SCALING INTERACTIVE OMNIMODAL WORLD MODELS FROM HETEROGENEOUS SUPERVISION

Anonymous authors
Paper under double-blind review

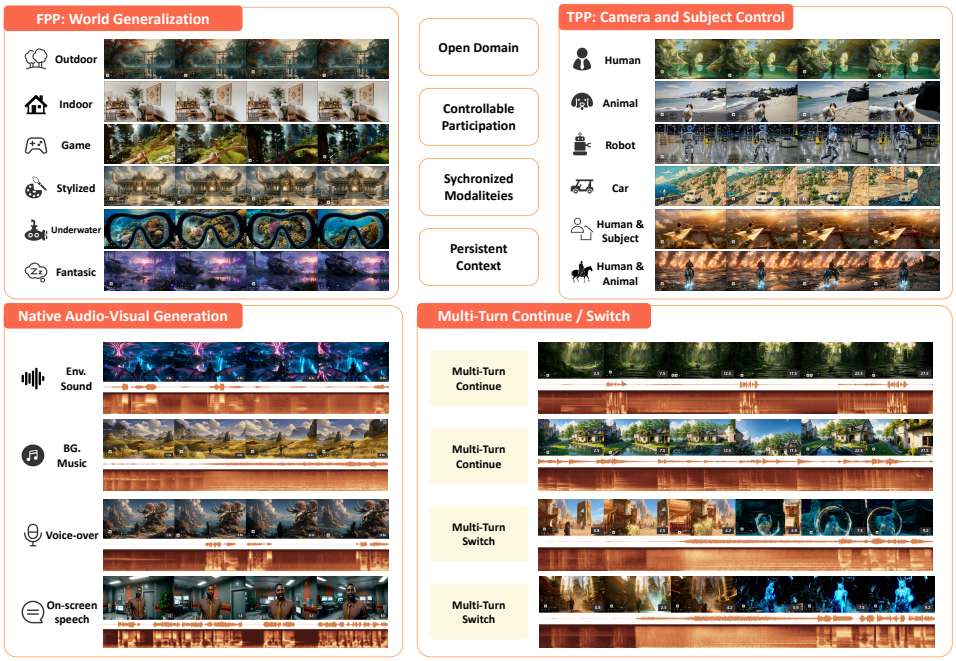

Figure 1: **Our model turns audio-visual generation into an enterable world.** Given a reference observation, a structured world description, and a user control sequence represented as a relative 6-DoF trajectory, one model supports first- and third-person interaction, synchronized environmental sound and speech, and multi-turn continuation.

ABSTRACT

Interactive omnimodal world models should respond to continuous control while generating the synchronized video, environmental sound, speech, and music that make a world feel alive. Existing approaches typically use source-specific action spaces or focus on visual prediction, making it difficult to combine heterogeneous interaction and audiovisual supervision. We present ECHOWM, an interactive omnimodal world model that canonicalizes interaction as a dataset-calibrated relative 6-DoF camera trajectory. The same camera-intent interface supports first-person observer motion and third-person camera–subject–world evolution, with viewpoint semantics learned from data. To reconcile complementary supervision, we progressively train an audiovisual prior, trajectory conditioning, and their joint realization on AV-rich, control-clean, and balanced data mixtures. We then convert the model for causal few-step deployment through audiovisual teacher forcing, self-generated distribution matching, and long-horizon rollouts, exposing it to the errors that accumulate during interactive generation. On WBench Navigation and SANA-WM-Bench, ECHOWM achieves leading interactive and visual-quality performance, while its four-step causal variant largely preserves these capabilities over long-horizon rollout.

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

Generative video models are rapidly evolving into expressive audiovisual media engines (Polyak et al., 2024; Google DeepMind, 2025b; HaCohen et al., 2026), jointly synthesizing visual content, environmental sound, speech, and music. Yet these models remain largely prompt-driven and passive, while interactive world models primarily focus on visual prediction under user or agent actions (Yang et al., 2023; Bruce et al., 2024; Hu et al., 2023; Wang et al., 2024a; He et al., 2024; Ren et al., 2025; Zhu et al., 2026a). A general interactive audiovisual world model should support both: continuous control over how the world evolves and native generation of the sounds, speech, and music that accompany it. Reaching this goal, however, requires learning from heterogeneous sources whose supervision is complementary but not directly compatible.

The first challenge is *interaction heterogeneity*. Raw keyboard and mouse actions are device- and environment-specific, while camera trajectories provide a more transferable geometric representation. Yet geometry alone remains ambiguous across viewpoints: the same trajectory may represent ego-motion in first person, but coordinated subject motion and camera following in third person. A second challenge is *audiovisual heterogeneity*. Different sources provide uneven combinations of video, speech, music, environmental sound, and reliable interaction supervision, and the sources with the cleanest controls are often not those with the richest audiovisual content. Naively mixing them therefore entangles incompatible interaction semantics and uneven supervision.

We address both challenges with a common principle: *canonicalize what should be shared while preserving what should remain distinct*. Heterogeneous controls are mapped into a shared relative metric 6-DoF trajectory space, with global scale calibration preserving displacement magnitude and viewpoint context determining how the same geometry is realized. For audiovisual learning, we instead exploit complementary supervision: AV-rich data establish visual and acoustic priors, control-clean data ground trajectory-conditioned interaction, and balanced refinement reconciles the two. This enables a shared interaction interface across first- and third-person scenes without explicit actor trajectories or hand-designed camera rigs.

Deployment introduces a further distribution shift: interactive generation must condition on its own previously generated video and audio rather than clean training histories. We therefore progressively convert the bidirectional audiovisual generator into a causal few-step model using audiovisual teacher forcing followed by self-generated rollout training and distribution matching. Extending this training to longer synchronized rollouts exposes the model to the accumulated errors encountered during sustained interaction. Based on these principles, we develop ECHOWM, an *interactive omnimodal world model* that combines continuous camera interaction with native generation of video, environmental sound, speech, and music. Our experiments evaluate the resulting interaction canonicalization, viewpoint disambiguation, metric calibration, factorized audiovisual learning, and long-horizon causal rollout. Our main contributions are:

- **Unified interaction across viewpoints.** We canonicalize heterogeneous controls into a shared relative metric 6-DoF trajectory space while preserving viewpoint semantics, enabling a single interface for both first-person ego-motion and third-person camera–subject–world evolution.
- **Scalable learning from heterogeneous supervision.** We organize audiovisual-rich and control-clean data according to their most reliable supervision and progressively learn audiovisual generation, trajectory-conditioned interaction, and their joint realization within a single model.
- **Long-horizon interactive audiovisual generation.** We adapt the model to causal deployment through trajectory-conditioned audiovisual teacher forcing and self-generated rollouts with few-step distribution matching, improving robustness under sustained interactive generation.

2 RELATED WORK

2.1 INTERACTIVE GENERATIVE WORLD MODELS

World models have evolved from latent dynamics for planning (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020) toward action-conditioned visual simulation and open-ended interactive generation (Yang et al., 2023; Bruce et al., 2024; Li et al., 2025a; Google DeepMind, 2025a; Sun et al., 2025; Wang et al., 2026b; Robbyant Team et al., 2026; Zhu et al., 2026a; DreamX Team et al., 2026).

Recent systems further extend real-time interaction, domain coverage, and long-horizon persistence through scalable training and explicit or implicit history mechanisms (Gao et al., 2026; Jiang et al., 2026; AlayaWorld Team et al., 2026; Chen et al., 2026; Huang et al., 2026; Riemann Dynamics, 2026; Yin et al., 2026). Most remain primarily visual simulators and represent interaction through environment-specific actions or camera controls. We instead study a general interactive audiovisual world model with a shared camera-centric interface across first- and third-person scenes and native generation of synchronized video and audio.

2.2 Action and Camera Conditioning

Interactive generation commonly represents control through either discrete actions or continuous camera geometry. Action-conditioned models provide intuitive keyboard, controller, or embodied interfaces, but their semantics are inherently tied to the source environment (Jiang et al., 2026; Li et al., 2025a; Jiang et al., 2026). Camera-controllable generation instead conditions on poses, trajectories, or viewing rays (Wang et al., 2024b; He et al., 2024; 2025; Xu et al., 2024; Ren et al., 2025; Robbyant Team et al., 2026). Recent positional encodings incorporate relative or metric camera geometry directly into attention (Li et al., 2025b; Zhang et al., 2025a; Qiao et al., 2026). Our formulation is complementary to these encoding mechanisms: we canonicalize heterogeneous controls and recovered poses into a shared metric relative 6-DoF trajectory while explicitly retaining viewpoint semantics, so that the same geometry can represent first-person observer motion or third-person camera–subject–world evolution.

2.3 Audio-Visual and Long-Horizon Generation

Recent work has progressed from video-conditioned sound synthesis toward native joint audio-visual generation, multisensory world prediction, and causal audiovisual streaming (Luo et al., 2023; Zhang et al., 2026; Cheng et al., 2025; Polyak et al., 2024; Google DeepMind, 2025b; Liu et al., 2024; HaCohen et al., 2026; Wang et al., 2025; Su et al., 2026; NVIDIA, 2026). These models establish strong joint audio-video generation, but largely target media generation or embodied prediction rather than continuous open-domain geometric interaction across viewpoints. Long-horizon generation has meanwhile been explored through causal diffusion, rollout-aware training, bounded context, and persistent memory (Liu et al., 2025; Zhu et al., 2026b; Riemann Dynamics, 2026; Zhu et al., 2026b; Zhao et al., 2026; Zhang et al., 2025b; Yang et al., 2025; Xiao et al., 2025; Yu et al., 2025; Wu et al., 2026; Chen et al., 2026; Riemann Dynamics, 2026). Our work builds on causal post-training but studies trajectory-conditioned joint audiovisual world generation: the model is adapted to its own synchronized audio-video histories under continuous interaction and further trained over extended self-generated rollouts.

3 Method

3.1 Overview and Problem Formulation

Given media context $\mathcal{M}$, structured text $\mathcal{C}$, and a user control sequence $U_{1:T}$, we generate synchronized video and audio conditioned on the requested camera motion. The media context is either empty, a reference observation I_{ref}, or a synchronized prefix $(V_{1:\tau}, A_{1:\tau})$. We convert the control sequence and training poses to a common calibrated relative trajectory $P_{1:T}$ and model

$$p_{\theta,\phi}(V_{1:T}, A_{1:T} \mid \mathcal{M}, \mathcal{C}, P_{1:T}), \tag{1}$$

where θ is the pretrained audio-visual backbone and ϕ is the trajectory-conditioning branch. Static text describes the initial scene and viewpoint; $P_{1:T}$ specifies subsequent camera evolution. Our formulation centers on a shared geometric interaction space: heterogeneous controls and training poses are canonicalized into calibrated relative 6-DoF trajectories, which are injected through relative camera geometry and interpreted according to viewpoint semantics. We describe these components below and defer data construction details to section A.1.

3.2 Metric Camera-Intent Conditioning

Unified Camera-Intent Interface. We encode interaction as a camera trajectory, providing one geometric interface for instrumented controls, simulation, and poses recovered from Internet video.

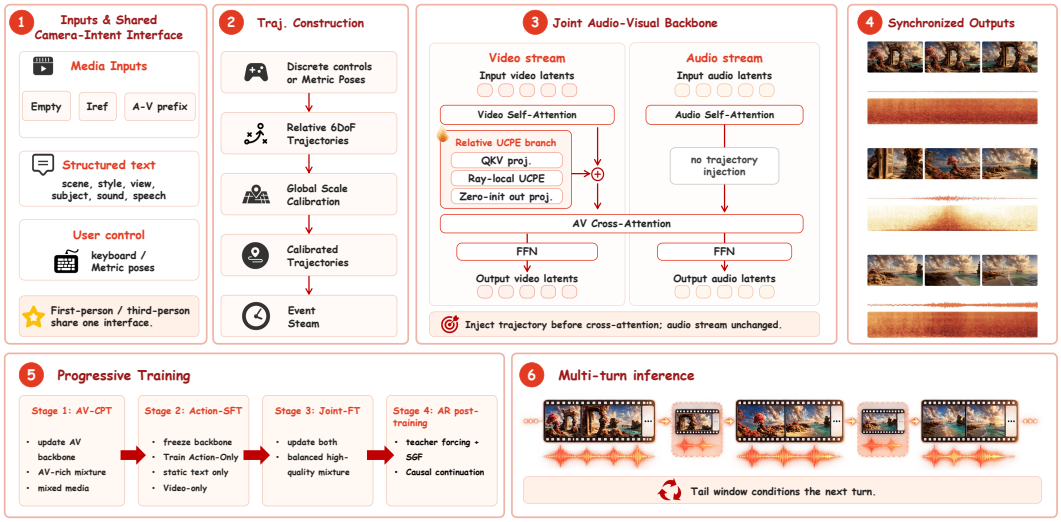

Figure 2: **Overview of our method.**User controls are converted into globally calibrated relative camera trajectories and injected into the video stream through a relative UCPE branch. Progressive training yields synchronized audio-visual generation and supports multi-turn inference through tail-window conditioning.

In first-person scenes, the trajectory specifies observer motion; in third-person scenes, it specifies camera motion, while character locomotion and camera–subject coupling are learned from data. Let $T_t \in \mathrm{SE}(3)$ be the camera-to-world pose at frame t. We express it relative to the initial camera pose as $\Delta T_t = T_0^{-1} T_t$, with $\Delta T_0 = I$. This removes dependence on the global coordinate frame while retaining continuous 6-DoF motion. At inference, keyboard or joystick controls are mapped to local increments $\boldsymbol{\xi}_t \in \mathfrak{se}(3)$ and composed as $\Delta T_t = \Delta T_{t-1} \operatorname{Exp}(\boldsymbol{\xi}_t)$. Training poses bypass this control mapping and use the same initial-relative representation.

Global Translation Scale Calibration. Relative poses may still have different translation scales across sources. Normalizing each clip independently would erase cross-clip motion magnitude, so we use one robust scale estimated from metric training trajectories. For clip i of length L_i, let $\Delta \mathbf{t}_{i,k}$ denote the translation component of its initial-relative pose, and let $m_i = \max_{1 \le k \le L_i} \|\Delta \mathbf{t}_{i,k}\|_2$ be the clip's maximum displacement. We set the global scale to the empirical 90th percentile:

$$s_{\text{global}} = Q_{0.9}(\{m_i \mid i \in \mathcal{D}_{\text{train}}\}), \tag{2}$$

For each frame, we set $\widehat{\mathbf{t}}_{i,k} = \Delta \mathbf{t}_{i,k} / s_{\text{global}}$ and replace the translation component of $\Delta T_{i,k}$ with $\widehat{\mathbf{t}}_{i,k}$ to obtain the calibrated pose $\widehat{\Delta T}_{i,k}$. The resulting sequence is $P_{i,1:L_i} = \big(\widehat{\Delta T}_{i,k}\big)_{k=1}^{L_i}$. The same s_{global} is used during training and inference, preserving relative displacement magnitudes across clips and continuation boundaries.

Trajectory Conditioning. We condition video tokens on $P_{i,1:L_i}$ and the corresponding camera intrinsics using Unified Camera Positional Encoding (UCPE) (Zhang et al., 2025a). UCPE encodes camera motion through relative ray geometry, avoiding dependence on absolute ray coordinates. We implement it as a zero-initialized residual camera branch shared by first- and third-person data. Audio receives no direct trajectory input; any coupling between camera motion and generated audio is mediated by the backbone's learned audio–visual interactions.

3.3 Factorized Audio-Visual and Interaction Learning

Reliable audio-visual and trajectory supervision are concentrated in different subsets of our heterogeneous data. We therefore train EchoWM progressively: AV-CPT establishes the audio-visual prior, Control-SFT learns trajectory conditioning while preserving this prior, and Joint-FT finally reconciles both capabilities on high-quality paired data.

Training Data Mixtures. We construct synchronized clips from gameplay, simulation, and Internet video, using an audio-visual path to select reliable appearance, speech, and sound

and a geometry path to recover and filter metric camera trajectories. Persistent text fields $(\text{scene}, \text{style}, \text{viewpoint}, \text{subject})$ are separated from temporal fields $(\text{narrative}, \text{speech}, \text{sound})$; motion-bearing narrative is removed whenever the trajectory is the control condition. Detailed source filtering and clip construction are provided in section A.1. The resulting corpus provides three complementary mixtures: AV-rich data for AV-CPT, control-clean data with reliable trajectories for Control-SFT, and a smaller high-quality intersection for Joint-FT. Action logs describe issued controls, whereas realized camera trajectories provide the motion supervision.

AV-CPT. We first adapt the audio-visual backbone θ on the AV-rich mixture using joint video and audio objectives with the full structured annotation. Training alternates among empty context, a reference frame, and synchronized audio-visual prefixes, yielding the adapted initialization θ_{AV}.

Control-SFT. We then freeze θ_{AV} and train only the zero-initialized trajectory branch ϕ on control-clean data. This stage uses the video objective and static scene conditions, while removing motion-bearing narrative and audio descriptions, such that the calibrated trajectory $P_{1:T}$ uniquely specifies camera evolution.

Joint-FT. Finally, we jointly update (θ, ϕ) at a reduced learning rate on the balanced mixture. Audio supervision and semantic audio conditions are restored, while motion-bearing narrative remains excluded. This stage consolidates trajectory control and native audio-visual generation after each capability has been learned from its most reliable supervision.

3.4 Streaming Audio-Visual Post-Training

The bidirectional diffusion backbone must be converted into a causal, few-step generator for streaming inference. We first use teacher forcing to establish causal attention, then apply short-horizon SGF with DMD, and finally extend SGF to long self-rollouts. The trajectory branch remains active and the generator parameters (θ, ϕ) are updated unless stated otherwise.

3.4.1 Causal Audio-Visual Teacher Forcing

We first convert the bidirectional audio-visual generator into a causal chunk-wise model through teacher forcing. Aligned video and audio latents are partitioned into N macro-chunks, with the two modalities sharing the same temporal chunk boundaries and diffusion level within each chunk. During training, chunk i is predicted from clean preceding chunks and its own noisy state, while the clean current target and all future chunks are masked out. Formally, letting $\mathcal{B}_{i,\text{clean}}$ and $\mathcal{B}_{i,\text{noisy}}$ denote the clean and noised states of chunk i, respectively, the causal visibility is

$$\mathcal{S}(\mathcal{B}_{i,\text{clean}}) = \{\mathcal{B}_{j,\text{clean}} : j \le i\}, \mathcal{S}(\mathcal{B}_{i,\text{noisy}}) = \{\mathcal{B}_{j,\text{clean}} : j < i\} \cup \{\mathcal{B}_{i,\text{noisy}}\}. \tag{3}$$

The same causal structure is applied to video self-attention, audio self-attention, and both cross-modal directions. This preserves synchronized audio-visual context while allowing all target chunks to be trained in parallel without leaking the clean current target or future information. We optimize the standard flow-matching objective for both modalities, conditioned on structured text and the trajectory available up to each chunk. This stage initializes causal audio-visual generation from clean histories before exposing the model to its own predictions in the subsequent SGF stage.

3.4.2 Self-Gradient Forcing and Long-Horizon Streaming

Teacher forcing conditions on ground-truth histories, whereas autoregressive inference conditions on the model's own predictions. We adopt self-gradient forcing (SGF) (Zhuang et al., 2026) to reduce this train–test mismatch. SGF first performs a gradient-free rollout with the inference-time few-step sampler to obtain self-generated audio-visual histories, and then replays the same trajectory differentiably while treating the sampled rollout states as stop-gradient. This exposes the model to its own accumulated errors without backpropagating through the full sequence of denoising decisions.

We combine SGF with distribution matching distillation (DMD) (Yin et al., 2024). For a recovered latent $\bar{x}_i$, we add diffusion noise at level s and match the student distribution to the frozen multi-step teacher:

$$\mathcal{L}_{\text{DMD}} = \mathbb{E}_{i,s}\left[w(s)\left\langle \text{sg}\big(s_{\text{fake}} - s_{\text{real}}\big), \bar{x}_i \right\rangle\right]. \tag{4}$$

Here, the real score is produced by the frozen teacher, while the fake score is updated on detached student samples. To extend this training to long-horizon streaming, we roll out multiple consecutive

Table 1: Comparison on the WBench Navigation split (Ying et al., 2026). Results are sorted by Average.

Model	Avg.	Quality	Setting	Interaction	Consistency	Physical
EchoWM	**81.6**	81.5	79.4	86.6	89.8	70.6
EchoWM-Flash	81.0	81.1	88.3	**87.9**	77.5	70.1
HiDream-O1-World (Ying et al., 2026)	80.9	81.0	82.2	80.0	88.0	**73.3**
Alaya-EVOKE (Ying et al., 2026)	80.8	**82.8**	83.8	78.6	86.9	72.1
LingBot-World (fast v2) (Robbyant Team et al., 2026)	79.4	81.8	76.8	82.8	86.5	69.1
Kling 3.0 (Ying et al., 2026)	79.0	81.4	91.0	69.4	83.7	69.3
LingBot-World (Robbyant Team et al., 2026)	78.5	78.9	72.6	80.1	**89.9**	71.2
HY-World 1.5 (Sun et al., 2025)	78.1	78.1	72.2	86.8	86.9	66.3
LingBot-World (fast) (Robbyant Team et al., 2026)	77.4	79.4	77.9	79.2	84.9	65.7
HappyOyster (Ying et al., 2026)	76.8	77.3	74.2	84.9	84.3	63.5
Lyra 2.0 (4-step AR) (Shen et al., 2026)	76.4	77.1	73.2	85.6	79.3	66.7
Cosmos3-Super (NVIDIA, 2026)	76.0	76.4	**91.6**	58.0	85.0	69.2
SANA-WM (4-step AR) (Zhu et al., 2026a)	76.0	79.3	76.1	82.2	80.7	61.9
DreamX-World (5B AR) (DreamX Team et al., 2026)	75.0	77.5	80.8	78.6	74.9	63.3
ABot-World (Jiang et al., 2026)	74.7	76.8	71.4	84.0	79.5	61.7

segments under the same causal inference procedure. For each modality, we maintain a bounded sink-plus-FIFO context consisting of a small number of persistent early chunks together with the most recent chunks. The same cache policy is used during training and inference. We apply the DMD objective to every rollout segment, so later segments are optimized under histories that already contain accumulated visual and acoustic prediction errors. This directly trains the model for the distribution encountered during sustained autoregressive generation.

4 Experiments

We evaluate EchoWM on WBench Navigation (Ying et al., 2026) and SANA-WM-Bench (Zhu et al., 2026a), which measure interactive navigation, trajectory following, visual quality, and long-horizon consistency. We report the multi-step model and, where applicable, its four-step causal variant EchoWM-Flash. We also conduct pairwise human evaluation on a 200-case interactive world-model benchmark and evaluate native image-conditioned audio-visual generation on our benchmark. Full protocols, extended results, and qualitative comparisons are provided in the supplement.

4.1 Benchmark Results

WBench Navigation. On the official 158-case split, EchoWM achieves the highest Average score (81.7) among the compared methods, with 89.8 in Consistency and 87.2 in Interaction (table 1). EchoWM-Flash retains an Average of 81.0 and attains the highest Interaction score (87.9), showing that the four-step causal model preserves most of the multi-step model's benchmark performance. The full protocol and qualitative comparisons are in section B.1.

SANA-WM-Bench. On the 241-frame benchmark, EchoWM achieves the highest VBench Overall on both Simple and Hard splits (83.91 and 83.96), while also obtaining the lowest translation and camera-motion errors on Simple trajectories (table 3). At 961 frames, EchoWM maintains strong visual quality and revisit consistency, but accumulated rotation drift becomes more pronounced, especially on Hard trajectories. EchoWM-Flash further achieves the best VBench Overall among the evaluated causal models on both splits. Complete long-horizon results are provided in section B.2.

Table 2: Native audio-visual generation under matched reference images and semantic conditions.

Model	PQ ↑	AQ ↑	AIF ↑	AVAlign ↑	AVSync (s) ↓
HappyOyster	6.84	3.16	2.72	0.22	0.96
Cosmos 3	6.71	3.08	3.01	0.33	0.81
LTX-2.3	**7.08**	**3.31**	2.94	0.34	0.74
EchoWM	6.96	3.27	**3.19**	**0.36**	**0.68**

Native audio-visual generation. To complement visual interaction benchmarks, we construct a 120-case image-conditioned benchmark, where each case pairs a reference image with visual and audio conditions for a 10-second clip. We compare EchoWM, HappyOyster, Cosmos 3, and LTX-2.3 under matched semantic conditions using metrics adapted from StreamAV-Bench (Liu et al., 2026):

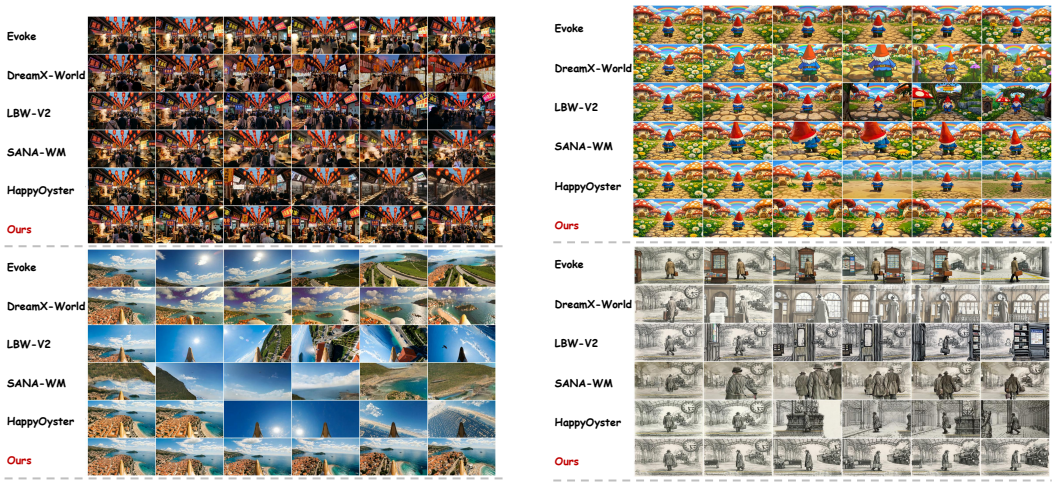

Figure 3: Qualitative comparison on Wbench Navigation subset.

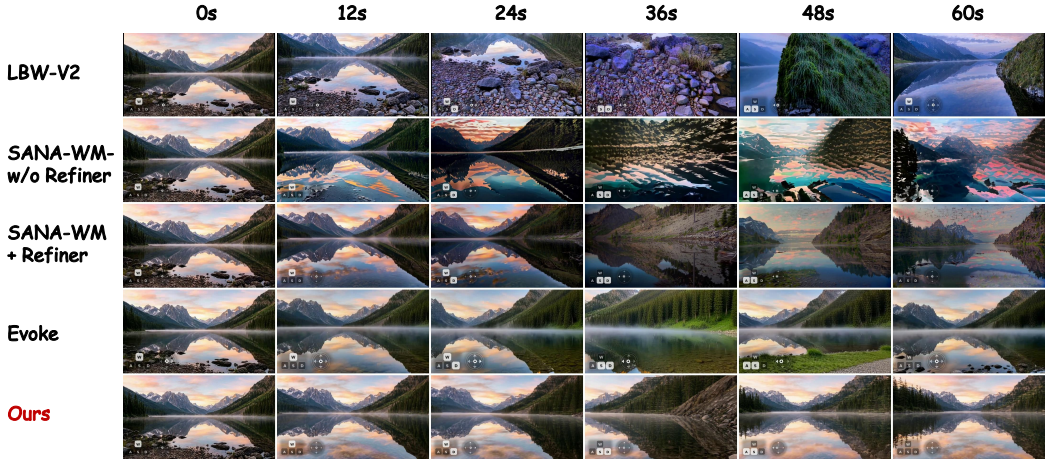

Figure 4: Qualitative comparison of causal streaming on SANA-WM-Bench. Under matched initial image and camera traj., our model shows less drift and better preserves trajectory adherence, scene identity, and spatial structure than the compared causal baselines.

Table 3: **SANA-WM-Bench comparison.** The left panel reports the 241-frame multi-step evaluation; the right panel reports the 961-frame causal evaluation.

Short horizon (241 frames)

Method	VBench ↑	R ↓	T ↓	CMC↓
Simple				
Matrix-Game 3 (Wang et al., 2026b)	79.41	4.474	0.406	0.453
LingBot-World-v2 (Robbyant Team et al., 2026)	81.13	3.092	0.282	0.310
HY-WorldPlay (Sun et al., 2025)	82.05	2.256	0.284	0.298
DreamX-World (DreamX Team et al., 2026)	82.42	1.839	0.261	0.273
SANA-WM (Zhu et al., 2026a)	81.75	**0.489**	0.194	0.195
EchoWM	**83.91**	0.523	**0.160**	**0.162**
Hard				
Matrix-Game 3 (Wang et al., 2026b)	79.37	12.819	0.583	0.717
LingBot-World-v2 (Robbyant Team et al., 2026)	81.67	6.367	0.357	0.426
HY-WorldPlay (Sun et al., 2025)	81.24	6.346	0.387	0.437
DreamX-World (DreamX Team et al., 2026)	82.03	8.804	0.467	0.556
SANA-WM (Zhu et al., 2026a)	81.79	**1.248**	**0.199**	**0.206**
EchoWM	**83.96**	1.697	0.256	0.266

Causal long horizon (961 frames)

	Trajectory and visual quality				Revisit consistency			Temporal
Method	VBench↑	R ↓	T ↓	CMC↓	PSNR↑	SSIM↑	LPIPS↓	ΔIQ↓
Simple								
Evoke (Yin et al., 2026)	80.11	9.859	2.159	2.190	13.50	0.2774	0.5010	-0.49
LingBot-World-v2 (Robbyant Team et al., 2026)	79.20	22.516	2.391	2.541	11.90	0.1715	0.5906	2.59
SANA-WM + Causal Refiner (Zhu et al., 2026a)	79.55	17.875	2.222	2.338	**14.41**	0.2660	0.5748	**-0.75**
SANA-WM w/o Refiner (Zhu et al., 2026a)	78.21	15.359	2.288	2.393	9.19	0.1661	0.6321	-0.12
EchoWM-Flash	**80.13**	**5.009**	**1.592**	**1.611**	**14.41**	**0.2998**	**0.4802**	2.09
Hard								
Evoke (Yin et al., 2026)	80.75	20.097	1.924	2.060	13.06	0.2592	0.5292	0.30
LingBot-World-v2 (Robbyant Team et al., 2026)	78.74	34.416	2.125	2.413	12.34	0.2183	0.5732	3.84
SANA-WM + Causal Refiner (Zhu et al., 2026a)	79.60	23.794	1.932	2.124	13.89	0.2605	0.5696	**-0.24**
SANA-WM w/o Refiner (Zhu et al., 2026a)	78.27	18.717	2.006	2.142	9.26	0.1729	0.6077	2.06
EchoWM-Flash	**81.06**	**14.221**	**1.735**	**1.829**	**14.04**	**0.2910**	**0.4893**	1.30

production quality (PQ), perceptual audio quality (AQ), audio instruction fulfillment (AIF), semantic audio-video alignment (AVAlign), and temporal synchronization error (AVSync). Together, these metrics separate audio fidelity and instruction following from semantic and temporal audio-video correspondence. Benchmark construction and evaluation details are provided in section B.3.

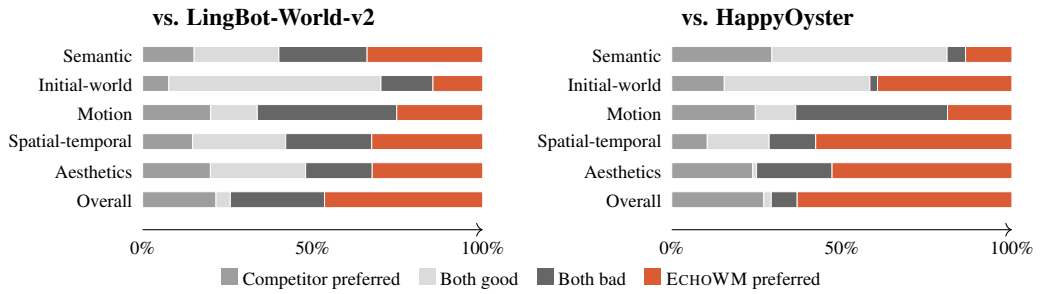

Figure 5: **Aggregate pairwise human evaluation.** Each bar reports the percentage of judgments for one dimension; segments distinguish competitor preference, equal-quality outcomes, and EchoWM preference. The orange segment denotes EchoWM preference.

Table 4: Preliminary ablation layout aligned with each design claim. PQ, AQ, AIF, and AVSync denote production quality, audio quality, instruction fulfillment, and temporal synchronization; the suffix −D denotes long-horizon drift, and ASR denotes audio-state retention.

(a) Viewpoint semantics

Viewpoint conditioning	Quality ↑	Setting ↑	Interaction ↑	Consistency ↑	Physical ↑
w/o viewpoint semantics	80.8	77.8	84.8	86.7	68.7
w/ viewpoint semantics	**81.5**	**79.4**	**87.2**	**89.8**	**70.6**

(b) Progressive audio-visual/control training

Training strategy	Inter. ↑	Cons. ↑	PQ ↑	AQ ↑	AIF ↑	AVSync ↓
Naive Joint	83.2	84.7	6.55	2.78	2.58	1.04
Control-SFT only	85.4	86.9	6.39	2.60	2.44	1.12
AV-CPT → Control-SFT	86.6	88.1	6.76	3.03	2.83	0.89
+ Joint-FT	**87.2**	**89.8**	**6.87**	**3.14**	**2.91**	**0.82**

(c) Metric-scale calibration

Scale normalization	Corr. ↑	Slope ↑	Scale MAE ↓	Boundary jump ↓
Per-clip	0.56	0.49	0.218	0.176
Global maximum	0.84	0.81	0.103	0.071
Global $Q_{0.9}$ (ours)	**0.92**	**0.94**	**0.064**	**0.043**

(d) Causal post-training

Post-training	Inter. ↑	Cons. ↑	PQ-D ↓	AQ-D ↓	AVSync-D ↓	ASR ↑
Causal AV TF	83.8	73.9	0.44	0.56	0.56	2.92
+ Short SGF/DMD	86.1	76.1	0.35	0.49	0.50	3.19
+ Long SGF/DMD	**87.9**	**77.5**	**0.29**	**0.43**	**0.45**	**3.48**

Human evaluation. In pairwise judgments over 200 cases, EchoWM receives the largest share of overall-preference votes against both LingBot-World-v2 (46.50%) and HappyOyster (63.13%). The detailed results also reveal a limitation: against HappyOyster, semantic following is preferred more often for HappyOyster (29.44%) than for EchoWM (13.63%). Per-dimension and per-scene results, along with the evaluation protocol, are provided in section B.4.

4.2 Ablation Studies

Viewpoint semantics and metric scale. Explicit viewpoint semantics improves interaction and consistency, supporting our hypothesis that camera geometry alone is insufficient to disambiguate first- and third-person realization. For metric control, global $Q_{0.9}$ calibration preserves the relationship between requested and realized displacement more reliably than per-clip or maximum-based normalization (table 4 and fig. 19).

Progressive audio-visual/control training. Naive joint optimization yields a weaker balance between interaction and audio-visual generation. AV-CPT establishes a strong audio-visual prior, Control-SFT acquires trajectory control while preserving this prior, and Joint-FT best reconciles the two capabilities. The full schedule therefore improves controllability while maintaining audio quality, instruction following, and audio-video synchronization (table 4).

Causal post-training. Causal teacher forcing establishes autoregressive generation but does not expose the model to the errors accumulated from its own predictions. Short-horizon SGF/DMD reduces this train–test mismatch, while long-horizon self-rollout training further improves interaction and audio-visual stability under sustained generation (table 4).

Training-scale trends. To separate data scaling from optimization effects, we vary the fraction of the full trajectory-conditioned training set while keeping the remaining training settings fixed. As shown in fig. 6, global trajectory normalization scales more consistently than per-clip normalization, and the robust global percentile reaches the full-model Interaction anchor of 87.2 at 100% data. The progressive schedule also maintains visual quality as the training set grows, while the correspond-

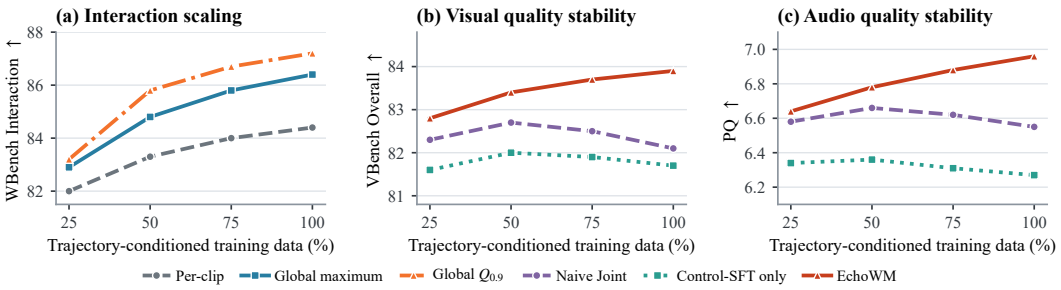

Figure 6: **Scaling trends under heterogeneous trajectory-conditioned supervision.** As trajectory-conditioned training data increases, robust global calibration continues to improve interaction while preserving visual quality. The progressive schedule further maintains stable audio-visual quality and improves production audio quality at larger training scales.

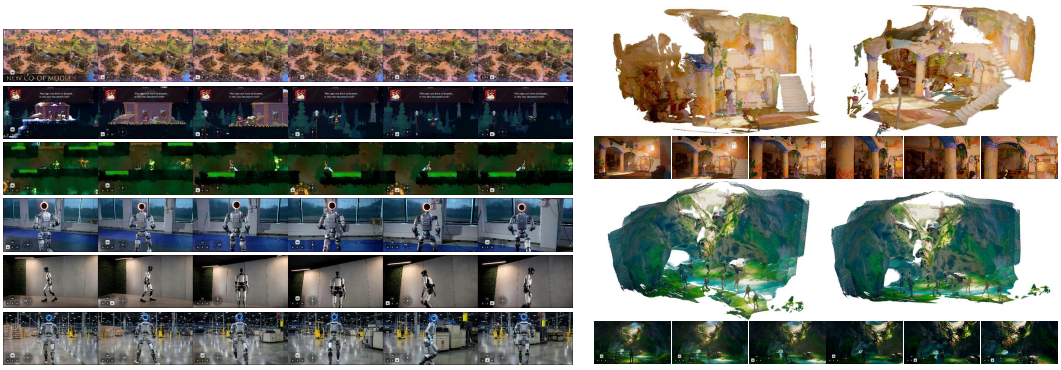

Figure 7: **Qualitative diagnostics.** (left) On unseen 2D/2.5D game and robot domains, EchoWM maintains coherent viewpoint evolution and preserves scene or subject structure under navigation. (right) Reconstructed scenes from generated trajectories remain recognizable across views, providing qualitative evidence of multi-view spatial coherence.

ing PQ curve increases from 6.64 to 6.96, supporting the claim that adding heterogeneous control supervision need not reduce native audio quality.

4.3 Qualitative Diagnostics

We provide qualitative diagnostics for two design choices: global trajectory scale calibration and explicit viewpoint semantics. Under matched navigation conditions, the examples show more gradual displacement response and improved preservation of first-person viewpoint and third-person subject visibility (figs. 19 and 21). These comparisons are qualitative; the full examples and protocol are provided in section B.5. The combined qualitative panel in fig. 7 also shows that the same camera intent generalizes to unseen 2D/2.5D game and robot domains, while trajectory-conditioned generation preserves recognizable scene layouts in qualitative 3D reconstructions.

5 Conclusion and Future Directions

We presented EchoWM, an interactive omnimodal world model that unifies first- and third-person interaction through relative metric camera intent while natively generating synchronized video and audio. By learning from complementary audiovisual and control supervision and adapting the model for causal rollout, EchoWM supports continuous interaction and multi-turn generation across diverse worlds. Looking forward, we aim to extend the interaction space beyond navigation and viewpoint control toward richer agent behaviors, and to incorporate stronger persistent world representations for more reliable long-horizon simulation. These directions would further move generative media from passive content creation toward continuously explorable and interactive worlds.

AI USE STATEMENT

We used generative AI tools for language polishing, brainstorming, and limited code assistance during the preparation of this work. All scientific hypotheses, model designs, experiments, analyses, and reported results were developed and verified by the authors. Generative AI was not used to fabricate experimental measurements or scientific evidence, and all AI-assisted content was manually reviewed and revised by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work studies interactive audio-visual world modeling for research purposes. As the model can generate realistic and controllable audio-visual content, it may potentially be misused to fabricate events or create misleading synthetic media. We therefore encourage clear disclosure of generated content and caution against deployment in safety-critical applications without appropriate human oversight and validation. Our training data are collected or processed in accordance with their applicable licenses, and any public release will document intended use, limitations, and potential misuse risks.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, camera-intent representation, metric trajectory calibration, heterogeneous data construction, progressive training schedule, and causal post-training procedure in the main paper and appendix. The supplementary material further specifies data processing, training and inference configurations, rollout construction, benchmark protocols, and evaluation settings. We plan to release the model code, data-processing and trajectory-annotation scripts, and evaluation tools to facilitate reproduction of the reported results.

REFERENCES

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

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

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.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *CVPR*, 2025.

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*, 2026.

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

Google DeepMind. Genie 3: A new frontier for world models, 2025a. URL https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/.

Google DeepMind. Veo 3, 2025b. URL https://deepmind.google/technologies/veo/veo-3/.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

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

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. CameraCtrl II: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*, 2025.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

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.

Junchao Huang, Guian Fang, Shengju Qian, Xianghao Kong, Zhuoran Zhao, Wei Huang, Yihua Du, Zixin Zhang, Justin Cui, Yuchao Gu, Yukang Chen, Xinting Hu, Tianyu He, Shaoshuai Shi, Zhuotao Tian, Xin Wang, Mike Zheng Shou, and Li Jiang. SolarWM: Open data and scalable training for long-horizon video world models. *arXiv preprint arXiv:2609.02886*, 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*, 2025a.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-0540.

Haohe Liu, Gael Le Lan, Xinhao Mei, Zhaoheng Ni, Anurag Kumar, Varun Nagaraja, Wenwu Wang, Mark D. Plumbley, Yangyang Shi, and Vikas Chandra. SyncFlow: Toward temporally aligned joint audio-video generation from text. *arXiv preprint arXiv:2412.15220*, 2024.

Kaiqi Liu, Haoxuan Zeng, Jingqi Liu, Jiacong Fang, Ziqi Cai, Yunyao Mao, Henglin Liu, Yu Sheng, Shuchen Weng, and Boxin Shi. StreamAV-Bench: A comprehensive benchmark for streaming audio-video generation. *arXiv preprint arXiv:2608.26336*, 2026.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-Foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*, 2023.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. Technical report, NVIDIA, 2026.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie Gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Zhijian Qiao, Xinjiang Wang, Jiajie Chen, Haoming Huang, Meng Li, Chih-Chung Chou, Jing Wang, and Shaojie Shen. MeRoPE: Metric rotary position embedding for camera-controlled video generation. *arXiv preprint arXiv:2609.01252*, 2026.

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 *CVPR*, 2025.

Riemann Dynamics. Matrix-Game 3.5: Enhancing real-time streaming interactive world models with patch memory. Technical report, Riemann Dynamics, 2026. URL `https://matrix-game-v3-5.github.io/`.

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*, 2026.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable generative 3D worlds. *arXiv preprint arXiv:2604.13036*, 2026.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Ying Li, Zezhong Qian, Haoyang Huang, and Nan Duan. OmniForcing: Unleashing real-time joint audio-visual generation. *arXiv preprint arXiv:2603.11647*, 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. *arXiv preprint arXiv:2512.14614*, 2025.

Jiahua Wang, Leqi Zheng, Jialong Wu, Yaoxin Mao, and Shijie Cheng. Audio-visual world models: Learning physically grounded multisensory dynamics. *arXiv preprint arXiv:2512.00883*, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-ω. *arXiv preprint arXiv:2605.15195*, 2026a.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-driven world models for autonomous driving. In *ECCV*, 2024a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*, 2024b.

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*, 2026b.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, and Ming-Ming Cheng. Infinite-World: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems*, 2025.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. CamCo: Camera-controllable 3D-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. UniSim: Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

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

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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 2025 Conference Papers*, 2025.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. *arXiv preprint arXiv:2512.07237*, 2025a.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems*, 2025b.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. FoleyCrafter: Bring silent videos to life with lifelike and synchronized sounds. *IJCV*, 2026.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal Forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026.

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. *arXiv preprint arXiv:2602.02214*, 2026b.

Junhao Zhuang, Shiyi Zhang, Yuxuan Bian, Yaowei Li, Yawen Luo, Yijun Liu, Weiyang Jin, Songchun Zhang, Xianglong He, Xuying Zhang, Haoran Li, Haoyang Huang, Zeyue Xue, and Nan Duan. Self Gradient Forcing: Native long video extrapolation. *arXiv preprint arXiv:2607.20368*, 2026.

A Additional Data Details

This appendix gives the implementation details omitted from the main data section.

A.1 Source-Specific Capture Details

Internally collected gameplay. Our internally collected recordings span first- and third-person play across open-world exploration, urban driving, and equestrian traversal. Programmatic scripts control the characters and provide explicit control inputs together with cleaner action–observation correspondence than uncontrolled Internet video. A ReShade-based plug-in suppresses heads-up displays and interface overlays before capture, preserving the original field of view without post-hoc cropping. We synchronously record high-resolution RGB video, the native stereo game mix, script-issued action logs, and environment metadata. The native mix retains dialogue, music, ambience, and interaction sounds such as footsteps, vehicles, and collisions. Because camera poses are not exported from these games, metric trajectories are recovered from video.

Human-played Internet gameplay. These recordings contain no source-native action logs, but exhibit natural control timing and context-dependent reactions to scene events. They include in-game dialogue, player commentary, and more diverse speech and acoustic variation than the internally scripted recordings. Continuous clips with reliable recovered trajectories contribute to trajectory-conditioned training, while the broader pool strengthens audio-visual and speech modeling.

Unreal Engine simulation. UE provides ground-truth metric poses, synchronized action logs, and controlled translation and rotation coverage, but no natural audio. We drive characters in the Play-In-Editor world using local-frame navigation commands and engine physics rather than navigation-mesh paths, allowing realized motion to reflect collisions and environmental constraints. When a character encounters an obstacle, the runtime controller adjusts its heading and continues the rollout. Each character carries one head-mounted first-person camera and four character-relative third-person cameras. At every step, we export synchronized frames, keyboard states, ground-truth camera poses and intrinsics, and scene metadata for all five views. Training randomly selects one view with its corresponding pose and viewpoint metadata; the capture rig is never exposed to the model.

General Internet video. General Internet video contributes realistic appearance, environmental sound, speech, richer composition, and camera–subject relationships such as tracking, orbiting, and reframing. Since these clips provide neither source-native action logs nor standardized navigation, they primarily expand the audio-visual prior. A smaller subset that passes metric pose recovery and motion-quality filtering is additionally included in trajectory-conditioned training.

A.2 Temporal Segmentation and Metric Pose Recovery

For audio-visual training, shot-boundary detection isolates continuous spans in Internet and cinematic footage before direct slicing into short clips. For non-UE data used for trajectory-conditioned training, we first preserve continuous windows of approximately one minute. Reconstructing the long window before slicing keeps neighboring clips in a shared coordinate system and provides broader temporal support for geometric optimization. All retained media are re-encoded into canonical video and audio formats while preserving stereo channels when present.

UE provides camera-to-world poses directly in metric units. For other trajectory-eligible sources, we use ViPE (Huang et al., 2025) with VGGT-Omega (Wang et al., 2026a) and MoGe-2 as complementary geometric backends. ViPE jointly optimizes temporally consistent metric-scale camera poses and intrinsics; we additionally allow per-frame intrinsic optimization to accommodate zoom, stabilization, cropping, and imperfectly known intrinsics. Following the robust annotation principle of SANA-WM (Zhu et al., 2026a), pose estimation is separated from subsequent trajectory quality control.

Because processing every frame of a long sequence is expensive, we temporally subsample each window before pose estimation. The recovered sparse trajectory is aligned to the original times-

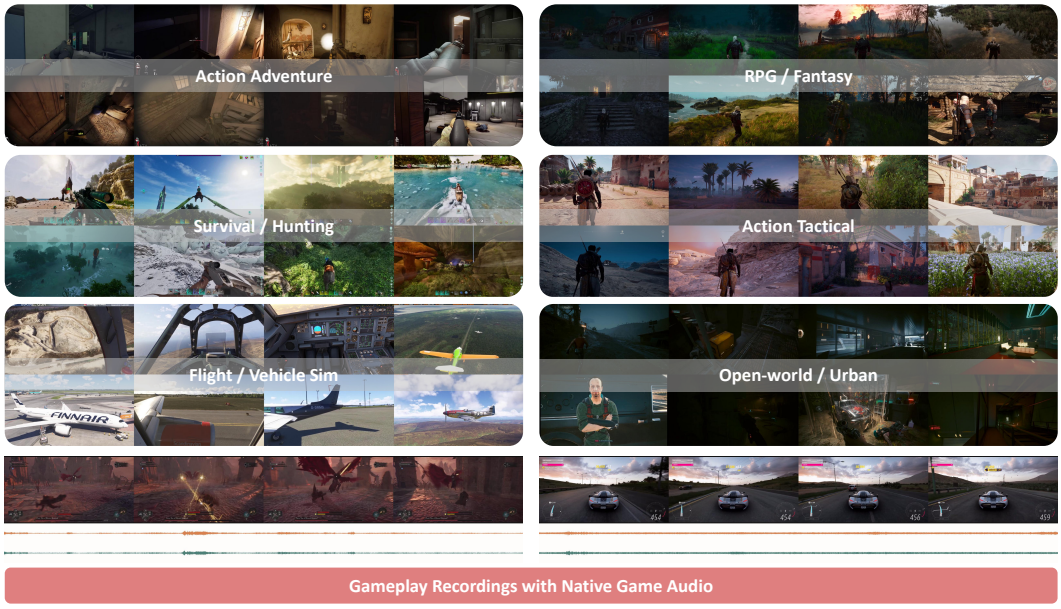

Figure 8: **Gameplay recordings with native game audio.** Representative frames from the internally collected gameplay corpus span six activity categories, with synchronized native-game audio waveforms shown below the visual sequences.

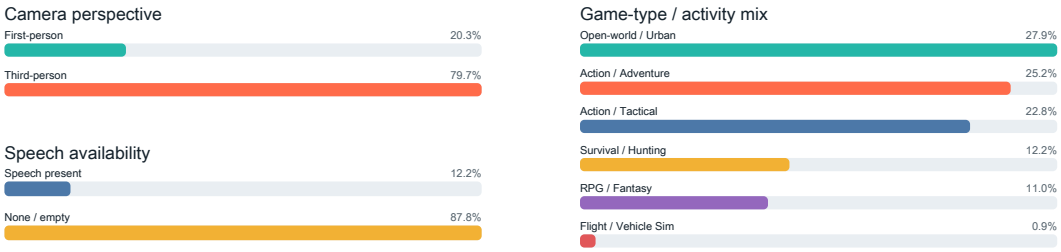

Figure 9: **Game-video dataset statistics.** The combined training corpora contain predominantly third-person clips, a broad mix of game activities, and a smaller speech-present subset. Percentages are normalized within the reported collection.

tamps using spherical linear interpolation for rotations and linear interpolation for translations. This restores frame-aligned annotations without modifying the video observations. The dense trajectories and intrinsics are converted to a common camera-to-world convention, axis orientation, metric length unit, and temporal sampling rate. Each long sequence is then sliced into short examples using identical temporal boundaries for video, audio, pose, intrinsics, and available control streams. UE rollouts bypass reconstruction and interpolation but follow the same final alignment and clip-construction pipeline.

We apply trajectory-level quality control before admitting reconstructed sequences to trajectory-conditioned training. The checks cover reconstruction reliability and intrinsic stability, temporal consistency including excessive translation or rotation jitter, and motion plausibility including implausible displacement, degenerate trajectories, and extreme scale outliers. Ground-truth UE trajectories undergo the same convention conversion and motion-range checks while retaining their original metric values.

A.3 Filtering and Structured Annotation

We remove visually degenerate clips using perceptual quality, aesthetic quality, exposure, and brightness criteria. For sources with audio, we reject invalid tracks and clips with unusable loudness

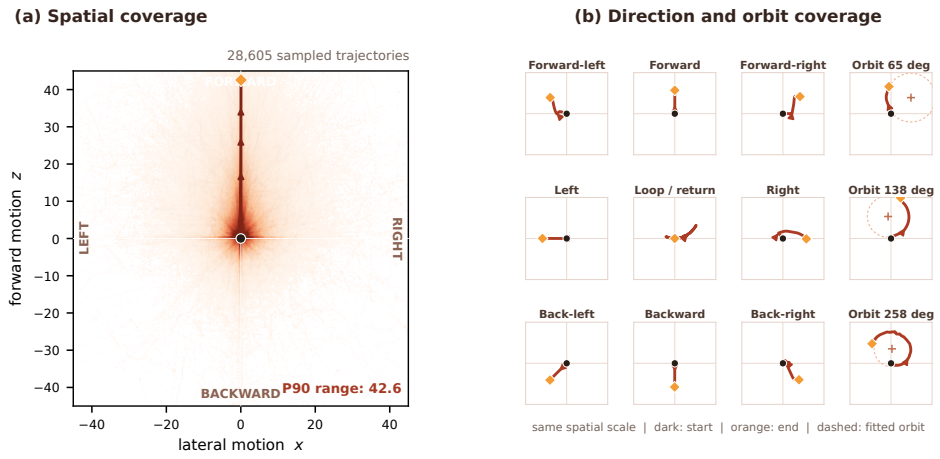

Figure 10: **Trajectory coverage of the control-clean mixture.** (a) Aggregate lateral–forward motion coverage over 28,605 sampled trajectories; the marked P90 range summarizes the robust spatial extent. (b) Representative straight, diagonal, return-loop, and orbit trajectories shown at a common spatial scale. Black and orange markers denote trajectory start and end points, respectively, and dashed circles show fitted orbits.

statistics, while separately identifying speech-containing examples for stage-wise sampling. Sources without audio, such as UE renders, are treated as missing-audio examples rather than semantic silence.

We use Qwen3-Omni (Xu et al., 2025) and Gemini-3-Pro to process synchronized video and audio. Their outputs are normalized to the following schema:

- **Persistent context: scene** describes the environment and persistent geometry; **style** describes rendering, lighting, and color treatment; **viewpoint** specifies the initial first- or third-person observer–subject relationship; and **subject** describes the primary subject's identity and appearance.
- **Temporal events: narrative** records chronological scene evolution, including subject behavior and camera motion; **speech** records spoken content and speaker role when identifiable; and **sound** records non-speech ambience, music, and interaction effects.

Source-specific names such as `static_scene_caption`, `character`, `perspective`, `narrative_caption`, and `sounds` are normalized to the canonical fields. Generation metadata and raw tagged responses are retained for provenance but are not exposed as semantic conditioning fields.

A.4 Mixture Construction and Motion Coverage

The AV path supplies the broad AV-rich pool, the geometry path supplies the control-clean pool, and their high-quality intersection forms the balanced Joint-FT mixture. For Action-SFT, the motion-bearing narrative field and other time-varying descriptions are removed so that static fields and the reference observation define the initial scene while the trajectory specifies subsequent camera evolution. For AV-CPT, the full annotation is retained. For Joint-FT, static fields are combined with speech and sound while narrative remains excluded.

For the coverage analysis, we sample 28,605 trajectories from the control-clean mixture. Their aggregate distribution spans lateral and forward–backward displacement, with representative straight, diagonal, return-loop, and orbiting motion. This exposes Action-SFT to varied realized camera evolution rather than a narrow set of discrete command templates.

B Additional Experimental Details

This section reports evaluation protocols and supporting results omitted from the main text.

B.1 WBench Navigation

We evaluate EchoWM and EchoWM-Flash on the official 158-case Navigation split of WBench (Ying et al., 2026), covering first- and third-person multi-turn interaction. We map supported control interfaces to our shared trajectory condition and follow the official evaluation protocol. The reported dimensions are Quality, Setting, Interaction, Consistency, and Physical, along with their Average. Scores use the official 0–100 normalization, with higher values indicating better performance; the main-text table reports all scores.

Qualitative comparisons. Under matched initial observations and navigation conditions, competing models can exhibit abrupt viewpoint changes, unintended zooming, subject-scale variation, or scene-layout drift. The figures below show representative first-person and third-person comparisons, as well as multi-turn rollouts from the causal model.

B.2 SANA-WM-Bench

We evaluate SANA-WM-Bench (Zhu et al., 2026a) at 241 frames for the multi-step model, at 961 frames using four sequentially chained overlapping 241-frame generations, and at 961 frames using four-step causal streaming inference for EchoWM-Flash. We report rotation error R (degrees), relative translation error T, camera-motion consistency error (CMC), and VBench Overall. At 961 frames, we additionally report same-pose revisit consistency using PSNR, SSIM, and LPIPS, and temporal imaging-quality change $\Delta\mathrm{IQ} = q_1 - q_6$, where q_1 and q_6 are the Imaging Quality scores of the first and final 10-second windows.

Table 5: Long-horizon comparison on the full SANA-WM-Bench trajectory. Revisit consistency is measured on same-pose pairs. VBench Overall is the normalized Quality component. Temporal degradation is $\Delta\mathrm{IQ} = \mathrm{IQ}_{0\text{–}10} - \mathrm{IQ}_{50\text{–}60}$ in percentage points; lower is better, and a negative value indicates improvement in the last window. Best and second-best results within each split are shown in bold and underlined, respectively.

	Trajectory and visual quality				Revisit consistency			Temporal
Method	VBench↑	$R\downarrow$	$T\downarrow$	CMC↓	PSNR↑	SSIM↑	LPIPS↓	ΔIQ↓
Simple-Trajectory Split								
Infinite-World (Wu et al., 2026)	79.18	16.55	1.98	2.08	12.60	0.284	0.595	6.72
LingBot-World (Robbyant Team et al., 2026)	**81.82**	10.47	2.01	2.05	<u>14.59</u>	**0.366**	**0.394**	<u>0.04</u>
HY-WorldPlay (Sun et al., 2025)	68.82	17.89	2.36	2.45	12.83	0.321	0.616	23.59
Matrix-Game 3 (Wang et al., 2026b)	78.53	12.96	1.83	1.92	12.29	0.326	0.553	2.41
SANA-WM (Zhu et al., 2026a)	79.29	<u>7.59</u>	<u>1.59</u>	<u>1.63</u>	14.16	0.333	0.458	3.79
EchoWM	<u>81.36</u>	**3.22**	**0.918**	**0.927**	**15.10**	<u>0.335</u>	<u>0.451</u>	**0.02**
Hard-Trajectory Split								
Infinite-World (Wu et al., 2026)	79.51	41.31	2.49	2.84	12.04	0.248	0.617	4.16
LingBot-World (Robbyant Team et al., 2026)	**81.89**	18.99	<u>1.65</u>	1.81	14.08	<u>0.332</u>	**0.436**	<u>0.58</u>
HY-WorldPlay (Sun et al., 2025)	70.46	35.46	2.34	2.64	13.72	0.328	0.654	25.88
Matrix-Game 3 (Wang et al., 2026b)	78.79	18.79	1.67	1.82	12.17	0.317	0.556	**0.32**
SANA-WM (Zhu et al., 2026a)	79.60	**10.02**	1.66	<u>1.72</u>	<u>14.10</u>	0.327	<u>0.469</u>	3.09
EchoWM	<u>81.72</u>	<u>12.05</u>	**1.265**	**1.358**	**14.58**	**0.333**	0.482	0.93

The IQ summary in table 6 shows that EchoWM-Flash has the highest mean and minimum IQ on both trajectory splits. Some baselines exhibit a smaller first-to-final change, but their consistently lower IQ indicates that a smaller drop alone does not imply stronger long-horizon quality preservation.

B.3 Native Audio-Visual Generation

We construct a 120-case image-conditioned benchmark to measure native audio generation alongside visual continuation. Each case contains a reference image, a visual continuation description,



Figure 11: WBench Navigation comparison, case 1.

and an audio condition. GPT-6 is used only to construct a scene-consistent visual prompt and a plausible audio description; it is not used as the primary evaluator. The audio descriptions cover environmental ambience, foreground events, speech, music, and mixed conditions when appropriate.

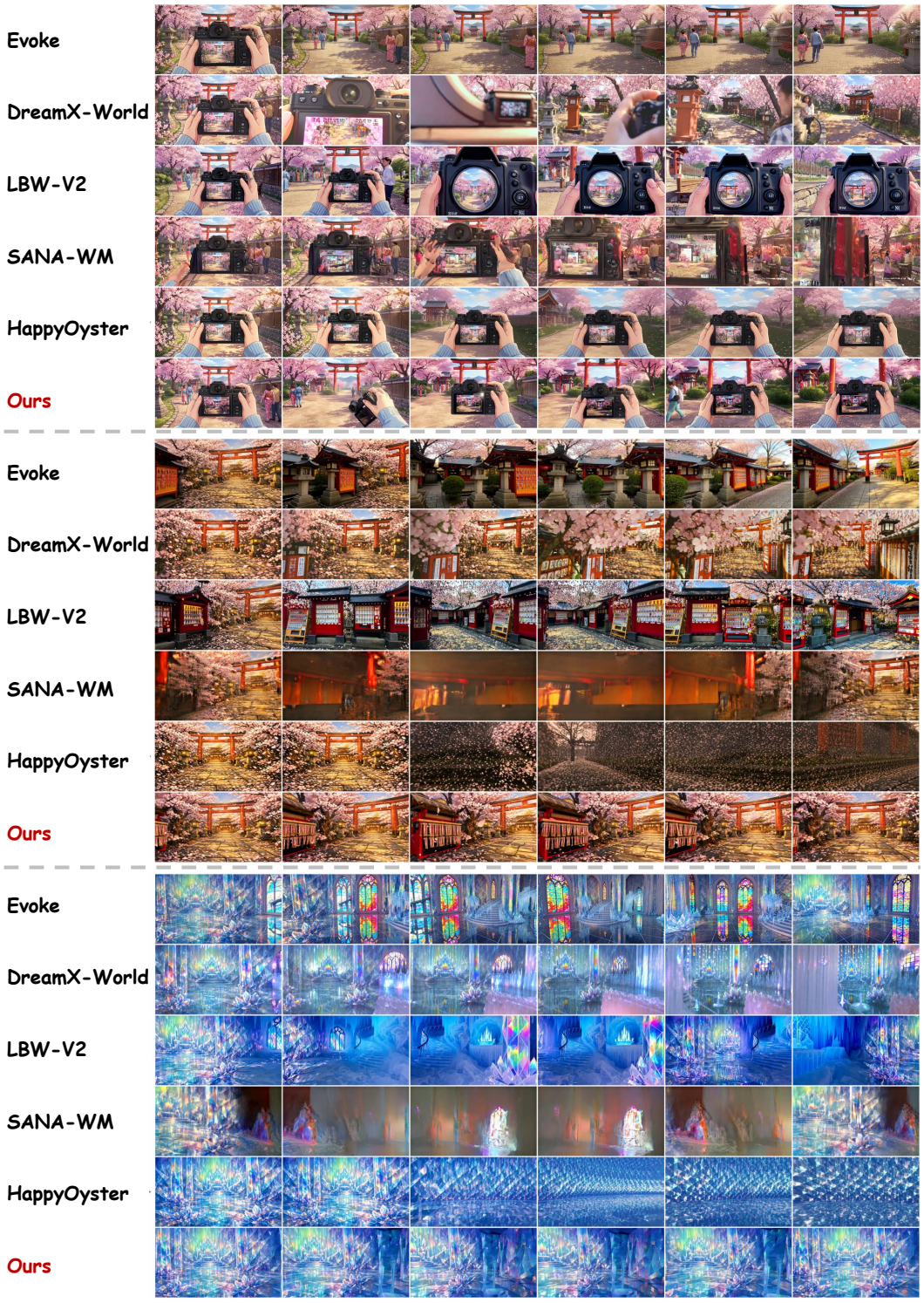

Figure 12: Additional WBench Navigation comparison.

They specify plausible acoustic content rather than ground-truth audio associated with the reference image.

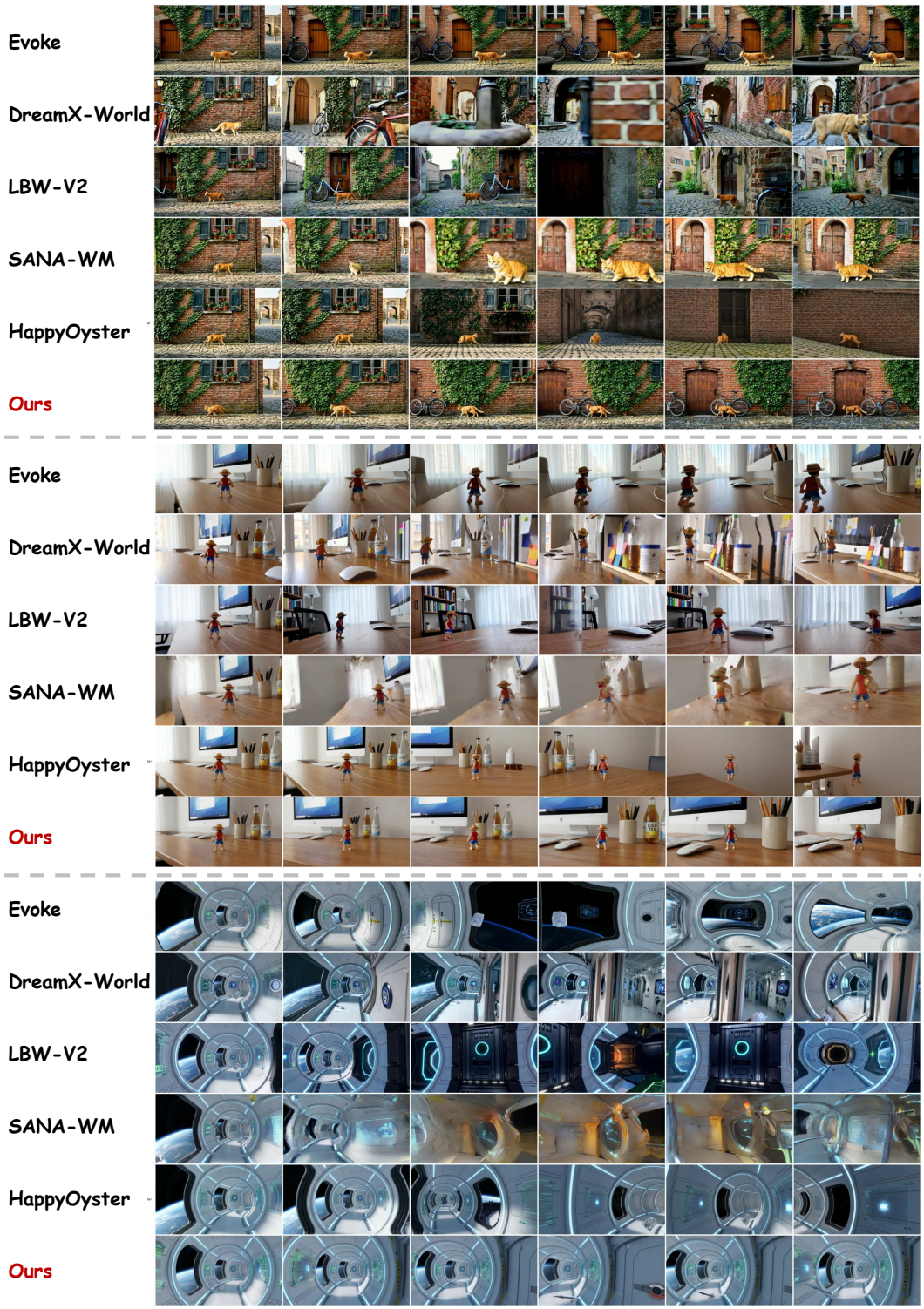

Figure 13: Additional WBench Navigation comparison.

The prompts are fixed before evaluation and manually reviewed to remove visually unsupported, ambiguous, or overly specific sound events. Each model receives the same reference image and semantic conditions and generates a 10-second clip through its native image-conditioned interface.



Figure 14: **Qualitative comparison of causal world models on WBench Navigation.** EchoWM-Flash shows better action adherence and multi-turn consistency in scene structure, subject identity, and visual style.

Table 6: Imaging quality (IQ) on the 961-frame Simple- and Hard-Trajectory splits. We report the first-window, mean, minimum, and final-window IQ; higher is better. EchoWM-Flash maintains higher mean and minimum IQ than the compared causal baselines.

Split	Method	First IQ	Mean IQ	Min IQ	Final IQ
Simple	**EchoWM-Flash**	**73.04**	**71.67**	**70.94**	70.94
	Evoke Stage3 Post-Distill	69.07	68.30	66.96	69.56
	SANA-WM + Causal Refiner	70.27	69.01	67.08	**71.02**
Hard	**EchoWM-Flash**	**72.96**	**71.36**	**70.60**	**71.65**
	Evoke Stage3 Post-Distill	69.20	68.69	65.86	68.90
	SANA-WM + Causal Refiner	69.95	68.28	65.86	70.20

Table 7: Scene-level user-study results against LingBot-World-v2 and HappyOyster. All values are percentages. The LingBot-World-v2 comparison uses seven annotators, while the HappyOyster comparison uses eight annotators.

		vs. LingBot-World-v2				**vs. HappyOyster**			
Scene	Dimension	LBW-v2	Both good	Both bad	ECHOWM	HappyOyster	Both good	Both bad	ECHOWM
Autonomous driving	Semantic following	30.48	29.52	17.14	22.86	18.33	44.17	13.33	24.17
	Initial-world consistency	3.81	66.67	15.24	14.29	10.83	50.00	0.00	39.17
	Motion	21.90	6.67	42.86	28.57	5.83	10.83	35.83	47.50
	Spatial-temporal consistency	10.48	31.43	29.52	28.57	10.00	24.17	12.50	53.33
	Visual aesthetics	11.43	34.29	21.90	32.38	19.17	3.33	17.50	60.00
	Overall preference	24.76	1.90	30.48	**42.86**	14.17	3.33	7.50	**75.00**
Embodied robot	Semantic following	12.90	9.45	42.17	35.48	42.94	41.53	6.85	8.67
	Initial-world consistency	7.37	60.60	21.66	10.37	21.37	42.14	2.02	34.48
	Motion	21.20	7.83	53.46	17.51	31.45	7.46	52.62	8.47
	Spatial-temporal consistency	15.90	25.58	34.10	24.42	10.89	17.74	18.15	53.23
	Visual aesthetics	22.81	28.34	24.42	24.42	30.44	1.01	29.64	38.91
	Overall preference	22.35	3.00	39.63	**35.02**	40.93	1.81	11.69	**45.56**
Multi-social	Semantic following	21.43	28.57	14.29	35.71	12.50	62.50	6.25	18.75
	Initial-world consistency	21.43	64.29	7.14	7.14	6.25	62.50	0.00	31.25
	Motion	14.29	42.86	14.29	28.57	0.00	31.25	43.75	25.00
	Spatial-temporal consistency	21.43	42.86	0.00	35.71	0.00	12.50	6.25	81.25
	Visual aesthetics	28.57	28.57	7.14	35.71	18.75	0.00	12.50	68.75
	Overall preference	28.57	7.14	7.14	**57.14**	12.50	0.00	6.25	**81.25**
Natural outdoor	Semantic following	17.86	16.43	37.86	27.86	46.88	43.75	6.88	2.50
	Initial-world consistency	6.43	67.14	13.57	12.86	21.88	34.38	1.25	42.50
	Motion	15.71	7.86	51.43	25.00	38.75	10.00	41.25	10.00
	Spatial-temporal consistency	15.71	21.43	26.43	36.43	13.13	14.37	8.13	64.38
	Visual aesthetics	14.29	28.57	22.86	34.29	28.75	0.63	21.88	48.75
	Overall preference	17.86	4.29	31.43	**46.43**	41.25	1.25	8.75	**48.75**

If separate visual and audio prompt fields are unavailable, we concatenate the two descriptions while preserving their semantic content.

We report production quality (PQ), perceptual audio quality (AQ), audio instruction fulfillment (AIF), semantic audio-video alignment (AVAlign), and temporal synchronization (AVSync). Higher is better for all metrics except AVSync, which is expressed in seconds. This 10-second image-conditioned evaluation is distinct from a long-horizon streaming benchmark; scores from different durations and prompt sets are not directly comparable.

B.4 HUMAN EVALUATION

Our World-Model Benchmark contains 200 cases: 100 game scenes, 50 humanoid robot scenes, 20 natural outdoor scenes, 15 household scenes, and 15 driving scenes. It includes 132 third-person and 68 first-person cases. We use the WBench control convention (Ying et al., 2026): in first person, WASD moves the camera and the arrow keys rotate the view; in third person, WASD moves the visible subject and the arrow keys orbit the camera around it. The same command semantics are used for each compared model.

Annotators judge semantic following, initial-world consistency, spatial-temporal consistency, motion quality, and visual aesthetics. The The LingBot-World-v2 comparison uses seven annotators and the HappyOyster comparison uses eight; the scene-level analysis covers 99 cases in autonomous driving, embodied robot, multi-social, and natural outdoor scenes. These breakdowns show that

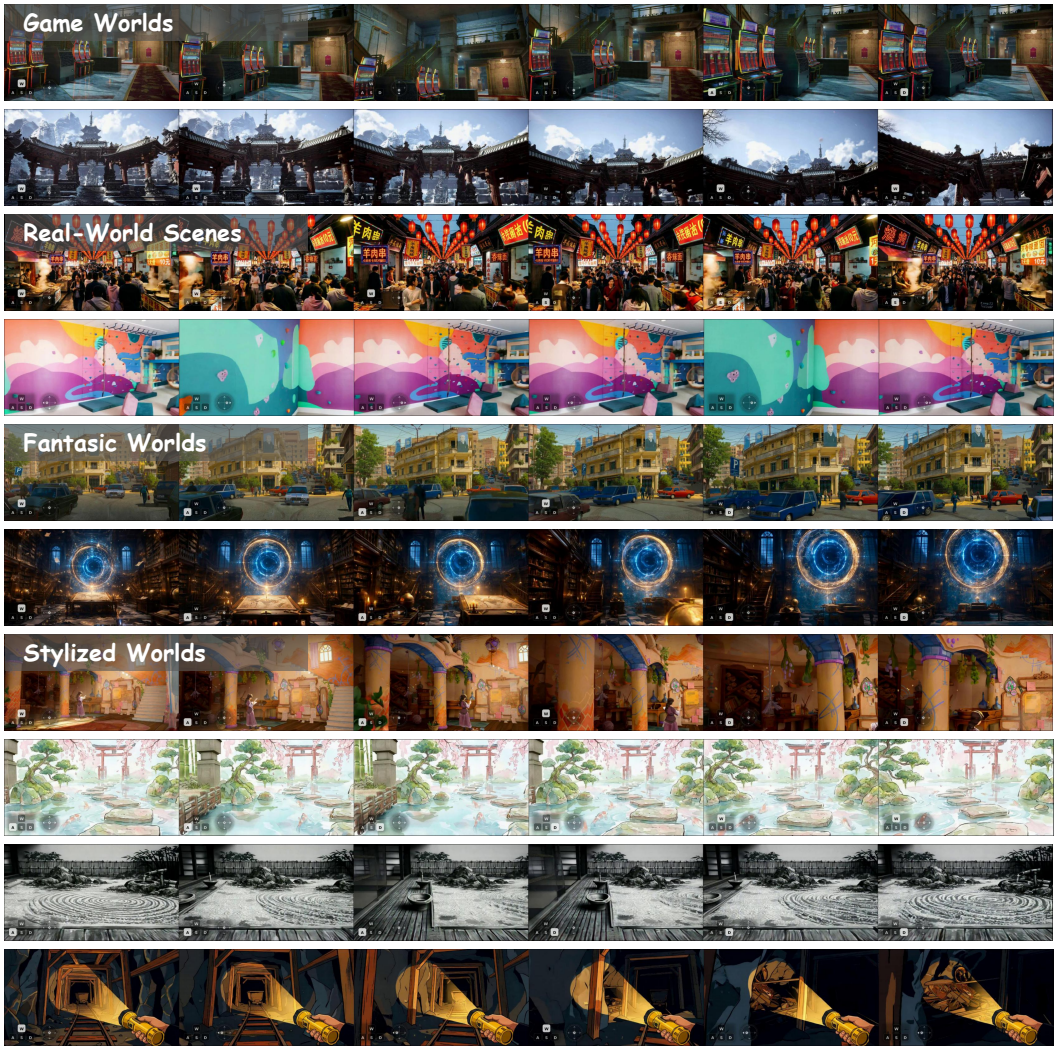

Figure 15: **First-person world generalization.** Temporally ordered rollouts cover game worlds, real-world scenes, fantastic worlds, and stylized worlds. Across varied layouts and rendering styles, ECHOWM uses the same camera-intent interface for observer-centric navigation.

overall preference can coexist with localized weaknesses, particularly in semantic following for some non-game scenes.

B.5 QUALITATIVE ANALYSIS AND DESIGN DIAGNOSTICS

The following examples complement the quantitative benchmarks with first-person generalization, third-person subject tracking, multi-view spatial coherence, synchronized audio-visual generation, and long-horizon continuation. The 3D reconstructions are qualitative evidence of scene-level coherence and do not establish metric geometric accuracy, since they depend on the downstream reconstruction procedure.

We also include matched comparisons for trajectory scale calibration and viewpoint semantics. The speed-control figure is qualitative: it visualizes realized displacement over matched rollouts and should not be read as a quantitative response-curve evaluation.

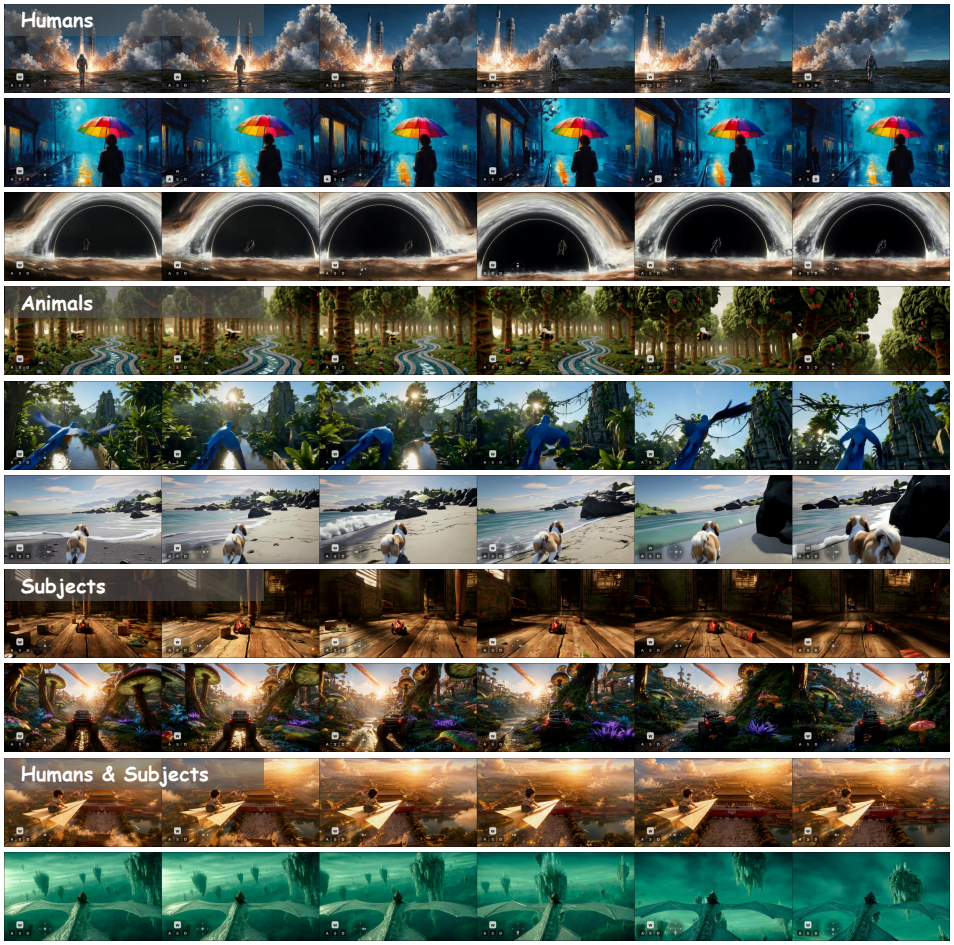

Figure 16: **Third-person generalization across visible subjects.** Representative rollouts span humans, animals, other foreground subjects, and scenes containing both humans and additional subjects. The shared camera-intent interface supports learned camera–character evolution across these subject configurations without a subject-specific controller.



Figure 17: **Native synchronized audio-visual generation.** Each example pairs its world and audio description with time-stamped video frames, the generated waveform, and a log-frequency spectrogram. The cases contain environmental sound, background music, and character speech within the same generated audio-visual sequence, rather than adding audio as post-processing.

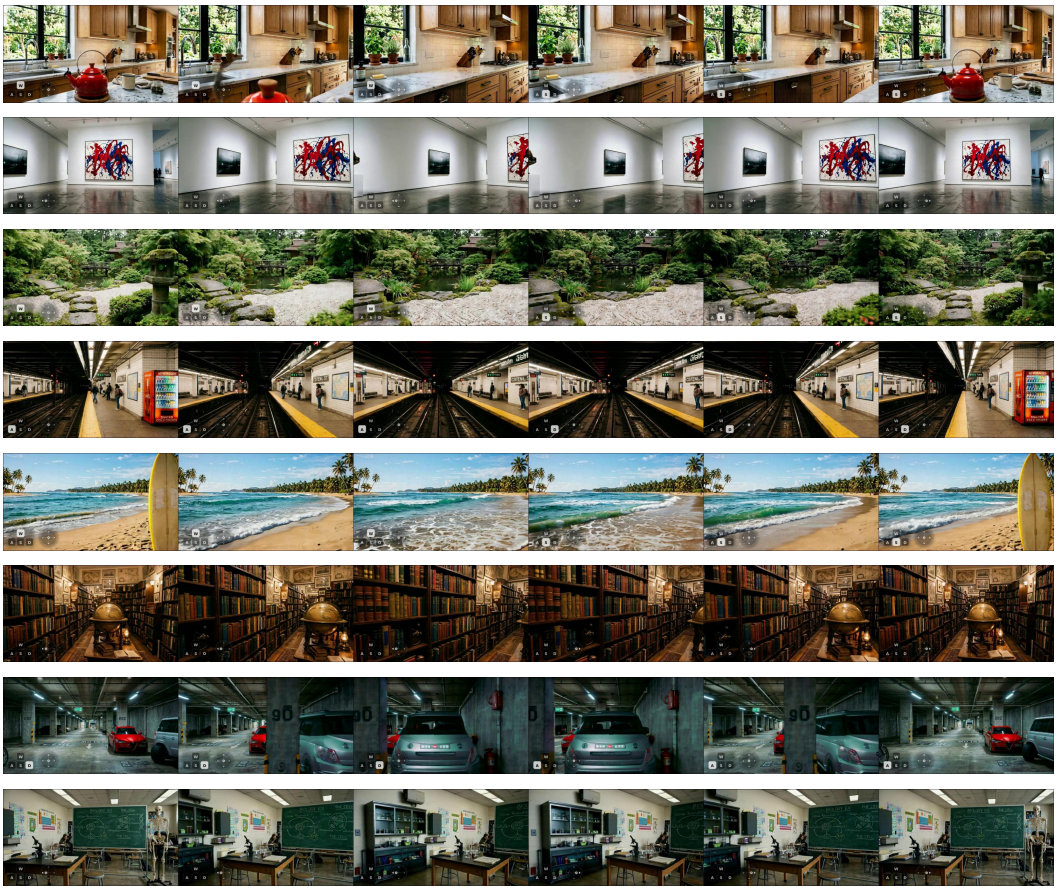

Figure 18: **Visual memory during long-horizon continuation.** Each row samples a temporally ordered trajectory that moves through and revisits a scene. Returning to previously observed views tests whether layout, appearance, and scene identity remain coherent over continued generation.

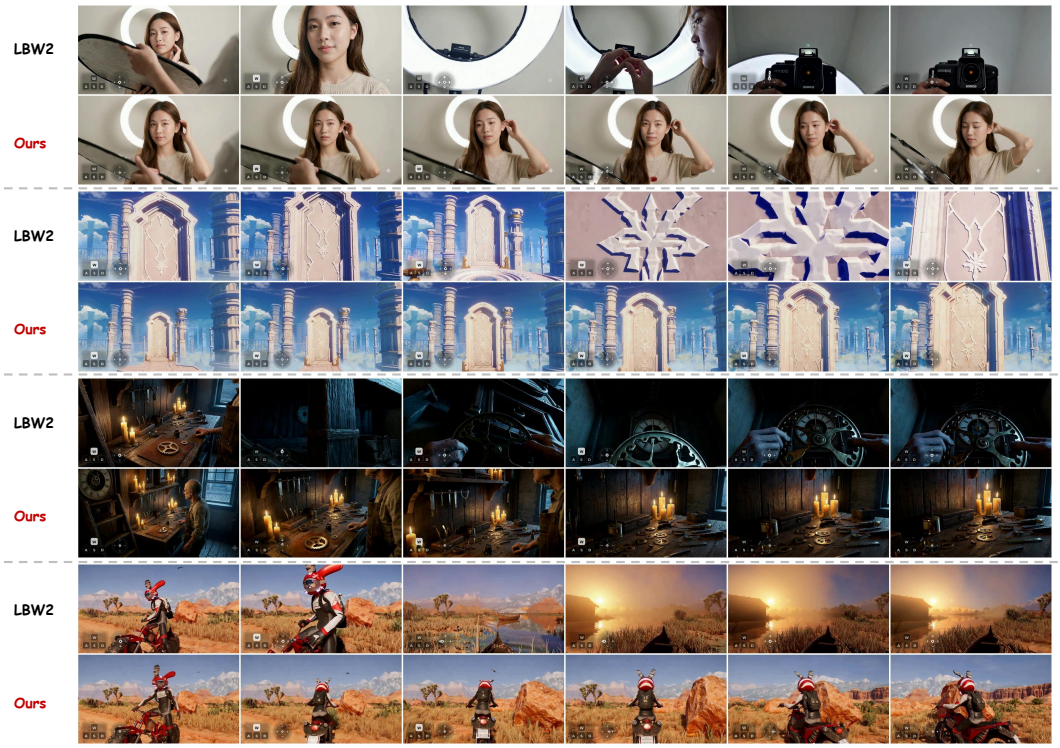

Figure 19: **Qualitative scale-control comparison.** LingBot-World-v2 (Robbyant Team et al., 2026) and ECHOWM receive matched reference observations and navigation conditions. When the camera approaches a foreground character or scene structure, LingBot-World-v2 exhibits larger changes in apparent subject scale and less predictable displacement response, whereas ECHOWM preserves a more gradual scale change across the trajectory. This qualitative comparison illustrates the effect of using one global translation scale, which keeps requested displacement amplitudes comparable across clips and makes approach behavior more controllable.

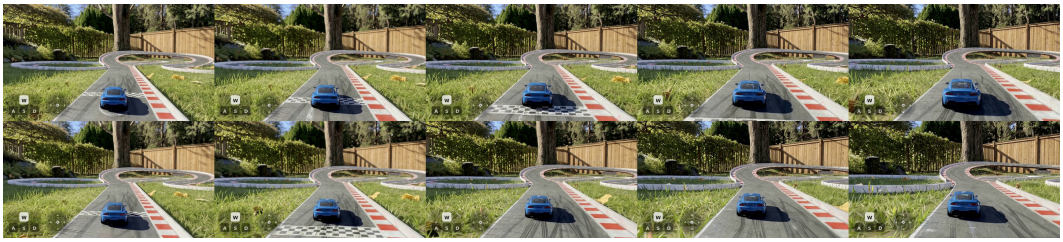

Figure 20: **Qualitative inference-time speed control.** Each row shows five temporally ordered frames from a navigation rollout under the same scene and control interface. The changing vehicle position illustrates the realized displacement over the rollout and complements the quantitative response curves described above.

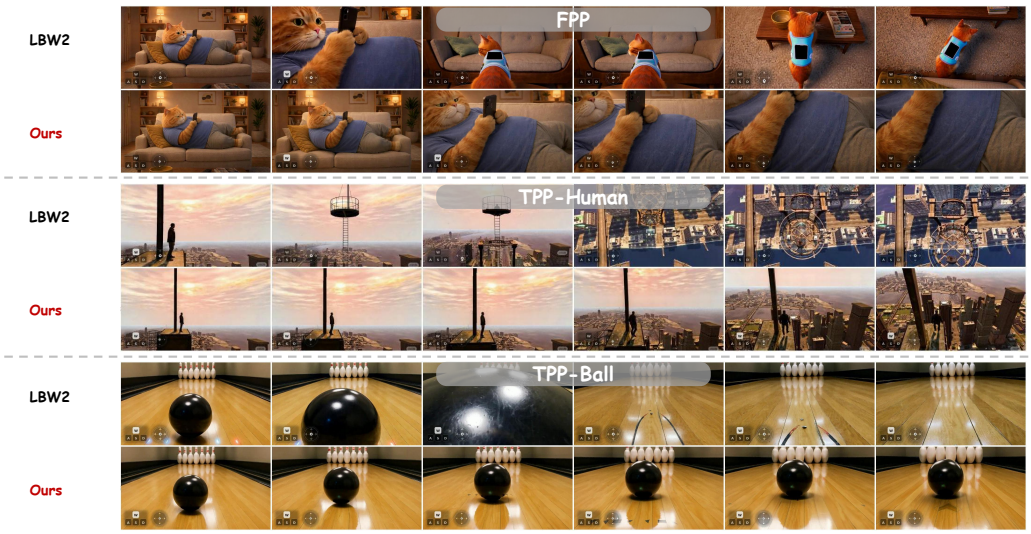

Figure 21: **Qualitative viewpoint semantics and third-person tracking.** LingBot-World-v2 (LBW2) (Robbyant Team et al., 2026) and EchoWM receive matched reference observations and navigation conditions in first-person (FPP), third-person human (TPP-Human), and third-person object (TPP-Ball) cases. In the FPP example, LBW2 changes to an exocentric observation as the rollout proceeds, whereas EchoWM retains the requested first-person semantics. In both TPP examples, LBW2 loses the controlled subject from view, while EchoWM keeps the human or ball visible and maintains a more stable camera–subject relationship across the sequence.