REWORLD: AN INTERACTIVE WORLD MODEL WITH LONG-HORIZON MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive world models must follow actions, recall past views, and stream in real time. Control needs a short attention horizon; memory needs a long one. Jointly learning these abilities can strengthen action following while weakening revisit consistency. ReWorld separates these during training through mixed per-head attention windows and random head routing. Routing exposes each head to both horizons, allowing all heads to share a bounded cache at deployment without additional model parameters. Chunk-drop training prepares the model for sparse histories, while a bounded KV cache and pose-indexed landmark bank support long-horizon recall at inference. The bank preserves selected past chunks and retrieves them by camera pose, retaining spatially relevant history beyond the recent window. A metric-aligned data pipeline unifies eight synthetic, game, and real-world sources, and LoRA-confined distillation enables both high-fidelity and real-time 704×1280 generation from one backbone. Against six recent interactive world models, ReWorld achieves the lowest rotation error ($11.95°$), best camera-motion consistency, and highest mean score across the seven evaluated VBench dimensions. On 64 s out-and-back rollouts (384 latents), its fixed 12-chunk cache still regenerates the starting view.

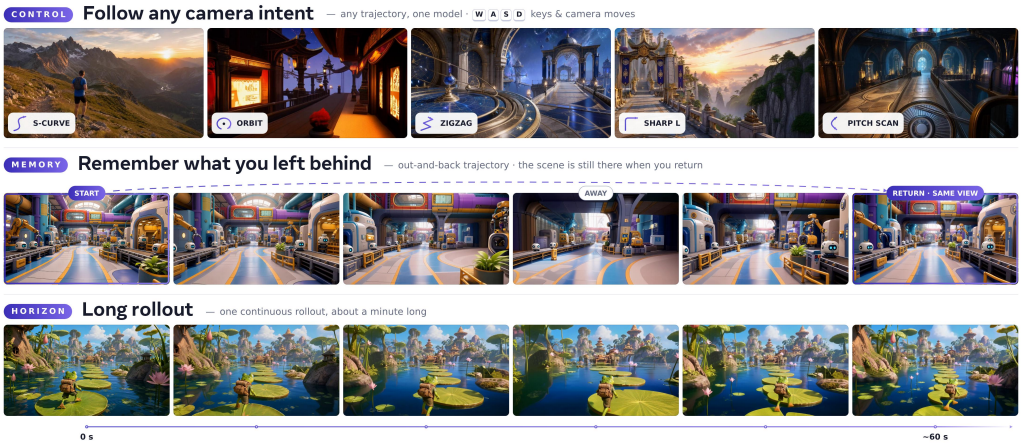


Figure 1: **ReWorld**: camera following (top), revisit regeneration after an out-and-back excursion (middle), and long-rollout streaming (bottom).

1 INTRODUCTION

A world model simulates an environment an agent can act in (Ha & Schmidhuber, 2018). Recent interactive systems turn a stream of user actions into video of a coherent, explorable environment (Bruce et al., 2024; Parker-Holder & Fruchter, 2025; Sun et al., 2025; Wang et al., 2026; Robbyant Team et al., 2026). Expectations have converged: the model should react, reflecting a key press in the very next frames; remember, so that a revisited place looks the way it did before; and stream, generating at interactive rates over unbounded rollouts.

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.

Current systems pursue these goals along two axes. For control, one route folds camera pose into attention, so attention logits depend on relative pose (Xu et al., 2026; DreamX Team, 2026); the other directly injects an explicit action signal (Sun et al., 2025; Wang et al., 2026). For memory, camera pose is the retrieval key of choice, in three forms: selecting which past frames re-enter the context (Yu et al., 2025), indexing an external memory bank (Xiao et al., 2025; Wu et al., 2025), or entering the attention index itself (Xu et al., 2026).

Learning a world model comes down to learning two abilities: control, so that the next frames follow the current action, and memory, so that a revisited place looks the way it did before. When the two are trained together, control learns well but memory does not: adding direct action injection improves every control metric while revisit fidelity drops (Sec. 3.5, Table 5). The two abilities also ask for different attention windows. Control should not depend on the window size: at inference the model must respond to the current action correctly whatever window it is given, and what the response requires—the current scene and the current command—sits inside a short window. Memory is the opposite: retrieval can only be learned under a long window, because a model that cannot see the far past has nothing to retrieve from. The learning of the two abilities can therefore be split by window: control learned under short windows, memory under long ones.

This paper presents ReWorld, an action-controllable streaming world model designed in two steps: split the training of control and memory by window, then consolidate memory under a fixed KV budget at inference. We keep both conditioning channels: pose-indexed attention (MRoPE) finds cached content by pose, and direct action injection states the command. What we separate is the window each attention head trains under. Mixed attention windows give the model both window lengths in every iteration: most heads are *local* and attend only to a short recent window, while a small set of *global* heads attends over the entire causal past. Random head routing then switches which heads are global at every step, cycling through a fixed pool of random partitions. The routing is what makes the split workable. The split itself cannot be kept at inference: the full history is gone, and all that remains is one bounded cache that every head reads, so there is no way to give some heads a long window and others a short one. A fixed partition would train heads to specialize in windows that deployment cannot provide. Routing removes this dependence: which heads are global changes at every step, every head trains under both windows, and neither ability binds to particular heads—at inference any head can read the shared bounded cache. The recipe adds no parameters, losses, or modules, and it makes a testable prediction: swapping cache-compression policies at inference should leave control error unchanged (Sec. 3.5).

With control insulated, the second step strengthens memory at inference: the rollout is unbounded but the cache is not. ReWorld therefore *consolidates*, keeping few chunks but keeping them intact: a chunk aging out of the recent window is stored in a bounded landmark bank only if the camera has travelled far enough since the last stored landmark—one full-resolution snapshot per stretch of camera travel—and once the bank is full, each admission evicts the member most spatially redundant with the rest. *Retrieval* then fills the fixed cache with a sink chunk (Xiao et al., 2024), a recent window, and the landmarks closest to the current camera pose. Inference thus reads a sparse, non-contiguous history, while standard training sees only a complete, contiguous prefix—a train–test mismatch (Huang et al., 2025c). *Chunk-drop training* closes the gap: a random subset of past chunks is masked at every step, teaching the model to reconstruct scene state from incomplete memory and making the spliced caches of deployment in-distribution (Secs. 2.3 and 2.3).

Two components complete the system. A metric-scale-aligned pipeline places synthetic, real, and game footage on a single physical action scale, so the same key press moves the camera the same physical distance in every source, with palindrome trajectories supplying the revisit evidence memory training needs (App. D). Distribution-matching distillation with self-forcing rollouts (Yin et al., 2024b;a; Huang et al., 2025c) compresses sampling into a few denoising steps inside a LoRA adapter (Hu et al., 2022), so one backbone serves both a high-fidelity multi-step operating point and a real-time interactive one (Sec. 2.4).

Our contributions are:

- **Window-split training of control and memory.** Mixed per-head attention windows train control under short windows and memory under long ones, and random head routing keeps control in every head complete within the short window, so cache compression at deployment does not disturb action following (Secs. 2.2 and 3.5).

- **Chunk-drop training with consolidation-based inference.** A training augmentation that makes sparse, non-contiguous KV caches in-distribution, paired with an inference mechanism that consolidates aged chunks into a landmark bank kept bounded by redundancy-based eviction and retrieves them by pose proximity, so spatial memory persists over unbounded rollouts at a fixed KV budget (Secs. 2.3 and 2.3).
- **A metric-aligned multi-source data pipeline.** Synthetic, real, and game footage aligned to one physical action scale, with palindrome augmentation providing revisit supervision (App. D).
- **Real-time deployment and evaluation.** LoRA-confined few-step distillation that gives a single backbone both a high-fidelity and a real-time operating point, together with an evaluation protocol covering action following, long-horizon recall, and general video quality (Secs. 2.4 and 3).

2 METHOD

2.1 OVERVIEW

ReWorld is a causal flow-matching diffusion transformer (Lipman et al., 2023; Esser et al., 2024) that streams video from a text prompt, an optional reference image, and per-chunk 6-DoF camera actions. Figure 2 summarizes its components: pose-indexed attention and action injection, mixed attention windows with random head routing, a bounded landmark cache with chunk-drop training, and four-step LoRA distillation. Training uses eight metrically aligned data sources (App. D).

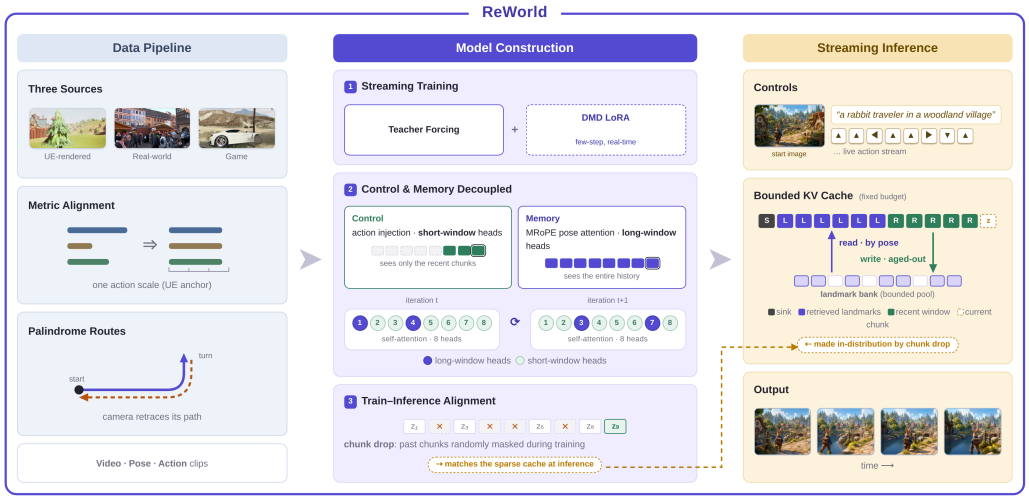


Figure 2: **Overview of ReWorld.** *Left:* metrically aligned UE, real, and game footage with palindrome revisit supervision (App. D). *Middle:* action injection and MRoPE learn control and memory through randomly routed local/global attention windows; chunk drop prepares the model for sparse caches. A DMD LoRA enables few-step inference. *Right:* a fixed cache combines a sink, recent chunks, and pose-retrieved landmarks from a bounded bank.

2.2 CAMERA-CONTROLLABLE STREAMING GENERATION

Backbone. We make Wan2.2-TI2V-5B (Wan et al., 2025; Peebles & Xie, 2023) causal over chunks of four latent frames, with $L{=}12$ chunks per training window. At inference, each generated chunk appends its keys and values to the cache $\mathcal{C}$ (App. A.1).

Pose-indexed attention as implicit memory. Under standard time-and-space RoPE, a revisited location is positionally just a distant timestamp, leaving retrieval to content alone. We instead build spatial memory into the attention index, adopting Memory-RoPE (MRoPE) after PM-RoPE's

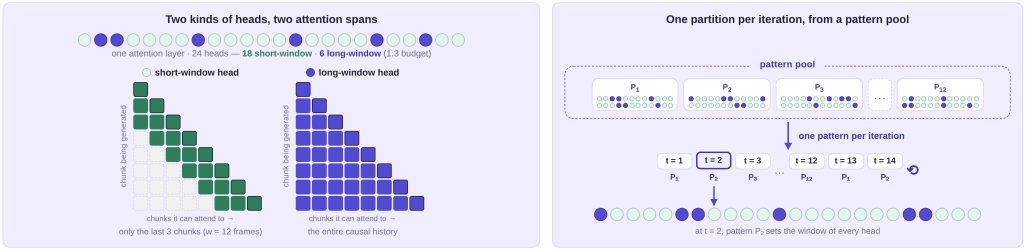


Figure 3: **Randomly routed attention windows.** Local heads see the last w=12 frames; global heads see the full history. Each step selects one of $|\mathcal{P}|$=12 partitions, preventing fixed head specialization.

camera-phase design (Xu et al., 2026): each latent frame f carries a relative camera-to-world pose $P_f \in \mathrm{SE}(3)$, which a zero-initialized MLP maps to a phase offset on RoPE queries and keys, so attention depends on pose difference rather than temporal distance. The cache thus becomes an implicit spatial memory—chunks stored with the pose they were seen from, retrieved by pose proximity on revisits—for the cost of one small MLP plus two linears (formulation and two-pass comparison in App. A.2, Table 6). **MRoPE alone gives strong revisit memory but imprecise control** (Sec. 3.5, Table 5).

Action injection. For control authority we therefore inject the command directly: each latent frame's commanded pose becomes a Plücker ray map, in the same first-frame coordinates as the MRoPE poses, projected by an MLP and added token-wise to the patch embeddings (construction in App. A.3). The ray map tells each token of the *current* chunk where its camera should look; MRoPE poses tell attention where every *cached* token was seen. Both encode the same trajectory, and when every head receives both, the action signal crowds out pose-keyed retrieval: **control improves, but long-horizon memory degrades** (Sec. 3.5).

Mixed per-head attention windows. Control requires the current scene and command; memory requires access to distant content. We split each block's H=24 heads into 18 *local* heads attending to the last w=12 latent frames (three chunks) and six *global* heads attending to the full causal history (Fig. 3; mask in App. A.4). This trains control and retrieval under their respective horizons.

Random head routing. A fixed partition would not survive deployment, where all heads read the same bounded cache (Sec. 2.3). We therefore cycle through a pool $\mathcal{P}$ of $|\mathcal{P}|$=12 random six-head global sets, switching at every optimizer step (Fig. 3, right). A head is local on most steps, so its control must be complete within the short window, while retrieval lives in whichever heads currently see far; changing the inference cache policy should then leave control error unchanged—a prediction Sec. 3.5 tests.

Training uses teacher-forcing diffusion at 480p, followed by a 720p warm start; configuration details are in App. A.5.

2.3 Memory Consolidation under a Bounded KV Budget

An unbounded KV cache grows with rollout length and exhausts device memory at 720p (Yu et al., 2025; Sun et al., 2025; Wang et al., 2026; Wu et al., 2025). ReWorld separates *consolidation*, which selectively stores aged chunks intact in a bounded landmark bank, from *retrieval*, which fills a fixed B=12-chunk cache. Chunk-drop training prepares the model to read this sparse history.

Landmark consolidation and retrieved. At chunk k, the active cache is

$$\mathcal{C}_k = \underbrace{\{1\}}_{\text{sink}} \cup \underbrace{\{k-5,\ldots,k-1\}}_{\text{recent window}} \cup \underbrace{\mathcal{T}_k}_{\text{retrieved landmarks}}, \qquad |\mathcal{T}_k| = 6, \quad |\mathcal{C}_k| = B = 12. \tag{1}$$

The sink anchors scene layout (Xiao et al., 2024); all chunks retain full resolution. A chunk leaving the recent window enters the bank $\mathcal{M}$ ($|\mathcal{M}| \le K$, K=30) only after camera travel exceeds a stride

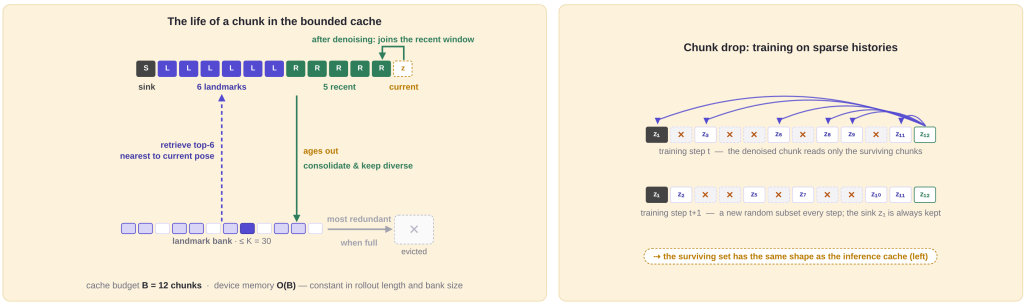


Figure 4: **Memory consolidation and chunk-drop training.** *Left:* one sink, six pose-retrieved landmarks, and five recent chunks fill the $B{=}12$ cache. Aged chunks enter a bounded landmark bank and return on revisits. *Right:* training masks random past chunks to match sparse inference histories.

δ since the last stored landmark. This odometer rule admits revisit passes as well as first visits. A full bank evicts its most pose-redundant member. After each generated chunk, the six landmarks nearest the current pose form $\mathcal{T}_k$. The full bank resides in host memory, keeping device storage at $O(B)$ regardless of rollout length (App. B.1).

Chunk-drop training. Standard teacher forcing exposes a complete, contiguous prefix, whereas deployment supplies a sparse history. We bridge this gap by retaining the sink and five randomly selected chunks from each $L{=}12$-chunk training window. Queries are never dropped: each chunk receives its denoising loss and attends to surviving earlier clean keys and its own chunk (App. B.2). Section 3.5 compares this design with sliding-window, mean-pooled, and retrieval-free caches under the same $B{=}12$ budget, plus an unbounded full-KV upper bound.

2.4 Real-Time Distillation with a Lightweight Adapter

To run in real time, ReWorld generates with only four denoising steps, obtained by step distillation confined to a plug-in LoRA in the AR-train-then-distill route of the LongLive series (Yang et al., 2025a; Chen et al., 2026) (Fig. 5): the multi-step model is the teacher, and the student is the *same backbone, frozen*, plus a rank-128 LoRA on the attention and feed-forward layers—the LoRA is the only thing trained (the DMD critic is also a LoRA over the frozen weights). Training follows distribution-matching distillation with self-forcing rollouts (Huang et al., 2025c; Yin et al., 2024b;a): the student streams chunks from its own KV cache, just as at deployment, so distillation needs only text prompts—no video. The prompts are generic long captions from another domain (248K, extended from VidProM), and the action and pose injection path carries no LoRA, so the model's control is untouched by distillation.

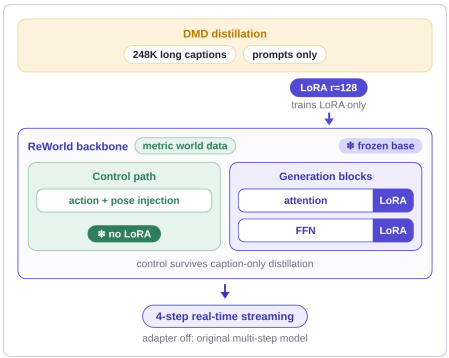


Figure 5: **Real-time distillation.** Caption-domain DMD LoRA plugs into the frozen backbone; the control path carries no LoRA.

One backbone thus has two modes: adapter off is the original multi-step model; adapter on streams in real time at four steps. The 2.6 GB adapter also transfers across base checkpoints, re-distilling in 2k–4k steps when needed. Two practical rules: distill at the deployment resolution, with prompts that resemble deployment prompts; the production student uses block size 16 with a single rollout and folds classifier-free guidance into the adapter.

Table 1: Camera controllability on 240 clips per method (40 images $\times$ 6 trajectories), after $\mathrm{Sim}(3)$ and orientation alignment. Lower is better; column best in bold.

	Overall			RotErr° ↓ per trajectory					
Method	**RotErr° ↓**	**TransErr↓**	**CamMC↓**	**dolly**	**strafe**	**arc yaw**	**arc pitch**	**s curve**	**palindrome**
SANA-WM	13.02	0.123	0.388	4.95	1.51	32.01	21.33	17.54	0.80
DreamX	13.10	0.114	0.381	2.50	1.21	33.00	20.71	18.06	3.13
HY-WorldPlay	14.66	0.114	0.427	**0.26**	**0.24**	43.53	22.00	21.73	**0.20**
Matrix-Game 3.0	15.45	**0.075**	0.394	8.19	1.00	39.93	**20.59**	17.75	5.24
LingBot-World	12.59	0.107	0.354	3.10	6.01	**26.21**	26.83	**12.23**	1.16
Yume-1.5	14.24	0.131	0.428	0.84	2.01	29.05	24.57	25.60	3.38
ReWorld (ours)	**11.95**	0.102	**0.332**	2.69	1.07	27.63	23.61	16.06	0.64

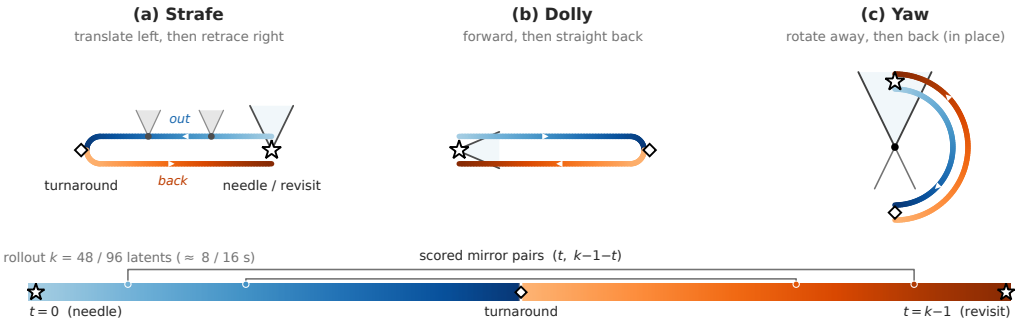


Figure 6: **Constructed palindrome trajectories.** Outbound (blue), retraced (orange): the needle view at $t{=}0$ must be regenerated at $t{=}k{-}1$; scored mirror pairs span at least half the rollout.

3 Experiments

3.1 Setup

Implementation. We train Wan2.2-TI2V-5B (Wan et al., 2025) at 480p, then warm-start at 720p, using the routing and chunk-drop recipe of Sec. 2. Unless stated otherwise, deployment uses a four-step DMD LoRA at 704×1280, a $B{=}12$-chunk cache, and a $K{=}30$ landmark bank (App. C.1).

Evaluation suite. We evaluate (i) *camera controllability* against six interactive baselines (Sec. 3.2); (ii) *long-horizon memory* with a needle-in-a-haystack (NIAH) protocol up to 384 latents (Sec. 3.3); (iii) *video quality* on seven video-intrinsic VBench (Huang et al., 2024) dimensions (Sec. 3.4); and (iv) *ablations* of recipe, cache policy, and fusion (Sec. 3.5).

3.2 Camera Controllability

The benchmark pairs 40 start images with the six canonical trajectories of Table 1; clips are re-tracked with ViPE (Huang et al., 2025a) and aligned so the comparison measures trajectory *shape* (App. C.2). We report RotErr (geodesic), TransErr, and CamMC against SANA-WM (Zhu et al., 2026), DreamX (DreamX Team, 2026), HY-WorldPlay (Sun et al., 2025) (its lightweight Wan-based variant on this benchmark), Matrix-Game 3.0 (Wang et al., 2026), LingBot-World (Robbyant Team et al., 2026), and Yume-1.5 (Mao et al., 2025).

ReWorld attains the best overall RotErr and CamMC and stays competitive on the rotation-heavy arcs—the hardest regime for every method.

Table 2: Palindromic revisit benchmark at k=48/96 latents. Path length (outbound optical flow, px) contextualizes similarity and is not ranked. Best in **bold**; second best underlined.

	k=48 (≈8 s)					k=96 (≈16 s)				
Method	**Path (px)**	**SSIM↑**	**LPIPS↓**	**DINO↑**	**ORB↑**	**Path (px)**	**SSIM↑**	**LPIPS↓**	**DINO↑**	**ORB↑**
SANA-WM	296	0.318	0.433	0.810	0.174	499	0.313	0.528	0.700	0.161
DreamX	279	0.214	0.553	0.740	0.180	454	0.194	0.627	0.603	0.159
HY-WorldPlay	113†	**0.448**	**0.163**	**0.969**	**0.464**	210†	**0.427**	**0.247**	**0.942**	**0.460**
Matrix-Game 3.0	312	0.318	0.369	0.891	0.292	724	0.275	0.478	0.850	0.247
LingBot-World	352	0.274	0.557	0.663	0.177	834	0.251	0.635	0.509	0.162
Yume-1.5	165	0.268	0.533	0.747	0.179	332	0.269	0.586	0.661	0.152
ReWorld (ours)	286	0.349	0.282	0.913	0.325	615	0.384	0.332	0.932	0.379

† HY-WorldPlay has the shortest paths (113/210 px), which can favor revisit similarity.

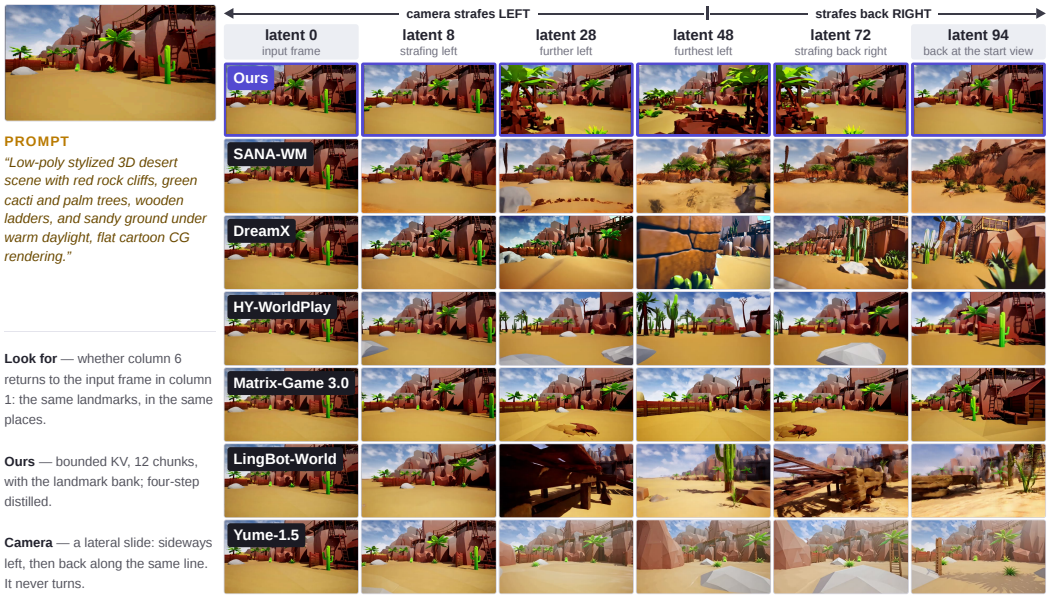


Figure 7: **Memory on a strafe-and-return rollout.** All methods share the input and trajectory; columns show latents 0/8/28/48/72/94 (k=96). The last view should match the first. ReWorld (top; four-step, B=12 with landmark bank) preserves the returning layout.

3.3 Long-Horizon Memory

Protocol and metrics. We probe spatial memory with NIAH probes on *palindrome* trajectories (Fig. 6): views generated early—the needles—must be reproduced after a long gap. The benchmark pairs 12 start images with 3 palindromes (*strafe*, *dolly*, *yaw*) at $k \in \{48, 96\}$ latents (≈8/16 s; 36 clips per method and length); our own arms additionally run nested lengths $k \in \{96, 192, 288, 384\}$ for the ablations (Sec. 3.5). Revisit pairs are scored by SSIM, LPIPS, DINO, and ORB, plus executed *path length*—a rollout that barely moves scores high on any similarity metric (App. C.3).

Comparison to baselines. Table 2 reports both lengths; ReWorld runs its default deployment (consolidation with landmark bank, B=12), HY-WorldPlay its flagship autoregressive model. k=48 is a sanity check several baselines can still solve from context (App. C.3); k=96 forces recall from beyond every bounded window. HY-WorldPlay posts the strongest baseline scores but also moves the least—210 vs. 332–834 px median outbound path at k=96. Figure 7 shows the contrast qualitatively.

Table 3: VBench quality on the seven video-intrinsic dimensions (higher is better), computed on the shared benchmark clips after normalization (App. C.4). Best per column in bold.

Method	Mean ↑	Imaging	Aesthetic	Subject cons.	Background cons.	Temporal flick.	Dynamic degree	Motion smooth.
SANA-WM	0.835	0.668	0.537	0.908	0.919	0.942	0.900	0.969
DreamX	0.828	0.670	0.542	0.880	0.904	0.909	0.946	0.946
HY-WorldPlay	0.842	0.670	**0.601**	**0.976**	**0.957**	**0.973**	0.733	**0.987**
Matrix-Game 3.0	0.836	**0.713**	0.461	0.894	0.919	0.921	0.983	0.962
LingBot-World	0.841	0.692	0.577	0.905	0.915	0.918	0.933	0.947
Yume-1.5	0.844	0.669	0.565	0.878	0.911	0.934	**0.988**	0.964
ReWorld (ours)	**0.850**	0.665	0.579	0.929	0.929	0.952	0.912	0.979

Table 4: Revisit SSIM↑ on long rollouts. *Top:* training recipe (CD: chunk drop; RH: random head routing). *Bottom:* cache policy; full-KV is the unbounded upper bound.

	Revisit SSIM↑			
Method	k=96	k=192	k=288	k=384
Training recipe (consolidation arm)				
base	0.4019	0.3747	0.3698	0.3387
+ CD	0.4043	**0.3941**	0.3846	0.3463
+ RH	0.4319	0.3787	0.3836	0.3565
+ CD + RH (ours)	**0.4358**	0.3929	**0.3943**	**0.3752**
Inference cache policy				
full-KV	0.4463	0.4231	(OOM)	(OOM)
window	0.4129	0.3687	0.3505	0.3476
naive-merge	0.3741	0.3371	0.3467	0.3541
consolidation w/o bank	**0.4376**	0.3758	0.3743	0.3629
consolidation w/ bank	0.4358	**0.3929**	**0.3943**	**0.3752**

Table 5: Control–memory decoupling. Fusion variants use the same training protocol (App. C.5).

Method	RotErr ° ↓	Trans Err↓	Cam MC↓	Revisit SSIM↑
MRoPE only	17.66	0.204	0.394	**0.3898**
Action + MRoPE	13.21	0.107	0.371	0.3376
Routing (ours)	**12.94**	**0.093**	**0.346**	0.3752

Note: Table 5 tests whether routing decouples action-following from long-range recall, scoring the Sec. 3.2 control metrics (variants in App. C.5). Action injection improves every control metric over MRoPE only but drops revisit SSIM (0.3898 to 0.3376)—the interference motivating the window split; routing matches it on control while restoring SSIM (0.3752).

3.4 Generation Quality

We assess the seven video-intrinsic VBench (Huang et al., 2024) dimensions (no text prompt) on the controllability clips, normalized to 32 frames, 1280×704, 16 fps (App. C.4); the comparison is internally controlled but not comparable to native-length numbers. For HY-WorldPlay, this benchmark evaluates its flagship autoregressive model.

ReWorld attains the highest mean (Table 3). HY-WorldPlay leads several consistency dimensions but has lower Dynamic Degree; read these scores alongside motion and revisit results (Sec. 3.3).

3.5 Ablations

Training recipe and cache policy. In Table 4, the *upper block* varies the training recipe—chunk drop (CD), random head routing (RH), and both (ours)—all trained from scratch (720p, 20k steps); the *lower block* varies the cache policy at the same B=12, with *full-KV* as the undeployable upper bound (OOM beyond k=192; arm details in App. C.5). The far-gap lengths $k \in \{288, 384\}$, beyond any sliding window, are discriminative (Fig. 8).

4 Related Works

Streaming video generation. Autoregressive reformulations (Chen et al., 2024; Teng et al., 2025; Zhang & Agrawala, 2025; Kodaira et al., 2026) of bidirectional video diffusion (Wan et al., 2025; Kong et al., 2024; Yang et al., 2025b) restore a temporal arrow, made real-time by few-step distillation (Yin et al., 2024b;a; Wang et al., 2024; Lv et al., 2025), attention reuse (Zhao et al., 2025), self-rollout

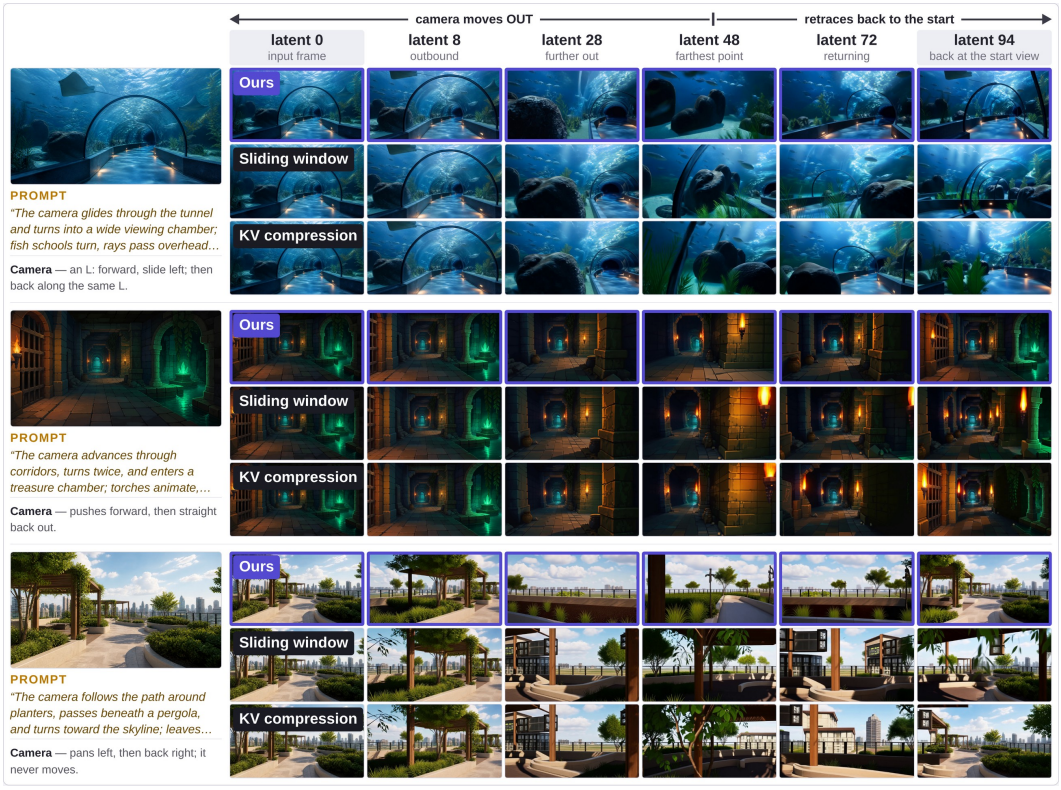


Figure 8: **Inference cache policies:** sliding window and KV compression drift or reinvent the revisited scene; the landmark cache (ours) returns to the start view.

training (Huang et al., 2025c; Liu et al., 2025), and AR-train-then-distill systems (Chen et al., 2026). ReWorld follows this recipe but confines DMD distillation to a LoRA (Hu et al., 2022), preserving the high-fidelity multi-step mode.

Interactive world models. Interactive generation spans latent simulators (Ha & Schmidhuber, 2018), latent-action pretraining (Bruce et al., 2024; Parker-Holder & Fruchter, 2025), action-conditioned diffusion (Sun et al., 2025; Wang et al., 2026; Robbyant Team et al., 2026; Zhu et al., 2026; Huang et al., 2025b), explorable scenes (Mao et al., 2025; Team et al., 2025; Cheng et al., 2026; Liu et al., 2026), and streamed deployments (Chen et al., 2025; DreamX Team, 2026). ReWorld keeps both control channels—pose folded into attention and an injected action command—but treats their interference as a training problem, splitting each head's training window.

Memory in world models. To keep long rollouts consistent (Gu et al., 2025), existing designs store growing frame archives (Yu et al., 2025; Xiao et al., 2025), a 3D reconstruction (Wu et al., 2025), or pose-dependent attention over retained history (Xu et al., 2026), while sinks and sliding windows (Xiao et al., 2024) bound the cache but forget the rest. ReWorld takes the bounded route without the forgetting: aged chunks consolidate into a fixed-capacity landmark bank in KV space—kept diverse by eviction, retrieved by pose proximity, read via chunk-drop training.

5 Conclusion

We present ReWorld, which combines window-split training and random head routing with a bounded, pose-indexed landmark cache. Metric-aligned data and LoRA distillation enable real-time generation with strong control, visual quality, and recall beyond the sliding-window horizon. Future work includes pose-free memory and richer interaction.

AI USE STATEMENT

AI assistance was used for manuscript editing and LaTeX preparation. The authors retain responsibility for ReWorld's methodology, experiments, and conclusions.

REPRODUCIBILITY STATEMENT

Appendices A–D provide architectural details, training and inference settings, memory-bank procedures, data preparation, and evaluation protocols. They specify the configurations used for ReWorld's camera-control, long-horizon memory, and video-quality experiments.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Jake Bruce, Michael 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 *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.15391.

Boyuan Chen, Diego Marti Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, 2024.

Ming Chen, Liyuan Cui, Wenyuan Zhang, Haoxian Zhang, Yan Zhou, Xiaohan Li, Songlin Tang, Jiwen Liu, Borui Liao, Hejia Chen, et al. Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation. *arXiv preprint arXiv:2508.19320*, 2025.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, Yuyang Zhao, Huizi Mao, Ying-Cong Chen, Enze Xie, Xiaojuan Qi, and Song Han. Longlive-2.0: An nvfp4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026.

Xinhua Cheng, Haiyang Zhou, Wangbo Yu, Tanghui Jia, Bin Lin, Yunyang Ge, Weiqi Li, and Li Yuan. 360explorer: Exploring 4d controllable world in panoramic videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3300–3308, 2026.

DreamX Team. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025.

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

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 (ICLR)*, 2022. arXiv:2106.09685.

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*, 2025a.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv: 2505.14357*, 2025b.

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*, 2025c.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

InternRobotics Team. OmniWorld: A multi-domain and multi-modal dataset for 4D world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

Akio Kodaira, Tingbo Hou, Ji Hou, Markos Georgopoulos, Felix Juefei-Xu, Masayoshi Tomizuka, and Yue Zhao. Streamdit: Real-time streaming text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29200–29210, 2026.

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

Zhen Li, Chuanhao Chen, Haiyang Yang, Yang Han, Bo Zhu, Zhuangzi Zhou, et al. Sekai: A video dataset towards world exploration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a. arXiv:2506.15675.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast, and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

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 (CVPR)*, 2024.

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.

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.

Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026.

Zhengyao Lv, Chenyang Si, Tianlin Pan, Zhaoxi Chen, Kwan-Yee K. Wong, Yu Qiao, and Ziwei Liu. Dual-expert consistency model for efficient and high-quality video generation. *https://arxiv.org/abs/2506.03123*, 2025.

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.

Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. Google DeepMind Blog, `https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/`, 2025. Announced August 5, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, 2023.

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

HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

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

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency model. *arXiv preprint arXiv:2405.18407*, 2024.

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

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.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

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.

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

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025b.

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*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. *arXiv preprint arXiv:2311.18828*, 2024b.

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. *ICCV*, 2025.

Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025.

Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast. In *ICLR*, 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics (SIGGRAPH)*, 37(4), 2018.

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

APPENDIX

A ARCHITECTURE AND TRAINING DETAILS

This appendix expands Sec. 2.2 with the chunked causal attention layout, the full MRoPE formulation and its comparison against two-pass pose conditioning, the action-injection construction, the per-head window mask, and training details.

A.1 CHUNKED CAUSAL ATTENTION LAYOUT

We turn the bidirectional Wan2.2-TI2V-5B backbone into a streaming generator by imposing chunk-wise causality: video is generated as a sequence of latent chunks (L=12 chunks of four latent frames per training window), with full attention within a chunk and causal attention across chunks, so at inference the model emits one chunk per denoising pass and appends its keys and values to the cache $\mathcal{C}$.

A.2 MROPE FORMULATION AND DESIGN COMPARISON

Each latent frame f carries a relative camera-to-world pose $P_f \in \mathrm{SE}(3)$ (anchored to the first frame) with descriptor $c_f = [\mathrm{vec}(R_f);\, t_f] \in \mathbb{R}^{12}$, which a zero-initialized MLP maps to a phase offset δ_f applied on top of RoPE to queries and keys within the same attention pass,

$$\tilde{q} \;=\; \mathrm{RoPE}(q)\, e^{\,i\,\delta_{f(q)}}, \qquad \tilde{k} \;=\; \mathrm{RoPE}(k)\, e^{\,i\,\delta_{f(k)}}, \qquad \langle \tilde{q}, \tilde{k} \rangle \;\propto\; e^{\,i\,(\delta_{f(q)} - \delta_{f(k)})}, \tag{2}$$

so the attention logit depends on the pose difference $\delta_{f(q)} - \delta_{f(k)}$ rather than temporal distance. Zero-initialized $\mathrm{SE}(3)$ residuals on the values and outputs ($\tilde{v} = v + W_v(P_f^{-1} \circ v)$, $\mathrm{out} = W_o y + W_p(P_f \circ y)$) complete the conditioning. Table 6 contrasts this single-pass design with two-pass pose conditioning: designs such as HY-World 1.5 add a separate camera-aware attention pass and fuse the outputs, so that pose can *steer* generation, whereas MRoPE uses pose as a *retrieval index*, folding it into the existing pass—the cost is one small MLP plus two linears, and the attention kernel, mask layout, and per-head windows stay untouched.

Table 6: Pose conditioning in attention: design comparison. Two-pass designs such as HY-World 1.5 add a camera-aware attention pass next to the temporal-RoPE pass and fuse the outputs; E-PRoPE trims the second pass to spatially reduced tokens; MRoPE folds pose into the same pass, leaving the attention kernel, mask layout, and per-head windows unchanged.

Design	**Attention index**	**Where pose enters**	**Attn. passes**	**Extra cost**
Temporal RoPE	time & space	—	1	—
HY-World 1.5 (two-pass) (Sun et al., 2025)	time + projective pose	separate camera-aware pass, fused	2	$\approx 2\times$ attention FLOPs
E-PRoPE (DreamX Team, 2026)	time + projective pose	separate reduced-token camera pass, added back	2	one reduced pass ($\sim 4.5\times$ fewer tokens)
MRoPE (ours)	time, space & pose	Q/K phase + zero-init V/O residuals	1	1 small MLP + 2 linears

A.3 ACTION INJECTION CONSTRUCTION

The commanded pose of each latent frame is expanded into a Plücker ray map: every spatial position receives the 6-D Plücker coordinates $[d,\ o \times d]$ of its viewing ray (direction d, camera center o) under that pose, expressed in the same first-frame coordinate system as the MRoPE poses. An MLP projects the ray map to the model width, and the result is added token-wise to the patch embeddings at the transformer input.

A.4 Per-Head Window Mask

With $H{=}24$ attention heads per block, the context of head h is

$$\text{context}(h) \;=\; \begin{cases} \text{full causal history}, & h \in \mathcal{G} \text{ (6 global heads)}, \\ \text{last } w{=}12 \text{ frames}, & \text{otherwise (18 local heads)}. \end{cases} \tag{3}$$

The 18 local heads attend only to the last $w{=}12$ latent frames (three chunks), and the set $\mathcal{G}$ of $|\mathcal{G}|{=}6$ global heads attends to the entire causal history—a fixed 1:3 global-to-local budget, with $\mathcal{G}$ resampled by the routing pool of Sec. 2.2.

A.5 Training Details

ReWorld is trained in two stages. Stage one trains at 480p (384×640; 24×40 latent grid) on the metric-aligned multi-source mixture (App. D). Stage two warm-starts at 720p (704×1280; 44×80 latents, 880 tokens per frame), where naively reusing 480p weights leaves cold rows in the spatial RoPE table and dilutes global-head softmax over the $3.67\times$ longer per-frame token count; we therefore rescale the spatial RoPE position axes by the ratio of the 480p to 720p post-patch grids (position interpolation on height and width only, leaving the temporal axis untouched), which preserves the spatial frequency range the model was trained on. Both stages use teacher-forcing diffusion training with timestep shift 5.0 over 1000 timesteps, classifier-free training guidance 3.0, AdamW with learning rate 10^{-5} and $(\beta_1, \beta_2) = (0.0, 0.999)$, EMA decay 0.99, and hybrid-sharded FSDP with sequence parallelism of degree 2. For real-time deployment, the trained model is further distilled into a 4-step generator with DMD-style self-forcing distillation (Yin et al., 2024b; Huang et al., 2025c) using LoRA adapters (rank 128); we defer these details to Sec. 2.4.

A.6 Distillation and Adapter Deployment

The 2.6 GB rank-128 adapter transfers across base checkpoints, with re-distillation taking 2k–4k steps when needed. Distillation uses the deployment resolution and deployment-like prompts. The production student uses block size 16, a single rollout, and classifier-free guidance folded into the adapter. Disabling the adapter recovers the high-fidelity multi-step backbone (Sec. 2.4).

B Memory System Details

This appendix expands Sec. 2.3 with the landmark-bank management rules and storage layout, and the implementation details of chunk-drop training.

B.1 Landmark Bank: Pose Parameterization, Admission, Eviction, and Tiered Storage

Pose parameterization. Each chunk carries a single camera pose $P_k \in \mathrm{SE}(3)$, taken as the first-frame rotation with block-mean translation, and the bank is managed by pose alone.

Odometer admission. A chunk aging out of the recent window is stored only if the camera has travelled at least a stride δ (measured in median step lengths) since the *last stored* landmark; the chunks in between are the densest, most redundant samples along a path, and are dropped. Crucially, the rule measures distance travelled—never whether a place is already banked—so a revisit pass is stored just like the first visit, whereas gating against the whole bank would discard exactly the chunk a later revisit needs (only $\sim$64% far-revisit coverage in simulation).

Redundancy-based eviction. Once the bank is full ($|\mathcal{M}| = K$), each admission evicts the member most redundant in pose space—the one whose nearest neighbour is closest. The first two members are never evicted, since the earliest region is the one revisited across the longest gaps.

Tiered storage and cache assembly. The bank is held in a tiered store: a pinned host-memory master copy of all K members, with only the retrieved working set staged on device. After every generated chunk, the six landmarks nearest to the current pose re-enter the cache as $\mathcal{T}_k$ of Eq. (1),

where they are read through pose-indexed attention. Device memory thus stays at $O(B)$ chunks regardless of rollout length, with the full bank in host memory; landmark transfers are prefetched on a separate stream and overlap with denoising.

Chunk-denominated budgets. The budget is counted in *chunks* (equivalently frames), not tokens: the per-frame token count scales with spatial resolution ($3.67\times$ from 480p to 720p under our latent grids), so a token-denominated budget would silently retain $3.67\times$ fewer frames at 720p and evict distant observations earlier at higher resolution. Chunk-denominated budgets keep the retained temporal horizon resolution-invariant, which we found essential for fair cross-resolution comparison.

Relation to prior memory designs. Pose- and geometry-keyed memory is well explored. Frame-archive methods store every generated frame and de-duplicate only among the subset retrieved at query time, so storage grows linearly with the rollout (Xiao et al., 2025; Yu et al., 2025); geometry-grounded methods consolidate observations into an explicit 3D structure, at the cost of an inference-time reconstruction-and-rendering stack whose resolution degrades as the scene grows (Wu et al., 2025). Our consolidation instead operates in the model's own KV space under a hard budget: chunks are admitted intact, the budget is enforced by redundancy-based eviction once the bank saturates, and members are re-attached by plain pose-proximity retrieval—no reconstruction, no auxiliary conditioning branch, and cost independent of rollout length.

B.2 Chunk-Drop Training: Masking and Parallelism

Under chunk drop, the attention mask is restricted so that the query for chunk z_k attends only to the clean keys of surviving earlier chunks and to its own chunk. Queries are never dropped, meaning that every chunk still receives its denoising loss. The keep-set is resampled independently at every optimization step and broadcast so that all data- and sequence-parallel ranks apply an identical mask.

Why chunk drop induces redundant scene storage. Because the visible history differs at every training step, the model can never rely on any particular chunk still being present when it is needed. Storing a piece of scene state—say, the layout of a room—in only one chunk therefore does not work: that chunk is dropped in roughly half of the training steps, and the loss on later chunks rises whenever it is missing, so the only way to keep the loss low under every random subset is to spread important scene information over several chunks and to reconstruct the scene from whatever gapped subset happens to survive. The compressed cache of Eq. (1) is exactly such a gapped subset, so reading from it at inference is a situation the model has seen throughout training, not something new at deployment; Sec. 3.5 ablates this component jointly with random head routing.

C Experimental Details

This appendix expands Sec. 3 with the full training and inference configuration and the operational details of the evaluation protocols.

C.1 Training and Inference Configuration

Training. The two resolution stages on top of the Wan2.2-TI2V-5B backbone (Wan et al., 2025) are a 480p (384×640) pre-training stage, followed by a 720p (704×1280) warm-start with interpolated spatial RoPE positions; optimizer, guidance, and parallelism settings follow the training details of Sec. 2.2. Each training window covers $L{=}12$ latent chunks z_k, conditioned on per-chunk actions a_k and poses P_k as defined in Sec. 2.1. Of the $H{=}24$ attention heads, the global set $\mathcal{G}$ ($|\mathcal{G}|{=}6$) attends over the full causal history while the remaining heads use a local window of $w{=}12$ frames; during training, $\mathcal{G}$ cycles through a fixed pool of 12 random six-head partitions, switching every optimizer step (random head routing, Sec. 2.2), and KV chunks are randomly dropped down to 6 kept chunks plus one sink chunk (chunk drop, Sec. 2.3).

Inference. The real-time student is a 4-step DMD-distilled (Yin et al., 2024b;a) LoRA (rank 128, block size 16, single-rollout student) applied to the EMA weights of the multi-step model, following the self-forcing style distillation recipe (Huang et al., 2025c). Inference runs at 720p with 4 denoising

steps, CFG scale 1, and 16 latent frames per block; per-head attention windows are a training-time construct and are not applied at evaluation—all heads attend over the bounded cache. Bounded-memory arms operate under a KV-cache chunk budget of $B{=}12$ on the cache $\mathcal{C}$; the landmark bank $\mathcal{M}$ is capped at $|\mathcal{M}| \leq K$ with $K{=}30$ (Sec. 2.3).

C.2 Camera-Controllability Protocol

Benchmark construction. The 40 start images are drawn from the six baselines' official repositories at 704×1280, so that no single method is favored. The six canonical trajectories are *dolly* (sustained forward), *strafe* (lateral left–right), *arc_yaw* (forward with yaw), *arc_pitch* (forward with pitch), *s_curve* (forward with alternating yaw), and *palindrome* (forward then exact return); every trajectory contains forward translation, since pure-rotation ground truth has zero displacement and degenerates the similarity alignment.

Interface adaptation and registration. Each method receives the same camera intent translated into its native control interface (action keys, pose sequences, or text, as appropriate). Generated clips are re-tracked with ViPE (Huang et al., 2025a), and the estimated trajectory is registered to the intended trajectory by a $\mathrm{Sim}(3)$ position alignment together with a global orientation alignment (restricted to $\det = +1$), which absorbs per-method conventions in step magnitude, handedness, and coordinate frame so that the comparison measures whether the trajectory *shape* follows the intent. RotErr is the geodesic rotation error in degrees, reported per trajectory and overall; TransErr and CamMC are the aggregate translation error and camera motion consistency.

C.3 NIAH Construction and Memory Metrics

Scored revisit pairs. Every method receives the same camera intent through its native control interface, with per-image matched prompts and a fixed seed. Scored revisit pairs are the mirror pairs of the palindrome, anchored in the earliest fifth of the clip and kept only when their gap spans at least half the rollout, up to five pairs per clip; the revisit gap thus reaches 47 latents at $k{=}48$ and 95 latents at $k{=}96$. At $k{=}48$ several baselines can therefore still hold the start view in context—LingBot-World keeps a 52-latent window, while SANA-WM (linear attention) and Yume-1.5 are unbounded—whereas the 95-latent gap at $k{=}96$ lies beyond every bounded window.

Longer memory test. The longer version of the memory test, run for our own arms, uses grouped-action explorations completed into palindromes at nested lengths $k \in \{96, 192, 288, 384\}$ latents; each length is a prefix of the same trajectory, so scores are comparable across k.

Metric definitions. *SSIM* measures pixel-level structure. *LPIPS* is a perceptual distance between deep features (lower is better). *DINO* is the cosine similarity of DINO ViT features and asks whether the two frames show the same scene layout. *ORB* is the fraction of matched ORB keypoints and asks whether specific landmarks reappear in place. Each score is averaged over revisit pairs, then over clips, and reported as an absolute value. *Path length* is the accumulated median optical flow of the outbound half of the rollout, i.e., how much visual ground the rollout actually covers before turning back.

C.4 VBench Protocol

The seven video-intrinsic VBench (Huang et al., 2024) dimensions requiring no text prompt are Imaging Quality, Aesthetic Quality, Subject Consistency, Background Consistency, Temporal Flickering, Dynamic Degree, and Motion Smoothness, computed on the same clip set as the controllability benchmark. Native outputs differ across methods in length, resolution, and frame rate, so all clips are normalized to a common specification before scoring: 32 uniformly sampled frames spanning the full clip, resized to 1280×704, and re-encoded at 16 fps.

C.5 Ablation Arms

Cache-policy arms. Every bounded arm in Table 4 keeps the same sink chunk and differs only in how the rest of the $B{=}12$ budget is filled: a sliding window of the most recent chunks, mean-pooled

naive-merge compression of aged chunks, and consolidation without and with the landmark bank (the deployment policy of Sec. 2.3); *full-KV* retains the entire history, is the undeployable upper bound, and runs out of memory beyond $k{=}192$.

Fusion variants. *MRoPE only* uses no action embedding and sets all $H{=}24$ heads global, leaving control to pose-indexed positional encodings alone; *Action + MRoPE* adds the additive action embedding with all heads still global—fusion without routing; *Routing* is the full mixed global/local configuration with per-step routing. All three variants are trained under the same protocol as the main model.

D Data Pipeline

ReWorld is trained on an eight-source joint corpus of 220,724 pose-annotated clips, each an RGB video with a per-frame camera trajectory and a text caption (Fig. 9). Table 7 groups the sources into UE-rendered footage (a control and an expansion set from our Unreal Engine (UE) pipeline, App. D.1), real-world footage (DL3DV (Ling et al., 2024), RealEstate10K (Zhou et al., 2018), Sekai real-walking-hq (Li et al., 2025a)), and game footage (game roaming, OmniWorld-Game (InternRobotics Team, 2025), Sekai game-walking). The UE sources anchor the metric scale for all others (per-source construction in App. D.4).

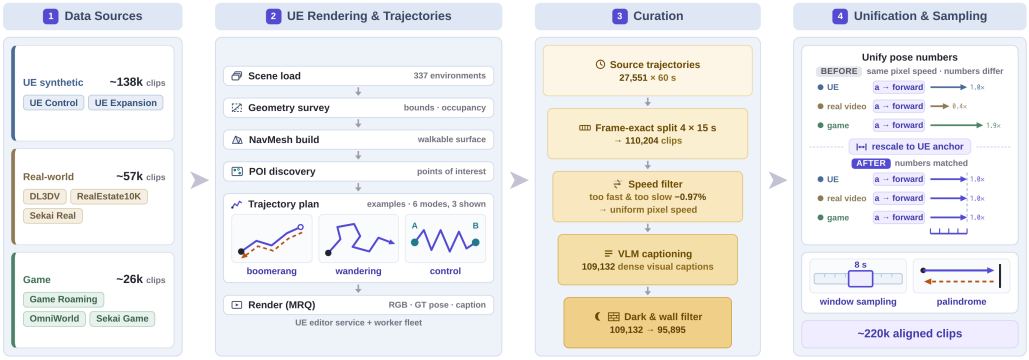


Figure 9: **The four-stage data pipeline.** Eight UE, real-world, and game sources; an automated UE service renders 337 environments with NavMesh-planned trajectories; a curation funnel filters each pool and evens out on-screen speed; unification rescales all poses to one shared metric scale before clip-level sampling.

Table 7: The eight-source joint training corpus. #Clips are post-filtering counts (App. D.3); Pose and Caption give annotation provenance.

Source	#Clips	Duration	Pose	Caption
UE control (ours)	95,895	15 s	engine GT (c2w, cm)	pipeline-generated
UE expansion (ours)	42,536	30–60 s	engine GT (c2w, cm)	scene summary + camera
DL3DV	29,864	15 s	VIPE	Qwen-VL
RealEstate10K	12,065	$\geq$ 8 s	VIPE (re-estimated)	pre-computed
Sekai real-walking-hq	14,730	60 s	MegaSaM (c2w)	official CSV
Game roaming (79 games)	18,387	15 s	VIPE (metric)	Qwen-VL
OmniWorld-Game	5,629	$\sim$12 s	engine (w2c$\rightarrow$c2w)	per-clip JSON
Sekai game-walking	1,618	60 s	engine GT (UE5, c2w)	official CSV
Total	220,724			

D.1 UE Rendering Pipeline

An automated service inside the UE editor plans and renders trajectories over 337 environments, exporting RGB frames, exact engine poses, per-frame action labels for control clips, and scene

metadata (App. D.7). Trajectories are planned for control, not just coverage: the control mode anchors each path at two points of interest and fills the segment with randomized motion, sampling backward, strafing, and diagonal moves as evenly as forward (Fig. 9). **The action-following accuracy of Sec. 3 owes as much to these direction-balanced trajectories as to the model design**.

D.2 Metric-Scale Alignment

The action $a_k \in \mathbb{R}^6$ conditioning each latent chunk is a frame-to-frame 6-DoF camera increment, so the same commanded action must mean the same motion in every source, on screen and in the pose numbers. The curation speed filter settles the on-screen half; the numbers still disagree — estimated poses (VIPE, MegaSaM) are defined only up to scale, engine units differ — so we rescale the translations of each source s by a single divisor, the ratio of pooled per-latent-step translation-increment medians against the UE anchor:

$$\sigma_s \;=\; \frac{\mathrm{med}_s[\|\Delta t\|]}{\mathrm{med}_{\mathrm{UE}}[\|\Delta t\|]}, \qquad \mathrm{med}_{\mathrm{UE}}[\|\Delta t\|] = 0.3667, \qquad t \;\leftarrow\; t/\sigma_s. \tag{4}$$

A single global scalar suffices because per-clip scale spread within each source is bounded; all sources are unified to camera-to-world matrices in a common camera basis (divisors and audits in App. D.5).

D.3 Filtering and Sampling

Each source passes the curation funnel of Fig. 9 — splitting, the speed filter, captioning, darkness/completeness checks — yielding the counts of Table 7 (per-source thresholds in App. D.4). Sources are concatenated and sampled uniformly at the clip level, so the mixture equals the clip-count proportions — 63% UE, 26% real, 11% game — independent of clip duration. Each access draws a random 189-frame window resampled to 24 fps ($\approx$ 8 s, one training window; Sec. 2.1); with probability 0.2 on the two UE sources the window becomes a palindrome, supplying revisit evidence (App. D.6).

D.4 Per-Source Construction

UE-rendered fly-throughs (metric anchor). Our Unreal Engine rendering pipeline produces camera fly-throughs with exact ground-truth trajectories: each clip directory contains the rendered video, a `camera.json` with per-frame absolute 4×4 camera-to-world matrices (right-handed, Y-up, centimeters) together with intrinsics, and a `summary.json` recording the scene and trajectory metadata (plus an overhead trajectory visualization). The full rendering run yields 130,986 raw clips over 337 purchased and in-house environments (indoor, urban, and landscape scenes), drawn from six trajectory modes planned at generation time — *local_explore* ($\sim$29%), *reveal_pan* ($\sim$26%), *aerial* ($\sim$15.5%), *control* ($\sim$15%), *wandering* ($\sim$8%), and *boomerang* ($\sim$6%) — each walking at a constant, mode-specific speed (180 cm/s for control, 120 cm/s for the exploratory modes). Because raw mode and scene frequencies are skewed, we curate a balanced subset by scene round-robin: for each mode we target $\sim$8,000 clips while cycling over environment and level identifiers, which yields 47,726 clips spanning 248–324 environments per mode; darkness filtering (App. D.3) reduces this to the 42,536-clip expansion set of Table 7, whose captions combine the pipeline scene summary with a camera-motion sentence. The 95,895-clip control set of Table 7 comes from an earlier, control-heavy rendering run (27,551 renders of 60 s, split into 110,204 clips) and is obtained from 109,132 captioned clips by the same darkness filter (removing black-frame and wall-facing clips); its trajectories are generated from the explicit discrete action space of the control mode, with per-frame action labels kept consistent with the realized motion, and its captions are generated by the rendering pipeline itself. Because UE poses are exact and metric, these sources serve as the scale anchor in App. D.2.

UE luminance and motion filtering. UE renders fail in a characteristic way: when the camera clips into geometry or faces an unlit surface, frames are near-black, so both UE sets pass through a luminance filter. Each clip is probed with 16 uniformly sampled grayscale frames and dropped if its mean luminance is below 35 (on 0–255) or if more than half of the sampled frames are dark; a lightweight motion probe during captioning separately removes frozen clips whose frames barely change. This filter reduces the control set from 109,132 captioned clips to 95,895 (-12.1%) and the round-robin-selected expansion set from 47,726 clips to 42,536 (-10.9%).

DL3DV. From DL3DV-10K real-scene walkthroughs (Ling et al., 2024) we cut frame-accurate 15 s slices whose pose sub-segments are aligned to the video slice with zero drift, yielding 29,864 clips with no missing modality. Poses are estimated with VIPE (Huang et al., 2025a) (near-metric). Each clip is captioned with Qwen-VL (Bai et al., 2025) (average 149 words per caption, no duplicates or empties).

RealEstate10K. We use RealEstate10K (Zhou et al., 2018) indoor/real-estate footage (watermark-inpainted video), retaining clips of at least 8 s, which leaves 12,065 of the original 22,275. Rather than the released SLAM trajectories, poses are re-estimated with VIPE so that pose provenance matches DL3DV; captions come from a pre-computed caption store.

Game roaming. The game-roaming source is curated from a 494.7-hour collection of gameplay recordings spanning 168 games. A roaming whitelist keeps 88 titles suited to free exploration and walking (pure shooters are excluded), which are sliced into 24,801 clips of 15 s at 1920×1080, 30 fps. Each clip receives VIPE metric camera-to-world poses (OpenCV convention, first frame anchored to the identity) and a Qwen-VL caption; successive cleaning passes remove pose-quality outliers, dark clips, and full-screen menu or loading content, leaving 18,387 clips across 79 games (nine whitelisted titles yield no surviving clips). Because the VIPE poses of this source are already metric, it enters App. D.2 with $\sigma = 1$.

OmniWorld-Game. From the simulator/game domain of OmniWorld (InternRobotics Team, 2025) (479 scenes) we normalize the released quaternion extrinsics, invert the world-to-camera matrices to camera-to-world at build time, and merge only frame-contiguous reconstruction splits into runs (splits across index gaps are independent reconstructions — largely teleports — and are never merged). Runs are cut into 12 s windows with 9 s stride (25% overlap), discarding runs shorter than 8.5 s, and each clip is re-anchored so that its first pose is the identity. This yields 5,629 clips (median 12.0 s) with per-clip JSON captions.

Sekai real-walking-hq and game-walking. From the Sekai corpus (Li et al., 2025a) of first-person walking videos we intersect available 720p/60 s videos with released pose files, obtaining 14,730 real-walking clips; clips are kept as uncut 60 s segments. Poses are per-frame MegaSaM (Li et al., 2025b) camera-to-world estimates (the c2w convention was verified empirically by forward-direction alignment, and pose–video synchronization was verified by cross-correlating optical-flow speed with pose angular speed, with median lag 0). The companion game-walking split contributes 1,618 clips rendered in UE5 with engine ground-truth c2w trajectories, downscaled to 720p without re-cutting. Both splits use the official CSV captions (caption plus location, weather, crowd, and time-of-day tags).

D.5 Alignment Details and Audits

Per-source divisors. The divisor of Eq. (4) is measured by replaying the exact training-time windowing (a 189-frame window resampled to 24 fps, i.e. 0.1640625 s per latent step, with all trajectory augmentation disabled), converting poses to per-latent translation increments Δt, and pooling the increment magnitudes over sampled clips. The resulting divisors are: UE $\sigma = 100$ (exact cm→m conversion; anchor), DL3DV $0.0709/0.3667 = 0.1923$, RealEstate10K $0.0825/0.3667 = 0.2273$, OmniWorld-Game $0.641/0.3667 = 1.75$ (pooled over the locomotion subset, since the UE anchor is locomotion-dominated and including the 29% vehicle clips would inflate the divisor), Sekai real-walking $0.0032/0.3667 = 0.0086$, and Sekai game-walking $0.0039/0.3667 = 0.0107$; the game-roaming source is already metric under VIPE and uses $\sigma = 1$. The divisor is applied to the translation column of every pose in the loader. The UE expansion set moves genuinely more slowly than the control set (its raw per-latent-step translation median is ≈ 20 cm versus ≈ 37 cm); we deliberately do not renormalize it separately, since the speed difference is real rather than a scale error.

Spread and post-hoc audits. A single global scalar per source suffices because the per-clip scale spread is bounded: the p_{90}/p_{10} ratio of per-clip median increments is 3.65 (DL3DV), 2.66 (RealEstate10K), and 2.86/2.67 (Sekai real/game) — under one order of magnitude and attributable to genuine capture-speed variation — so neither per-scene normalization nor metric-depth rescaling

is needed. A post-hoc audit of the model-facing increments after rescaling confirms alignment: the per-source median $\|\Delta t\|$ relative to UE is 1.00 (UE, by construction), 0.92 (DL3DV), 1.00 (RealEstate10K), 1.21 (OmniWorld-Game, a vehicle-tail effect), 0.96 (Sekai real), and 0.94 (Sekai game).

Coordinate-convention normalization. Coordinate conventions are normalized jointly with scale. All sources are brought to camera-to-world matrices in the OpenCV camera basis (forward $+z$, down $+y$): the UE camera basis (forward $+x$, up $+y$) is conjugated by a non-trivial 90° rotation about the vertical axis (a naive $\mathrm{diag}(1,-1,-1)$ sign flip is incorrect and was ruled out by an empirical derivation against engine trajectories); OmniWorld world-to-camera extrinsics are inverted at clip-build time; Sekai extrinsics are already c2w and are ingested without inversion; the legacy camera-referenced (w2c) RealEstate10K mode is not used in the joint configuration. The audit further verifies that forward motion maps to a dominant $+t_z$ in every source (DL3DV is t_x-dominated in aggregate, but this reflects its orbital/strafing capture style rather than a convention error) and that yaw maps to the same Euler index with consistent sign. Rotations are already dimensionless (radians) and are never rescaled; only translation passes through Eq. (4).

D.6 Sampling and Conditioning Details

Window extraction. Within a long clip the 189-frame window — the native training-window length of 48 latent frames, i.e. $L{=}12$ chunks of 4 — starts at a uniformly random frame, and frames are resampled to 24 fps by uniform striding (clips shorter than the required span are stretched by linear index interpolation). A 60 s Sekai clip therefore contributes a different random ≈ 8 s excerpt (7.88 s) on every epoch, which converts long uncut footage into diverse window-level supervision without offline re-slicing.

Palindrome augmentation. The palindrome augmentation of App. D.3 replaces the training window by a random half of the window concatenated with its temporal reverse; this forces the camera to retrace its own path within a single training window and additionally provides reverse-motion coverage. Under this augmentation, pose downsampling is SE(3)-aware (rotation slerp), since arithmetic averaging across the palindrome mirror point produces singular matrices. A whole-window reversal probability exists in the loader but is disabled in the joint configuration.

Emitted conditioning. From each window the loader emits the RGB frames, the per-chunk actions, and the pose context consumed by MRoPE. The 189 frame poses are downsampled to the 48 latent steps, re-anchored so the first camera is the identity, and differenced into per-step 6-DoF increments (three translations plus three Euler angles), which are grouped per latent chunk into the actions $a_k \in \mathbb{R}^6$. The MRoPE context is the relative camera-to-world trajectory $P_k \in \mathrm{SE}(3)$ itself, with translations additionally normalized by the window's maximum radius (a pure rescaling that preserves the identity anchor), so that pose-indexed attention operates on a bounded, source-agnostic coordinate frame.

D.7 UE Rendering Service

The UE service automates scene loading, geometry survey, navigation-mesh construction, point-of-interest discovery, trajectory planning, and Movie Render Queue rendering across 337 environments. Exports include RGB frames, pinhole intrinsics, per-frame ground-truth camera-to-world poses, and scene/trajectory metadata. The six trajectory modes, constant speeds, and resulting control and expansion sets are detailed in App. D.4.