SEERWORLD: TOWARDS OUT-OF-SIGHT WORLD EVOLUTION IN GENERATIVE VIDEO MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive video world models offer a promising route toward world simulation by enabling camera-controlled exploration of evolving visual environments, while aiming to maintain consistency across changing viewpoints. Faithful world simulation, however, requires more than preserving what has previously been observed: entities should remain participants in the world's dynamics, interacting with events and changing in response even outside the observer's field of view. Current models struggle throughout this process: a mid-rollout prompt switch may fail to instantiate the new event; when the camera later revisits an earlier region, the subject and background may drift or disappear; and even when the subject is preserved, it often remains behaviorally static rather than reacting to or interacting with the event, making the resulting world evolution behaviorally and causally implausible. We formalize this challenge as *Out-of-Sight Reactive Evolution* and introduce the SeerWorld-Data along with SeerWorld-Bench. As a baseline reference solution, we develop SeerWorld, a segment-autoregressive video model trained on our proposed data, together with a reasoning-augmented generation framework, which autonomously reasons about how the off-screen world evolves and renders the resulting dynamics. Evaluation results show that our approach outperforms recent state-of-the-art models in event-conditioned interaction rendering while preserving revisit consistency, establishing a reference point for autonomous world evolution beyond the current field of view. The model and data will be made publicly available.

1 INTRODUCTION

Building models that can simulate evolving environments has become an important research direction in generative AI. By learning world dynamics, a world model uses observed history and input actions to predict plausible future states (Team et al., 2026b; Gao et al., 2026; Zhang et al., 2026; Assran et al., 2025). Such predictive representations have enabled progress in decision making and autonomous driving (Feng et al., 2025; Zhou et al., 2025; Liang et al., 2025; Gao et al., 2024b;a), agent training and embodied intelligence (Li et al., 2025; Wang et al., 2025; Hafner et al., 2025), and synthetic environment generation (Bruce et al., 2024; Brooks et al., 2024).

Generative video models (Wan et al., 2025; HaCohen et al., 2026; Yang et al., 2025c; Wu et al., 2025) provide a direct approach to visual world modeling (Team et al., 2026b; He et al., 2025; Team et al., 2026a): their predictions are immediately observable, while text and camera controls offer intuitive interaction. Yet locally plausible frames alone do not yield a coherent long-horizon world. A model must preserve the environments it has traversed and the events that have occurred; otherwise, later states become disconnected from the history that should have produced them. Accordingly, recent video world models have explored various mechanisms for long-horizon memory and consistency, including recurrent state-space memory (Zhu et al., 2026), memory retrieval or dynamic frame eviction (Yu et al., 2025; Yang et al., 2026), and geometry-grounded spatial memory (e.g., 3D point clouds) (Wu et al., 2026; Team et al., 2026a; Ma et al., 2026a).

Most existing approaches nevertheless tend to equate successful memory with restoring a previously observed state. Such persistence is necessary but insufficient: a world does not freeze outside the observer's field of view. New entities may enter, environmental conditions may change, and intervening events may cause an off-screen subject to flee, recoil, defend itself, or seek cover. Revisiting

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

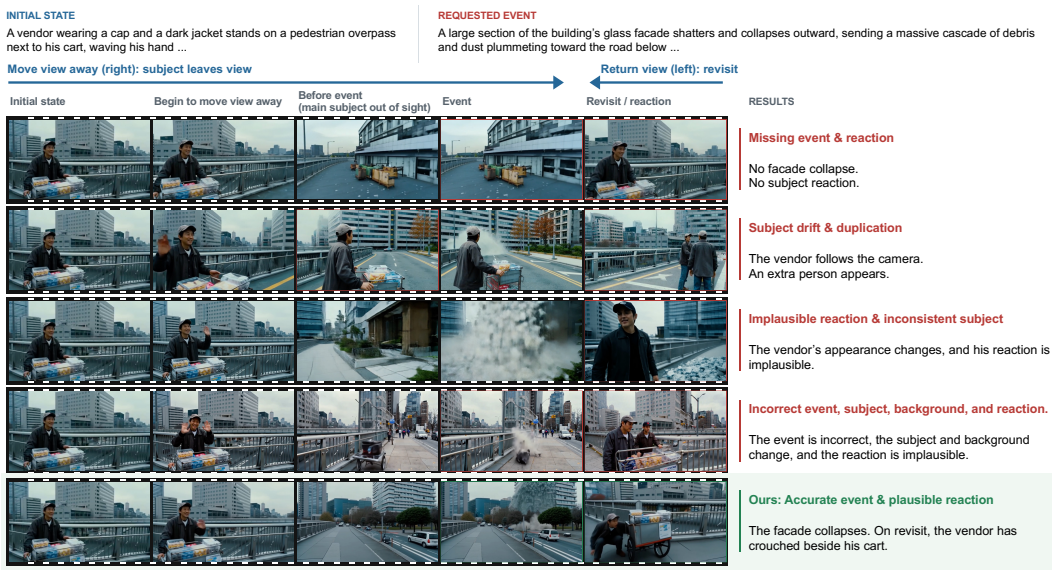

Figure 1: Challenges in *Out-of-Sight Reactive Evolution*. The camera first observes a subject (a vendor and his handcart), moves away to reveal a newly requested event (a collapsing building facade), and then moves back to the original scene. Previous methods (rows 1–4) exhibit various failures. In contrast, SeerWorld (row 5) realizes the event, revisits the original subject, and renders a causally plausible reaction while preserving subject and scene consistency.

that subject should therefore reveal not only consistent appearance but also the consequences of what occurred while it was unseen. We define this capability as *Out-of-Sight Reactive Evolution*, a concrete event-conditioned setting for studying the problem of world evolution. The camera leaves a subject, an event occurs elsewhere in the evolving world, and a later revisit must reveal a causally plausible change that is inferred from the realized rollout rather than prescribed in the initial prompt.

Out-of-Sight Reactive Evolution couples three challenges that existing models rarely address simultaneously: preserving subject and scene consistency across camera motion; realizing an event introduced by a mid-rollout prompt switch; and rendering a causally plausible subject response when the camera returns. As illustrated in Fig. 1, HY-World 1.5 (HunyuanWorld, 2025), Echo-WM (Zhang et al., 2026), LingBot-World (Team et al., 2026b), and SANA-WM (Zhu et al., 2026) exhibit various failures: the event or reaction may be absent, the subject may follow the camera or become duplicated, appearance and background may drift, or the subject may exhibit implausible behavior. Appearance persistence, instruction following, and event-conditioned behavior must therefore be measured jointly. Therefore, we introduce SeerWorld-Data, a controlled generative dataset of 12k+ long videos ($\geq$ 10 seconds). The training set spans 19 event families and 65 fine-grained sub-event types, 16 scene families and 63 sub-scene types, and 14 reaction types across photorealistic and game-world domains. Segment-level prompts and temporal annotations provide direct supervision for event realization, long-range visual continuity, and behavioral change.

As a reference solution, we develop SeerWorld (**S**ubject-and-**E**vent **E**volution via **R**easoning), a reasoning-and-retrieval-augmented framework that separates world-state reasoning from visual rendering. A VLM serves as a *brain* that continuously perceives and records the evolving world state, while the *renderer* generates the next segment based on the reasoning from the *brain*. We train the renderer on SeerWorld-Data with Identity-Anchored Attention and a two-stage strategy that first learns reactive rendering from clean data and then reduces exposure bias through online autoregressive rollout training. We further introduce SeerWorld-Bench, which evaluates whether the camera revisits the original region, the switched prompt realizes the requested event, the subject responds plausibly, and the subject and background remain consistent. These coupled capabilities are not captured by existing video benchmarks such as VBench (Huang et al., 2024), which primarily emphasize perceptual quality and short-range temporal behavior.

Our contributions are threefold: (i) We formulate *Out-of-Sight Reactive Evolution*, which requires revisit consistency, mid-rollout event realization, and causal reactive evolution. (ii) We introduce SeerWorld, a brain-renderer collaboration framework aimed to continuously perceive and render

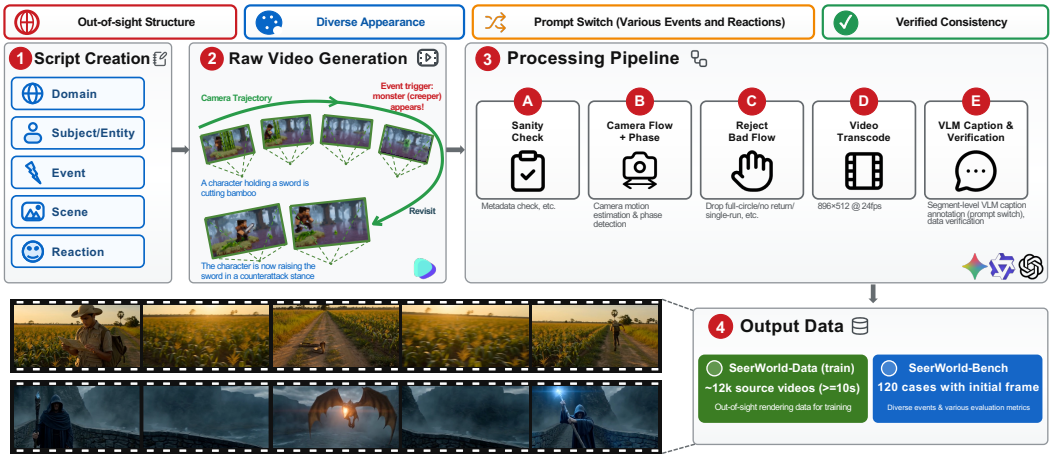

Figure 2: Overview of the data engine for SeerWorld. Structured scenarios are rendered into single-take out-of-sight interactions, verified against visual evidence, temporally annotated, and converted into a training supervision dataset and evaluation benchmark.

the out-of-sight world dynamics. (iii) We propose SeerWorld-Data, which contains more than 12k videos spanning diverse scenes, events, and reactions, and introduce the SeerWorld-Bench that measures the *Out-of-Sight Reactive Evolution* capability. We hope this work holds the potential to advance the research of out-of-sight dynamics and causal evolution in video world modeling.

2 Related Work

Video world models & autoregressive video generation. Interactive video world models extend generative video models into controllable, long-horizon simulators. Representative systems include LingBot-World (Team et al., 2026b; Gao et al., 2026), Echo-WM (Zhang et al., 2026), and Alaya-EVOKE (Yin et al., 2026). A closely related training direction is segment-autoregressive generation, which preserves bidirectional attention within each generated segment while maintaining autoregressive dependence across segments (Rui et al., 2026). We build on this formulation and focus on whether the generated world continues to evolve beyond the current field of view. Additional background is provided in Appendix Section A.

Out-of-sight dynamics. Recent work has begun to test whether unobserved states continue evolving across camera movement, occlusion, or other observation gaps (Ma et al., 2026b; Xu et al., 2026). ReMind (Xu et al., 2026), LiveWorld (Duan et al., 2026), and Hybrid Memory (Chen et al., 2026) address this problem through learned dynamic memory or explicit representations of static environments and dynamic entities. These methods mainly study persistence or continuation of an existing state or process. Our setting additionally introduces new events after the subject leaves view and requires the revisited subject to exhibit an inferred event-conditioned change, jointly testing prompt switching, long-range consistency, and world-evolution reasoning.

3 SeerWorld: Data and Method

3.1 Problem & Task Setting

A world evolves continuously rather than only where and when it is being observed. Its entities should therefore remain active participants in that evolution: as new events occur and environmental conditions change, they should autonomously interact with those changes and exhibit appropriate responses, including while they are outside the current field of view. Leaving the frame should not freeze an entity or remove it from the unfolding dynamics—it is *out of sight, but not out of the world*. We therefore formulate *Out-of-Sight Reactive Evolution* as a task for measuring whether a video world model or generation system can preserve an off-screen entity, propagate the consequences of intervening events to it, and render the resulting reaction when the entity is revisited.

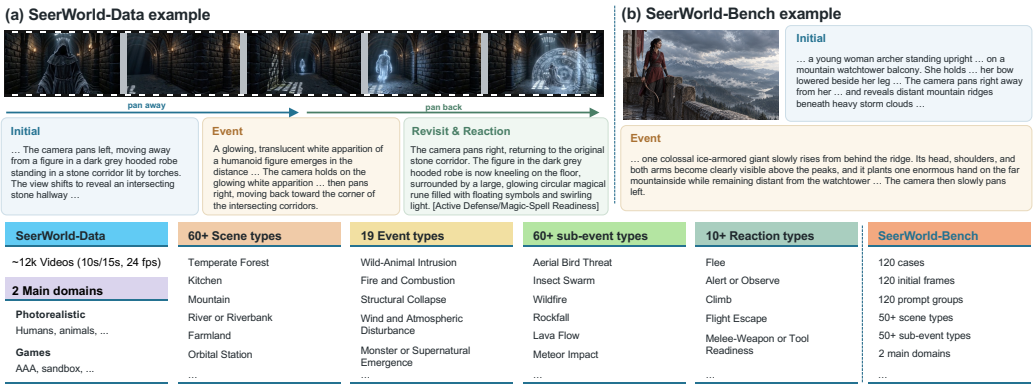

Figure 3: Examples and statistics of SeerWorld-Data and SeerWorld-Bench. The upper part illustrates representative training and benchmark examples, while the lower part summarizes their scale and scenario diversity.

Specifically, as shown in Fig. 1, an *Out-of-Sight Reactive Evolution* rollout can be abstracted into at least three semantic parts (each corresponding to at least one prompt): (1) $\mathcal{P}_{\text{init}}$ (*initial*): the part that covers the initial state of the scenario and the main subjects. Condition p_{init} specifies the subjects and scene, and the camera remains static or begins to move away from the subject. (2) $\mathcal{P}_{\text{event}}$ (*event*): the part that introduces an event after the initial state, where the subject has left the frame because the camera view has moved away. Condition p_{event} introduces an event (e.g., "something is on fire"), and the camera may remain static, continue moving away to show the event, or begin moving back toward the original region. (3) $\mathcal{P}_{\text{revisit}}$ (*revisit*): the part in which the camera returns to the initial region and the subject reappears with a reaction to the event. The ground-truth condition $p_{revisit}$ specifies this reaction, whereas the benchmark provides a blind revisit instruction without prescribing the concrete behavior. These parts describe semantic roles rather than a fixed number of generated segments. Each part may span one or more autoregressive chunks/segments. We use this three-part abstraction to isolate the minimal causal chain while allowing longer rollouts and additional intermediate observations in real-world scenarios. A system may internally infer a detailed control condition from the generated history, as our *brain* does instead of pre-specifying the eventual reaction in the initial condition. The evaluation targets the complete event-to-reaction chain: (1) the switch from p_{init} to p_{event} actually instantiates the requested event; (2) the camera leaves and later revisits the original subject and its corresponding region (not necessarily the exact initial frame or viewpoint, since the subject may have moved during the out-of-sight interval); (3) at revisit, the subject preserves its identity and core visual appearance without severe drift, and the surrounding scene and background remain reasonably consistent; (4) the subject exhibits a causally plausible state change or reaction to the realized event, with natural motion and visual quality.

3.2 SeerWorld-Data & Bench Construction

Fig. 2 shows the overall data construction pipeline of SeerWorld. We here introduce the main components of the pipeline. Additional construction details are provided in Appendix Section B.

SeerWorld-Data. (1) **Scenario design and video data generation.** Existing real-world video datasets rarely contain samples that align with our experimental setting. We therefore construct a controlled generative dataset. The curated data aims to make the event-to-reaction causal chain observable and separately auditable at each stage. Scenarios are represented by structured tuples containing subjects and their appearance, scenes, events, compatible reactions, and camera movements. We compose tuples under semantic compatibility constraints—for example, the selected events must admit a visually observable response from the chosen subjects—and render them into a single-take generation script with five beats: initial scenario with the main subject $\rightarrow$ move the camera view away $\rightarrow$ show event $\rightarrow$ move the camera view back $\rightarrow$ revisit and reveal reaction. The full scripts are written by a VLM (Gemini3.1-Pro (Team & DeepMind, 2026)). Videos are then generated by state-of-the-art video foundation models (Seedance et al., 2026; Team, 2026) based on the designed script. (2) **Verification and annotation.** Generated videos do not necessarily obey their scripts, so we verify visual evidence after generation. Specifically, we use Gemini3.1-Pro (Team & DeepMind,

2026) to check that each clip is a continuous shot, the designated subject remains visually consistent, the subject leaves and reappears, the event occurs in the intended view, the reaction is visible, and no severe hallucination invalidates the example. We further estimate global camera motion from Lucas–Kanade feature tracks using a RANSAC-based affine fit (Fischler & Bolles, 1981). To ensure the camera motion correctness, we reject full-circle camera motions, missing returns, single-direction motion, and unreliable direction estimates. We also ask the VLM to estimate when the designated subject becomes fully invisible and reappears, providing fine-grained metadata. For segment-wise supervision, we partition each video into multiple intervals and describe them jointly so that adjacent captions form a coherent progression. We audit these captions for future-event spoilers, references to an absent subject, malformed camera instructions, and excessive adjacent-caption overlap.

SeerWorld-Bench. (1) **Benchmark Construction**. We construct SeerWorld-Bench as a 120-case evaluation set spanning photorealistic (69 cases) and game/virtual (51 cases) domains. Its initial frames have two complementary sources: 66 are extracted from held-out videos disjoint from the training set, while 54 are generated with GPT-Image-2 (OpenAI, 2026). Each case is organized into three semantic parts: an initial state, an event, and a revisit. After semantic deduplication, the benchmark covers 50+ fine-grained scene types and sub-event types, as summarized in Fig. 3. Moreover, we conduct dedicated similarity screening against all training videos. For every benchmark case, we retrieve potentially similar training samples using both frame visual similarity and the lexical and embedding-based semantic similarity of the initial-event prompt pair. We then inspect retrieved candidate pairs with a VLM-based two-pass audit and human adjudication of the ambiguous cases. We remove any candidate judged highly similar, leaving no source-clip overlap, initial-event prompt-pair overlap, or human-confirmed high-similarity pair in the final benchmark. (2) **Evaluation Metrics.** We evaluate the complete rollout using Event Success, Revisit Success, Causal Plausibility, Reaction Quality, Visual Integrity, and revisit subject and background consistency. Detailed inputs, scoring criteria, and aggregation rules are provided in Appendix Section B.

3.3 METHOD

3.3.1 OVERVIEW

Most existing video world models provide little explicit support for interpreting what occurred during generation, maintaining a persistent account of entities outside the current view, or inferring how those entities should evolve before they reappear. We therefore introduce SeerWorld, a framework for *Out-of-Sight Reactive Evolution* that explicitly separates *world reasoning* from *visual rendering* (Fig. 4). A VLM *brain* observes every generated segment and records the realized entities, scene, and events. After an event occurs, it reasons over this history and predicts how the off-screen subject should have evolved. A segment-autoregressive *visual renderer* then generates the next segment from the inferred instruction, retrieved entity evidence, and recent generated latents. We equip the renderer with complementary entity and temporal anchors for long-range identity and short-range continuity. Identity-Anchored Attention (IAA) strengthens information flow from the entity anchor when the subject reappears. Our two-stage training first learns subject preservation and reactive rendering from clean ground-truth windows, and then performs online autoregressive rollout training from self-generated predecessor context. This Self-Forcing-style adaptation (Huang et al., 2026) narrows the train-inference gap and refines model performance.

3.3.2 ENTITY AND TEMPORAL ANCHORS

The entity anchor carries subject appearance across a long out-of-view interval. During training, it is selected from ground-truth frames in which the target subject is clearly visible. At inference time, the VLM inspects sampled frames from the generated history, identifies a clear and unoccluded observation of the target subject, and retrieves the corresponding latents. This helps the renderer maintain subject identity and appearance when the camera revisits. For temporal anchor, we adopt the four latents immediately preceding the target window, which participate in within-window full attention. This anchor is selected deterministically since recency is precisely the desired signal for view and motion continuity. Stage 1 takes these latents from the ground-truth videos, whereas stage 2 replaces them with the tail of the autoregressive rollout. At inference time they are always taken from the latest generated segment. The two anchors are therefore complementary: entity retrieval spans the long-term identity gap, while the temporal anchor bridges the local autoregressive boundary.

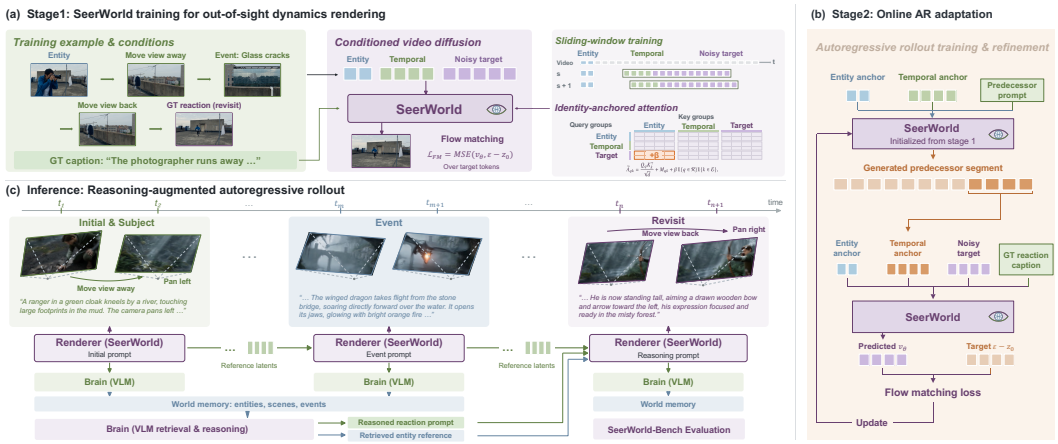

Figure 4: Overview of the SeerWorld framework. (a) Stage 1 learns out-of-sight reactive rendering. (b) Stage 2 refines the renderer through online autoregressive adaptation. (c) Inference uses brain–renderer collaboration for *Out-of-Sight Reactive Evolution*.

3.3.3 Identity-Anchored Attention

Although an appended entity reference participates in self-attention, it may not receive sufficient attention when the subject reappears: its contribution can be diluted by the much larger target and temporal context. We therefore introduce Identity-Anchored Attention (IAA), which makes the retrieved entity tokens a soft attention sink only for revisit samples. Let $\mathcal{R}$ denote target-query tokens in a revisit window and $\mathcal{E}$ the key tokens belonging to the entity anchor. For an attention head, IAA changes the pre-softmax score from A to

$$\widetilde{A}_{qk} = \frac{Q_q K_k^\top}{\sqrt{d}} + M_{qk} + \beta\, \mathbb{1}\{q \in \mathcal{R}\}\mathbb{1}\{k \in \mathcal{E}\}, \tag{1}$$

where M_{qk} is the base additive self-attention mask, $\beta > 0$ controls identity preference, and $\mathbb{1}\{\cdot\}$ denotes the indicator function. In full-attention video sequence, $M_{qk} = 0$ for valid token pairs and takes a large negative value only for masked or padded pairs. The IAA bias is directional: revisit tokens read from the entity anchor, while unrelated query-key pairs are unchanged. In implementation, we multiply the mask entries corresponding to the target-query–entity-key block by e^β before LTX converts the mask to log space, which is exactly equivalent to adding $+\beta$ in Eq. (1). IAA is enabled only for samples marked as revisit during training and for the revisit chunk during inference.

3.3.4 Two-Stage Autoregressive Training

Stage 1: Learning reactive evolution with clean anchors. For each sliding training window, the loader supplies both the entity anchor and the temporal anchor as conditioning. The training sequence is organized as [entity anchor | temporal anchor | target]. Both anchors remain clean at timestep zero and carry zero loss weight; only the target latents are noised and supervised by the flow-matching objective. IAA is activated on revisit windows. This stage teaches the renderer core capabilities such as event execution, event-conditioned reactions, and the roles of long-range identity and short-range temporal conditioning.

Formally, let $\mathbf{e}$, $\mathbf{c}$, and $\mathbf{z}$ denote the clean entity-anchor, temporal-anchor, and target latents, respectively; let p be the target prompt and $r \in \{0, 1\}$ the revisit indicator. For $\epsilon \sim \mathcal{N}(0, I)$ and a sampled noise level σ, we construct

$$\mathbf{x}^\sigma = [\mathbf{e} \mid \mathbf{c} \mid (1-\sigma)\mathbf{z} + \sigma\epsilon], \qquad \mathbf{s}^\sigma = \left[\mathbf{0}_{|\mathbf{e}|} \mid \mathbf{0}_{|\mathbf{c}|} \mid \sigma \mathbf{1}_{|\mathbf{z}|}\right], \tag{2}$$

where $\mathbf{s}^\sigma$ is the per-token timestep input, and $\mathbf{0}_n$ and $\mathbf{1}_n$ denote length-n all-zero and all-one vectors, respectively. Thus the two anchors remain clean at timestep zero, whereas every target token receives timestep σ and follows the flow-matching interpolation. The renderer v_θ predicts the target velocity

Table 1: Comparison results (%) on the SeerWorld-Bench. Best results are shown in bold and second-best results are underlined. *Core* denotes the proportion of cases concurrently achieving successful revisit, event realization, and full causal plausibility.

	Task metrics					Visual quality, consistency, and motion			
Method	Core	Revisit	Event	Causal	React.	Visual	Subj. Cons.	Bg. Cons.	Motion
LingBot-World (Team et al., 2026b)	15.8	92.5	77.5	29.2	31.3	68.8	87.6	94.6	98.0
LingBot-World-v2 (Gao et al., 2026)	20.0	76.7	68.3	27.1	25.4	76.7	82.6	92.5	98.4
YUME-1.5 (Mao et al., 2026)	15.8	62.5	86.7	22.5	20.0	65.4	72.9	88.7	–
SANA-WM (Zhu et al., 2026)	13.3	88.3	67.5	26.7	24.2	53.8	67.7	90.8	98.7
Echo-WM (Zhang et al., 2026)	16.7	60.8	66.7	22.5	21.7	<u>82.9</u>	<u>97.6</u>	89.2	98.7
Evoke (Yin et al., 2026)	20.0	95.8	80.0	<u>32.9</u>	<u>36.3</u>	78.3	93.5	95.4	99.0
Evoke-Turbo (Yin et al., 2026)	<u>21.7</u>	<u>98.3</u>	<u>89.2</u>	32.5	29.6	65.4	97.0	<u>96.2</u>	<u>99.1</u>
SeerWorld (ours)	**90.8**	**100.0**	**97.5**	**94.2**	**85.0**	**83.8**	**99.2**	**98.3**	**99.5**

$\epsilon - \mathbf{z}$, and stage 1 training minimizes

$$\mathcal{L}_{\mathrm{S1}} = \mathbb{E}_{(\mathbf{e},\mathbf{c},\mathbf{z},p,r),\,\sigma,\epsilon}\left[\frac{1}{|\Omega|D}\sum_{i\in\Omega}\left\|v_\theta(\mathbf{x}^\sigma,\mathbf{s}^\sigma,p,r)_i-(\epsilon_i-\mathbf{z}_i)\right\|_2^2\right], \tag{3}$$

where Ω contains target tokens and D is the latent channel dimension. The indicator r controls whether the IAA bias in Eq. (1) is enabled.

Stage 2: Online autoregressive adaptation. Stage 1 is trained with a clean ground-truth temporal anchor, but deployment conditions each new segment on latents produced by the model itself. Small errors in appearance, motion, and geometry can therefore accumulate at precisely the event-to-revisit transition. We reduce this exposure gap with a lightweight second stage initialized from stage 1. For every selected event-to-revisit pair, the currently updated renderer performs an online one-hop rollout of the immediately preceding segment. This predecessor rollout is also conditioned on a clean entity anchor, its ground-truth temporal anchor, and the predecessor's conditioning prompt. We take the final four latents of this self-generated predecessor as the temporal anchor $\widetilde{\mathbf{c}}_t$ for the revisit target, replacing its ground-truth temporal anchor. The resulting Stage-2 input is

$$\widetilde{\mathbf{x}}_t^\sigma = [\mathbf{e}_t \mid \widetilde{\mathbf{c}}_t \mid (1-\sigma)\mathbf{z}_t + \sigma\epsilon], \tag{4}$$

with the per-token timestep input $\mathbf{s}_t^\sigma$ defined as in Eq. (2). We optimize the same target-only velocity-matching objective as in Stage 1, now conditioned on the self-generated temporal anchor:

$$\mathcal{L}_{\mathrm{S2}} = \mathbb{E}_{t,\sigma,\epsilon}\left[\frac{1}{|\Omega_t|D}\sum_{i\in\Omega_t}\left\|v_\theta(\widetilde{\mathbf{x}}_t^\sigma,\mathbf{s}_t^\sigma,p_t,r_t)_i-(\epsilon_i-\mathbf{z}_{t,i})\right\|_2^2\right], \tag{5}$$

where Ω_t contains only target tokens, p_t is the target prompt, and r_t is the revisit indicator controlling IAA. The entity and temporal anchors remain loss-free conditioning. Additional training details for the two-stage construction are provided in Appendix Section D.

3.3.5 VLM Brain: Perception and Reasoning

At inference time, our framework adopts a VLM as *brain* to perceive and reason over the world evolution continuously. After every completed segment, the renderer decodes sampled frames for the *brain*, which identifies and records the visible entities, appearances, object states, scenes, and any realized changes. These observations accumulate into a persistent world memory. Once an injected event has been rendered and observed, the *brain* records what actually occurred, retrieves the identity evidence, and reasons over the accumulated context to infer how an off-screen subject would have causally responded. It then expresses the inferred state as a concrete generation prompt $\hat{p}$ for the relevant subsequent segment. Additional details are provided in Appendix Section E.

4 Experiments

4.1 Experiment Setup

We compare SeerWorld with recent interactive video world models that support autoregressive generation: LingBot-World (Team et al., 2026b), LingBot-World-v2 (Gao et al., 2026), YUME-1.5 (Mao et al., 2026), SANA-WM (Zhu et al., 2026), Echo-WM (Zhang et al., 2026), and the standard

Table 2: Component ablation of SeerWorld. Cumulative effects of IAA and Stage-2 online rollout training on SeerWorld-Bench. All values are percentages, and column-wise best results are bold.

Setting	Core	Revisit	Event	Causal	React.
Stage 1 w/o IAA	82.5	**100.0**	96.7	88.8	80.8
Stage 1 + IAA	84.2	99.2	**98.3**	88.7	83.9
Stage 1 + Stage 2	**90.8**	**100.0**	97.5	**94.2**	**85.0**

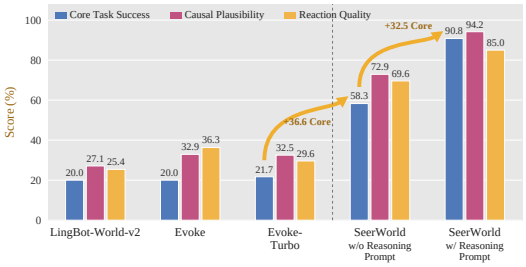

Figure 5: Effect of event-conditioned reasoning prompts on Out-of-Sight Reactive Evolution.

and Turbo variants of Evoke (Yin et al., 2026). All baselines receive the same initial, event, and blind revisit prompts without a prescribed reaction, switched at matched semantic boundaries through their native autoregressive interfaces. Implementation details and additional general video models with one-shot results under a different, composed-prompt setting are provided in Appendix Section G. We evaluate Out-of-Sight Reactive Evolution on SeerWorld-Bench with the following metrics: *Event Success*, *Revisit Success*, *Causal Plausibility*, *Reaction Quality*, and *Visual Integrity*. We additionally assess subject and background consistency between the opening and revisited views while allowing causally appropriate changes in pose, state, and viewpoint, and report *Motion Smoothness* using VBench (Huang et al., 2024). Our primary task-level measure, *Core Task Success*, requires Event Success, Revisit Success, and full Causal Plausibility (i.e., causal plausibility score=1) to hold simultaneously. Appendix Section B provides the complete evaluation protocol.

4.2 Implementation Details

The VLM brain is instantiated with Qwen3.8-27B (Yang et al., 2025a), served through vLLM (Kwon et al., 2023), while the visual renderer is initialized from LTX-2.5 (HaCohen et al., 2026). In Stage 1, we train a rank-128 LoRA for 8,000 steps with a learning rate of 1×10^{-4}, using a per-GPU batch size of 1 on eight H200 GPUs and eight gradient-accumulation steps, resulting in an effective batch size of 64. We set the IAA bias to $\beta = 3.0$. In Stage 2, we initialize from the Stage-1 checkpoint and conduct online autoregressive rollout training for 200 steps at 2×10^{-5}. Inference uses the official distilled LoRA and eight-step sampler to generate videos at 896×512 resolution and 24 fps. Complete training and inference settings are provided in Appendix Sections C and D.

4.3 Results of Out-of-Sight Reactive Evolution

Out-of-sight reactive evolution. As shown in Table 1, SeerWorld achieves 90.8% Core Task Success, exceeding the second-best baseline by 69.1 percentage points. Existing world models can often execute individual parts of the rollout: for example, Evoke-Turbo reaches 98.3% Revisit Success and 89.2% Event Success. However, its Core Task Success and Causal Plausibility result reveal that successful camera control and event rendering do not by themselves produce a coherent event–reaction chain. In contrast, SeerWorld obtains promising Revisit Success, Event Success, and Causal Plausibility. Its Causal Plausibility and Reaction Quality outperform other methods with a significant margin, demonstrating a substantially stronger ability to render causally plausible interaction evolution in the out-of-sight setting. **Consistency and generation quality.** The improvement in task success does not come at the cost of appearance preservation or visual quality. SeerWorld achieves promising Subject Consistency, Background Consistency, Visual Integrity, and Motion Smoothness. Its consistently high VLM-based scores and VBench (Huang et al., 2024) motion evaluation indicate strong *Out-of-Sight Reactive Evolution* performance while preserving generation quality.

4.4 Ablation Study and Additional Analysis

In this section, we adopt a research-question-driven analysis to analyze and validate the effectiveness of our method, including ablation studies and other analyses.

Do IAA and Stage-2 online rollout training improve out-of-sight reactive evolution? As shown in Table 2, IAA increases Core Task Success from 82.5% to 84.2% and Reaction Quality from 80.8% to 83.9%, demonstrating the benefit of strengthening attention to retrieved entity evidence. Stage 2 further raises Core Task Success to 90.8% and Causal Plausibility to 94.2%. These

Table 3: Transfer of the VLM-brain reasoning prompt (%).

Method	Core	Causal	React.
Evoke-Turbo	21.7	32.5	29.6
+ Brain prompt	45.0	53.3	50.8
SeerWorld (ours)	**90.8**	**94.2**	**85.0**

Table 4: Event–prompt alignment based on CLIP score.

Method	EPA-Abs	EPA-TopK
LingBot-World	0.2716	0.2827
LingBot-World-v2	0.2740	0.2860
SANA-WM	0.2789	0.2872
Echo-WM	0.2769	0.2896
Evoke	0.2893	0.3043
Evoke-Turbo	0.2885	0.3025
SeerWorld (ours)	**0.2942**	**0.3126**

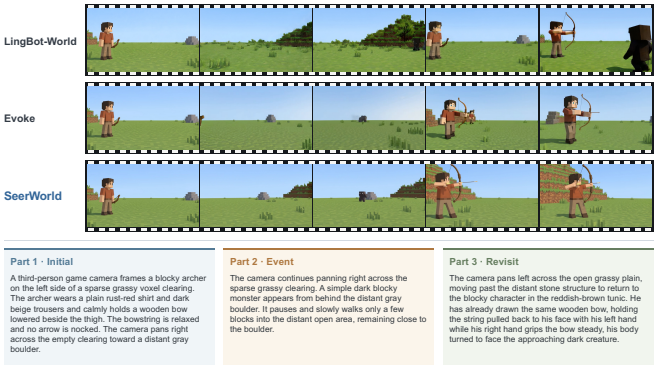

Figure 6: Qualitative comparison under the same VLM-brain reasoning prompt. More video demos are in the supplement.

results indicate that both IAA and stage-2 online rollout training improve the model capability for *out-of-sight reactive evolution*.

How do event-conditioned reasoning prompts affect different visual renderers? We replace the brain-generated reasoning prompt with the blind revisit prompt while retaining the trained SeerWorld renderer and visual conditioning. As shown in Fig. 5, performance drops substantially without event-conditioned reasoning, confirming its importance for inferring an appropriate response from the realized event. This variant nevertheless remains stronger than the baselines, indicating that the proposed data and renderer training already provide a strong reactive-evolution prior. Moreover, applying the same reasoning prompts to Evoke-Turbo consistently improves its task metrics (Table 3), demonstrating that the benefit transfers across visual renderers.

How do models compare in CLIP-based event–prompt alignment? As an additional reference to the VLM-based Event Success metric, we compare the generated event segment with its event prompt using CLIP ViT-B/32 (Radford et al., 2021). We recover each model's actual event-prompt interval, uniformly sample eight frames, and report their mean similarity (EPA-Abs) and the mean of the two highest similarities (EPA-TopK). As shown in Table 4, SeerWorld obtains the highest scores. These results should be interpreted only as a relative comparison: event prompts contain substantial shared scene and camera descriptions, while CLIP similarity cannot precisely determine whether a fine-grained event action has been correctly executed.

Does a shared reasoning prompt alone ensure faithful reactive evolution? We provide LingBot-World, Evoke, and SeerWorld with the same VLM-inferred revisit prompt to isolate renderer capability. As shown in Fig. 6, the baselines exhibit hallucinated event entities, subject duplication, or coarse revisit frames, whereas SeerWorld preserves the subject and scene while rendering a coherent defensive response. Thus, an appropriate reasoning prompt must be paired with a renderer capable of preserving history and realizing the inferred evolution.

Do the conclusions hold beyond the main automatic evaluation? We conduct a pairwise human study on benchmark cases. SeerWorld achieves win rates of 80% for task completion and 59% for visual quality, supporting the advantages observed under the automatic metrics. Appendix Section H provides the complete protocol, outcome breakdown, agreement analysis, and results.

5 Conclusion

In this paper, we formulate *Out-of-Sight Reactive Evolution* as a concrete setting for investigating autonomously world evolution and construct SeerWorld-Data/Bench to support learning and evaluating event realization, revisit, event-conditioned reaction, and visual consistency. We further present the SeerWorld framework, a reasoning-augmented brain-renderer collaboration framework that continuously perceives and renders the world dynamics. The renderer is trained with identity-anchored attention and a two-stage procedure that combines supervised learning of reactive rendering with autoregressive rollout training. Experiments show that SeerWorld substantially improves *Out-of-Sight Reactive Evolution* capabilities over recent video world models while preserving generation quality and consistency. We hope this study can provide a reference for future research in reasoning-enhanced world evolution. Limitations and future directions are discussed in Appendix Section I.

AI USE STATEMENT

Generative AI tools assisted with language refinement and manuscript organization. Generative AI was also used in constructing the dataset, and all AI-generated samples were screened to exclude harmful content. The authors reviewed all AI-assisted material and take full responsibility for the final dataset and manuscript.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the method, with implementation details and hyperparameter settings in the main paper and appendix. We plan to make the code and data publicly available.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

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

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

Zicheng Duan, Jiatong Xia, Zeyu Zhang, Wenbo Zhang, Gengze Zhou, Chenhui Gou, Yefei He, Feng Chen, Xinyu Zhang, and Lingqiao Liu. Liveworld: Simulating out-of-sight dynamics in generative video world models. *arXiv preprint arXiv:2603.07145*, 2026.

Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.

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

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *International Conference on Learning Representations*, volume 2024, pp. 22841–22860, 2024a.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024b.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, et al. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Team HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. *arXiv preprint*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2510.16732*, 2025.

Dingkang Liang, Dingyuan Zhang, Xin Zhou, Sifan Tu, Tianrui Feng, Xiaofan Li, Yumeng Zhang, Mingyang Du, Xiao Tan, and Xiang Bai. Unifuture: A 4d driving world model for future generation and perception. *arXiv preprint arXiv:2503.13587*, 2025.

Jinpeng Lu, Dexu Zhu, Haoyuan Shi, Linghan Cai, Guo Tang, Yinda Chen, Jie Cao, Duyu Tang, Yi Zhang, Yong Dai, et al. Current world models lack a persistent state core. *arXiv preprint arXiv:2606.20545*, 2026.

Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pp. 674–679, 1981.

Wenchao Ma, Changran Liu, Sharon X Huang, and Haomiao Jiang. Closing the loop: Training-free revisit consistency for autoregressive generative rendering. *arXiv preprint arXiv:2607.21848*, 2026a.

Ziqi Ma, Mengzhan Liufu, and Georgia Gkioxari. Out of sight, out of mind? evaluating state evolution in video world models. *arXiv preprint arXiv:2603.13215*, 2026b.

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

OpenAI. Introducing chatgpt images 2.0. `https://openai.com/index/introducing-chatgpt-images-2-0/`, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Shaohao Rui, Xiaofeng Mao, Zhanyu Zhang, Peijia Lin, Yansong Zhu, Yibo Zhang, Haibin Wan, Zhangrui Zhao, and Weijie Ma. Biwm: Advancing open-source interactive video world models with bidirectional autoregression. *arXiv preprint arXiv:2606.10135*, 2026.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

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

ByteDance Seed Team. One-take creation, flexible referencing: Introducing seedance 2.5. https://research.doubao.com/en/blog/one-take-creation-flexible-referencing-introducing-seedance-2-5, 7 2026.

Gemini Team and Google DeepMind. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2 2026.

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

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.

Xinjie Wang, Liu Liu, Yu Cao, Ruiqi Wu, Wenkang Qin, Dehui Wang, Wei Sui, and Zhizhong Su. Embodiedgen: Towards a generative 3d world engine for embodied intelligence. *arXiv preprint arXiv:2506.10600*, 2025.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems*, 38:49371–49393, 2026.

Tianshuo Xu, Yichen Xie, Depu Meng, Chensheng Peng, Quentin Herau, Bo Jiang, Yihan Hu, and Wei Zhan. Teaching video generators to remember: Eliciting dynamic memory for out-of-sight state evolution. *arXiv preprint arXiv:2605.25333*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

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*, 2025b.

Ying Yang, Zhengyao Lv, Yujia Zeng, Tianlin Pan, Haofan Wang, Yueming Lyu, Binxin Yang, Hubery Yin, Chen Li, Jing Lyu, et al. Stableworld: Towards stable and consistent long interactive video generation. *arXiv preprint arXiv:2601.15281*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025c.

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

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Songchun Zhang, Yaowei Li, Junhao Zhuang, Weiyang Jin, Haoyu Wang, Xin Lu, Yilang Sun, Shiyi Zhang, Haoran Li, Xiaoxiao Ma, et al. Echowm: Open and enterable omnimodal world models. *arXiv preprint arXiv:2608.23189*, 2026.

Xin Zhou, Dingkang Liang, Sifan Tu, Xiwu Chen, Yikang Ding, Dingyuan Zhang, Feiyang Tan, Hengshuang Zhao, and Xiang Bai. Hermes: A unified self-driving world model for simultaneous 3d scene understanding and generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27817–27827. IEEE, 2025.

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

APPENDIX

A EXTENDED RELATED WORK

Interactive and autoregressive video world models. Generative video backbones (Wan et al., 2025; HaCohen et al., 2026) provide the visual synthesis foundation for interactive world simulation. LingBot-World (Team et al., 2026b) and LingBot-World-v2 (Gao et al., 2026) extend this capability toward long-horizon, real-time interaction, with the latter combining bidirectional and autoregressive attention. Echo-WM (Zhang et al., 2026) targets open and enterable omnimodal world simulation, while Alaya-EVOKE (Yin et al., 2026) combines persistent world-state memory, mid-rollout reprompting, and few-step generation. Segment-autoregressive methods preserve full attention within each generated segment to reduce the quality degradation associated with fully causal adaptation. YUME-1.5 (Mao et al., 2026) combines long-context generation with bidirectional-attention distillation; Matrix-Game 3.0 (Wang et al., 2026) uses memory-augmented multi-segment autoregressive distillation; and BiWM (Rui et al., 2026) provides a general recipe for adapting bidirectional video backbones into controllable segment-autoregressive world models.

Memory and out-of-sight state evolution. Long-horizon consistency does not ensure that an unobserved world continues to evolve. Recent evaluations expose this limitation through camera look-away, occlusion, or illumination changes (Ma et al., 2026b; Lu et al., 2026). ReMind (Xu et al., 2026) trains a causal video generator to recover evolving states across interruptions. LiveWorld (Duan et al., 2026) maintains a persistent global state by separating a static 3D environment from dynamic entities advanced by an evolution engine. Hybrid Memory (Chen et al., 2026) similarly distinguishes static-background preservation from dynamic-subject tracking across exit–reentry intervals. These works primarily examine whether an active process or motion continues through an observation gap. In contrast, our setting requires a newly introduced event to be realized after the subject leaves view and its consequences to be inferred when the subject is revisited, combining mid-rollout instruction following (Yang et al., 2025b), long-range entity consistency, and reasoning over the realized visual history.

B SEERWORLD-DATA AND SEERWORLD-BENCH DETAILS

B.1 SEERWORLD-DATA CONSTRUCTION

Structured scenario generation. Each training example begins from a structured tuple specifying the domain, scene, subject and appearance, event, compatible reaction, and camera trajectory. Compatibility constraints exclude combinations whose consequence would be visually ambiguous or implausible for the designated subject. Gemini3.1-Pro expands the tuple into a continuous-shot script covering the initial observation, camera move-away, event reveal, camera move-back, and reacted subject. We render these scripts with Seedance 2.0 and 2.5 (Seedance et al., 2026; Team, 2026). The retained training set contains around 12k videos, each at least ten seconds long.

Verification and temporal annotation. Generated candidates are checked for continuous-shot structure, subject identity, complete subject exit and re-entry, visible event realization, visible reac-

tion, and severe rendering failures. Camera direction is verified independently from sparse Lucas–Kanade feature tracks (Lucas & Kanade, 1981) and a RANSAC affine fit (Fischler & Bolles, 1981); clips with a missing reversal, full-circle motion, only unidirectional motion, or unreliable estimates are rejected. For each retained video, we record the subject-disappearance and reappearance times, camera phase boundaries, event interval, reaction interval, and generation provenance. Segment captions are produced jointly across adjacent intervals and audited to remove future-event spoilers, references to an absent subject, malformed camera commands, and redundant neighboring descriptions.

B.2 Reviewed semantic taxonomy

Raw generation labels contain aliases and style prefixes and therefore overestimate semantic diversity. We consolidate them into a two-level taxonomy: a coarse family and a fine-grained semantic type. Synonyms are merged, and rendering styles such as AAA, voxel, fantasy, or photorealistic do not create a new type unless the underlying event, environment, or reaction is semantically distinct. Every source entry is assigned exactly once. Table 5 reports the resulting human-reviewed statistics.

Table 5: Reviewed semantic coverage of SeerWorld-Data and SeerWorld-Bench. Benchmark reactions summarize private reference actions used during construction; evaluation remains open-ended and accepts any causally plausible response.

Split	Videos	Events		Scenes		Reactions	
	/ cases	Fam.	Types	Fam.	Types	Fam.	Types
SeerWorld-Data	12k	19	65	16	63	5	14
SeerWorld-Bench	120	17	55	16	56	6	14

The event taxonomy includes, for example, fire and combustion, structural collapse, wild-animal intrusion, supernatural emergence, flooding, atmospheric disturbance, and vehicle failure. Reaction families include evasive relocation, active defense, etc.

B.3 SeerWorld-Bench construction and train–test screening

The 120 benchmark cases comprise 69 photorealistic and 51 game/virtual scenarios. Sixty-six initial frames come from held-out source videos and 54 are generated with GPT-Image-2 (OpenAI, 2026). We first remove exact source overlap and then screen every benchmark case against the complete training split at both the visual and semantic levels. Visual retrieval uses the benchmark initial frame to obtain its nearest training frames, while semantic retrieval operates on the paired initial and event prompts using lexical overlap and text-embedding similarity and requires both parts to come from the same training video. The union of these candidates is passed to a first VLM comparison. Pairs with substantial similarity in both the initial state and event are re-evaluated by a second multimodal pass that jointly receives the two initial frames and all four prompt fields. Ambiguous cases are manually adjudicated, and any benchmark candidate with a confirmed identical or near-identical training counterpart is removed or revised. A high first-frame similarity alone, or membership in the same scene or event category, is not treated as evidence of duplication.

B.4 Prompt protocol and reaction-intensity calibration

We label every event by its required reaction intensity. A *strong* event requires an urgent response such as fleeing or taking cover, whereas a *weak* event permits either a mild or strong response. This calibration prevents a visually present but implausibly weak response to a dangerous event from being counted as correct.

At evaluation time, models receive the initial prompt, the event prompt at the appropriate switch point, and a blind revisit instruction that asks the subject to react without specifying how. The model must therefore infer an appropriate response from the event realized during its own rollout; the benchmark neither exposes nor requires a predefined reaction type, and any causally plausible response is considered valid. We audit all three prompts to ensure that the initial prompt does not

reveal the event, the event prompt does not reveal the eventual response, and the blind revisit prompt contains neither a specific reaction nor future visual details.

B.5 Evaluation protocol

Our evaluation combines complementary sources of evidence rather than relying on a single judge. A strong VLM Qwen3.8 (Yang et al., 2025a) assesses task execution, causal plausibility, visual integrity, and revisit consistency. Established tool-based metrics provide independent diagnostics: VBench (Huang et al., 2024) measures motion smoothness, while CLIP ViT-B/32 (Radford et al., 2021) supplies the auxiliary event–prompt alignment comparison in Table 4. We further complement these automatic measures with the human study in Appendix Section H.

For the VLM-based metrics, we temporally stratify each generated rollout into its beginning, middle, and ending portions and uniformly sample 30 frames across the three portions. The judge receives these frames, the initial and event prompts, and the strong/weak event-intensity label used to calibrate reaction strength, but not a target reaction or ground-truth revisit description. It jointly inspects whether the camera returns, whether the requested event is realized, what response the subject exhibits, whether that response is causally plausible at the required intensity, and whether the revisited subject and environment remain consistent. Event and revisit success are binary; the remaining perceptual and behavioral criteria use the three-level scale $\{0, 0.5, 1\}$.

Event and revisit success. *Event Success* is 1 only if the event described by the event prompt visibly occurs in the middle portion of the generated rollout; otherwise it is 0. *Revisit Success* is 1 only if the camera first leaves the original subject and region and subsequently returns to the same location with the subject visible again. A reaction may cause the subject to leave the frame after being revisited without invalidating the revisit. A cut to an unrelated scene, or a return without the original subject, receives 0.

Causal plausibility. This metric evaluates whether the subject's changed behavior or state is a plausible consequence of the event that was actually rendered. The judge may use changes relative to stable landmarks—for example, retreating behind a doorway, moving behind cover, or relocating to higher ground—as evidence of an off-screen response. A score of 0 indicates no reaction when one is required, or a response that contradicts the event's causal logic; 0.5 indicates a plausibly related but insufficiently strong response to a strong event; and 1 indicates a causally appropriate response at the expected intensity. We do not require one prescribed action: escape, cover, safe relocation, or defensive readiness may all receive full credit when appropriate. If either Event Success or Revisit Success is 0, Causal Plausibility is set to 0 because the generated video provides no valid event–reaction chain to assess.

Reaction quality and visual integrity. *Reaction Quality* isolates the visual realization of the response from its semantic appropriateness. It receives 0 for broken, distorted, or incoherent reaction motion; 0.5 for a visible but stiff or mildly artifacted response; and 1 for a natural and coherent action or changed state. If no changed reaction state is visible, the per-case value is null and is treated as 0 in the all-case mean reported for cross-model comparison. *Visual Integrity* evaluates the complete rollout independently of task correctness: 0 denotes severe structural breakdown, hallucination, or melting; 0.5 denotes a watchable video with visible local artifacts, instability, or flicker; and 1 denotes a clean and coherent rollout without obvious breakdown, with particular attention paid to prompt-switch boundaries.

Revisit consistency. *Subject Consistency* compares the opening subject with the revisited subject while explicitly allowing its pose, behavior, and state to change. A score of 0 denotes a different or severely drifted identity, 0.5 a recognizable subject with noticeable identity or appearance inconsistencies, and 1 a clearly preserved identity. If the subject does not reappear and therefore cannot be compared, the case receives a score of 0. The reported Subject Consistency is averaged over all 120 benchmark cases. *Background Consistency* compares the opening and revisited locations while allowing modest viewpoint changes: 0 indicates a different scene, 0.5 the same general place with a major object or layout inconsistency, and 1 the same location with its principal structures and objects preserved. Minor viewpoint changes or incidental small-object differences do not reduce the score.

Motion smoothness. We report the Motion Smoothness dimension of VBench (Huang et al., 2024) over all 120 generated videos. This metric is independent of the VLM judge and measures local temporal smoothness rather than completion of the event–revisit–reaction chain. Because it is highly saturated for modern video generators, we treat it as a secondary quality diagnostic rather than a primary task metric.

Core success and aggregation. Let r_i and e_i denote binary Revisit and Event Success for case i, and let $c_i \in \{0, 0.5, 1\}$ denote Causal Plausibility. We define

$$\text{Core} = \frac{100}{N} \sum_{i=1}^{N} \mathbb{1}[r_i = 1 \ \wedge \ e_i = 1 \ \wedge \ c_i = 1]\,. \tag{6}$$

Thus Core Success requires all three task-critical conditions in the same case. Reaction Quality and Visual Integrity are reported separately and are intentionally excluded from Core, allowing task completion and rendering quality to be analyzed independently. Unless otherwise stated, three-level metrics are averaged over all 120 benchmark cases and reported as percentages.

C Training hyperparameters

Table 6: Training and inference settings for SeerWorld. Stage 2 inherits the Stage-1 LoRA architecture and IAA configuration.

Setting	Value
Visual renderer	LTX-2.5 (HaCohen et al., 2026)
VLM brain	Qwen3.8-27B (Yang et al., 2025a)
Output resolution / frame rate	896×512 / 24 fps
Entity-reference length (V2V train / I2V benchmark)	2 / 1 temporal latent steps
Temporal-anchor length	4 temporal latent steps
LoRA rank	128
IAA bias β	3.0
Stage-1 optimizer / precision	AdamW / bfloat16
Stage-1 learning rate / steps	1×10^{-4} / 8,000
Stage-1 effective batch size	64 (8 GPUs, accumulation 8)
Stage-1 gradient clipping	1.0
Stage-2 learning rate / steps	2×10^{-5} / 200
Stage-2 online / GT-anchor ratio	75% / 25%
Online and test-time sampler	official distilled adapter, 8 steps

Training and benchmark inference modes. Both training stages optimize the renderer in a video-to-video (V2V) setting, using the first two temporal latent steps of the training clip as the entity reference. For benchmark evaluation, we use the same learned renderer directly in image-to-video (I2V) mode: the provided initial frame is encoded into one entity-reference latent step, after which generation proceeds autoregressively. Each later segment is conditioned on the final four latent steps of its generated predecessor. We also experimented with training the renderer directly in I2V mode and observed comparable benchmark performance; we therefore retain the unified V2V training configuration for the reported model.

D Two-stage training procedure

Stage 2 uses a *one-hop* rollout rather than regenerating the entire history chain: it exposes the renderer to direct predecessor errors while keeping training tractable. Because the predecessor is regenerated by the current model as training progresses, the conditioning distribution co-evolves with the renderer rather than remaining fixed to a single offline snapshot. To prevent this focused adaptation from overwriting the initial, event, and clean-revisit behavior learned in Stage 1, we mix 25% deterministic samples with ground-truth temporal anchors drawn across all phases and 75% samples with online self-generated temporal anchors. At deployment, the resulting Stage-2 LoRA is used together with the same official distilled LoRA and eight-step sampler employed during online rollout.

Algorithm 1 Two-stage anchor and online-rollout construction

```
def stage1_sample(clip, target):
    entity = opening_entity_anchor(clip)
    prefix = gt_predecessor_tail(target, count=4)
    return entity, prefix, target.is_revisit

def stage2_sample(clip, target, use_online):
    if not use_online:  # 25% Stage-1 rehearsal
        clip, target = sample_stage1_window()
        return stage1_sample(clip, target)
    assert target.is_revisit and prompt_changed(target)
    entity = opening_entity_anchor(clip)
    temporal_anchor = gt_temporal_anchor(target.prev, count=4)
    with no_grad():
        prev = rollout_one_hop(current_model,
                entity, temporal_anchor, target.prev.prompt,
                sampler="distilled-8step")
    prefix = stop_grad(prev.latents[-4:])
    return entity, prefix, True
```

E VLM BRAIN PROTOCOL

The VLM brain is called after every generated segment rather than only at the final revisit. Each call receives temporally sampled frames from the newly generated segment together with the accumulated textual world record. Its first responsibility is observational: it records visible entities and their attributes, persistent scene landmarks, object states, camera movement, and any newly realized event. Before an event has been observed, the brain updates this record but does not propose a reaction. This separation prevents the initial observation from leaking future event or reaction information.

After the event segment, the brain reasons from the event that was actually rendered, not merely from the requested event prompt. It identifies the affected off-screen subject, infers one causally plausible changed state or behavior, and converts that inference into a concrete renderer prompt. The prompt specifies the camera return, preserved identity cues, the subject's post-event state, and visually realizable behavior, while avoiding hidden mental states or explanations that cannot be rendered. The benchmark remains open-ended: this inferred behavior need not match an exact reaction if it is causally plausible for the realized event.

For long-range identity conditioning, the framework can select a clear prior observation of the target entity and map it to the corresponding latent span. In the reported controlled I2V evaluation, the opening frame is deliberately constructed to show the target clearly and is used as the persistent one-step entity reference. Content-based retrieval is retained for longer or less constrained rollouts in which the clearest observation need not be the first frame. In either case, retrieval determines visual evidence, whereas the VLM-generated prompt determines the subsequent semantic state; the four most recent generated latent steps provide short-range temporal continuity independently of this decision.

Reasoning case study. Fig. 8 summarizes one reference brain-renderer interaction. The brain first records the knight's identity, equipment, and cathedral environment before the camera moves away. It then observes the event actually realized by the renderer: although the requested gargoyle was intended to remain near the gate, the generated creature begins advancing into the nave. Based on this observed threat and the accumulated subject memory, the brain produces a revisit prompt in which the knight has raised the same shield and adopted a braced defensive stance. The final observation verifies that the rendered response is causally appropriate while preserving the subject, equipment, and scene. This example illustrates how the brain conditions its reasoning on the realized visual history rather than relying only on the requested event prompt.

```
1. Observation update (after every segment):
Record visible entities, appearance, carried objects and their
states, scene context, and newly realized changes. Before an
event occurs, update memory only; do not generate a reaction
instruction.

2. Event-conditioned reasoning (after an event is observed):
Using the accumulated world record, identify the event that was
actually rendered and infer a causally plausible changed state for
the off-screen subject.
Output example: {"revisit_prompt": "The camera returns to the
original subject, who has already retreated behind cover."}

3. Entity retrieval (before a revisit):
Select the clearest prior observation of the target subject and
return its frame index for latent retrieval.

4. Post-generation update:
Observe and record the generated revisit segment so that the
updated world state is available to later rollout steps.
```

Figure 7: The staged VLM calls used for continual world observation, event-conditioned reasoning, and entity retrieval (§3.3.5).

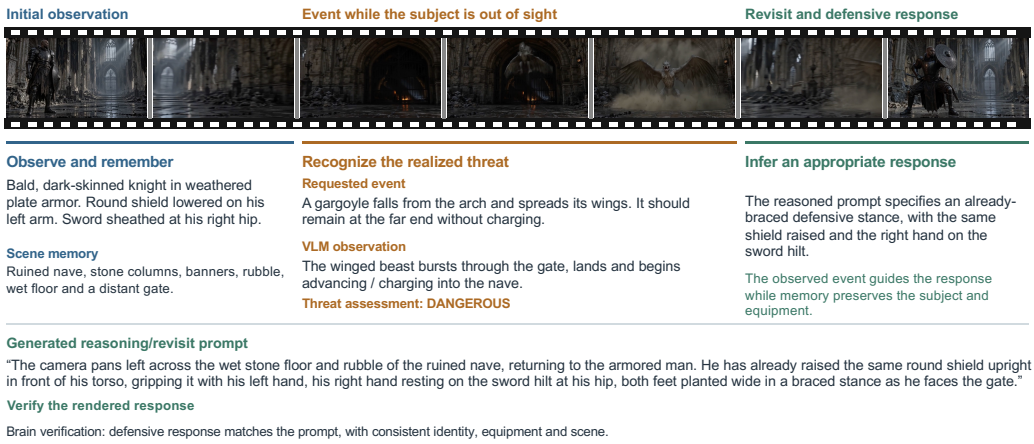

Figure 8: Case study summarizing the VLM brain's observation, event-conditioned reasoning, prompt generation, and verification over an out-of-sight rollout.

F Anchor layout and IAA implementation

Entity-reference and target-window tokens are combined through the native LTX conditioning path. Here, the target window denotes the complete generation-side latent sequence, including the clean temporal anchor and the target latents being denoised. Their physical order differs across code paths: training forms [entity reference | target window], whereas inference appends the reference and forms [target window | entity reference]. IAA therefore derives the entity-key range from the actual reference length and layout instead of assuming a fixed prefix position. Runtime checks verify the reference length, the contiguous denoised-query span, and the target–reference boundary.

IAA is enabled only for revisit windows, which are oversampled by a factor of three in Stage 1. If an entity reference contains n_e temporal latent steps and each step contains $H_\ell W_\ell$ spatial tokens, its key block contains $n_e H_\ell W_\ell$ tokens. For each denoised target query, we multiply the corresponding entity-key entries of the LTX attention mask by $\exp(\beta)$. LTX subsequently converts this mask to an additive log-space bias, producing the $+\beta$ term in Eq. (1). We use $\beta = 3.0$; other valid pairs are unchanged, and originally masked pairs remain masked.

G BASELINE REPRODUCTION DETAILS

Common prompt protocol. For every benchmark case, all autoregressive models receive the same three user-facing conditions: (i) an initial prompt describing the starting subject and scene, (ii) an event prompt introduced only after the camera has moved away, and (iii) a blind revisit prompt that asks the camera to return and the subject to react without revealing a particular reaction. The corresponding camera schedule moves the view away from the subject, exposes the event region, and then moves the view back. We inject each prompt at the nearest native chunk boundary and preserve the model's recurrent state, latent history, or memory cache across switches. Thus, no model is reset at the event or revisit boundary. In the main comparison, baselines directly execute the blind revisit prompt and receive neither VLM-generated reaction reasoning nor entity retrieval; SeerWorld internally interprets the same blind request using its VLM brain. For a more detailed ablation, the prompt-transfer and shared-prompt analyses in Table 3 and Fig. 6 also provide baseline renderers with the brain-generated reasoning prompt, allowing us to separate the contribution of event-conditioned reasoning from that of the proposed data and renderer training.

Compatibility of additional recent world models. We also conducted preliminary protocol-compatibility trials with HY-World 1.5 (HunyuanWorld, 2025), Matrix-Game 3.0 (Wang et al., 2026), and AlayaWorld (Team et al., 2026a), but do not include them in the quantitative comparison because their released models did not reliably support the complete evaluation protocol. In particular, HY-World 1.5 frequently continued the preceding content while weakly realizing or entirely ignoring the event introduced by the mid-rollout prompt switch. Matrix-Game 3.0 and AlayaWorld likewise did not consistently complete the combined move-away, newly injected event, blind revisit, and event-conditioned reaction sequence under their available inference interfaces. This mismatch is unsurprising: these systems are trained primarily for their respective action-conditioned interaction, spatial persistence, or general world-generation objectives, rather than on data that explicitly couples mid-rollout event injection with off-screen subject evolution. Their behavior in our setting should therefore not be interpreted as a general assessment of their visual quality or intended capabilities. We leave systematic adaptation and controlled comparison of these models to future work.

Evaluated baselines. For LingBot-World v1 (Team et al., 2026b) and v2 (Gao et al., 2026), we use the official interactive interfaces, translate the benchmark trajectory into their native camera controls, and schedule all prompts within one persistent rollout. We evaluate the released YUME-1.5 5B checkpoint (Mao et al., 2026) at 720p through its long-context autoregressive interface. For SANA-WM (Zhu et al., 2026), we use the released streaming checkpoint and assign the three prompts to a $3/2/3$ schedule over its eight autoregressive blocks, with camera motion supplied through its native trajectory interface. Echo-WM (Zhang et al., 2026) is evaluated with sequential prompt switching while retaining its persistent autoregressive world state. Evoke and Evoke-Turbo (Yin et al., 2026) use the official image-to-video segment mode with matched inputs and prompt boundaries, differing only in their released variants and native sampling configurations.

One-shot references. We additionally evaluate Wan2.2 TI2V-5B (Wan et al., 2025), LTX-2.3, and LTX-2.5 (HaCohen et al., 2026) as general one-shot video generators. Each model receives a single composed prompt containing the initial, event, and blind revisit instructions and generates the complete video in one pass. These models therefore observe the intended event and camera-return instruction before generation begins, unlike the main autoregressive setting in which the event is introduced only during rollout. We report their results in Table 7 for completeness but do not include them in the main comparison.

H HUMAN STUDY

We conduct a pairwise human study comparing SeerWorld with Evoke-Turbo on 50 cases randomly sampled from SeerWorld-Bench and proportionally stratified by domain and event strength. Two human raters independently evaluate all cases. To control for presentation bias, one rater receives Form A and the other receives Form B, which reverses the A/B assignment for every comparison. For each case, raters separately judge (i) completion of the event-revisit-reaction chain and (ii) visual quality, with particular attention to subject identity and appearance drift. Each judgment is decoded

Table 7: One-shot general video generators on SeerWorld-Bench, with SeerWorld included as an autoregressive reference below the separator. Best results among the one-shot models are bold (%).

	Task metrics					Visual quality, consistency, and motion			
Method	Core	Revisit	Event	Causal	React.	Visual	Subj. Cons.	Bg. Cons.	Motion
Wan2.2 TI2V-5B (Wan et al., 2025)	8.3	40.0	87.5	17.1	18.3	77.9	81.7	95.4	98.8
LTX-2.3 (HaCohen et al., 2026)	27.5	**97.5**	**94.2**	43.3	42.1	88.8	**96.7**	**99.2**	99.2
LTX-2.5 (HaCohen et al., 2026)	**34.2**	95.0	90.8	**50.8**	**42.9**	**90.4**	95.8	98.8	**99.4**
SeerWorld (ours; autoregressive)	90.8	100.0	97.5	94.2	85.0	83.8	99.2	98.3	99.5

into one of four outcomes: preference for SeerWorld, preference for Evoke-Turbo, a tie, or neither system succeeding. Thus, each criterion contains 100 judgments in total.

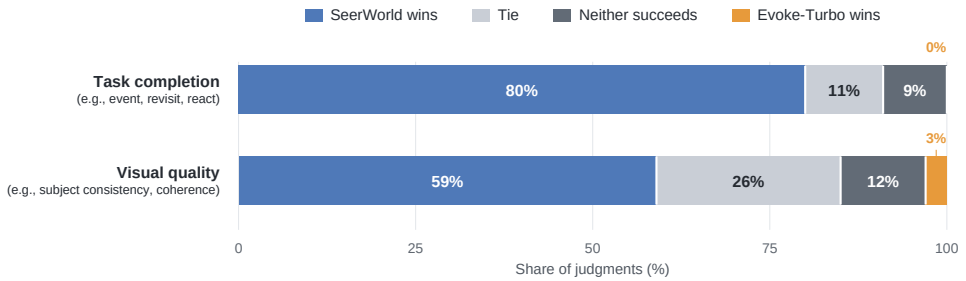

Figure 9: Pairwise human-study outcomes averaged across two raters, with 50 cases evaluated by each rater.

As summarized in Fig. 9, SeerWorld is preferred in 80% of all task-completion judgments, while Evoke-Turbo is never preferred; the remaining outcomes are 11% ties and 9% cases in which neither model succeeds. For visual quality, SeerWorld is preferred in 59% of judgments, compared with 3% for Evoke-Turbo, with 26% ties and 12% neither-success outcomes. Among direct preferences, SeerWorld accounts for 100% of task-completion choices and 95.0% on average across raters for visual quality. The two raters reach exact agreement on 84% of cases for task completion and 78% for visual quality (Cohen's $\kappa = 0.538$ and 0.616, respectively). Although the study contains only two raters, the reversed presentation and consistent preference patterns provide complementary human evidence for the improvements observed under the automatic metrics.

I Limitations and Future Work

Although the segment-autoregressive design can in principle continue the world rollout for an unbounded number of segments, the present work focuses on a single event followed by a single subject reaction. This controlled scope isolates the minimal out-of-sight event–reaction chain, but does not yet capture long-horizon worlds in which multiple events interact, earlier reactions alter later conditions, or several entities respond over time. Extending the data, reasoning memory, training procedure, and evaluation protocol to multi-event and multi-reaction evolution is therefore an important direction for future work. In addition, our current scenarios primarily employ horizontal move-away and move-back camera trajectories because they represent the most common and directly interpretable camera motion for revisit scenarios. We do not systematically evaluate more complex camera motion, such as large vertical displacement, depth-wise movement, orbiting, or compound six-degree-of-freedom trajectories, which may make scene correspondence, entity retrieval, and revisit consistency substantially more difficult. Future versions of SeerWorld-Data and SeerWorld-Bench should incorporate these trajectories and evaluate whether the framework remains robust under more general interactive camera control.

NOTE: Additional video demos are provided in the supplementary material.