WORLDPLAY2: EXTENDING REAL-TIME INTERACTIVE WORLD MODELS IN CONTROL AND HORIZON

Anonymous authors
Paper under double-blind review

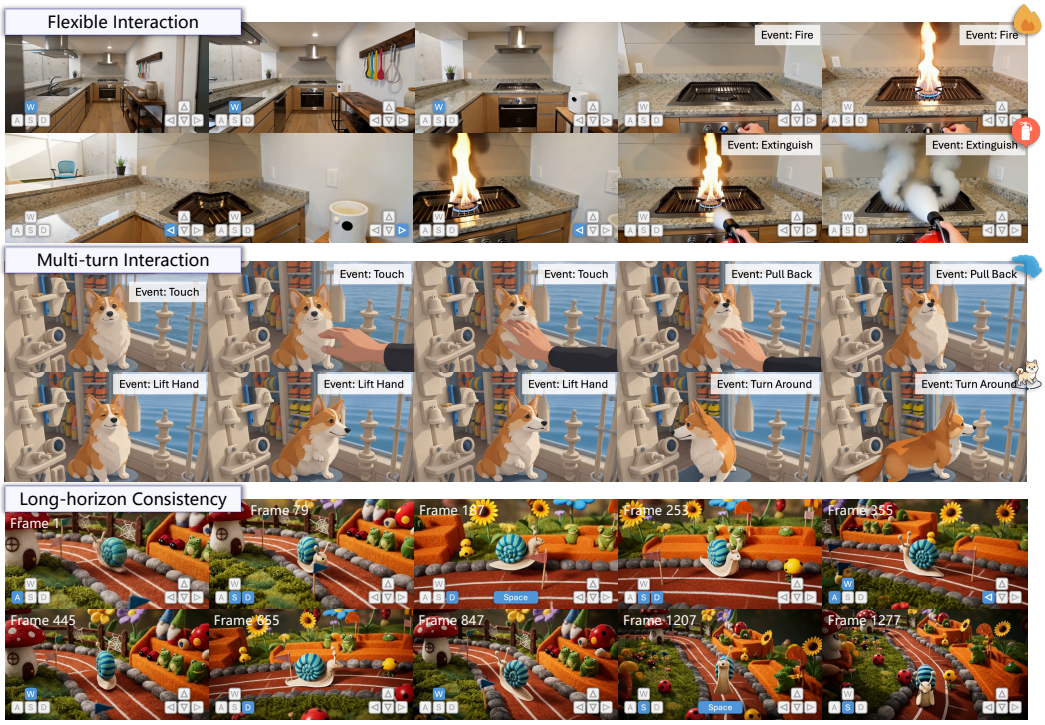

Figure 1: **WorldPlay2 is a real-time interactive world model enabling versatile controls with long-horizon consistency. Top:** It supports flexible interactions, including navigation controls and complex semantic events. **Middle:** It executes multi-turn interactions to achieve coherent storytelling. **Bottom:** It maintains long-horizon consistency under complex navigation controls.

ABSTRACT

Interactive world models require responding in real time to versatile controls and maintaining long-horizon consistency. However, modeling heterogeneous controls remains difficult, while explosive contexts and unstable distillation impede achieving both long-horizon consistency and real-time responsiveness. In this paper, we present WorldPlay2, an interactive world model that couples a factorized hybrid control interface with a co-design of compressed memory and stable distillation. 1) Our factorized hybrid control interface integrates frame-aligned action control with structured semantic control that explicitly disentangles scene appearance, character identity, and dynamic semantic events, thereby facilitating effective control learning. 2) To achieve efficient long-horizon modeling, we compress historical contexts into compact memory tokens shared by the autoregressive student and the bidirectional teacher. This design enables clip-wise, memory-conditioned score evaluation instead of jointly processing an entire long rollout, substantially reducing distillation overhead. 3) We further propose Stable Forcing, which initializes the autoregressive student via a few-step strategy and leverages full-rollout replay to preserve the quality of long-horizon rollouts, ensuring robust and stable distillation. Extensive experiments demonstrate the strong generalizability of our model and its superior performance compared to existing methods.

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 world models (Parker-Holder et al., 2025; Sun et al., 2026; He et al., 2025; Team et al., 2026d; Alibaba, 2026; Hong et al., 2025; Xu et al., 2026; Jiang et al., 2026; Zhu et al., 2026a) are beginning to transform video generators (Wan et al., 2025; Wu et al., 2025; Deepmind, 2025) from passive content-creation systems into interactive environments that evolve in response to user input. Such models have the potential to serve as general-purpose simulators (Agarwal et al., 2026; Brooks et al., 2024), allowing users to explore, interact with, and reshape the generated environments while providing scalable data generators and policy evaluators for embodied agents (Wiedemer et al., 2025; Team et al., 2025). Realizing these applications requires world models to support versatile controls, preserve coherent world states over long horizons, and operate in real time.

Controllability represents a primary challenge for world models. Early efforts (Sun et al., 2026; He et al., 2025; Team et al., 2026d) focused on navigation-oriented controls, lacking richer mechanisms for interacting with the environment. Although recent works (Team et al., 2026a; Gao et al., 2026) attempt to accommodate increasingly diverse interactive events, effectively integrating these heterogeneous inputs remains challenging because these control signals inherently operate across disparate semantic granularities and temporal horizons. Specifically, camera motion and character locomotion demand precise, frame-aligned control, whereas complex interactive events are more naturally expressed via high-level semantic commands spanning longer temporal horizons. Moreover, control signals must be disentangled from content factors such as scene appearance and character identity. Otherwise, the model may conflate visual appearance, spatial movement, and occurring events, leading to ambiguous supervision and unreliable responses.

The second challenge lies in co-designing memory and distillation. Existing methods (Team et al., 2026d; Hong et al., 2025; Xu et al., 2026; Zhu et al., 2026a) often rely on full-context bidirectional models as teachers. However, computational costs scale quadratically with video length, which not only hinders long-horizon modeling but, more crucially, makes score evaluation during distillation prohibitively expensive. Moreover, preserving long-horizon consistency during distillation presents additional challenges. This process typically requires the student model to autoregressively generate long video sequences (Hong et al., 2025; Sun et al., 2026; Xu et al., 2026; Team et al., 2026d), but due to error accumulation and few-step sampling, the student's generation distribution diverges significantly from that of the teacher, thereby rendering distillation unstable and severely degrading generation quality. Therefore, world models require a co-design in which memory is sufficiently compact to enable efficient long-horizon modeling and distillation is sufficiently stable to preserve long-horizon consistency.

In this paper, we introduce WorldPlay2, a real-time interactive world model that combines factorized hybrid control interface, distillation-oriented compressed memory, and stable long-horizon distillation. Specifically, we propose a factorized hybrid control interface that explicitly disentangles low-level movements, high-level semantic interactions, and visual content factors. Low-level movements, *e.g.*, camera motion and third-person character locomotion, are precisely modulated via frame-aligned action control. Concurrently, to govern high-level semantic interactions and content factors, we design a structured semantic control that partitions signals into three decoupled fields, *i.e.*, scene, character, and event, where each field governs a distinct concept within the world. By disentangling these signals, the model learns reusable combinations across the heterogeneous controls while retaining the precise responsiveness required for world models.

Then, we co-design the memory and distillation. For the memory mechanism, we compress the generated history into compact memory tokens, substantially reducing the computational cost for long-horizon modeling. Crucially, this design bypasses score evaluations across the full-resolution rollouts during distillation. By partitioning long-horizon rollouts into local temporal clips conditioned on compact memory tokens, we can compute scores independently per clip, ensuring scalable and computationally tractable distillation. For distillation, we propose Stable Forcing, a stable framework tailored for long-horizon distillation. We first warm-start the autoregressive student via a few-step initialization scheme inspired by PDD (Shaul et al., 2026), yielding a well-behaved few-step student. This initialization ensures that the student's rollout distribution closely aligns with that of the teacher over long horizons, thereby stabilizing subsequent distribution distillation. Building on this starting point, we perform distribution-matching distillation to enhance long-horizon consistency and mitigate exposure bias. In this stage, we introduce full-rollout replay to decouple

long-horizon rollouts from gradient backpropagation. During the forward rollout phase, each chunk undergoes full few-step sampling to maintain fidelity and bolster stability. Meanwhile, a random intermediate step is recorded and replayed with gradients during the backward pass. Combining our initialization with full-rollout replay preserves rollout quality over long horizons, thereby achieving stable and robust distillation.

Taken together, our model demonstrates remarkable generalization across different scenes and characters. As shown in Fig. 1, it not only supports versatile, multi-turn interactive controls, but also preserves geometric consistency over long horizons. Moreover, extensive quantitative and qualitative experiments validate the effectiveness of our methods, demonstrating superior performance compared to existing methods.

2 Related Work

Interactive World Models. Interactive world models generate future visual frames conditioned on previous observations and current actions, enabling users or embodied agents to interact with the environment. Recent world models have substantially expanded environmental diversity (Zhang et al., 2025; He et al., 2025; Li et al., 2025; Team et al., 2026b; Mao et al., 2025; Jiang et al., 2026; Team, 2025; 2026), long-horizon consistency (Sun et al., 2026; Hong et al., 2025; Xu et al., 2026; Team et al., 2026d; Wang et al., 2026c; Team et al., 2026c), and the range of supported controls (Parker-Holder et al., 2025; Gao et al., 2026; Team et al., 2026a; Mao et al., 2026; Alibaba, 2026; Tang et al., 2025). WorldPlay (Sun et al., 2026), Lingbot-World (Team et al., 2026d), and Wonder (Xu et al., 2026) utilize camera poses, discrete keyboard inputs, or pixel-space coordinate field as control signals to govern viewpoint transformation and character movement. Lingbot-World-V2 (Gao et al., 2026) and AlayaWorld (Team et al., 2026a) introduce language-driven events to further enable richer interactive controls. However, these control signals are inherently heterogeneous, making unified and effective representation particularly challenging. Moreover, existing methods model long-horizon consistency via retrieval (Yu et al., 2025; Xiao et al., 2026), sparse attention (Xu et al., 2026), or explicit 3D representations (Team et al., 2026c;a), while treating distillation as an isolated module. In contrast, our method co-designs the memory mechanism and distillation, improving both training efficiency and distillation stability.

Distillation. Distillation accelerates diffusion models by reducing the number of function evaluations. One representative line of studies aggregates multi-step instantaneous velocity into single-step average velocity. For instance, MeanFlow (Geng et al., 2026) derives the relationship between instantaneous and average velocities, whereas rCM (Zheng et al., 2026b), AnyFlow (Gu et al., 2026), and TiM (Wang et al., 2026b) implement a parallelism-compatible JVP kernel or differential derivation to scale this approach to large-scale models. PiFlow (Chen et al., 2026) and PDD (Shaul et al., 2026) further optimize trajectory learning to estimate average velocities more efficiently, achieving strong performance in bidirectional model distillation. Another major paradigm performs distribution matching distillation (Yin et al., 2024b;a; 2025; Zheng et al., 2026a; Zhu et al., 2026b; Huang et al., 2026), aligning the generation distribution of a few-step student with that of a multi-step teacher. This paradigm is widely adopted for interactive world models because it accelerates sampling, mitigates exposure bias, and inherits desirable properties from the teacher, such as long-horizon consistency. However, when the student performs few-step long-horizon rollouts, its generation distribution diverges significantly from that of the teacher, making distillation highly unstable.

3 Method

Our goal is to construct a real-time interactive world model $N_\theta(x_t|x_{<t}, A_{\leq t})$ parameterized by θ that supports versatile controls while maintaining long-horizon consistency. The model generates next chunk $x_t \in \mathbb{R}^{T\times H\times W}$ based on past observations $x_{<t} = \{x_{t-1}, ..., x_0\}$, controls $A_{<t} = \{A_{t-1}, ..., A_0\}$, and current control signal A_t. We first introduce our factorized hybrid control interface in Sec. 3.1, which disentangles heterogeneous inputs to support diverse interactions. In Sec. 3.2, we present our distillation-oriented compressed memory mechanism, enabling efficient long-horizon modeling and teacher supervision. Finally, we detail Stable Forcing in Sec. 3.3, a stable long-horizon distillation framework that distills a many-step autoregressive model into a few-step model while preserving long-horizon consistency. Fig. 2 illustrates the overview of our model.

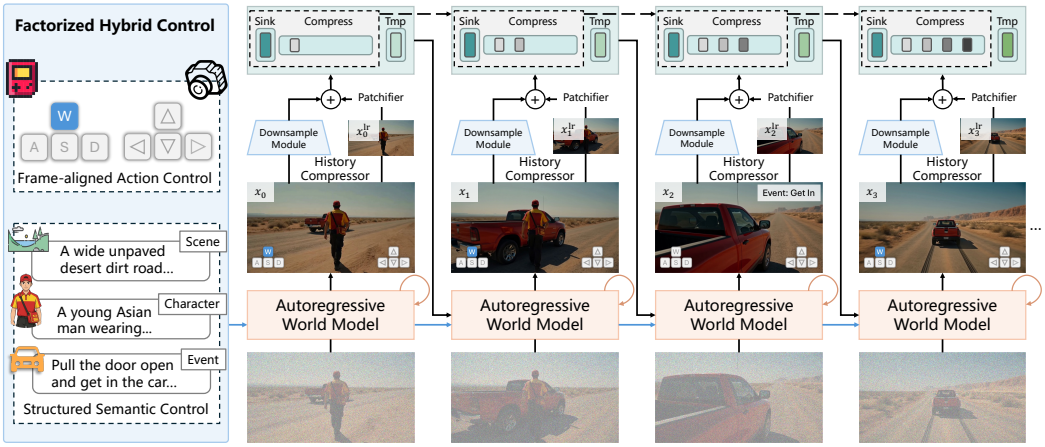

Figure 2: **Overview of WorldPlay2.** We employ factorized hybrid control interface to decouple heterogeneous inputs into frame-aligned action control and structured semantic control, ensuring accurate interactive responses. Concurrently, a history compressor abstracts past contexts into compact memory tokens to achieve efficient long-horizon inference.

3.1 Factorized Hybrid Control Interface

Interactive world models require versatile responsiveness to heterogeneous controls, which often exhibit varying levels of abstraction. Camera motion and character locomotion demand precise, frame-aligned signals, whereas complex interactions and content factors are more naturally expressed via high-level semantic instructions. Therefore, we propose a factorized hybrid control interface that disentangles low-level movements, high-level semantic interactions, and visual content factors. Specifically, each control signal is represented as $A_t = \{a_t, d_t\}$, where a_t denotes frame-aligned action control and d_t denotes structured semantic control.

Frame-aligned Action Control. Our low-level movements a_t consist of continuous camera pitch and yaw angles, discrete longitudinal and lateral movements, the camera perspective, and a special action (*i.e.*, jumping). We separately embed the continuous and discrete components and combine them into a unified action representation,

$$e_t = E(\text{continuous}) + \mathbf{e}[\text{discrete}], \tag{1}$$

where E is an MLP for continuous signals and $\mathbf{e}$ denotes the learnable embeddings for the discrete signals. The resulting action representation is aligned with the corresponding visual tokens and injected before the feed-forward network (FFN) in each Transformer block,

$$\tilde{h}_t = h_t + F\left([h_t \oplus e_t]\right), \tag{2}$$

where F is an auxiliary MLP, h_t denotes the hidden states, and $\oplus$ means channel concatenation.

Structured Semantic Control. High-level semantic control involves diverse interactions and visual content factors that often span longer temporal horizons, making it challenging to represent with low-dimensional vectors. Therefore, we utilize structured caption d_t as the semantic control signals. Specifically, d_t encapsulates visual content factors (*i.e.*, scene and character identity) as well as dynamic semantic events,

$$d_t = (d_{\text{scene}}, d_{\text{character}}, d_{\text{event}}). \tag{3}$$

Here, the scene field d_{scene} describes the environmental content, including spatial layout, objects, illumination, and visual style. $d_{\text{character}}$ specifies the persistent identity and appearance of the controlled entity. d_{event} describes the semantic change associated with the current event, such as object manipulation, environmental changes, and object appearance.

3.2 Distillation-Oriented Compressed Memory

Long-horizon world modeling requires access to historical information beyond a limited local context. A straightforward approach is to retain the entire full-resolution contexts (Team et al., 2026d;

Hong et al., 2025). However, this causes context length to scale linearly during autoregressive rollout, rapidly increasing inference latency and complicating long-horizon modeling. Moreover, this computational burden is further amplified during distillation, where the full-context bidirectional teacher is used to evaluate long student rollouts to provide supervision. To this end, we design a compressed memory mechanism inspired by Zhang et al. (2026a), shared across long-horizon modeling and distillation. This enables efficient context conditioning while keeping teacher evaluation during distillation computationally tractable.

Instead of conditioning the model directly on the full-resolution contexts, we employ a learnable history compressor $\mathcal{C}_\phi$ to encode the history context into a compact sequence of memory tokens:

$$m_{<t} = \left[x_{\text{sink}}; x_{\text{cmp}} = \mathcal{C}_\phi(x_{<t}, x^{\text{lr}}_{<t}); x_{\text{tmp}}\right], \tag{4}$$

where $[;;]$ denotes sequence concatenation, x_{sink} denotes sink tokens providing a stable reference, x_{tmp} represents adjacent temporal tokens that enforce temporal consistency, and $x^{\text{lr}}_{<t} \in \mathbb{R}^{\frac{T}{l}\times\frac{H}{s}\times\frac{W}{s}}$ denotes the low-resolution, low-frame-rate video latent encoded by the VAE, with $l = 2$ and $s = 4$ representing the temporal and spatial downsampling factors, respectively. For the history compressor $\mathcal{C}_\phi$, we adopt a dual-branch design (Zhang et al., 2026a). Specifically, the coarse branch processes $x^{\text{lr}}_{<t}$ through the DiT's patchifier to produce coarse features, while the fine branch passes $x_{<t}$ through a downsample module to yield residual fine features. This dual-branch representation provides both high-level semantic information and fine-grained visual details for subsequent generation. Compared to full-resolution contexts, our memory compression mechanism reduces the sequence length by a factor of approximately $l \cdot s^2 = 32$, significantly reducing computational overhead.

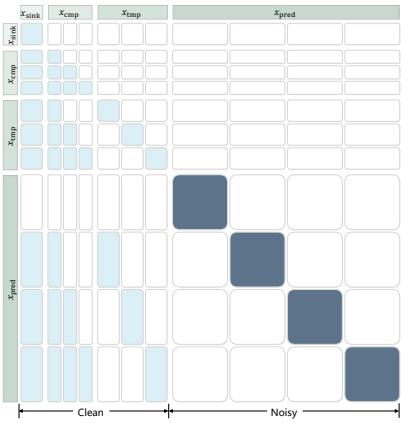

Figure 3: Causal attention mask for compressed memory training.

Given a long training video, we partition it into a compressed historical context and a target clip $x_{[t:t+L]}$. For the causal autoregressive student N_θ, we employ teacher forcing with a causal attention mask (illustrated in Fig. 3) to maintain temporal causality and follow the flow matching objective (Lipman et al., 2023),

$$\mathcal{L}_{\text{student}} = \mathbb{E}_{x,\boldsymbol{\epsilon},\sigma}\left\| N_\theta\left(x^\sigma_{[t:t+L]}, \sigma, m_{<t+L}, A_{\le t+L}\right) - (\boldsymbol{\epsilon} - x_{[t:t+L]})\right\|^2, \tag{5}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ denotes Gaussian noise and σ represents the diffusion noise level. Concurrently, the bidirectional teacher model is trained without the causal attention mask. By applying our proposed memory mechanism to both student and teacher models, we can efficiently model long-horizon consistency. Moreover, it enables long student rollouts to be partitioned into smaller clips for individual teacher evaluation, significantly reducing computational overhead during distillation.

3.3 Stable Forcing

Self Forcing (Huang et al., 2026) has emerged as an effective approach for distilling autoregressive video diffusion models, as it simultaneously reduces sampling steps and mitigates error accumulation. However, extending it to long-horizon rollout introduces new challenges. Performing long-horizon student rollouts with few sampling steps causes errors to compound rapidly across chunks, driving the student's generation distribution away from the teacher's and leading to unstable training. Additionally, evaluating scores over long rollouts using full-context teacher models demands substantial computational resources and GPU memory. To address these challenges, we introduce Stable Forcing as shown in Fig. 4, a long-horizon distillation framework designed to achieve stable training via few-step initialization, full-rollout replay, and efficient score evaluation.

Few-step Initialization. Stable long-horizon distillation requires the student to produce meaningful rollouts under few-step sampling. To establish a reliable few-step initialization, we extend PDD (Shaul et al., 2026) to our memory-augmented autoregressive student model. Specifically, PDD discretizes the diffusion noise schedule into K blocks, where each block i contains C subintervals $\{\sigma^i_0, \ldots, \sigma^i_{C-1}\}$. A parallel decoder then predicts the mean velocities $u_0, ..., u_{C-1}$ across

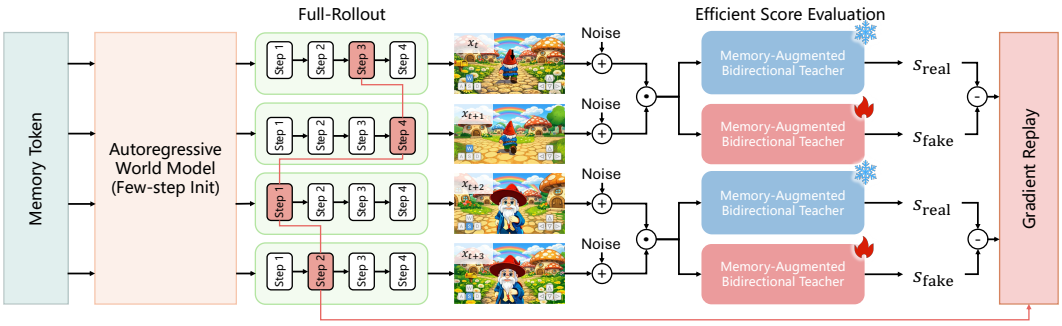

Figure 4: **Overview of Stable Forcing.** It integrates few-step initialization, full-rollout replay, and efficient score evaluation to achieve efficient and stable long-horizon distillation.

adjacent intervals within a block in a single forward pass. The training objective is,

$$\mathcal{L}_{\text{PDD}} = \mathbb{E}_{k\in[0,C-1]} \left\| u_k - \text{sg}(N_\theta(x_t^{\sigma_k^i}, \sigma_k^i, m_{<t}, A_{\le t})) \right\|^2, \tag{6}$$

where $x_t^{\sigma_k^i}$ is computed via parallel decoding sampling and $\text{sg}(\cdot)$ denotes the stop-gradient operator. By predicting multiple consecutive denoising intervals in parallel, it reduces the number of network evaluations and provides a reliable few-step initialization to stabilize subsequent long-horizon distribution matching distillation.

Full-rollout Replay. To further enhance the stability of distribution matching distillation, we decouple the rollout phase from gradient backpropagation. Specifically, during the rollout phase, each chunk performs full few-step sampling, and only its final prediction is incorporated into subsequent chunk generation and score evaluation. Simultaneously, we cache a randomly selected intermediate denoising timestep for each chunk and replay it with gradients after computing the score. This strategy not only improves rollout quality but also ensures supervision across denoising timesteps.

Efficient Score Evaluation. After the student model generates a long rollout $x_{[0:BL]}$, we perform an efficient score evaluation to obtain the distribution-matching signal. Specifically, the rollout is partitioned into B clips. For each clip $x_{[iL:(i+1)L]}$, the preceding chunks are encoded into compact memory tokens $m_{<iL}$ and the real and fake scores are evaluated using the teacher model v as follows,

$$s_{\text{fake/real}} = v(x^{\sigma}_{[iL:(i+1)L]}, \sigma, m_{<iL}, A_{\le(i+1)L}). \tag{7}$$

In this manner, we preserve long-horizon supervision while reducing the sequence length processed by the score model, thereby achieving more efficient score evaluation.

4 Experiments

Dataset. Our training corpus comprises two distinct subsets: a spatial navigation dataset and an interactive event dataset. Our navigation dataset aggregates various sources, including SpatialVID (Wang et al., 2026a), Sekai (Li et al., 2026), ABot-World (Jiang et al., 2026), internal gameplay recordings, and Unreal Engine (UE) rendering sequences, totaling 700K video clips (lasting 30s to 60s). To endow the model with flexible interactive capabilities, we construct an interactive event dataset comprising 10K clips (lasting 10s to 30s). This subset covers three categories, *i.e.*, environmental transition, object addition/removal, and complex interaction. Although the interactive event dataset is relatively small, the pretrained model inherently exhibits strong instruction-following capability. Therefore, it suffices to unlock this capability. Details are provided in the Appendix.

Implementation Details. Our training follows a multi-stage curriculum. Specifically, the base video diffusion model is first trained for navigation controls on our spatial navigation dataset. Subsequently, utilizing the full dataset, we integrate the memory compressor following the two-stage training regime of Zhang et al. (2026a) to improve long-horizon geometric consistency. Next, the bidirectional model is adapted into a chunk-wise autoregressive model via teacher forcing, initialized for few-step generation via PDD (Shaul et al., 2026). Finally, we leverage distribution matching distillation to obtain the target world model. Furthermore, we deploy inference optimizations covering computation graph fusion, low-bit quantization, KV caching, and a lightweight VAE to achieve real-time generation at 16 FPS on 8 H20 GPUs. See Appendix for more details.

Table 1: **Quantitative comparisons.** We benchmark our approach against recent interactive world models on both WBench and RevisitBench to systematically evaluate controllability, long-horizon consistency, and visual fidelity. *Abbreviations: Avg: Average, Qua: Quality, Set: Setting, Int: Interaction, Con: Consistency, Phy: Physical.*

	WBench						RevisitBench			
	Avg. ↑	**Qua.** ↑	**Set.** ↑	**Int.** ↑	**Con.** ↑	**Phy.** ↑	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	**MEt3R** ↓
WorldPlay (Sun et al., 2026)	78.1	78.1	72.2	86.8	86.9	66.3	17.05	0.553	0.416	0.179
AlayaWorld (Team et al., 2026a)	76.3	79.3	69.7	80.0	89.5	63.1	13.61	0.399	0.569	0.244
Lingbot-World-V2 (Gao et al., 2026)	79.4	81.8	76.8	82.8	86.5	69.1	14.46	0.469	0.506	0.301
EchoWM (Zhang et al., 2026b)	81.0	81.1	77.5	87.9	88.3	70.1	14.83	0.513	0.477	0.283
Alaya-Evoke-Turbo (Yin et al., 2026)	82.0	**81.9**	**82.1**	83.9	88.1	74.0	16.33	0.501	0.412	0.236
Ours (w/o Stable Forcing)	79.0	79.1	77.8	82.4	87.4	68.2	16.52	0.535	0.439	0.215
Ours (full)	**83.1**	81.8	81.5	**88.3**	**90.0**	**74.0**	**19.71**	**0.613**	**0.318**	**0.105**

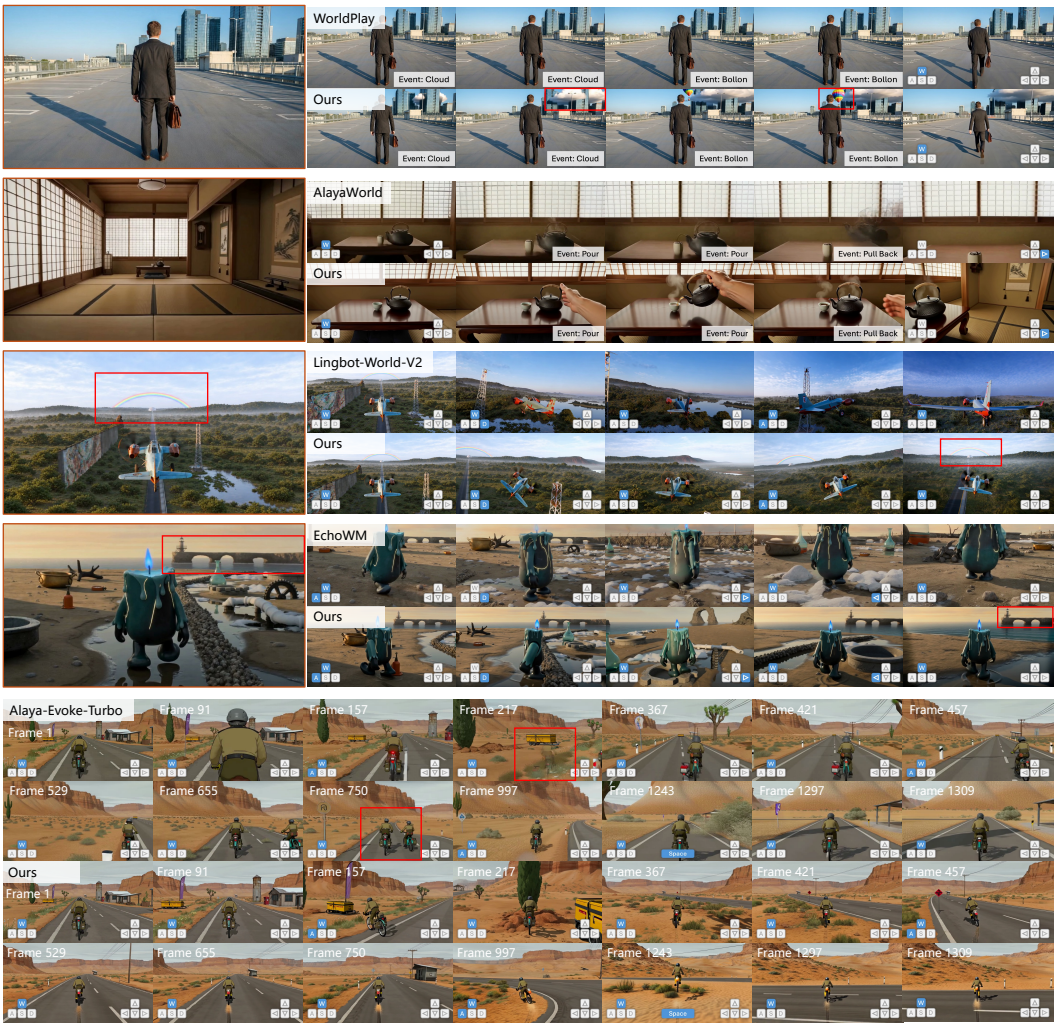

Figure 5: **Qualitative comparisons with existing methods.** WorldPlay2 supports versatile interactive events while demonstrating superior generalization across diverse characters and maintaining long-horizon geometric consistency.

Evaluation Details. To assess controllability, consistency, and visual fidelity, we benchmark all models on WBench (Ying et al., 2026). To systematically assess long-horizon geometric consistency, we introduce RevisitBench, an evaluation benchmark comprising 200 revisit trajectories following the protocol in Sun et al. (2026), which are curated from WBench and our self-collected validation sets. We quantify 2D visual consistency via LPIPS, PSNR, and SSIM, as well as 3D spatial consistency via MEt3R (Asim et al., 2025). Additionally, we curate 145 samples from WBench

and our interactive event dataset as test cases, which are held out from training, to evaluate model responsiveness to diverse interactive events. We deploy a vision-language model (VLM) (Seed, 2025) as an automated evaluator to score instruction adherence and execution accuracy. We compare our approach against five interactive world models: WorldPlay (Sun et al., 2026), AlayaWorld (Team et al., 2026a), Lingbot-World-V2 (Gao et al., 2026), EchoWM (Zhang et al., 2026b), and Alaya-Evoke-Turbo (Yin et al., 2026).

4.1 Comparisons with Existing Methods

Quantitative Results. Tab. 1 compares our method with five representative interactive world models on the navigation split of WBench and RevisitBench. WorldPlay2 achieves the highest overall average score of 83.1, surpassing the previous state-of-the-art baseline, Alaya-Evoke-Turbo (Yin et al., 2026), by 1.1 points. Notably, WorldPlay2 demonstrates a pronounced advantage in Interaction and Consistency. This empirically validates that our factorized hybrid control interface, coupled with the co-design of memory and distillation, significantly enhances long-horizon consistency while maintaining precise navigation controllability. RevisitBench further evaluates long-horizon geometric consistency under loop-closure trajectories. Constrained by fixed context window lengths, Lingbot-World-V2 (Gao et al., 2026) and EchoWM (Zhang et al., 2026b) suffer from memory degradation, failing to preserve geometric consistency. Although WorldPlay (Sun et al., 2026) incorporates camera poses for historical context retrieval, it is inherently susceptible to compounding camera pose drift as rollouts progress, making it challenging to retrieve accurate context. AlayaWorld (Team et al., 2026a) and Alaya-Evoke-Turbo (Yin et al., 2026) rely on explicit 3D representations to maintain memory, but suffer from metric scale ambiguities across different chunks. Such scale discrepancies severely impede fine-grained control and inevitably induce geometric inconsistencies. In contrast, WorldPlay2 maintains robust long-horizon geometric consistency without relying on error-prone retrieval or sensitive explicit 3D representations. Furthermore, as evidenced by the results, Stable Forcing substantially improves performance, directly demonstrating its effectiveness.

As shown in Tab. 2, we evaluate responsiveness across diverse interactive events. Although Lingbot-World-V2(Gao et al., 2026) and AlayaWorld (Team et al., 2026a) leverage chunk-wise captions to accommodate diverse controls, they fail to disentangle underlying world states. This entangles content factors with interactive events, making it challenging for them to capture precise correspondences between textual descriptions and visual dynamics, particularly in complex interactions. In contrast, our model explicitly factorizes the world state via the factorized hybrid control interface, yielding superior interactive fidelity.

Table 2: Quantitative comparison on responsiveness to interactive events. *Abbreviations: EO: Environment and Object change, CI: Complex Interaction.*

Method	Avg. ↑	EO. ↑	CI. ↑
WorldPlay (Sun et al., 2026)	43.3	47.9	38.7
AlayaWorld (Team et al., 2026a)	39.4	48.3	30.5
Lingbot-World-V2 (Gao et al., 2026)	52.5	61.5	43.4
EchoWM (Zhang et al., 2026b)	43.8	56.3	31.2
Alaya-Evoke-Turbo (Yin et al., 2026)	40.7	45.7	35.7
Ours	**74.7**	**79.2**	**70.1**

Qualitative Results. Fig. 5 presents the qualitative comparisons with baselines. Due to the scarcity of interactive event datasets and reliance on entangled control conditioning, WorldPlay (Sun et al., 2026) and AlayaWorld (Team et al., 2026a) fail to respond accurately to diverse interactive commands. Meanwhile, Lingbot-World-V2 (Gao et al., 2026) and EchoWM (Zhang et al., 2026b) struggle to maintain long-horizon geometric consistency as a consequence of memory decay inherent in their sliding-window designs. Furthermore, their lack of factorized representations between background scenes and foreground characters impedes precise and smooth locomotion, such as airplane turning and entity centering. Although Alaya-Evoke-Turbo (Yin et al., 2026) can generate long sequences, its dependence on explicit 3D representations often introduces severe temporal flickering and visual artifacts, such as ghosting and duplicate entity appearances. In contrast, our approach reliably executes diverse interaction events, enables fine-grained and fluid control over different characters, and preserves long-horizon consistency, highlighting the efficacy and superiority of our method. Please refer to the supplementary videos for comprehensive visualizations.

4.2 Ablation Studies

Controllability. Fig. 6(a) presents the ablation study on controllability. When trained without the interactive event dataset, the model's interactive capabilities are constrained, failing to respond to

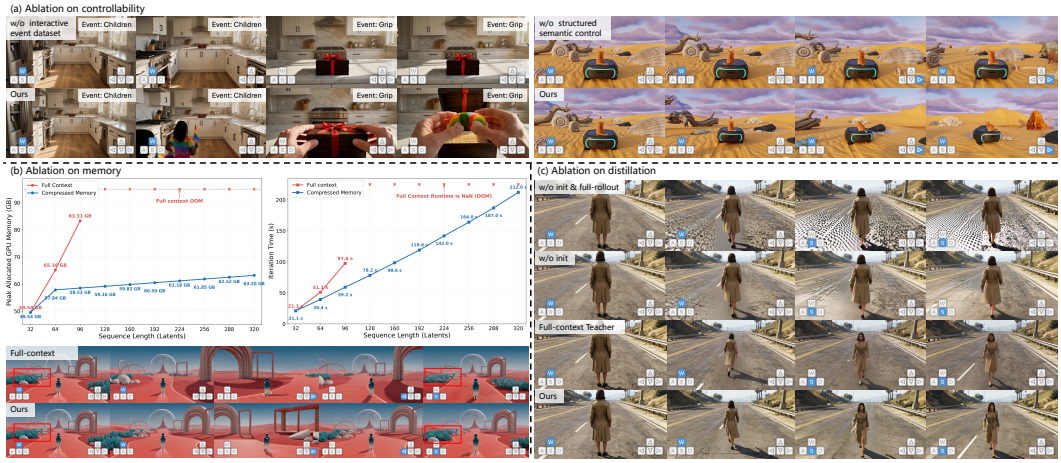

Figure 6: **(a) Ablation on controllability:** We verify the impact of interactive data scaling and structured semantic control. **(b) Ablation on memory:** We compare the GPU memory footprint and training time with the full-context baseline. For our method, we partition long sequences into multiple clips and compute sequentially within a single iteration. **(c) Ablation on distillation:** We analyze the components of Stable Forcing. Zoom in for details.

semantic events. Incorporating even a small fraction of interactive data unlocks these capabilities, enabling coherent interactions spanning dynamic semantic events and navigation controls. Furthermore, omitting our structured semantic control causes the model to conflate intricate foreground characters with background scenes, compromising precise navigation controllability.

Memory. To validate the efficiency of our compressed memory, we compare its GPU memory footprint and training time with the full-context baseline across various video lengths, as summarized in Fig. 6(b). At the same sequence lengths, compressed memory substantially reduces both memory consumption and iteration time, enabling scalable training and distillation on longer sequences. As shown in Fig. 6(b), our compressed memory achieves comparable long-horizon geometric consistency to its full-context counterpart when trained on 96 latents, demonstrating its effectiveness.

Distillation. Fig. 6(c) demonstrates the stability and efficacy of Stable Forcing. We first ablate the full-rollout replay, omitting it leads to progressive quality degradation during long-horizon generation and eventually triggers mode collapse, as evidenced by the ground artifacts. Furthermore, we evaluate the impact of the PDD initialization. Employing PDD initialization effectively mitigates blurry outputs and grid-like artifacts, confirming the robustness of our method. Additionally, while distilling with the full-context teacher achieves competitive performance, it demands substantially longer training wall-clock time (261s vs. 167s per iteration). Furthermore, extending the full-context teacher to longer horizon distillation (*e.g.*, 320 latents) triggers out-of-memory issues, whereas our method scales with superior efficiency.

5 Conclusion and Limitations

We present WorldPlay2, an interactive world model that achieves real-time responsiveness, versatile control, and long-horizon consistency. It significantly expands the interactive capabilities of world models, faithfully executing both navigation-oriented controls and semantic interactive events while maintaining geometric consistency over long horizons. Crucially, WorldPlay2 is designed with scalability at its core, which scales efficiently and stably as compute budgets and data volume expand. We hope WorldPlay2 serves as a crucial step toward advancements in embodied intelligence, spatial computing, and interactive entertainment.

Limitations. Despite the promising capabilities demonstrated by WorldPlay2, several challenges need further investigation. First, characters are still prone to gradual visual and semantic drift, occasionally failing to preserve strict identity consistency during long rollouts. Second, scaling our framework to infinite-horizon generation remains an open challenge. Maintaining both infinite rollout stability and long-term geometric consistency without error accumulation represents one of the most fundamental yet demanding frontiers in interactive world modeling.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets, implementing methods, assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, and interpreting results. We have not used generative AI tools for designing or providing feedback on research methodology or experiments. The rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code and editing the paper to improve readability. We have reviewed all AI-assisted work. Specifically, both the synthesized datasets and the LLM-generated code were verified and tested to guarantee data fidelity and execution correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken care to ensure that our results can be reproduced. The full model formulation and training objectives are presented in the main paper. Moreover, the complete training, evaluation details, and the composition and processing of our training dataset are provided in the Appendix.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Alibaba. HappyOyster, 2026. `https://www.happyoyster.com`.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. MEt3R: Measuring multi-view consistency in generated images. In *CVPR*, pp. 6034–6044, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Ollin Boer Bohan. TAEHV: Tiny autoencoder for hunyuan video. `https://github.com/madebyollin/taehv`, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Hansheng Chen, Kai Zhang, Hao Tan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. Pi-Flow: Policy-based few-step generation via imitation distillation. In *ICLR*, pp. 151521–151547, 2026.

Google Deepmind. Veo3 video model, 2025. `https://deepmind.google/models/veo/`.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex Attention: A programming model for generating optimized attention kernels. *arXiv preprint arXiv:2412.05496*, 2(3):4, 2024.

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.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *NeurIPS*, 38:75460–75482, 2026.

Yuchao Gu, Guian Fang, Yuxin Jiang, Weijia Mao, Song Han, Han Cai, and Mike Zheng Shou. AnyFlow: Any-step video diffusion model with on-policy flow map distillation. *arXiv preprint arXiv:2605.13724*, 2026.

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

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. RELIC: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

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

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. *NeurIPS*, 38:167283–167308, 2026.

Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, Wenlin Liu, Yun Wang, Xue Zheng, Rui Sun, et al. ABot-World-0: Infinite interactive world rollout on a single desktop gpu. *arXiv preprint arXiv:2607.19191*, 2026.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-GameCraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2(3):6, 2025.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *NeurIPS*, 38, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, and Kaipeng Zhang. Yume1.5: A text-controlled interactive world generation model. In *CVPR*, pp. 7752–7761, 2026.

OpenAI. Gpt 6 astra, 2026. `https://openai.com/index/gpt-6-astra/`.

Jack Parker-Holder, Shlomi Fruchter, et al. Genie 3: A new frontier for world models. *Google DeepMind Blog*, 2025.

ByteDance Seed. Seed 2.1, 2025. `https://seed.bytedance.com/en/seed2_1`.

Neta Shaul, Chao Liu, Arash Vahdat, and Julius Berner. Parallel decoding distillation for fast image and video generation. *arXiv preprint arXiv:2607.26004*, 2026.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling. In *ICML*, 2026.

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, et al. Hunyuan-GameCraft-2: Instruction-following interactive game world model. *arXiv preprint arXiv:2511.23429*, 2025.

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

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemini Robotics Team, Krzysztof Choromanski, Coline Devin, Yilun Du, Debidatta Dwibedi, Ruiqi Gao, Abhishek Jindal, Thomas Kipf, Sean Kirmani, Isabel Leal, et al. Evaluating gemini robotics policies in a veo world simulator. *arXiv preprint arXiv:2512.10675*, 2025.

InSpatio Team, Donghui Shen, Guofeng Zhang, Haomin Liu, Haoyu Ji, Hujun Bao, Hongjia Zhai, Jialin Liu, Jing Guo, Nan Wang, et al. Inspatio-world: A real-time 4d world simulator via spatiotemporal autoregressive modeling. *arXiv preprint arXiv:2604.07209*, 2026c.

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*, 2026d.

Tencent HY World Team. HunyuanWorld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.

Tencent HY World Team. HY-World 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv preprint arXiv:2604.14268*, 2026.

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.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. SpatialVID: A large-scale video dataset with spatial annotations. In *CVPR*, pp. 42592–42603, 2026a.

Zidong Wang, Yiyuan Zhang, Xiaoyu Yue, Xiangyu Yue, Yangguang Li, Wanli Ouyang, and Lei Bai. Transition models: Rethinking the generative learning objective. In *CVPR*, pp. 29178–29189, 2026b.

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*, 2026c.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

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.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. *NeurIPS*, 38:49632–49652, 2026.

Jiacong Xu, Hanwen Jiang, Zhixin Shu, Kalyan Sunkavalli, Vishal M Patel, and Yiqun Mei. Wonder: Video world model done better. *arXiv preprint arXiv:2607.26037*, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *NeurIPS*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, pp. 6613–6623, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, pp. 22963–22974, 2025.

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.

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 *SIGGRAPH Asia*, pp. 1–11, 2025.

Jintao Zhang, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, and Jianfei Chen. SageAttention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. *arXiv preprint arXiv:2411.10958*, 2024.

Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. TinyHistory: Lightweight video history embeddings via two-stage context learning. In *ECCV*, 2026a.

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*, 2026b.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-Game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

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*, 2026a.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *ICLR*, pp. 2582–2603, 2026b.

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*, 2026a.

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 *ICML*, 2026b.

APPENDIX

A DISCUSSION

A predominant paradigm among concurrent interactive world models (Gao et al., 2026; Jiang et al., 2026) relies on sliding-window mechanisms to facilitate autoregressive rollout over extended temporal horizons. While computationally tractable, this formulation fundamentally enforces a local assumption: the model's predictive distribution is heavily conditioned on adjacent frames, inherently ignoring distant observations. Consequently, such architectures struggle with long-horizon geometric consistency. In contrast, our framework departs from local receptive fields by designing an efficient memory mechanism with a stable distillation. This enables our model to anchor spatiotemporal invariants across long rollouts, maintaining high-fidelity geometric and semantic consistency. Despite these gains, we candidly acknowledge the boundaries of our approach. Scaling our framework to unbounded, infinite-horizon generation remains an open challenge. We posit that achieving both infinite-horizon rollout and long-term geometric persistence represents one of the most fundamental frontiers in interactive world modeling. Furthermore, recent coding agents (OpenAI, 2026) have demonstrated a remarkable ability to synthesize spatially coherent 3D environments via programmatic generation. This milestone signals a profound paradigm shift, transitioning agents from static textual domains to dynamic embodied multi-modal domains. We believe that integrating such intelligence into generative world models represents an exceptionally promising frontier.

B DATASET

B.1 SPATIAL NAVIGATION DATASET

As detailed in Tab. 3, our spatial navigation dataset comprises three complementary categories. First, to capture real-world physical dynamics and complex photorealistic textures, we curate high-quality subsets from SpatialVID (Wang et al., 2026a) and Sekai (Li et al., 2026). Specifically, we curate videos captured in open unobstructed environments with smooth navigation trajectories. Second, to reinforce long-term geometric consistency under complex camera trajectories, we design a dedicated loop-closure rendering pipeline in UE. The synthesized camera trajectories are strictly self-symmetric, *i.e.*, the second half of the sequence precisely reverses the camera path of the first half. During the first half exploration phase, multiple navigation actions are executed with randomized durations ranging from 1s to 5s. Finally, to broaden trajectory diversity and improve generalization, we incorporate open-source gameplay datasets (Jiang et al., 2026) and scale the simulation recording dataset. This substantially enriches the coverage of game genres, dynamic environments, and third-person characters.

B.2 INTERACTIVE EVENT DATASET

Our interactive event dataset comprises three categories: complex interaction, environmental transition, and object addition/removal, as detailed in Tab. 3 and Fig. 7. For complex interaction, we aim to jointly capture intricate behavioral interactions alongside spatial navigation, spanning 13 distinct action categories across both first-person and third-person perspectives. For environmental transition, we collect diverse videos capturing dynamic weather shifts (*e.g.*, clear skies transitioning to heavy rain, fog, or snow), seasonal evolutions (*e.g.*, summer lushness shifting to autumnal foliage or winter snowscapes), and holistic stylistic variations (*e.g.*, day-to-night lighting and artistic rendering styles). For object addition and removal, our dataset encompasses the emergence and disappearance of dynamic agents such as animals, flying crafts (*e.g.*, drones and aircraft), vehicles, tools, boxes, and handheld items.

B.3 DATA PROCESSING

Our data processing pipeline consists of three stages: structured semantic captioning, spatial navigation action extraction, and quality filtering.

Structured Semantic Captioning. In alignment with our structured semantic control, each video clip requires a disentangled textual description. To this end, we utilize VLMs (Team et al., 2023; Bai et al., 2025) to parse the visual information into structured captions using the following template.

Structured Semantic Captioning Template

Role: You are a professional video annotator specializing in interactive world models.

Core Task: You will be given a video. Your task is to describe it and output a single JSON object. Follow these rules exactly.

Important: This video has been pre-labeled as: PERSPECTIVE. Trust this label and follow the corresponding perspective handling rule below.

— Rules —

Language: Write all field values in English.

Output: Return ONLY one valid JSON object with exactly four keys: *scene*, *character*, *UI*, *dynamic_interactions*. No extra text, no markdown code fences, no commentary. Do not describe camera motion, camera rotation, camera shake, zooming, viewpoint movement, gameplay controls. Be specific, concrete, and visually grounded. Do not invent details that cannot be seen.

Perspective handling (affects the *character* field only): TPS: Describe the visual appearance of any visible player-controlled subject (humanoid, vehicle, or creature) under *character*. FPS: Set *character* to "None".

Scene: A unified descriptor encompassing visible terrain/objects, the environmental setting (e.g., forest, urban, room), and aesthetic styles (e.g., lighting, rendering scheme). Dynamic actions and transient events are prohibited.

Character: An appearance-only descriptor of the controlled agent, strictly omitting transient poses, gestures, or kinetic activities. It details four morphological dimensions: agent taxonomy & build, surface styling, equipped accessories, and active mounts.

UI: A unified spatial layout descriptor of all visible on-screen interface elements—including HUD components, status meters, minimaps, crosshairs, and text overlays—along with their designated screen positions, defaulting to "None" when no graphical interface is present.

dynamic_interactions: A chronological array of 5s intervals (start_time, end_time, event) cataloging visually grounded causal dynamics. It prioritizes salient physical engagements, hand gestures, contact transitions (approach, manipulation, release), and resulting state changes. Camera motions, static descriptions, and speculative actions are strictly excluded; inactive segments default to "None".

— Output Format —

```
{
    "scene": "...",
    "character": "...",
    "UI": "...",
    "dynamic_interactions": [
      {
        "start_time": 0,
        "end_time": 5,
        "event": "..."
      },
      {
        "start_time": 5,
        "end_time": 10,
        "event": "..."
      }
    ]
}
```

Spatial Navigation Action Extraction. We observe that directly applying ViPE (Huang et al., 2025) to long-horizon video clips frequently suffers from severe trajectory drift and scale collapse. To mitigate this degradation, we partition each long video into overlapping clips of 150 frames with a 10-frame temporal overlap. Camera poses are estimated independently for each sub-clip and

Table 3: **Data organization.** We detail the data type, category, data source, clip counts, and their corresponding proportions in the training corpus.

Data Type	Category	Data Source	Quantity	Ratio
Spatial Navigation	Real-World Dynamics	SpatialVID (Wang et al., 2026a)	30K	4.22%
		Sekai (Li et al., 2026)	20K	2.81%
	Synthetic 4D Scenes	UE Rendering	50K	7.04%
	Simulation Dynamics	ABot-World (Jiang et al., 2026)	30K	4.22%
		Gameplay Recording	570K	80.28%
Interactive Event	Complex Interaction	Internal Dataset	5K	0.70%
	Environmental Transition		3K	0.42%
	Object Addition/Removal		2K	0.28%

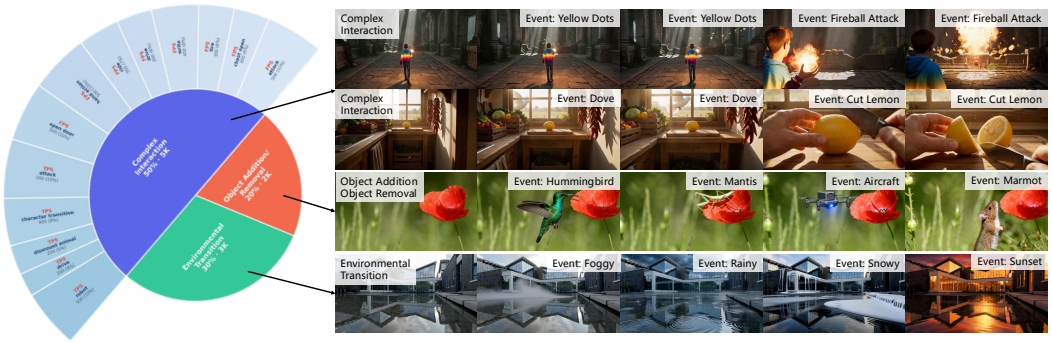

Figure 7: **Overview of our interactive event dataset.** The dataset encompasses three categories, including complex interaction, environmental transition, and object addition/removal. *Left*: The composition of the dataset. *Right*: Representative visual examples for each category.

subsequently stitched together by aligning metric scales using the overlapping windows. From the reconstructed camera poses, we derive frame-aligned action controls by decomposing the motion into discrete longitudinal and lateral translations alongside continuous pitch and yaw angles. Our internal gameplay recordings are captured under constant angular velocities per game. Leveraging this property, we compute the directional angular velocities offline for each game and adopt their empirical medians as the calibrated rotation rates along the corresponding axes.

Quality Filtering. Our quality filtering pipeline operates along two distinct dimensions: visual fidelity and action precision. Following Wu et al. (2025), we employ a comprehensive visual quality assessment model coupled with an aesthetic scoring operator to evaluate video clips across five perceptual dimensions, *i.e.*, sharpness, fine-detail retention, noise and compression artifacts, dynamic range, and aesthetic appeal, systematically filtering out low-quality candidates. Then, we evaluate the temporal smoothness and physical plausibility of the estimated camera poses for each video clip, filtering out abrupt camera jitter to ensure accurate action labels.

C More Implementation Details

C.1 Architecture Details

Frame-aligned Action Module. Since our frame-aligned actions contain both continuous camera rotations and discrete controls, we devise tailored encoding pathways as shown in Fig. 8. For continuous camera rotations, we employ an MLP to model varying angular velocities. For discrete controls, we adopt learnable embeddings to effectively capture concrete patterns. The two action embeddings are

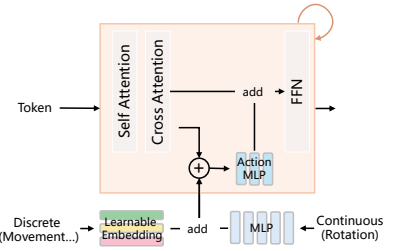

Figure 8: Illustration of our action module.

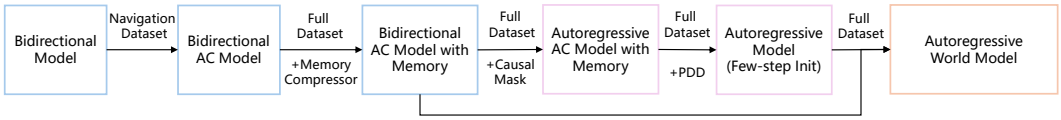

Figure 9: **Overview of our multi-stage training pipeline.** The bidirectional base model is progressively adapted into a streaming autoregressive model via action conditioning (AC), memory integration, and distillation.

Algorithm 1 Stable Forcing (Full-rollout Replay and Efficient Score Evaluation)

Require: Initialized student N_θ; real and fake score models $v_{\text{real}}, v_{\text{fake}}$; the number of rollout chunks M; denoising schedule $\{\sigma_d\}_{d=0}^{D}$;
Output: Few-step autoregressive student.

$\mathcal{H} \leftarrow \emptyset$; $\mathcal{X} \leftarrow \emptyset$
Stage 1: Full-rollout without gradients
for $i = 1, \ldots, M$ **do**
  $m_{<i} \leftarrow \text{CompressedMemory}(\mathcal{H})$
  $z_i \sim \mathcal{N}(0, I)$; $d_i \sim \text{Uniform}(\{0, \ldots, D-1\})$
  for $d = 0, \ldots, D-1$ **do**
    if $d = d_i$ **then**
      $\mathcal{R}_i \leftarrow \text{Snapshot}(z_i, \sigma_d, m_{<i}, A_{\leq i})$
    end if
    $z_i \leftarrow \text{Denoise}(z_i, N_\theta(z_i, \sigma_d, m_{<i}, A_{\leq i}))$
  end for
  Append $\text{sg}(z_i)$ to $\mathcal{X}$ and $\mathcal{H}$
end for
Stage 2: Efficient score evaluation without gradients
Partition $\mathcal{X}$ into B consecutive clips $\{x_b\}_{b=1}^{B}$
for $b = 1, \ldots, B$ **do**
  Sample score timestep σ and noise ϵ
  $x_b^\sigma \leftarrow \text{AddNoise}(x_b, \sigma, \epsilon)$
  $g_b \leftarrow \left[v_{\text{fake}}(x_b^\sigma, \sigma, m_{<b}, A_{\leq b}) - v_{\text{real}}(x_b^\sigma, \sigma, m_{<b}, A_{\leq b})\right]$
end for
Stage 3: Gradient replay with gradients
for $i = 1, \ldots, M$ **do**
  $\widetilde{z}_i \leftarrow \text{CleanPrediction}(z_i, \sigma_{d_i}, N_\theta(\mathcal{R}_i))$
  $\mathcal{L}_i \leftarrow \text{MSE}(\widetilde{z}_i, \text{sg}(\widetilde{z}_i - g_i))$
  Backpropagate $\mathcal{L}_i$
end for
Update θ once using the accumulated gradients
Update v_{fake} following DMD2 (Yin et al., 2024a)

then added and injected before the FFN in each Transformer block. Interestingly, we observe that omitting explicit camera poses incurs no visible degradation in navigation control. Moreover, enforcing rigid camera trajectories makes it difficult to simulate character-centric rotations. In contrast, our design unleashes smoother, more fluid third-person character locomotion and delivers stronger generalization across diverse characters.

History Compressor. Our history compressor builds upon the design paradigm of TinyHistory (Zhang et al., 2026a), combining 3D convolutions to compress spatiotemporal history with attention modules to enhance expressiveness. We further replace standard 3D convolutions with causal 3D convolutions. This design strictly prevents future information leakage, thereby adhering to the temporal causality of autoregressive generation.

C.2 Training Details

Our multi-stage training pipeline is summarized in Fig. 9. For the bidirectional model, each training sequence comprises 32 target latents with the associated memory context. For the AR model, the

Table 4: **Inference speed improvements from system optimizations.** Latency is benchmarked as the average time per chunk across multiple generated chunks. Latency reductions are reported relative to the baseline.

Optimization	Latency (s)	Speedup	Reduction
Baseline	4.138	1.00×	–
+ QKRoPE	4.103	1.01×	0.85%
+ QKVFusion	4.077	1.01×	1.47%
+ FFNCompile	4.074	1.02×	1.55%
+ Text Cache	3.955	1.05×	4.42%
+ FP8 Quantization	3.805	1.09×	8.05%
− FSDP	3.562	1.16×	13.92%
+ SageAttention (Zhang et al., 2024)	3.432	1.21×	17.06%
+ LightVAE (Final)	**0.998**	4.15×	75.88%

training sequence is formulated over 16 target latents, *i.e.*, 4 chunks, along with their corresponding memory context. To optimize AR training throughput, we pad the variable-length memory contexts to a uniform sequence length, which enables efficient kernel execution via FlexAttention (Dong et al., 2024). Furthermore, both bidirectional and AR models are trained using a progressive curriculum, increasing the maximum data length to facilitate smooth convergence. During Stable Forcing, we leverage the AR model as the student, distilling it into 4 steps under the supervision of the bidirectional teacher model. Specifically, the student model performs self-rollout over a horizon of 320 latents, which the teacher partitions into 10 clips to efficiently compute scores. To compute backward passes over these long-horizon sequences, we adopt a gradient replay strategy analogous to RELIC (Hong et al., 2025), effectively reducing peak GPU memory footprints during backpropagation. Alg. 1 outlines the pseudocode of Stable Forcing.

C.3 Inference Details

Complementing our algorithmic acceleration, we implement end-to-end systems optimizations across the entire inference pipeline, ultimately achieving a real-time streaming throughput of 16 FPS on 8 NVIDIA H20 GPUs as shown in Tab. 4. These optimizations include three complementary dimensions, *i.e.*, computational graph fusion, quantization coupled with caching reuse, and lightweight VAE.

Computational Graph Fusion. We fuse the rotary position embeddings with the query-key linear projections (QKRoPE), combine query-key-value linear projections into a single linear projection (QKVFusion), and apply block-level compilation to the feed-forward networks (FFNCompile).

Quantization Coupled with Caching Reuse. We incorporate FP8 quantization (FP8 Quantization) alongside SageAttention (Zhang et al., 2024) to further compress execution latency and peak memory footprints. Moreover, we cache the text embeddings (Text Cache), which bypasses redundant text encoding passes. Finally, we disable Fully Sharded Data Parallel (FSDP) during inference, thereby eliminating cross-device collective communication overheads.

Lightweight VAE. To overcome the severe latency bottleneck inherent in video VAE, we redesign and retrain a lightweight VAE grounded on TAE (Boer Bohan, 2025).

C.4 Evaluation Details

To systematically evaluate long-horizon geometric consistency, we construct RevisitBench, comprising 50 10s and 150 30s test cases with loop-closure trajectories. Specifically, for each test case, we first randomly sample discrete movement and continuous rotation to generate the trajectory for the first half of the sequence. For the remaining duration, we invert the preceding action sequence, making the model revisit its original path. These rigorous loop-closure cases reliably evaluate the geometric consistency of world models.

As described in the main paper, to evaluate the responsiveness to diverse interactive events, we leverage the VLM as an automated evaluator to quantitatively verify instruction adherence and execution

accuracy. Specifically, regarding environment and object change, we assess whether the expected state transitions faithfully occur and reach completion, as well as whether these visual transformations semantically correspond to the given interactive controls. For complex interactions, we further examine whether the fine-grained interactive details strictly align with the user-specified control inputs. We employ the following prompt template.

Interactive Evaluation Template

Role: You are an impartial evaluator of interactive world models. Evaluate how faithfully the video responds to the provided interactive control instructions, using only visual observations.

Inputs: Video: `<VIDEO>`; Interaction category: `<EO or CI>`; Control instructions: `<INSTRUCTIONS>`; Instruction timing: `<FRAME RANGES>`;

Evaluate the following dimensions:

A. Instruction Adherence. Assess whether the observed response matches the requested action, target, and attributes, without incorrect targets, reversed actions, or unintended substitutions. Assess whether the action or state transition reaches its intended endpoint. Distinguish completion from attempts or partial execution.

B. Execution Accuracy. Assess whether the requested action is visibly executed correctly. For *EO*, verify that the intended state transition occurs with the specified direction and attributes. For *CI*, verify the interaction and relevant details, including contact location, motion direction, manipulation, and the resulting relationship. Do not impose requirements beyond the instruction.

— Scoring —

Assign an integer from 0 to 4 to each dimension:
- 0: No fulfillment, or clear contradiction of the requirement.
- 1: Minimal fulfillment, with major errors or only a weak attempt.
- 2: Partial fulfillment, with substantial omissions or errors.
- 3: Mostly fulfilled, with minor errors or limited incompleteness.
- 4: Fully fulfilled, supported by clear visual evidence.

— Output Format —

```
{
    "category": "EO or CI",
    "evaluations": [
      {
        "instruction_id": 1,
        "instruction": "...",
        "scores": {
          "instruction_adherence": "3",
          "execution_accuracy": "4",
        },
        "reasoning": "Provide the evidence."
      }
    ]
}
```

D More Results

D.1 More Visualizations

Fig. 10, Fig. 11, and Fig. 12 illustrate the comprehensive qualitative performance of WorldPlay2 across diverse environments, different characters, and versatile controls. WorldPlay2 not only accommodates intricate semantic interactions (*e.g.*, dexterous grasping and box opening) alongside macro-level environmental transitions (*e.g.*, weather shift and explosion), but also exhibits exceptional third-person character controllability. Specifically, it faithfully synthesizes smooth character motions that adhere strictly to the physical and kinematic constraints of diverse entities. Furthermore, it sustains robust geometric consistency and structural fidelity even under compound, multi-stage interactive controls.

Table 5: Quantitative comparison on interactive event dataset.

Method	Avg. ↑	EO. ↑	CI. ↑
w/o interactive event dataset	46.7	57.5	35.8
Ours (Bidirectional)	**84.0**	**86.8**	**81.2**

Table 6: Quantitative comparison on structured semantic control.

Method	Rot. Err. ↓	Trans. Acc. ↑
w/o structured semantic control	0.478	88.9
Ours (Bidirectional)	**0.168**	**98.3**

Table 7: Quantitative comparison on memory. For fair comparisons, both variants utilize a bidirectional backbone and train under 96 latents.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	MEt3R ↓
Full-context	18.47	**0.688**	**0.406**	0.133
Ours	**18.80**	0.653	0.414	**0.128**

Table 8: Quantitative comparison on distillation.

	WBench					
Method	Avg. ↑	Qua. ↑	Set. ↑	Int. ↑	Con. ↑	Phy. ↑
w/o init & full-rollout	73.9	74.4	64.6	81.9	82.6	65.8
w/o init	75.2	76.2	68.0	82.6	83.3	65.9
Full-context Teacher	82.3	81.6	79.5	88.1	89.4	72.9
Ours	**83.1**	**81.8**	**81.5**	**88.3**	**90.0**	**74.0**

D.2 Quantitative Ablations

Controllability. To validate the factorized hybrid control interface, we perform ablations on our bidirectional teacher model. As presented in Tab. 5, we first probe the role of the interactive event dataset on semantic responsiveness. We observe that even in the absence of the interactive event data, the model already exhibits moderate responsiveness. Consequently, introducing a small fraction of interactive data is highly sample-efficient, yielding substantial gains in semantic interactivity. Moreover, we assess navigation performance regarding structured semantic control as in Tab. 6. Ablating this module causes navigation conditioning to couple with other world states, inducing semantic entanglement that degrades navigation accuracy.

Memory. As reported in Tab. 7, we evaluate long-horizon geometric consistency on revisit trajectories, comparing our compressed memory against the full-context baseline, where both variants are trained on 96 latents. While retaining uncompressed, full-resolution history endows high expressive capacity, training this model incurs prohibitive computational overhead and training time. In contrast, our compressed memory mechanism models long-horizon information in a resource-efficient manner, achieving highly competitive geometric consistency while drastically alleviating the training burden.

Distillation. To validate our distillation, we quantitatively benchmark different variants on WBench, as summarized in Tab. 8. Without few-step initialization and full-rollout replay, the student model suffers from severe mode collapse, leading to degraded visual quality (as shown in Quality dimension). While incorporating full-rollout replay partially mitigates training divergence, it remains visibly inferior to Stable Forcing across the metrics. Although distilling with the full-context teacher obtains competitive performance, scaling it to longer horizons suffers from prohibitive computational overhead, hindering efficient training. In contrast, our method achieves efficient distillation while maintaining generation quality, highlighting the effectiveness of Stable Forcing.



Figure 10: **Qualitative visualizations on versatile controls.** WorldPlay2 supports a broad range of open-ended interactions, such as fine-grained hand-object manipulations (dexterous grasping and box opening), dynamic environmental transitions and anomaly events (weather shifts and explosion), and complex full-body interactions (getting in vehicles and riding animals).



Figure 11: **Qualitative visualizations on long-horizon consistency.** WorldPlay2 generalizes robustly across different characters. Crucially, it can execute intricate interactions seamlessly alongside navigation controls (first case), while faithfully preserving spatiotemporal consistency over long horizons.

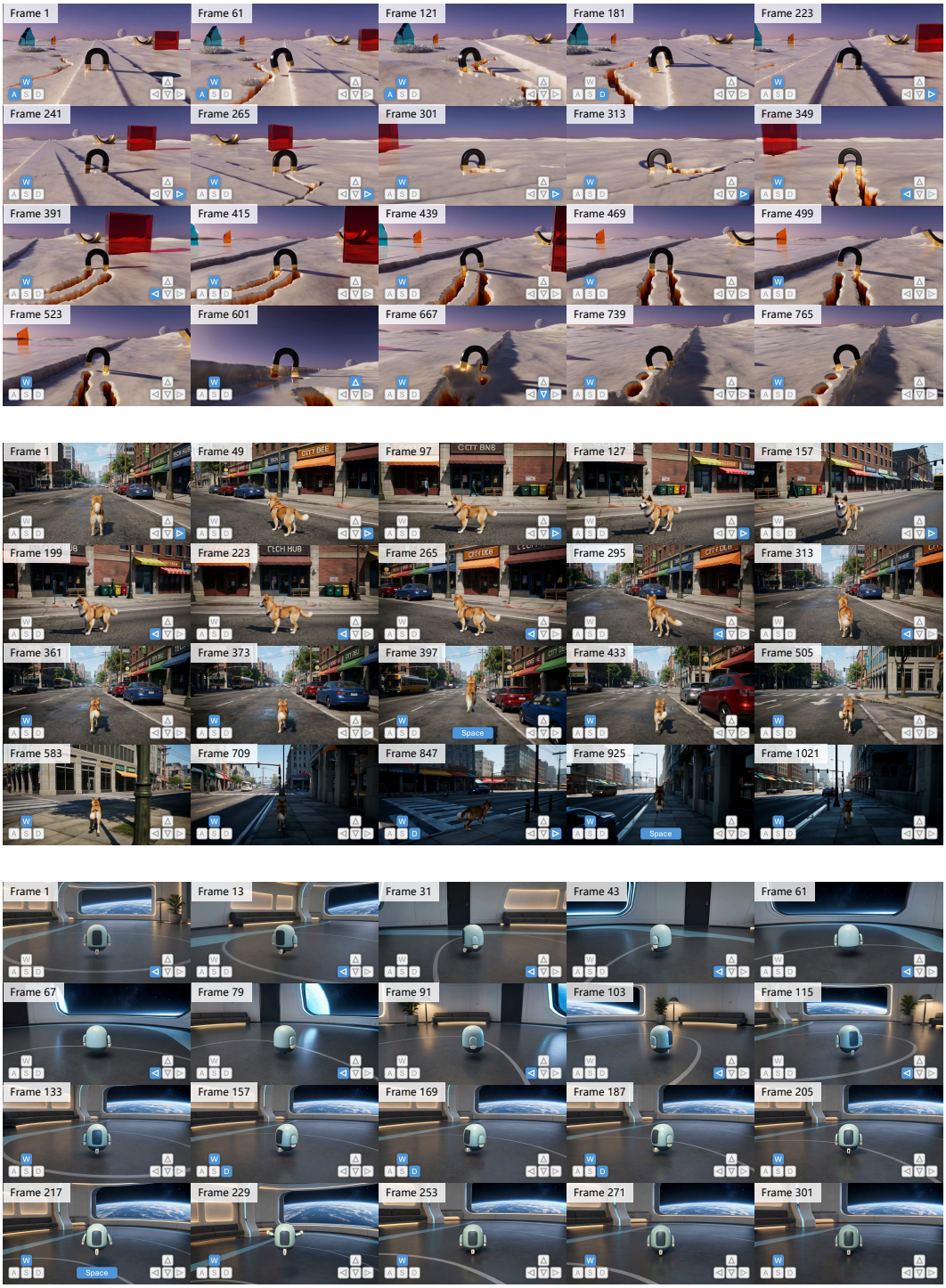

Figure 12: **Qualitative visualizations on long-horizon consistency. Top:** WorldPlay2 maintains geometric consistency under compound navigation controls. **Middle:** It generates structurally plausible scene layouts and preserves spatial coherence over long horizons. **Bottom:** It preserves geometric consistency under 360° camera rotations.