ALAYAWORLD: INTERACTIVE LONG-HORIZON AND PLAYABLE VIDEO WORLD GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Building playable virtual worlds traditionally requires substantial manual effort in asset creation and interaction design. Video world models offer a generative alternative, but sustained interaction requires models to respond to user controls and perform diverse actions while maintaining long-horizon scene consistency and generation stability. We present ALAYAWORLD, an open video world model for long-horizon interactive and playable world generation, supporting camera-controlled navigation and text-driven behaviors such as spell casting during continuous generation. ALAYAWORLD generates video autoregressively in chunks, coordinating persistent spatial memory, temporal history, and motion-aware conditions to support scene recall across viewpoints and continuous evolution across chunks. To support navigation and interactive behaviors across diverse scenes, we curate and annotate six real-world and gameplay datasets under a unified protocol and further construct GenEvent, a synthetic video dataset that supplements this mixture with supervision for instruction-driven behaviors such as spell casting. We combine training on imperfect histories with autoregressive distillation to improve long-horizon generation stability and enable four-step sampling per chunk. Experiments on iWorld-Bench, WorldMark, and WBench demonstrate the effectiveness of ALAYAWORLD in world consistency, generation stability, and interactive control. We provide an open research framework with reference implementations and reproducible pipelines for interactive and playable video worlds.



Figure 1: Interactive world simulation across diverse scenes. ALAYAWORLD synthesizes explorable worlds that span firstand third-person viewpoints, real-world, game, and synthetic domains, and both indoor and outdoor environments. Moreover, it accommodates open-ended actions such as spell-casting, weapon combat, and monster summoning.

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

With appropriate programming such a display could literally be the Wonderland into which Alice walked.

— Ivan E. Sutherland, The Ultimate Display (1965)

Building playable virtual worlds typically involves interconnected development stages, including asset creation, animation, rendering, and interaction design, requiring substantial time and manual effort. Video world models offer a generative approach to creating such environments: they continuously synthesize future observations from visual history and user controls, allowing users to explore scenes and direct actions and events within them. Sustained play imposes several requirements. Environments should retain recognizable structure and appearance as users move and return; changing instructions should elicit corresponding behaviors; and motion, scene evolution, and visual quality should remain coherent as interaction continues. Together, these requirements motivate interactive video world generation supporting long horizons, responsive control, and diverse events.

Recent advances in autoregressive video generation, interactive control, and memory support this goal. Action-conditioned and text-controlled world models support continuous exploration and changing user instructions (He et al., 2025; Mao et al., 2025; Tang et al., 2025). Spatial memory provides another foundation: historical observations can be retrieved using camera states (Xiao et al., 2025) or stored in evolving geometric representations and rendered from target viewpoints (Ren et al., 2025; Wu et al., 2025). These advances underpin continuous generation, user control, and cross-view scene recall. Building on them, we study how models preserve established scenes while responding to changing user instructions over extended generation.

Maintaining this continuity requires the model to draw on history at different temporal scales. When users return to a previously explored region, the model should generate observations consistent with the established scene, even when the relevant history falls outside its recent context. Meanwhile, new instructions may change ongoing actions and events, but these changes should connect naturally with the current scene and motion. Subsequent generation therefore depends on both long-term scene history and recent dynamics. As this history is continually updated from the model's own outputs, prediction errors can propagate through subsequent conditions and accumulate over time.

We present ALAYAWORLD, an open video world model for long-horizon interactive and playable world generation. Given an initial observation, ALAYAWORLD generates subsequent video autoregressively in chunks under user-specified camera trajectories and text instructions that can change at chunk boundaries. The model uses ViGeo (Yu et al., 2026b) to estimate geometry from newly generated observations and incorporates it into a persistent 3D point cache. Rendering this cache along target camera trajectories provides viewpoint-specific scene evidence for the next chunk. Temporal history and motion-aware conditions retain recent dynamics to support continuous actions and scene evolution. A unified causal encoding protocol coordinates these visual conditions and maintains consistent condition construction during training and inference. To prepare the model for repeatedly consuming its own outputs, we train with corrupted histories and prediction errors, followed by autoregressive distillation to enable four-step sampling per chunk.

To support navigation and interactive behaviors across diverse scenes, we curate and annotate six real-world and gameplay datasets under a unified protocol for training. We further construct GenEvent, a synthetic video dataset that complements the scene and motion information in these sources with explicit supervision for instruction-driven behaviors such as spell casting.

We evaluate ALAYAWORLD on iWorld-Bench (Fang et al., 2026), WorldMark (Xu et al., 2026b), and WBench (Ying et al., 2026), examining the complete system through world consistency, generation quality, control alignment, and revisitation memory. Our contributions are:

- We present ALAYAWORLD, an autoregressive video world model for long-horizon interactive and playable world generation that coordinates persistent spatial memory, recent dynamics, and user controls to support continuous exploration and instruction-driven events.
- We curate and annotate six real-world and gameplay datasets under a unified protocol and introduce GenEvent, a synthetic video dataset providing complementary supervision for diverse instruction-driven behaviors.
- We provide an open framework with reference implementations and reproducible pipelines for research on interactive and playable video worlds.

2 RELATED WORK

Recent work has advanced interactive video world generation, enabling users to explore generated environments and influence their evolution through control signals and language instructions (He et al., 2025; Mao et al., 2025; Tang et al., 2025; Parker-Holder & Fruchter, 2025; Sun et al., 2025; Robbyant Team et al., 2026; Qian et al., 2026). We review memory mechanisms and training strategies for long-horizon generation, followed by datasets for learning interactive behaviors.

2.1 LONG-HORIZON VIDEO WORLD GENERATION

One line of research preserves history through retained frames, learned compression, or recurrent propagation (Zhang et al., 2025; Zhu et al., 2025; Ding et al., 2026). These methods organize evidence by temporal order, balancing historical coverage against context cost and information loss. They support recent dynamics and forward continuity, but temporal retention alone does not ensure access to observations of a region revisited after a long detour. ALAYAWORLD combines compressed recent history with spatially indexed scene memory to serve these complementary needs.

A complementary direction organizes observations by camera state or scene geometry (Li et al., 2025a; Yu et al., 2026a; Qian et al., 2026). Pose-indexed retrieval selects relevant historical frames (Xiao et al., 2025; Chen et al., 2026), while explicit geometric memory accumulates scene evidence for target-view rendering (Ren et al., 2025; Wu et al., 2025). Building on this paradigm, ALAYAWORLD combines persistent scene recall with recent-motion conditioning and instruction-driven events in a chunk-wise autoregressive system.

Long-horizon training and few-step distillation address complementary aspects of sustained generation. Condition corruption, error recycling, and supervision on self-generated rollouts expose models to imperfect histories (Li et al., 2025b; Yuan et al., 2026; Huang et al., 2025b; Cui et al., 2025). Consistency learning and distribution matching provide principles for few-step generation (Song et al., 2023; Yin et al., 2024; Zheng et al., 2026). ALAYAWORLD adopts these principles for continuation from imperfect histories and four-step sampling, retaining spatial memory, recent-motion context, and user controls throughout autoregressive generation.

2.2 DATASETS FOR WORLD GENERATION

Scene-centric video collections provide appearance, scene-layout, and camera-motion coverage for world modeling (Zhou et al., 2018; Ling et al., 2024; Li et al., 2025c; Wang et al., 2025). Camera metadata may either be provided by the data source or be recovered through geometric estimation (Huang et al., 2025a). Game-oriented datasets complement these captures with interactive content and open-world gameplay (Che et al., 2024; Yu et al., 2025; Li et al., 2026). Our training mixture combines real-world and gameplay sources under a common annotation schema; GenEvent adds synthetic supervision for action-triggered events that are sparsely represented in these sources.

3 ALAYAWORLD

3.1 INTERACTIVE AUTOREGRESSIVE FORMULATION

Given an initial observation x_{init}, a user specifies a camera trajectory π_i and a text instruction y_i for chunk i. The model generates a latent video chunk z_i and decodes it to $\hat{x}_i$. Generation factorizes as

$$p_\theta(z_{1:N} \mid x_{\text{init}}, \pi_{1:N}, y_{1:N}) = \prod_{i=1}^{N} p_\theta(z_i \mid c_i, y_i), \qquad c_i = \mathcal{C}(x_{\text{init}}, \hat{x}_{<i}, \pi_{\leq i}), \tag{1}$$

where $\mathcal{C}$ constructs visual conditions from available history and the requested viewpoint. Future controls are unnecessary: a user can issue a new trajectory or instruction after each completed chunk. In this paper, *playable* refers to this combination of navigation and instruction-driven visual behaviors, without assuming a symbolic game state or an explicit game-rule engine.

The backbone is the video branch of LTX-2.3 (Lightricks, 2026). We remove its audio module and use the remaining video diffusion transformer. Each chunk contains $K = 4$ latent frames. With

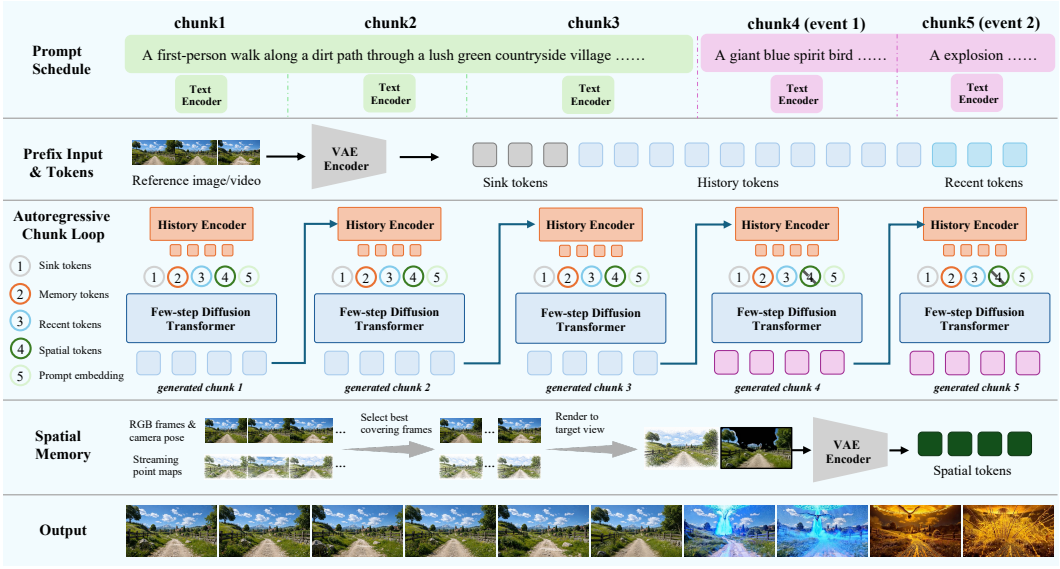

Figure 2: Autoregressive formulation of ALAYAWORLD. At each chunk, a persistent appearance anchor, compressed temporal history, spatial memory rendered at requested viewpoints, and recent-motion conditions jointly guide video generation with the text instruction. Completed chunks update history and spatial memory; text instructions can change between chunks to introduce events.

temporal stride $r = 8$, a continuation emits approximately 32 new RGB frames, excluding causal decoder prefixes. At 24 fps this represents approximately 1.33 seconds of video; the duration of generated content is distinct from the wall-clock time required to produce it.

The visual input to the transformer is

$$S_i^\tau = [\, s;\, h_i;\, g_i;\, n_i;\, z_i^\tau \,], \tag{2}$$

Figure 2 illustrates the generation loop. s is a persistent appearance anchor, h_i contains compressed temporal history, g_i contains target-view spatial evidence, n_i is a recent-motion condition, and z_i^τ is the noisy target. Conditions use zero diffusion timestep. The history embeddings enter directly; other visual latents pass through the patch embedder. Full attention operates within this sequence, and only the target segment is denoised. Autoregressive causality follows from using completed history when constructing conditions, while frames within a new chunk are generated jointly.

3.2 PERSISTENT SPATIAL MEMORY

Let $\mathcal{M}_{i-1}$ denote the point cache before chunk i. The cache stores estimated geometry with appearance from previous observations. We render it along the requested camera path, then register new observations after generation:

$$(R_i, V_i) = \mathcal{R}(\mathcal{M}_{i-1}, \pi_i), \qquad \mathcal{M}_i = \mathcal{U}\big(\mathcal{M}_{i-1}, \mathcal{G}(\hat{x}_i), \hat{x}_i\big), \tag{3}$$

where $\mathcal{G}$ is ViGeo (Yu et al., 2026b), $\mathcal{R}$ returns rendered RGB views and their validity, and $\mathcal{U}$ registers the estimated points into the cache. Once a region has been generated and registered, it can contribute to later views even after leaving the recent temporal window.

Rendering requires a common geometric convention. We scale-align the requested camera trajectory to the point cache using the median of pairwise displacement ratios over the shared prefix. A least-squares fit of a pinhole camera model to the point map estimates intrinsics once per clip; they remain fixed during rollout. These steps coordinate trajectory scale and projection across chunk boundaries.

To form g_i, an available RGB prefix frame is prepended to R_i. The causal VAE encodes the resulting sequence, and the prefix latent is discarded. Invalid rendered regions do not provide reliable scene evidence, so their tokens are removed from the visual condition sequence. The generator uses valid projections as historical evidence and its learned prior to synthesize newly exposed content. The

requested trajectory determines the rendered viewpoints and thus supplies camera guidance through g_i; a separate camera AdaLN branch is not used in the default model.

A fixed-length rendered sequence bounds the generator's context at fixed output resolution, even as exploration history grows. The external cache still requires storage and geometric processing, so bounded attention context alone does not imply constant total memory usage or end-to-end latency.

3.3 Temporal History and Motion-aware Conditions

Spatial memory recalls previously observed content at new viewpoints, while temporal conditions describe the dynamics leading into the next chunk. We use two temporal paths: h_i summarizes a longer recent window through learned compression, whereas n_i provides a direct VAE representation of recent dynamics at the chunk boundary. Although their source frames overlap, n_i bypasses the history encoder, complementing the compressed summary with a local continuation condition. Because the causal VAE's first latent has a different temporal receptive field from continuation latents, we define conditioning windows in RGB space before encoding.

For an anchor RGB frame at time t, the motion condition uses the second latent of a nine-frame window containing the anchor and its eight predecessors. Temporal history is encoded from a 25-frame window and compressed by a learned history encoder H_ϕ:

$$n_i = \mathcal{E}(x_{t-8:t})_2, \qquad h_i = H_\phi\big(\mathcal{E}(x_{t-24:t})\big), \tag{4}$$

where $\mathcal{E}$ is the causal VAE encoder. The $25 = 1 + (4-1) \times 8$ RGB frames produce exactly four history latents. The nine-frame window yields two latents, allowing n_i to carry recent temporal context. The anchor frame remains available as a decoding prefix.

Shared causal encoding protocol. The visual conditions follow a common rule: define RGB windows and causal prefixes before encoding, then retain the latents appropriate to their roles. Spatial renderings discard their prefix latent, n_i uses a continuation latent with preceding-frame context, and h_i compresses the encoded history window. The data loader, training pipeline, and autoregressive rollout use these same window and prefix conventions to coordinate temporal context at chunk boundaries. Hard memory dropout removes temporal-memory tokens and their sequence positions during training; unavailable temporal memory is likewise omitted at initialization. Chunk handoff retains the causal prefix required by direct latent continuation or RGB decode–re-encoding.

3.4 Text-driven Events during Continued Exploration

Text instructions specify scene semantics and requested events, while geometry-rendered conditions guide the viewpoint. A user can retain an instruction across chunks or replace y_i at a chunk boundary without resetting the visual history. Temporal history and the motion condition convey the state reached by the preceding chunk, allowing the next event to develop from the available scene.

For example, a rollout can begin with navigation, receive a spell-casting instruction, and continue under another movement command. GenEvent contributes event-centric videos to the training mixture (Section 4.2). At inference, the chunk-level text interface combines a new request with accumulated visual context. Interaction is expressed through generated visual consequences; these observations alone do not establish explicit damage, inventory, or game-rule state. Preserving changed object states through long occlusions remains a limitation of the observation-based representation.

4 Data and Training

4.1 Curated Data and Unified Annotations

We curate six real-world and gameplay datasets for training: Sekai (Li et al., 2025c), SpatialVID (Wang et al., 2025), RealEstate10K (Zhou et al., 2018), DL3DV (Ling et al., 2024), MUGEN (Tan et al., 2026), and GameVerse; the latter two are internally curated. A unified annotation schema combines global attributes with timestamped descriptions of scene content, dynamics, and camera behavior. Per-frame camera intrinsics and poses are included when available or recoverable. Together with GenEvent, these sources form a training mixture of 222,147 clips. Appendix A details the reported composition, filtering, and annotations.

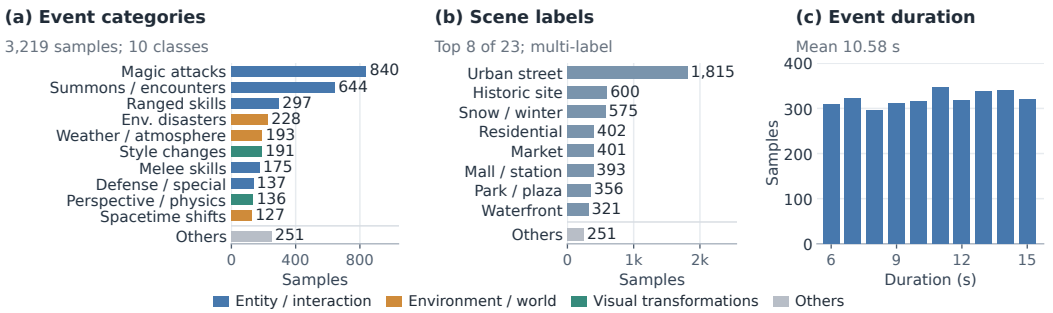

Figure 3: Distribution statistics of GenEvent. Event-category counts, scene/style coverage, and clip-duration distributions characterize the retained synthetic event videos.

4.2 GenEvent: Synthetic Data for Playable Behaviors

GenEvent complements the six curated datasets with synthetic examples of underrepresented instruction-driven behaviors, such as spell casting.

Dataset collection. We select first-person walking videos from Sekai-Real (Li et al., 2025c), converting them from 30 to 24 fps. Gemini 2.5 Flash (Google DeepMind, 2025) screens the 36-second frame for visual quality and checks viewpoint consistency across frames at 35, 36, and 37 seconds, using acceptance thresholds of 5.5 and 5.0, respectively. This helps preserve viewpoint continuity at the event transition. Caption-based scene classification supports balanced sampling across environments.

Gemini 3.1 Pro (Google DeepMind, 2026) generates scene-conditioned event prompts using the boundary frames and source captions. Preassigned event slots and cultural themes encourage diversity, while recent event labels reduce repetition. Grok (xAI, 2026) then generates the first event from the source frame at 36 seconds; a second event or aftermath starts from the preceding event's final frame. We concatenate the generated clips with the source prefix and manually review instruction alignment, interaction coherence, viewpoint preservation, and continuity. Appendix A.3 details prompting, synthesis, and filtering.

Dataset statistics. GenEvent contains 3,219 retained event examples at 1280×720 resolution and a nominal frame rate of 24 fps. The annotated event segments span 6–15 seconds, with a mean of 10.58 seconds and a total duration of 9.46 hours. These durations count only the first annotated event per sample, excluding the preceding source footage and any subsequent events. Figure 3 summarizes the event, scene, and duration distributions. The annotation labels span 10 categories organized into three groups: entity interactions, environmental and world changes, and visual transformations. Magic attacks (840 examples, 26.1%) and entity summons or encounters (644, 20.0%) are the largest categories, alongside ranged and melee skills, defensive actions, environmental disasters, and visual changes. Source-scene annotations cover 23 non-exclusive tags, including urban streets, historic sites, winter scenes, indoor malls or stations, and waterfronts. Event durations are relatively evenly distributed, with each one-second bin accounting for approximately 9–11% of the samples.

Training supervision. GenEvent contributes synthetic event videos and event descriptions to the shared training mixture. The construction pipeline produces detailed descriptions and short user instructions, either during prompt generation or through subsequent annotation. Appendix A.4 describes the annotation process and statistical conventions.

4.3 Training Stages

We train AlayaWorld in three stages using the curated datasets and GenEvent. First, we adapt the pretrained backbone to this mixture through full-parameter bidirectional training with image-, video-, and text-conditioned examples. Second, we pretrain the history encoder through masked-history reconstruction, then fine-tune the autoregressive generator with spatial memory, temporal

Table 1: Results on iWorld-Bench (Fang et al., 2026).

		Generation Quality				Trajectory Following		Memory Ability	
Model	**Avg.**	Image Quality	Brightness Consistency	Color Temp. Constraint	Sharpness Retention	Motion Smoothness	Trajectory Accuracy	Memory Symmetry	Trajectory Alignment
NVIDIA Cosmos	0.6285	0.6778	0.6952	0.7170	0.4363	0.9907	0.4955	0.3738	0.6419
HunyuanVideo-1.5	0.7089	**0.7128**	0.7027	0.7477	0.5545	0.9908	0.6844	0.6336	0.6449
WAN 2.2	0.5316	0.5545	0.3886	0.3411	0.3428	0.9557	0.6514	0.4480	0.5703
YUME 1.5	0.5797	0.6232	0.3810	0.4165	0.4023	0.9765	0.7113	0.5276	0.5988
Matrix-Game 2.0	0.5179	0.4851	0.2963	0.2937	0.4149	0.9848	0.7008	0.3311	0.6362
HY-World 1.5	0.7729	0.6675	0.8051	0.7819	0.6634	0.9921	0.7472	0.8481	0.6776
AlayaWorld	**0.8456**	0.6620	**0.9492**	**0.9379**	**0.8361**	**0.9924**	**0.7985**	**0.8871**	**0.7018**

Table 2: Results on WorldMark's First-Person Real split (Xu et al., 2026b). Trans. and rot. denote translation and rotation. DA: direction accuracy; DP: direction purity; MS: motion stability; RL: response latency; LM/GM/RM: local/global/revisit memory; PQ/AQ: perceptual/aesthetic quality.

		Action Dynamics								World Memory			Visual Quality	
Model	**Avg.**	DA Trans.	DA Rot.	DP Trans.	DP Rot.	MS Trans.	MS Rot.	RL Trans.	RL Rot.	LM	GM	RM	PQ	AQ
DreamX-World	72.20	96.91	76.60	78.18	70.38	65.41	27.71	96.10	97.00	93.29	34.04	79.89	71.16	51.98
HY-GameCraft 1.0	59.55	89.07	2.74	74.82	63.07	66.76	13.70	90.97	51.99	91.31	36.54	74.20	64.55	54.38
HY-World 1.5	77.00	92.55	85.05	81.18	80.31	31.66	86.39	85.44	85.13	94.17	58.88	87.52	76.15	56.59
LingBot-World	79.63	89.14	87.32	75.43	83.64	79.14	94.26	86.56	**99.20**	92.26	36.85	79.30	75.65	56.49
Matrix-Game 2.0	77.08	95.12	80.66	80.57	75.97	**79.89**	78.85	**98.23**	86.47	91.31	40.56	72.22	71.15	51.00
Matrix-Game 3.0	80.09	**98.37**	78.30	**84.83**	73.38	76.39	96.74	91.83	96.39	91.52	50.60	80.38	69.62	52.86
SANA-WM	75.19	88.74	83.08	65.51	78.55	46.79	83.39	76.83	97.41	91.70	53.45	82.05	76.76	53.16
Yume 1.5	67.62	51.85	59.47	75.16	52.53	55.35	79.45	40.66	77.11	**99.36**	55.25	80.96	**87.24**	**64.72**
AlayaWorld	**82.95**	78.67	**92.31**	72.55	**90.32**	68.05	**98.08**	75.48	96.76	92.53	**75.77**	**93.43**	85.94	58.50

history, and motion-aware conditions under camera and text control. Auxiliary next-chunk prediction, synthetic history corruption, and prediction-error replay prepare the model for continuous generation from imperfect context.

Finally, we distill the autoregressive teacher into a four-step student by combining distribution matching, supervision on student rollouts, and adjacent-noise consistency learning. The student retains the memory and control interface and the chunk-wise generation loop. Training objectives, parameter-freezing schedules, and implementation details are provided in Appendix B.

5 Experiments

We report results on three complementary benchmarks. iWorld-Bench (Fang et al., 2026) evaluates generation quality, trajectory following, and memory along return paths. WorldMark (Xu et al., 2026b) measures action dynamics, world memory, and visual quality under shared scene and action inputs. WBench (Ying et al., 2026) assesses eight aspects of world consistency on its 158-case navigation split. Evaluation details are provided in Appendix C.

5.1 Generation, Control, and Memory on iWorld-Bench

Table 1 shows that AlayaWorld achieves comparable image quality while leading on the other seven metrics. Its brightness consistency, color-temperature constraint, and sharpness retention indicate stable appearance over continued generation. AlayaWorld also obtains the highest trajectory accuracy (0.7985), memory symmetry (0.8871), and trajectory alignment (0.7018). The memory scores further indicate scene recall beyond local continuity. Together, these results support its ability to follow navigation controls while preserving visual content along return paths.

Table 3: WBench consistency scores (Ying et al., 2026). Bg.: background; Sp.: spatial; GSp.: gated spatial; Seg.: segment; Pers.: perspective; Subj.: subject; Geo.: geometric; Photo.: photometric.

Model	Avg.	Bg.	Sp.	GSp.	Seg.	Pers.	Subj.	Geo.	Photo.
Yume 1.5	80.1	90.3	71.5	71.4	99.4	48.0	88.8	88.0	83.3
Matrix-Game 3.0	74.5	85.7	81.0	80.4	89.8	13.3	83.0	87.6	75.3
HY-World 1.5	86.9	92.7	**90.6**	**84.9**	**100.0**	62.5	89.1	92.0	83.1
HY-GameCraft	72.6	86.5	60.5	60.5	99.4	17.9	82.6	88.3	**85.0**
LingBot-v2	86.5	92.5	82.3	78.7	98.1	84.5	88.9	87.1	79.8
Genie 3	82.6	90.7	79.9	78.4	93.6	54.5	90.4	88.6	84.5
Happy Oyster	84.3	91.4	77.7	75.8	96.2	75.0	91.5	87.2	79.8
DreamX-World	74.9	88.7	74.7	74.4	99.4	31.7	81.9	72.2	75.8
SANA-WM	80.7	90.5	77.1	76.4	97.5	49.0	85.4	88.7	81.2
JoyAI-Echo-1.5	88.3	93.6	86.8	79.8	98.7	83.1	90.0	92.5	82.0
XGEN-Jing	83.7	91.0	78.6	77.8	98.1	65.7	90.9	90.8	76.6
AlayaWorld	**89.5**	**94.1**	87.9	81.9	98.1	**86.6**	**93.4**	**94.1**	80.3

5.2 Action Dynamics and World Memory on WorldMark

On WorldMark's First-Person Real split (Table 2), AlayaWorld achieves the highest global and revisit memory scores (75.77 and 93.43), alongside the best rotational direction accuracy, direction purity, and motion stability. It ranks second in perceptual and aesthetic quality. These results highlight scene persistence and stable rotational control, while lower translational accuracy and response scores identify room for improvement in translation commands.

5.3 World Consistency on WBench

Table 3 reports a consistency aggregate of 89.5 for AlayaWorld, the highest among the compared systems. AlayaWorld leads in background, perspective, subject, and geometric consistency, and ranks second in spatial and gated spatial consistency. These results support AlayaWorld's ability to preserve scene appearance and structure across viewpoint changes.

5.4 Ablation Study

Table 4 compares variants without sink tokens s, spatial memory g_i, or temporal memory h_i, and one replacing ViGeo with Depth Anything 3 (DA3) (Lin et al., 2025). Removing spatial memory produces the largest drops in Interaction and Consistency, by 17.2 and 16.9 points, respectively. This variant controls the camera through text prompts instead of trajectory-conditioned renderings, losing explicit geometric guidance and persistent scene recall. Removing sink tokens and temporal memory reduces Consistency by 2.6 and 3.0 points, respectively, and lowers Video Quality. DA3 slightly improves Setting and Physical, but ViGeo performs better overall.

Table 4: Ablations on the WBench navigation split. VQ: Video Quality; Set.: Setting; Int.: Interaction; Cons.: Consistency; Phys.: Physical.

Variant	VQ	Set.	Int.	Cons.	Phys.
w/o sink token	77.6	69.1	78.4	86.9	62.4
w/o spatial memory	78.1	69.4	62.8	72.6	**63.5**
w/o temporal memory	77.3	69.1	76.5	86.5	61.4
DA3 instead of ViGeo	73.9	**70.0**	72.2	88.0	**63.5**
AlayaWorld (full)	**79.3**	69.7	**80.0**	**89.5**	63.1

5.5 Qualitative Results

Figure 4 illustrates how AlayaWorld combines navigation, scene recall, and interactive events during continuous generation. In (a), viewpoint changes reveal different parts of the environment while preserving recognizable subjects and scene appearance. In (b), returning views recover the kitchen layout and parked car, supporting scene recall beyond adjacent frames. In (c), the model trained with GenEvent produces fire, ice, and summoning effects across realistic and stylized scenes, integrating diverse events into the surrounding environment. The 60-second rollouts in (d) retain characteristic textures and scene styles as exploration progresses through natural, residential, and stylized environments, illustrating sustained visual coherence across multiple autoregressive chunks. Additional qualitative results are provided in Appendix D.



Figure 4: Qualitative results of ALAYAWORLD. (a) Camera-controlled generation along user-specified trajectories. (b) Leave-and-return consistency, comparing observations before departure and after revisiting the same region. (c) Text-driven spell casting after training with GenEvent. (d) Long-horizon autoregressive generation. Frames within each panel are ordered temporally.

6 CONCLUSION

We presented ALAYAWORLD, an open video world model for long-horizon interactive and playable world generation. It combines persistent spatial memory, temporal history, and motion-aware conditions within a chunk-wise autoregressive framework to support camera-controlled exploration and changing text instructions. We curate and annotate six real-world and gameplay datasets under a unified protocol and introduce GenEvent to provide complementary supervision for instruction-driven behaviors such as spell casting. Training on imperfect histories and autoregressive distillation support sustained generation with four-step sampling per chunk. Experiments on iWorld-Bench, WorldMark, and WBench demonstrate the model's effectiveness in interactive control, generation stability, and world consistency. The open framework provides a foundation for further research on interactive and playable video worlds.

Reproducibility statement

We describe the model interface, visual-condition construction, training stages, and data mixture in Sections 3 and 4. We provide implementation details, source counts, annotation procedures, and benchmark evaluation protocols in the appendices. Our framework provides reference implementations and reproducible pipelines for interactive and playable video world generation.

AI use statement

We used generative AI tools to assist manuscript organization, language drafting, and LaTeX preparation. Our training-data pipeline uses generative models for synthetic event videos and vision-language models for caption annotation. We take responsibility for the technical descriptions, data provenance, and empirical claims in this paper.

References

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. GameGen-X: Interactive open-world game video generation. *arXiv preprint arXiv:2411.00769*, 2024. URL https://arxiv.org/abs/2411.00769.

Zhifei Chen, Luozhou Wang, Guibao Shen, Dongyu Yan, Shuai Yang, Tianshuo Xu, Yihua Du, Wei Wang, Tianyi Gui, Lianghua Huang, and Yingcong Chen. ReWorld: An interactive world model with long-horizon memory. *arXiv preprint arXiv:2608.23565*, 2026. URL https://arxiv.org/abs/2608.23565.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-Forcing++: Towards Minute-Scale High-Quality Video Generation. *arXiv preprint arXiv:2510.02283*, 2025. URL https://arxiv.org/abs/2510.02283.

Yanbo Ding, Zhizhi Guo, Quanyue Song, Yishan He, Zhixiang He, Yongxiang Li, and Yali Wang. Ripple: Real-time streaming audio-video generation with cross-modal recurrent memory. *arXiv preprint arXiv:2607.26818*, 2026. URL https://arxiv.org/abs/2607.26818.

Jianjie Fang, Yingshan Lei, Qin Wan, Ziyou Wang, Yuchao Huang, Yongyan Xu, Baining Zhao, Weichen Zhang, Chen Gao, Xinlei Chen, and Yong Li. iWorld-Bench: A benchmark for interactive world models with a unified action generation framework. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.03941.

Google. Gemma Models Overview. Google AI for Developers documentation, 2026. URL https://ai.google.dev/gemma/docs. Updated April 2, 2026; accessed September 26, 2026.

Google DeepMind. Gemini 2.5 Flash Model Card. Official model card, 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf. Accessed September 26, 2026.

Google DeepMind. Gemini 3.1 Pro Model Card. Official model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Published February 19, 2026; accessed September 26, 2026.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025. URL https://arxiv.org/abs/2508.13009.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. ViPE: Video Pose Engine for 3D Geometric Perception. *arXiv preprint arXiv:2508.10934*, 2025a. URL https://arxiv.org/abs/2508.10934.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025b. URL https://arxiv.org/abs/2506.08009.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. VMem: Consistent Interactive Video Scene Generation with Surfel-Indexed View Memory. *arXiv preprint arXiv:2506.18903*, 2025a. URL https://arxiv.org/abs/2506.18903.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable Video Infinity: Infinite-Length Video Generation with Error Recycling. *arXiv preprint arXiv:2510.09212*, 2025b. URL https://arxiv.org/abs/2510.09212.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, Zizhen Li, Fanrui Zhang, Jiaxin Ai, Zhixiang Wang, Yuwei Wu, Tong He, Jiangmiao Pang, Yu Qiao, Yunde Jia, and Kaipeng Zhang. Sekai: A Video Dataset towards World Exploration. *arXiv preprint arXiv:2506.15675*, 2025c. URL https://arxiv.org/abs/2506.15675.

Zhen Li, Zian Meng, Shuwei Shi, Wenshuo Peng, Yuwei Wu, Bo Zheng, Chuanhao Li, and Kaipeng Zhang. WildWorld: A large-scale dataset for dynamic world modeling with actions and explicit state toward generative ARPG. *arXiv preprint arXiv:2603.23497*, 2026. URL https://arxiv.org/abs/2603.23497.

Lightricks. LTX-2.3 Model Card. Hugging Face model repository, 2026. URL https://huggingface.co/Lightricks/LTX-2.3.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the Visual Space from Any Views. *arXiv preprint arXiv:2511.10647*, 2025. URL https://arxiv.org/abs/2511.10647.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ling_DL3DV-10K_A_Large-Scale_Scene_Dataset_for_Deep_Learning-based_3D_Vision_CVPR_2024_paper.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A Text-Controlled Interactive World Generation Model. *arXiv preprint arXiv:2512.22096*, 2025. URL https://arxiv.org/abs/2512.22096.

Moonshot AI. Kimi-K2.6 Model Card. Official Hugging Face model repository, 2026. URL https://huggingface.co/moonshotai/Kimi-K2.6. Accessed September 26, 2026.

Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. Google DeepMind Blog, August 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, Mengyin An, Haobo Zhang, Biao Jiang, Jiahua Wang, Haofeng Sun, Yang Liu, and Yangguang Li. Matrix-Game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026. URL https://arxiv.org/abs/2608.29910.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.03751.

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

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models. *arXiv preprint arXiv:2303.01469*, 2023. URL https://arxiv.org/abs/2303.01469.

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

Ivan E. Sutherland. The ultimate display. In *Proceedings of the IFIP Congress*, pp. 506–508, 1965. URL https://worrydream.com/refs/Sutherland_1965_-_The_Ultimate_Display.pdf.

Jiaming Tan, Mingliang Zhai, Zhen Li, Yuwei Wu, Chuanhao Li, and Kaipeng Zhang. AlayaVista: Streaming world modeling from panoramic states to perspective video, 2026. URL https://arxiv.org/abs/2609.14462.

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, and Qinglin Lu. Hunyuan-GameCraft-2: Instruction-following interactive game world model. *arXiv preprint arXiv:2511.23429*, 2025. URL https://arxiv.org/abs/2511.23429.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. SpatialVID: A Large-Scale Video Dataset with Spatial Annotations. *arXiv preprint arXiv:2509.09676*, 2025. URL https://arxiv.org/abs/2509.09676.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025. URL https://arxiv.org/abs/2506.05284.

xAI. Grok Imagine Video. Official model documentation, 2026. URL https://docs.x.ai/developers/models/grok-imagine-video. Accessed September 26, 2026.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term Consistent World Simulation with Memory. *arXiv preprint arXiv:2504.12369*, 2025. URL https://arxiv.org/abs/2504.12369.

Gangwei Xu, Qihang Zhang, Jiaming Zhou, Xing Zhu, Yujun Shen, Xin Yang, and Yinghao Xu. Next Forcing: Causal World Modeling with Multi-Chunk Prediction. *arXiv preprint arXiv:2606.11187*, 2026a. URL https://arxiv.org/abs/2606.11187.

Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Kaipeng Zhang, and Yongtao Ge. WorldMark: A Unified Benchmark Suite for Interactive Video World Models. *arXiv preprint arXiv:2604.21686*, 2026b. URL https://arxiv.org/abs/2604.21686.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved Distribution Matching Distillation for Fast Image Synthesis. *arXiv preprint arXiv:2405.14867*, 2024. URL https://arxiv.org/abs/2405.14867.

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. URL https://arxiv.org/abs/2605.25874.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. GameFactory: Creating new games with generative interactive videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2501.08325.

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Sri Siddarth Chakaravarthy P, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, and Animesh Garg. MosaicMem: Hybrid Spatial Memory for Controllable Video World Models. *arXiv preprint arXiv:2603.17117*, 2026a. URL https://arxiv.org/abs/2603.17117.

Zhu Yu, Jingnan Gao, Runmin Zhang, Lingteng Qiu, Zhengyi Zhao, Rui Peng, Yichao Yan, Kejie Qiu, Siyu Zhu, Zilong Dong, Si-Yuan Cao, and Hui-Liang Shen. Towards Consistent Video Geometry Estimation. *arXiv preprint arXiv:2605.30060*, 2026b. URL https://arxiv.org/abs/2605.30060.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real Real-Time Long Video Generation Model. *arXiv preprint arXiv:2603.04379*, 2026. URL https://arxiv.org/abs/2603.04379.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025. URL https://arxiv.org/abs/2504.12626.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rCM: A Unified Teacher-Forcing and Self-Forcing Open Recipe for Autoregressive Diffusion Distillation in Streaming Video Generation and Interactive World Models. *arXiv preprint arXiv:2606.25473*, 2026. URL https://arxiv.org/abs/2606.25473.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo Magnification: Learning View Synthesis using Multiplane Images. *arXiv preprint arXiv:1805.09817*, 2018. URL https://arxiv.org/abs/1805.09817.

Tianrui Zhu, Shiyi Zhang, Zhirui Sun, Jingqi Tian, and Yansong Tang. Memorize-and-Generate: Towards long-term consistency in real-time video generation. *arXiv preprint arXiv:2512.18741*, 2025. URL https://arxiv.org/abs/2512.18741.

A TRAINING DATA CURATION, ANNOTATION, AND GENEVENT

Our training resources combine six real-world and gameplay datasets curated and annotated under a unified protocol with GenEvent, which adds targeted supervision for instruction-driven behaviors. This appendix details source preparation and annotation, followed by GenEvent construction and its integration into training.

A.1 CORPUS COMPOSITION AND PREPROCESSING

We summarize the training sources in Table 5. We report clip counts for the six real-world and gameplay sources and 3,219 retained event examples for GenEvent. We state the overall training volume of 222,147 clips separately from the retained-event statistics. Figure 5 shows temporally ordered examples from all seven sources.

Table 5: Dataset-level counts for the training sources. GenEvent reports retained event examples; the other sources report clip counts. The overall training volume is stated separately in the text.

Source	Content	Count
Sekai walking	Real, first-person navigation	21,561
SpatialVID	Real, indoor camera motion	23,210
RealEstate10K	Real, property walkthroughs	17,429
DL3DV	Real, multi-view scenes	7,905
MUGEN	Internally curated real videos	21,436
GameVerse	Gameplay recordings	124,116
GenEvent	Retained synthetic event examples	3,219

Video preprocessing. We use a shared filtering pipeline that ingests a manifest, computes cached features, and retains clips that pass every active gate. Technical validation checks decoding, a source resolution of at least 720p, a duration of at least three seconds, frame rates between 24 and 65 fps, and supported codecs. Common gates screen exposure and borders, shot changes, motion, text or interface overlays, and foreground-person occupancy; estimated trajectories also undergo pose-stability checks. Thresholds are calibrated per source rather than replacing the shared pipeline with separate source-specific procedures. The GameVerse clips are approximately 66 seconds long.

A.2 UNIFIED ANNOTATION PROTOCOL

We annotate all sources with a hierarchical schema that separates clip-level context from segment-level dynamics. A vision-language model consumes frames sampled at 1–2 fps with explicit timestamps. We use Kimi-K2.6 (Moonshot AI, 2026) as the default captioning backend, with Gemini and Gemma (Google, 2026) as alternatives.

Global and segment descriptions. Global attributes describe weather, time of day, location type, camera perspective, camera motion, and video style. Timestamped segments use `time_range_s` and separate `subject_motion`, `environment_motion`, `static_scene`, and `camera_description`. These tracks are fused into a `full_prompt` of 5–9 present-tense sentences and a `short_prompt` of 15–45 words for caption dropout and augmentation.

Camera information. The annotation includes a `camera_path` field with 16 discrete categories, distinct from subject motion. This semantic camera label is separate from continuous geometric metadata. For real-world sources without camera metadata, ViPE (Huang et al., 2025a) recovers per-frame intrinsics and poses. These offline data annotations are distinct from ViGeo geometry estimation in the generation loop. GenEvent uses synthetic stationary camera conditions during autoregressive training and distillation.

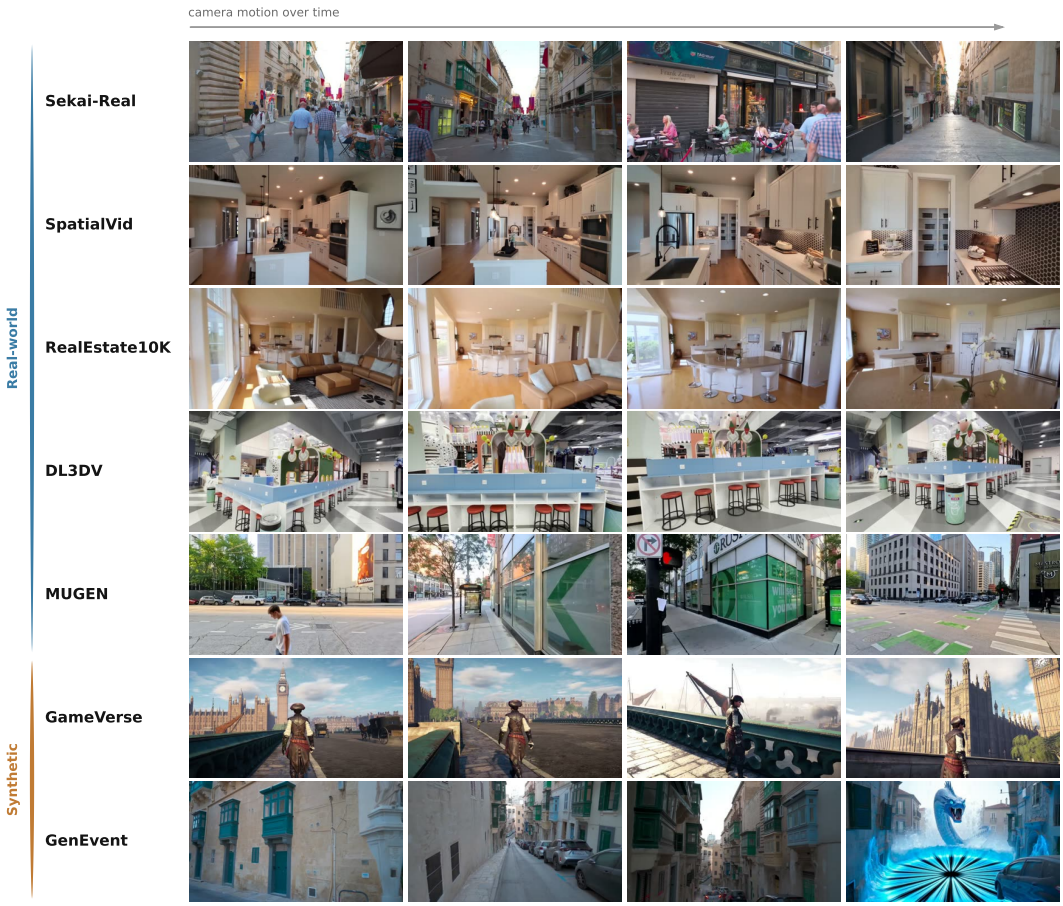

Figure 5: Examples from the seven training sources: Sekai-Real, SpatialVID, RealEstate10K, DL3DV, MUGEN, GameVerse, and GenEvent. Each row shows four temporally ordered frames from one video, illustrating scene appearance, camera motion, and event dynamics. The GenEvent row is a training example containing a synthetic event, distinct from the model outputs in Figure 4.

A.3 GenEvent Construction and Filtering

GenEvent combines real first-person walking footage with synthetic events. Construction comprises source screening, scene-balanced selection, event prompting, image-to-video synthesis, and manual review.

Source preparation and filtering. We resample Sekai-Real walking videos from 30 to 24 fps and extract frames at 35, 36, and 37 seconds. Gemini 2.5 Flash scores the 36-second frame for sharpness, exposure, scene clarity, visual-effects suitability, and crowd sparsity, with respective weights 0.4, 0.2, 0.2, 0.1, and 0.1. We retain candidates with a weighted score of at least 5.5. A separate comparison of the three frames requires a viewpoint-consistency score of at least 5.0 on a 0–10 scale, screening out substantial turns, pans, and shot changes near the transition.

Scene classification and sampling. Gemini 2.5 Flash analyzes source captions to identify scene categories and support scene-balanced selection. The retained dataset covers 23 non-exclusive scene tags, including urban streets, residential areas, historic sites, indoor environments, forests, waterfronts, and winter scenes. Selection prioritizes underrepresented environments.

Event prompt generation. Gemini 3.1 Pro receives the three boundary frames, source captions, and scene information. Each source is assigned an event type and a cultural theme before prompting. Events span 10 categories covering weapon skills, magic, summons, creature encounters, environ-

mental changes, and visual transformations; themes cycle through more than 80 entries. Labels from the 60 most recently generated samples form an avoidance list to reduce repetition. Structured prompting determines the scene and event identity, prepares user descriptions, considers scene interactions, and produces the generation prompt. The prompt requests a stable first-person viewpoint. Prompts for subsequent events use the preceding event's prompt and final frame as context.

Sequential video synthesis. We use `grok-imagine-video` in image-to-video mode, initializing the first event from the source frame at 36 seconds. Samples follow three structures: one event; two successive, causally related actions; or an event followed by its continuing aftermath. For the latter two, the preceding event's final frame initializes the next generated clip. The generated clips have variable durations and are concatenated with the 36-second source prefix using `ffmpeg`.

Human review. Reviewers assess the generated effects against their prompts and reject salient unintended object appearances, instruction–video mismatches, incoherent reactions or impacts, severe unintended changes to scene layout, and discontinuities at the source/event or event/event boundaries. They also screen for implausible motion reversals, unexplained speed changes, loss of first-person perspective, and subtitle or UI overlays. Prompted transformations, summons, and time-altering effects are evaluated in context rather than rejected simply for departing from ordinary physical behavior.

A.4 Event Supervision and Statistical Conventions

Text annotations. Event descriptions are produced either during structured prompt generation or through subsequent annotation. The prompt-generation schema includes literary descriptions, casual descriptions, and short user phrases. In the subsequent annotation path, Gemini 3.1 Pro Preview takes the generation prompt and a frame captured two seconds after event onset to produce a short user instruction. This instruction focuses on the skill and its visual form without specifying what it strikes or destroys. These descriptions provide event-focused text alongside the general scene and motion annotations.

Statistical scope. The statistics in Section 4.2 and Figure 3 describe 3,219 retained event examples. Duration counts the first annotated event per sample, excluding source footage and later events; the mean is 10.58 seconds and total is 9.46 hours. The annotation summary lists 10 event categories in three groups and 23 non-exclusive scene tags. Scene tags can overlap and do not form a mutually exclusive partition.

B Implementation and Training Details

B.1 Latent Timing and Initialization

Under the causal encoding convention used here, a temporal stride of $r = 8$ maps $1 + (L - 1)r$ RGB frames to L latent frames. The 25-frame history therefore yields four latents, while the nine-frame motion window yields two; the second motion latent supplies the recent-motion condition. A continuation chunk contains four new latents, corresponding to approximately 32 emitted RGB frames. Causal prefixes provide encoding context and are excluded from the emitted continuation. The first observation initializes the appearance anchor and available spatial evidence. The first rollout step omits temporal-memory tokens, matching memory-free cases during training.

B.2 Stage 1: Domain Adaptation

We adapt the pretrained backbone through full-parameter bidirectional training on clips of up to 20 seconds, at 24 fps and nominal 540p/720p resolutions. Image-, video-, and text-conditioned examples preserve multiple initialization modes. Real-world, gameplay, and event videos provide complementary scene, motion, and behavior coverage. Text selection depends on source-segment duration: segments longer than 20 seconds use a whole-clip summary, while segments of at most 20 seconds randomly alternate between the summary and a detailed segment caption. Textual camera descriptions are randomly dropped. Our schedule uses a clip-length-dependent sigma shift and a final low-noise refinement pass. Source proportions are retuned for autoregressive training.

B.3 Stage 2: Memory-Conditioned Autoregressive Training

Flow objective. For a clean target latent z_i, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, and noise level $\tau \in [0, 1]$, we construct the noisy target and train the velocity predictor (Liu et al., 2023; Lipman et al., 2023) as

$$\begin{aligned} z_i^\tau &= (1-\tau)z_i + \tau\epsilon, \\ \mathcal{L}_{\text{flow}} &= \mathbb{E}\Big[\|v_\theta(z_i^\tau, \tau \mid c_i, y_i) - (\epsilon - z_i)\|_2^2\Big], \end{aligned} \tag{5}$$

where c_i contains the visual conditions and y_i denotes the text instruction.

History-encoder pretraining. Let w be a history sequence and Ω a reconstruction target window. Every history frame is independently perturbed as $\tilde{w}_j = (1-\sigma_j)w_j + \sigma_j\epsilon_j$, with $\sigma_j \sim \mathcal{U}(0.2, 1)$. The history encoder compresses this corrupted sequence to condition reconstruction of the target window w_Ω using Eq. 5. The target-window length varies during training. The backbone is frozen while the history encoder and LoRA (Hu et al., 2022) adapters are trained; no camera-control signal is used in this pretraining phase.

Autoregressive fine-tuning. We subsequently unfreeze the backbone and jointly train it with the history encoder using spatial, temporal, and motion-aware conditions. Camera control is conveyed through spatial renderings along the requested trajectory, while text specifies the desired scene content and behaviors. Text conditions use the shared annotation interface. Condition construction follows the same causal encoding protocol during training and inference.

Auxiliary next-chunk prediction. Following Next Forcing (Xu et al., 2026a), intermediate backbone features are combined into F_i, and an auxiliary head f_ψ predicts the velocity of a noised next chunk. For independently sampled noise $\epsilon_{i+1} \sim \mathcal{N}(0, I)$, we use

$$\begin{aligned} \tilde{\tau} &= \frac{10\tau}{1+9\tau}, \qquad z_{i+1}^{\tilde{\tau}} = (1-\tilde{\tau})z_{i+1} + \tilde{\tau}\epsilon_{i+1}, \\ \mathcal{L}_{\text{NF}} &= \mathbb{E}\Big[\big\|f_\psi(F_i, z_{i+1}^{\tilde{\tau}}, \tilde{\tau}) - (\epsilon_{i+1} - z_{i+1})\big\|_2^2\Big]. \end{aligned} \tag{6}$$

The autoregressive fine-tuning objective is $\mathcal{L}_{\text{flow}} + 0.5\mathcal{L}_{\text{NF}}$. Future video supplies auxiliary training supervision only and is unavailable as an observed condition at inference.

Synthetic history corruption. To expose the model to imperfect histories, we adopt synthetic drift simulation following Helios (Yuan et al., 2026). Perturbations affect temporal, spatial, and recent-motion conditions while preserving the appearance anchor. They include Gaussian-noise mixing, downsampling and upsampling to emulate blur, and changes to deviations around the latent mean to emulate saturation variation.

Prediction-error replay. Following Stable Video Infinity (Li et al., 2025b), a one-step clean-target estimate defines the residual

$$\begin{aligned} \hat{z}_i^0 &= z_i^\tau - \tau v_\theta(z_i^\tau, \tau \mid c_i, y_i), \\ \delta_i &= \hat{z}_i^0 - z_i. \end{aligned} \tag{7}$$

An error bank stores residuals by noise level and chunk length for later replay. Residual perturbations are applied to selected historical conditions and target latents in the training implementation. Synthetic corruption supplies drift-like examples before the bank warms up, and error replay subsequently takes priority. The two mechanisms are mutually exclusive within a training step, and the persistent appearance anchor remains clean.

B.4 Stage 3: Four-Step Autoregressive Distillation

Student rollouts and distribution matching. The autoregressive teacher uses approximately 30 denoising steps per chunk. We train a four-step student using distribution matching (Yin et al., 2024) and supervision on student rollouts (Cui et al., 2025). Historical student predictions are detached during rollout. A frozen teacher estimates the target distribution, while a trainable critic estimates the student distribution. Teacher and critic share a backbone through critic-LoRA switching, with critic updates performed more frequently than student updates. Spatial memory, temporal conditions, user controls, and the chunk-wise memory-update loop retain the same interface.

Adjacent-noise consistency. We further regularize the student using adjacent noise levels on a 50-level grid, following consistency-learning principles (Song et al., 2023; Zheng et al., 2026). Let G_θ denote the student's clean-latent prediction and θ^- its exponential-moving-average target. For paired states on the distillation trajectory,

$$\mathcal{L}_{\mathrm{CM}} = \mathbb{E}\Big[d_{\mathrm{Huber}}\big(G_\theta(z_i^\tau, \tau \mid c_i, y_i), \\ \mathrm{sg}\Big[G_{\theta^-}(z_i^{\tau'}, \tau' \mid c_i, y_i)\Big]\big)\Big], \qquad \tau > \tau' \geq 0, \tag{8}$$

where sg stops gradients. The student objective is

$$\mathcal{L}_{\mathrm{distill}} = \mathcal{L}_{\mathrm{DMD}} + 0.5\mathcal{L}_{\mathrm{CM}}. \tag{9}$$

Student updates optimize LoRA parameters while the base backbone and memory modules remain frozen; critic parameters are updated separately. The 50-level training grid is distinct from the four-step inference schedule.

B.5 Training Configuration

We train at 24 fps with nominal 540p/720p resolutions. Each autoregressive chunk contains 4 latents. Our teacher uses approximately 30 denoising steps per chunk, and the distilled student uses four.

C Benchmark Evaluation Protocols

C.1 Navigation Benchmarks

iWorld-Bench. iWorld-Bench (Fang et al., 2026) evaluates image quality, brightness consistency, color-temperature constraint, sharpness retention, motion smoothness, trajectory accuracy, memory symmetry, and trajectory alignment. Scores lie in $[0, 1]$, with higher values better. Brightness, color-temperature, and sharpness metrics assess temporal appearance stability; motion smoothness uses frame-interpolation consistency, and trajectory accuracy evaluates motion directions. Memory symmetry and trajectory alignment assess return-path correspondence. These proxies do not directly evaluate hidden object state or physical interaction outcomes.

WorldMark. We report the First-Person Real split of WorldMark (Xu et al., 2026b), with 125 videos per model. Action dynamics comprise direction accuracy, direction purity, motion stability, and response latency, each evaluated separately for translation and rotation. World memory comprises local, global, and revisit memory, while visual quality comprises perceptual and aesthetic quality. Scores lie in $[0, 100]$, with higher values better. The comparison is restricted to this split.

WBench. We evaluate the complete Consistency group on the 158-case navigation split of WBench (Ying et al., 2026). Its eight metrics cover background, spatial, gated spatial, segment, perspective, subject, geometric, and photometric consistency; scores lie in $[0, 100]$, with higher values better. Return-view spatial similarity assesses scene recall, while the gated counterpart penalizes insufficient motion. Segment continuity measures the absence of detected hard cuts. Perspective consistency tracks subject-position stability and is not a direct camera-pose error. Geometric and photometric measures rely on estimated depth and poses for reprojection. We report the benchmark's Consistency aggregate.

C.2 Comparison Settings

We use ViGeo for spatial memory in the default model and compare AlayaWorld with complete systems that differ in their backbones and training resources. These comparisons assess overall performance. We examine the contributions of individual components through the ablations in Section 5.4, including a variant that replaces ViGeo with DA3.

D Additional Qualitative Results

We provide additional examples of visual diversity, camera control, scene revisitation, interactive prompt switching, spell casting, and long-horizon generation.

Diverse visual styles. Figure 6 shows navigation in six visual styles, including photorealistic, ink-wash, neon-lit, block-based, painterly, and illustrated appearances. The canal, bridge, and surrounding buildings remain recognizable as the viewpoint advances within each sequence, while the characteristic style is retained. These examples illustrate camera-controlled exploration across distinct visual domains.



Figure 6: Navigation in six visual styles. Each three-frame sequence progresses from left to right; highlighted keys indicate the displayed commands.

Camera-controlled generation. Figure 7 presents camera-controlled generation in nighttime streets, a fantasy waterfront, a lakeside landscape, a first-person game corridor, an illustrated train, and a historic courtyard. Translations and rotations reveal new portions of the environment while retaining characteristic scene content and visual style, supporting exploration across realistic and stylized settings.

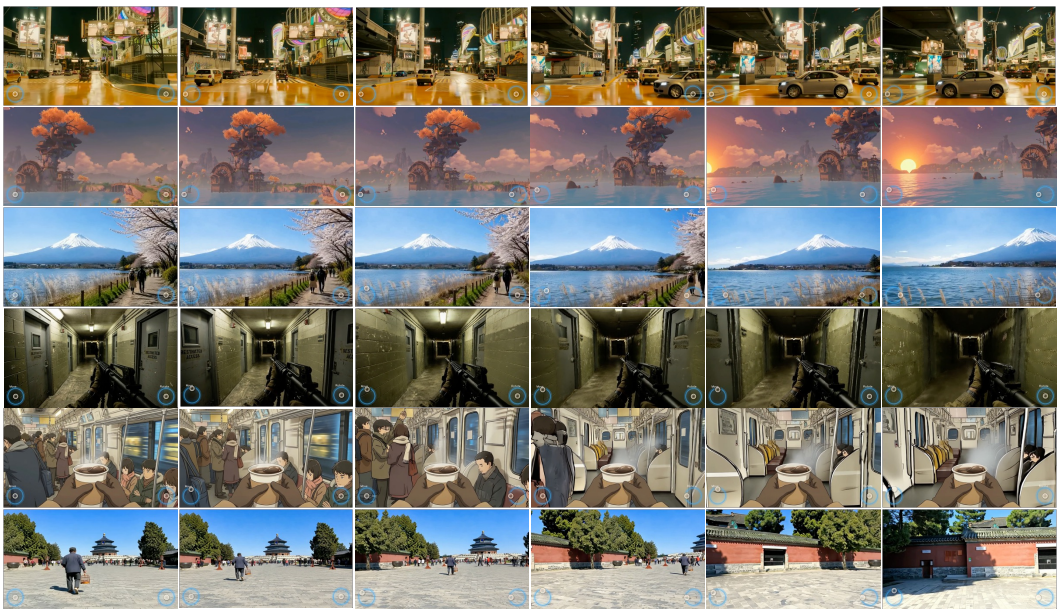

Figure 7: Camera-controlled generation across diverse environments. Each row shows six time-ordered observations, with movement and rotation controls overlaid.

Leave-and-return consistency. Figure 8 compares leave-and-return sequences in a bedroom and an art gallery. Our returning views recover the bed, desk, bookshelf, and prominent red-and-blue painting, retaining recognizable layout and appearance after intervening turns. In these examples, the compared methods exhibit changed furnishings or artworks, drift to different room layouts, or visible reconstruction artifacts.

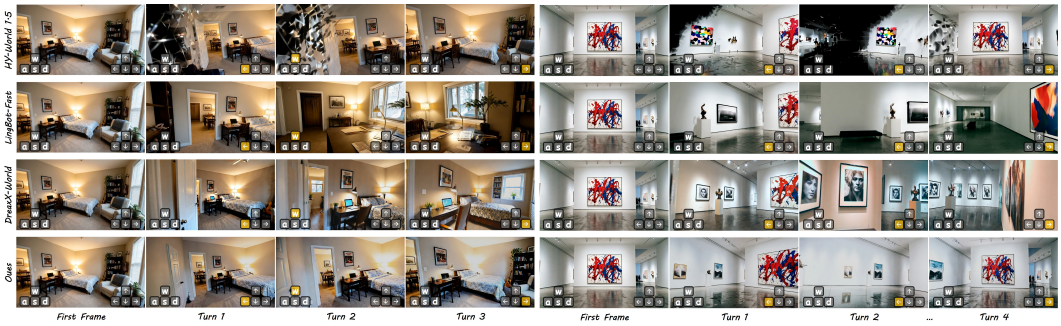

Figure 8: Leave-and-return consistency in two indoor scenes. Rows show HY-World 1.5, LingBot-Fast, DreamX-World, and ALAYAWORLD; columns within each scene follow the displayed turns.

Interactive navigation and events. Figure 9 demonstrates changes between navigation and event instructions within ongoing rollouts. Prompt switches produce a fiery apparition, water effects, a glowing owl, and explosions in the current environment. The owl sequence resumes walking before another event, while the last sequence returns to a stationary view. These transitions illustrate alternating exploration and user-directed events.



Figure 9: Interactive navigation and event control. Black and red arrows mark switches to ordinary and event prompts, respectively; highlighted keys show navigation inputs.

Spell casting from a shared scene. Figure 10 shows different spell-driven continuations of the same desert scene. Fire, water, ice, summoned entities, and terrain and lighting changes illustrate the range of event responses from a common starting view.

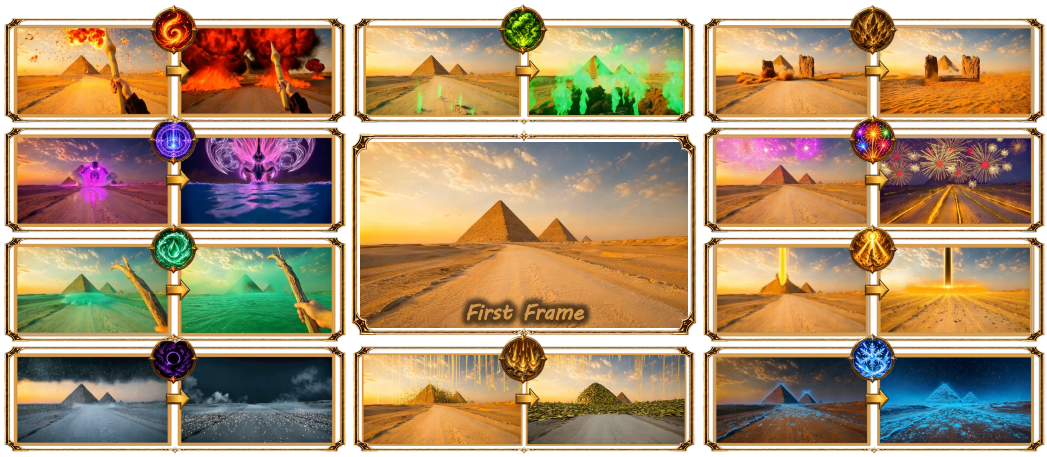

Figure 10: Different spell-driven continuations from a shared initial scene. The central image is the starting observation; surrounding frame pairs show the generated events.

Events across scenes and styles. Figure 11 shows events in autumn paths, snowy shrines, open landscapes, painted villages, fantasy forests, and coastal settings. Creature encounters, fire, ice, and luminous effects are integrated into both realistic and stylized scenes.



Figure 11: Spell casting and events across scenes and styles. Each three-frame sequence shows an event developing from its initial scene.

Long-horizon autoregressive generation. Figure 12 samples four rollouts at 0, 20, 40, and 60 seconds. Snowy streets, a rocky coast, illuminated buildings, and a green canal retain characteristic appearance as exploration continues across multiple autoregressive chunks.

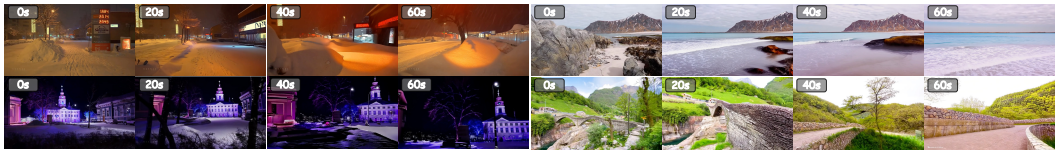

Figure 12: Long-horizon autoregressive generation. Four sequences are sampled at 0, 20, 40, and 60 seconds throughout continued exploration.