-5.6-Sol (OpenAI, 2026a) subsequently reviews each assignment using only the benchmark caption, retrieved training captions, similarity scores, and the predefined semantic-equivalence rubric. The reviewer does not receive generated videos, method identities, or evaluation results; disagreements are resolved by applying the conservative rule that an event remains Unseen unless semantic equivalence is established. This procedure produces 110 Seen and 190 Unseen events; retrieval scores, nearest neighbors, review rationales, and final labels are released with the benchmark.

C.3 Prompt Construction

Our prompt is assembled from three pieces in a fixed order: the current game control, the active event(s), and the clip caption. Keyboard and mouse input is first translated with a game-specific action lexicon. For example, a control piece may describe the character walking forward or the camera orbiting left; the idle input also has its own explicit description. An event annotation provides a category, a free-form caption, and its half-open time span $[s_i, e_i)$. Its text is formatted as `[category] caption` and is included only while the event is active.

Concretely, the assembled prompt has the following form (the bracketed labels are shown in English for readability):

```
[movement and camera] <current control phrases>, [<event
category>] <active event caption>, ....  <clip caption>
```

All simultaneously active control phrases are combined into the first piece, while each event keeps its own category–caption string. A missing control or event piece is simply omitted. The clip caption is appended to every version of the prompt.

We encode each distinct temporal version of the prompt once and reuse its embedding wherever that version recurs. During generation, the model uses the version valid for the current time span. Therefore, an event caption conditions only its annotated interval, whereas the clip caption remains present throughout the rollout.

D Additional Experiments

D.1 Quality Breakdown on Our Benchmark

Table 7 breaks down the quality scores for all 120 cases in our benchmark, including Wander, Event+Wander, and Event-only cases. OpenWorldAct obtains the highest dynamic score and the second-highest motion smoothness and aggregate quality scores; Alaya-Evoke has the highest aggregate quality score.

Table 7: **Detailed quality results on our 120-case benchmark.** All values are percentages; higher is better.

Method	Perceptual Quality						Mean ↑
	Aesthetic	Imaging	Temporal	Dynamic	Motion	HPSv3	
Yume-1.5	61.25	**69.85**	95.22	76.67	98.06	67.61	78.11
DreamX-World	61.48	66.84	94.75	75.83	97.66	68.77	77.56
AlayaWorld	61.71	67.89	**96.07**	73.33	97.54	66.41	77.16
LingBot-World-V2	63.53	66.89	94.74	76.67	97.35	73.08	78.71
Alaya-Evoke	**66.99**	68.03	95.48	90.00	**98.33**	**73.84**	**82.11**
OpenWorldAct	62.05	67.66	95.71	**90.83**	98.30	64.12	79.78

Table 8: **Detailed results on 67 third-person-view test cases, a subset of public WBench.** All values are percentages; higher is better.

Metric	Yume 1.5	DreamX-World	Alaya World	LingBot-World v2	Alaya-Evoke	OpenWorldAct (ours)
Video Quality						
Aesthetic	67.28	65.48	66.79	67.80	**69.08**	67.75
Imaging	**70.35**	66.99	67.60	68.06	68.45	70.32
Flickering	94.52	94.02	95.79	93.56	94.83	**95.92**
Dynamic	83.58	62.69	66.67	**85.07**	**85.07**	76.12
Smoothness	97.67	97.60	97.79	96.78	97.99	**98.56**
HPSv3-Norm	75.17	73.59	69.54	**78.98**	78.10	75.90
Quality avg. (6)	81.43	76.73	77.36	81.71	**82.25**	80.76
Setting						
Scene	65.81	**71.40**	53.49	60.23	70.70	56.98
Subject	**89.70**	84.93	88.25	85.97	87.76	84.18
Setting avg. (2)	77.75	78.16	70.87	73.10	**79.23**	70.58
Interaction						
Navigation	70.57	78.37	74.52	**84.43**	69.60	81.65
Event-Edit	58.86	50.29	25.83	61.81	54.29	**74.38**
Subject-Action	57.45	45.79	37.29	49.67	49.25	**58.73**
Interaction avg. (3)	62.29	58.15	45.88	65.30	57.71	**71.59**
Consistency						
Background	92.62	91.82	**94.29**	92.47	91.75	93.32
Spatial	80.27	78.86	**83.94**	81.73	79.36	80.91
Gated Spatial	78.60	78.07	**82.14**	79.48	79.16	79.06
Segment	**100.00**	**100.00**	98.41	92.54	**100.00**	98.51
Perspective	78.76	72.26	75.34	86.12	74.46	**90.52**
Subject	91.87	91.04	91.08	89.33	91.71	**92.14**
Geometric	88.86	86.64	93.39	88.61	89.09	**93.76**
Photometric	82.69	78.24	**84.56**	79.60	80.31	81.14
Consistency avg. (8)	86.71	84.62	87.89	86.23	85.73	**88.67**
Physical						
Causal Fidelity	60.77	60.32	60.62	63.40	60.11	**64.42**
Visual Plausibility	58.27	58.08	**61.69**	58.84	60.34	59.13
Physical avg. (2)	59.52	59.20	61.16	61.12	60.23	**61.78**
Overall (avg. of 5 dims)	73.54	71.37	68.63	73.49	73.03	**74.67**

D.2 Results on Public WBench

We additionally evaluate all six systems on 67 third-person-view test cases from the public WBench benchmark (Ying et al., 2026). These 67 cases form a subset of WBench, selected because OpenWorldAct is designed for third-person character control; we follow the public evaluation protocol for every method. This external evaluation is separate from our 120-case benchmark and contains none of its custom examples. We report the five evaluation dimensions and the overall score on this 67-case subset.

Table 9: **Event completion across semantic families and compositional strata.** All entries are mean completion on a 0–3 scale, regenerated from the same 300 per-event records used in Table 1. Both partitions cover all 300 events.

Breakdown	Groups	Yume-1.5	DreamX-World	AlayaWorld	LingBot-World-V2	Alaya-Evoke	OpenWorldAct
Event Family							
Environment change	60	0.3667	1.0667	0.0000	0.4667	0.3500	**1.3667**
Character action	180	0.8167	0.1722	0.3667	0.9556	0.5667	**1.9222**
Coupled action–effect	30	1.9667	0.8667	0.4667	1.1333	1.2000	**2.5667**
Target interaction	30	1.8667	0.1667	0.4000	1.5333	1.4333	**2.7333**
Compositional Stratum							
Simple non-interaction	45	0.5333	0.7556	0.5111	0.9778	0.4889	**1.8667**
Complex non-interaction	45	1.0444	0.8889	0.2444	1.2889	1.1556	**1.9556**
Interaction, one complex	45	1.0667	0.6444	0.2889	1.1333	0.7556	**2.4222**
Interaction, two complex	45	1.1556	0.1778	0.2444	0.8222	0.5333	**2.4444**
Event-only complex sequence	120	0.9417	0.1250	0.2833	0.7500	0.5833	**1.6333**

Table 10: **Matched inference-time control–event composition diagnostic.** All three settings use the same 60 Event+Wander cases. Event metrics cover the same 180 event groups; response rate, control accuracy, and IoU_{all} are percentages.

Inference Setting	Camera ↑	Character ↑	Response Rate	Completion (0–3) ↑	IoU_{all} ↑
Event only	–	–	90.5556	2.5167	72.9516
Wander only	76.1468	97.7114	–	–	–
Full: Event + Wander	75.2729	99.7647	78.3333	2.1722	56.2154

D.3 Event Breadth and Compositionality

The complete 300-event benchmark provides two complementary views of semantic breadth. The event-family partition separates environment changes, character actions, coupled action–effect processes, and target-conditioned character–object interactions. The compositional partition groups Event+Wander videos by whether their three-event schedule is simple, contains a complex event, or contains one or two complex target interactions, and includes the four-event Event-only sequences as a separate stratum. A complex event is decomposed into linked or overlapping atomic events; a target interaction additionally binds a character action to an object and an observable consequence.

As Table 9 shows, OpenWorldAct leads in every semantic family and every compositional stratum. The gain is especially pronounced for coupled action–effect processes (2.5667 vs. 1.9667) and target interactions (2.7333 vs. 1.8667), where a high completion score requires more than producing an isolated motion or visual effect. Its score remains 2.4444 when two complex interactions occur in the same rollout, compared with 1.1556 for the strongest baseline, and reaches 1.6333 on the newly added Event-only complex sequences, compared with 0.9417 for the strongest baseline. These results demonstrate breadth over held-out authored requests and robustness to composition; they do not assume that the underlying concepts were absent from generic video–language pretraining.

D.4 Matched Prompt-Composition Diagnostic

The central composition question is symmetric: an event prompt should not broadly override wander control, and concurrent wander should not broadly suppress event execution. We therefore evaluate the full model on the same 60 Event+Wander cases (180 event groups) under three inference settings. Event-only inference retains the authored event prompts and schedules but removes wander controls; Wander-only inference retains the authored keyboard and camera controls but removes event prompts; and full inference activates both prompt channels. The initial images, authored conditions, and evaluation protocol are otherwise unchanged.

Relative to Wander-only inference, the full setting changes camera and character accuracy by -0.8739 and $+2.0533$ percentage points, respectively. Relative to Event-only inference, it reduces event response rate by 12.2223 percentage points, mean completion by 0.3445, and IoU_{all} by 16.7362 points; equivalently, it retains 86.5030%, 86.3114%, and 77.0585% of the corresponding Event-only values. Because score-zero events receive zero completion and zero IoU, we can also separate



Figure 7: **Additional visual comparisons.** From top to bottom: character action and coupled action–effect process; character locomotion and environment-state change; character-material transformation and multi-stage action–effect event. Within each panel, rows show Yume-1.5, DreamX-World, AlayaWorld, LingBot-World-V2, Alaya-Evoke, and OpenWorldAct, respectively.

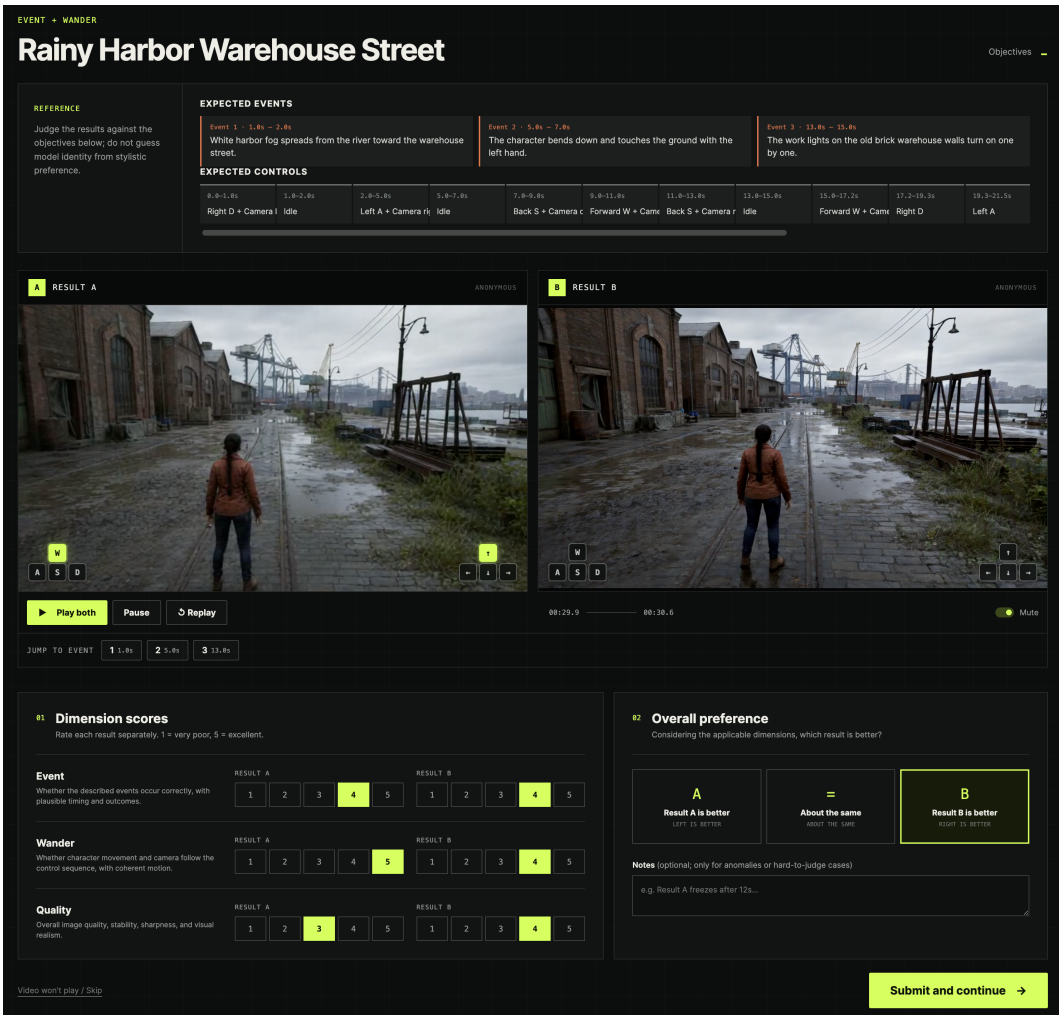

Figure 8: **User-study interface for an Event+Wander case.** Raters compare two anonymous videos against the same event and control schedule, score each video separately on Event, Wander, and Quality using five-point scales, and choose an overall preference or tie.

response frequency from the fidelity of nonzero responses. Conditional on a nonzero response, mean completion is 2.7792 for Event-only and 2.7730 for full inference, a difference of only -0.0062. Conditional temporal IoU is 80.5600% and 71.7644%, respectively, a difference of -8.7956 points. Thus concurrent wander primarily introduces an event-triggering and temporal-localization cost, while the semantic completeness of events that do occur is nearly preserved. This diagnostic establishes effective but imperfect composition without paired control–event training data, rather than claiming interference-free execution.

D.5 Additional Visual Comparisons

Figure 7 provides six further event-level comparisons from the benchmark. They span character actions, coupled action–effect processes, locomotion, an environment change, and a character-material transformation. Each event is shown at its start, midpoint, and late-stage sample using the native frame rate of each system. Prompt text is shortened for display.

Table 11: **Pairwise user-study comparison matrix.** Each off-diagonal cell reports win/tie/loss counts for the row method against the column method; the three counts sum to the number of judgments for that pair. Dashes mark self-comparisons. DreamX and LingBot abbreviate DreamX-World and LingBot-World-V2 in the column headers.

Row / Column	AlayaWorld	DreamX	LingBot	Alaya-Evoke	Yume-1.5	OpenWorldAct
AlayaWorld	–	7/4/4	0/0/15	0/4/11	6/2/4	0/0/15
DreamX-World	4/4/7	–	0/3/12	1/2/9	3/4/5	1/0/12
LingBot-World-V2	15/0/0	12/3/0	–	11/1/2	10/1/1	2/5/7
Alaya-Evoke	11/4/0	9/2/1	2/1/11	–	8/2/1	1/0/13
Yume-1.5	4/2/6	5/4/3	1/1/10	1/2/8	–	0/0/9
OpenWorldAct	**15/0/0**	**12/0/1**	**7/5/2**	**13/0/1**	**9/0/0**	–

Table 12: **Human ratings by evaluation dimension.** Entries are mean five-point ratings; n is the number of ratings for each method.

Method	n	Event ↑	Wander ↑	Quality ↑
Yume-1.5	56	2.393	2.232	2.429
DreamX-World	67	2.149	1.701	2.134
AlayaWorld	72	1.472	1.861	3.014
LingBot-World-V2	70	2.943	3.786	**3.714**
Alaya-Evoke	66	2.470	2.500	3.561
OpenWorldAct	**65**	**3.815**	**4.246**	3.708

E User Study Details

We collect 198 pairwise judgments from 13 raters. Each task presents an anonymous pair of system outputs under matched initial conditions and randomizes presentation side. Raters answer which video better realizes the complete requested interaction, considering both schedule following and visual quality; ties are allowed. Each video is also scored independently along three dimensions on a five-point scale: *Event* measures whether the requested semantic events occur with the correct entities, effects, and temporal order; *Wander* measures adherence to the requested character and camera controls throughout the rollout; and *Quality* measures visual fidelity, temporal coherence, and artifact severity. We compute Elo with an initial rating of 1000 and $K = 32$, and report a tie-aware win rate that counts a tie as one half win. Because the sampled comparisons are not balanced across system pairs, we report the number of judgments for every comparison and the number of ratings contributing to each system-level dimension score. Figure 8 shows the interface used for these pairwise and dimension-specific judgments.

Table 12 separates semantic execution and low-level control from general visual preference, making it possible to identify whether a pairwise advantage comes from interaction fidelity or appearance alone.