MOTIONMEM: FUTURE TRAJECTORY FORECASTING AS DYNAMIC MEMORY FOR VIDEO GENERATION

Anonymous authors
Paper under double-blind review

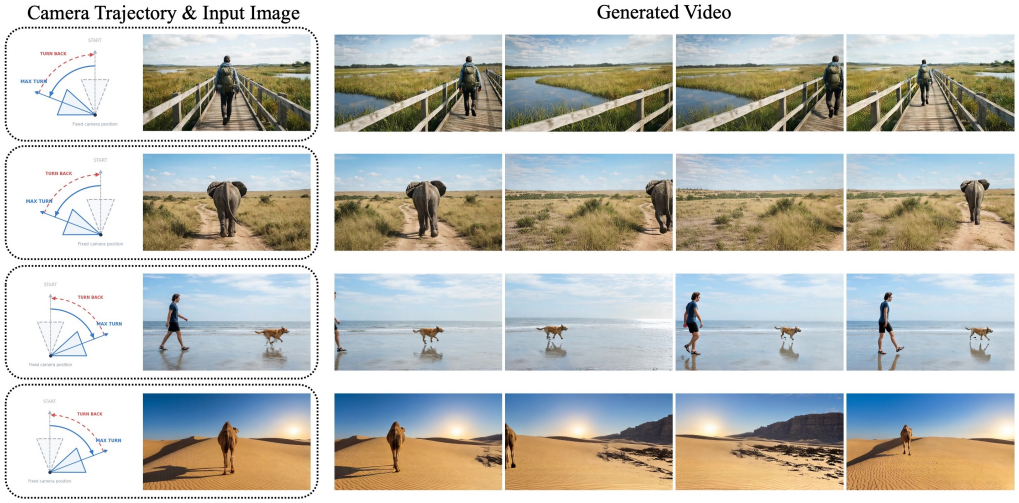

Figure 1: Examples synthesized by MotionMem. It maintains scene and object consistency over streaming generation, including cases where dynamic subjects leave and re-enter the view.

ABSTRACT

Long-horizon world generation requires models to both preserve past observations and reason about how the world evolves over time. Existing memory mechanisms are effective for recalling static scenes, but are less suited to dynamic entities whose future states must be inferred before historical information can be reused. We propose **MotionMem**, a unified *plan–retrieve–generate* framework that explicitly couples future-state prediction with memory retrieval. MotionMem jointly models RGB videos and instance-level track maps, predicting future trajectories with only a single denoising step as a lightweight motion plan, which is then used to retrieve and align relevant historical appearance for subsequent RGB generation. Trajectory-Guided Dynamic Memory uses the predicted motion to retrieve and place historical object appearance alongside camera-aligned scene memory. We further distill MotionMem into a causal few-step streaming generator for efficient interactive generation. Experiments demonstrate improved long-term consistency in both scene recall and dynamic evolution, while preserving these advantages under real-time streaming generation.

1 INTRODUCTION

Recent advances in video generation have enabled increasingly long, efficient, and controllable generation, pushing video models toward interactive world models that continuously synthesize future observations from an initial image under camera or user control (Team et al., 2026; Qian et al., 2026; Sun et al., 2025). Such models must generate successive video chunks in a causal streaming manner while maintaining a persistent world state throughout long-term exploration. A central challenge is therefore memory, which involves both static scenes and dynamic entities. When the camera leaves and later revisits a previously observed region, its geometry, appearance, and fine-grained details

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.

should remain consistent. When a dynamic object leaves the field of view and later reappears, however, the model must not only preserve its identity and appearance, but also reasonably account for how its motion and state may have evolved while unobserved. Streaming world modeling thus requires remembering what the world looked like and reasoning about how the world evolves without waiting to generate an entire sequence at once.

Existing long-horizon video and world models address memory primarily by preserving historical frames or KV caches and allowing the current generation to attend to them (Hong et al., 2025; Oshima et al., 2025; Wu et al., 2025; Yi et al., 2026; Tu et al., 2026; Meng et al., 2026). Such implicit memory establishes long-range connections through attention but leaves the model to determine which historical information corresponds to the current state. Recent methods have also introduced geometry-aware memory mechanisms, using reconstructed scene representations, or correspondence-aware latent features to retrieve and align past observations (Li et al., 2025; Wang et al., 2025; 2026b; Huang et al., 2025; Wang et al., 2026c; Yu et al., 2026b). These representations substantially improve static scene recall under camera revisits, yet dynamic memory poses a different challenge. Static content can largely be handled through camera motion, whereas an out-of-view object may continue moving or changing state before it reappears. Its historical appearance therefore cannot simply be retrieved and placed according to camera geometry; the model must first infer where the object will be before retrieving relevant history. Existing approaches to persistent dynamics either infer future states implicitly from historical features (Chen et al., 2026c), rely on restricted motion transitions (Lillemark et al., 2026), or introduce external planning or simulation systems that are less suitable for real-time applications (Duan et al., 2026; Wang et al., 2026a). This leaves a gap between explicit dynamic memory and causal streaming generation.

To bridge this gap, we propose MotionMem, a unified plan–retrieve–generate framework that explicitly couples future-state prediction with memory retrieval. MotionMem jointly models RGB videos and instance-level track maps within a single video diffusion transformer. During inference, it predicts future trajectories in a single denoising step as an explicit motion plan, which is then used to retrieve historical appearance before subsequent RGB generation. This structured process enables dynamic memory to explicitly reason about future object states before retrieving the relevant past.

Building on this motion plan, MotionMem uses camera-aligned scene memory to recall static content and Trajectory-Guided Dynamic Memory to handle moving entities. The dynamic memory retrieves each instance from the historical frame where it is most clearly observed and uses the predicted target track map to determine its future location and scale. The corresponding image patch is resized, encoded, and composited with scene memory at the target location. Together with global 2D memory, they link past observations to predicted future states for generation.

To make trajectory-guided memory practical for interactive generation, we distill the multi-step MotionMem diffusion model into a causal few-step streaming generator. We first initialize causal autoregressive generation through teacher-forcing (Zhu et al., 2026), then train with self-gradient forcing (Zhuang et al., 2026) incorporating distribution matching distillation (Yin et al., 2024). At inference, the model conditions each new chunk on its own generated history and updates memory online. Experiments show that MotionMem (see Fig. 1) improves long-term consistency in both scene revisitation and dynamic evolution while retaining these benefits in the streaming setting.

Our contributions are summarized as follows:

- We propose MotionMem, a unified plan–retrieve–generate framework that couples explicit future-state prediction with historical memory retrieval for long-term world generation.
- We introduce joint RGB–trajectory modeling, which enables a single video diffusion transformer to predict instance-level future track maps as an explicit motion plan and subsequently use them to guide RGB generation.
- We introduce Trajectory-Guided Dynamic Memory, which retrieves and places historical object appearance according to predicted trajectories to preserve dynamic entities across occlusion and out-of-view motion.
- We bring trajectory-guided dynamic memory to causal few-step streaming generation, enabling real-time world generation while preserving long-term consistency.

2 Related Work

2.1 Memory for Long-Horizon Video Generation

Long-horizon video generation requires preserving visual information beyond the local generation context. Existing approaches mainly address this challenge through *retrieval*, *compression*, or *explicit spatial memory*. Retrieval-based methods reuse historical frames or latents according to viewpoint overlap or learned similarity (Xiao et al., 2026; Yu et al., 2025; Sun et al., 2025; Nam et al., 2026), while compact-memory approaches summarize long histories into selected tokens, latent states, or KV caches to reduce computation (Hong et al., 2025; Oshima et al., 2025; Wu et al., 2025; Zhang et al., 2025; Yi et al., 2025; 2026; Tu et al., 2026; Meng et al., 2026). Learned memory mechanisms further extract or encode task-relevant history into fixed-size representations (Yu et al., 2026a; Wei et al., 2026; King et al., 2026).

A complementary line of work introduces geometry-aware memory to explicitly align past observations with the current view. These methods organize history using surfels, 3D reconstructions, depth-lifted latent features, geometric anchors, or correspondence-aware latent memories (Li et al., 2025; Wang et al., 2025; 2026b; Huang et al., 2025; Wang et al., 2026c; Zhang et al., 2026; Chen et al., 2026a; Ma et al., 2026; Yu et al., 2026b; Qian et al., 2026). Such representations substantially improve scene consistency under camera revisits. However, most existing memory mechanisms are primarily designed for *static scene recall*: geometry can determine where previously observed content should reappear, but does not specify how an unobserved dynamic object evolves before returning. In contrast, our method explicitly predicts the future trajectories of dynamic objects and uses them to retrieve and place instance-level historical appearance in the target view.

2.2 Persistent Dynamics in Video World Models

Recent work has begun to model dynamic entities that persist outside the current view. HyDRA retrieves compressed historical tokens to preserve moving subjects (Chen et al., 2026c), but future object states are still inferred implicitly through attention rather than explicit trajectories. FloWM evolves a structured latent world state through flow-equivariant transitions driven by camera and object motion (Lillemark et al., 2026); this geometric formulation is effective for predictable motion but less suited to complex non-rigid or semantic dynamics. LiveWorld reconstructs a static 3D scene and uses virtual monitors to independently evolve out-of-view dynamic entities (Duan et al., 2026), at the cost of a relatively heavy multi-stage perception, simulation, and rendering pipeline. WorldDirector uses an external LLM to plan 3D object trajectories before video generation (Wang et al., 2026a), introducing additional planning overhead that limits real-time generation.

In contrast, MotionMem predicts instance-level future trajectories within the video model itself and uses them to explicitly retrieve and place historical appearance before RGB generation.

3 Method

3.1 Overview

Long-term world generation requires preserving observed content while reasoning about how dynamic entities evolve. We propose **MotionMem** (Fig. 2), a unified *plan–retrieve–generate* framework that jointly models RGB videos and instance-level track maps with a single video diffusion transformer. At inference, MotionMem predicts future object trajectories, uses them to update explicit scene and dynamic patch memories, and generates the target video conditioned on the updated memory. Sec. 3.2 introduces our unified RGB–trajectory modeling, Sec. 3.3 presents the decoupled scene and dynamic memory, and Sec. 3.4 describes the complete plan–retrieve–generate procedure.

3.2 Joint RGB–Trajectory Modeling

We represent dynamic states using instance-level *track maps*, where each object is assigned a unique color that remains consistent across frames. RGB frames and track maps are encoded into latent representations using the video VAE and jointly modeled by a single video diffusion transformer.

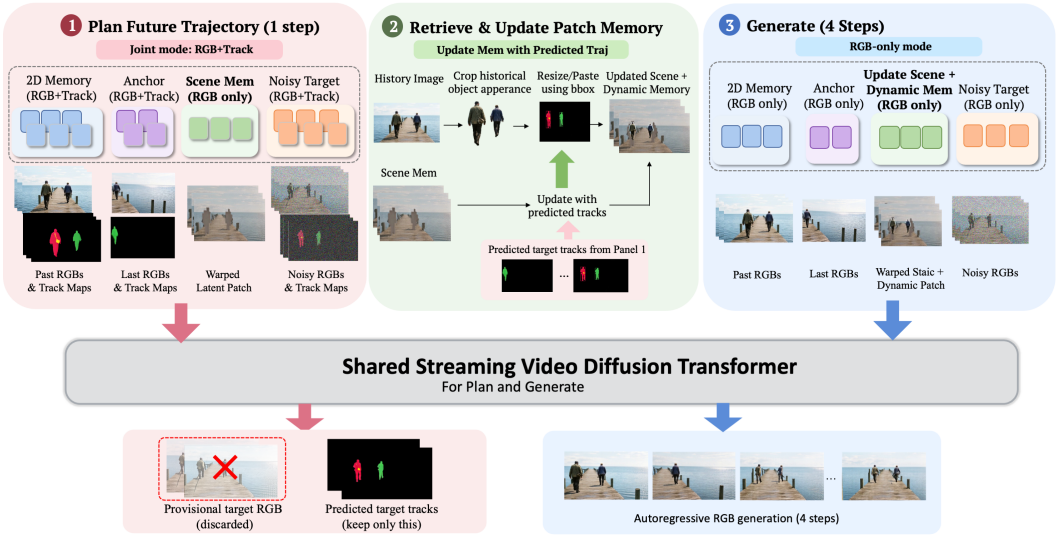

Figure 2: Overview of MotionMem's plan–retrieve–generate inference. The **single denoising step** jointly predicts provisional RGB frames and future instance trajectories, of which only the trajectories are retained as the motion plan. They guide the retrieval and placement of historical object appearance to update the scene and dynamic memory. The model then switches to RGB-only mode, using the remaining **four denoising steps** to generate the target video conditioned on the updated memory. Trajectory planning is therefore performed only once per video chunk, with the same diffusion transformer shared across planning and generation.

The transformer input consists of four components: retrieved 2D memory $\mathbf{X}_{\text{ret2D}}$, scene memory $\mathbf{X}_{\text{scene}}$, the latest observed frame as an anchor $\mathbf{X}_{\text{anchor}}$, and the target chunk $\mathbf{X}_{\text{target}}$. Here, $\mathbf{X}_{\text{scene}}$ is a patch-based memory obtained by warping static patches; its construction is detailed in Sec. 3.3.

For the retrieved 2D memory, anchor, and target, RGB and trajectory latents are paired as

$$\mathbf{X}_i = [\mathbf{X}_i^{\text{rgb}}; \mathbf{X}_i^{\text{traj}}], \qquad i \in \{\text{ret2D}, \text{anchor}, \text{target}\}. \tag{1}$$

The four components are then concatenated along the sequence dimension:

$$\mathbf{X} = [\mathbf{X}_{\text{ret2D}}; \mathbf{X}_{\text{scene}}; \mathbf{X}_{\text{anchor}}; \mathbf{X}_{\text{target}}]. \tag{2}$$

We jointly optimize RGB and trajectory prediction using flow matching. For each modality $m \in \{\text{rgb}, \text{traj}\}$, we interpolate between noise $\mathbf{x}_0^m$ and data $\mathbf{x}_1^m$ and minimize

$$\mathcal{L}_m = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[\|\mathbf{v}_\theta^m(\mathbf{x}_t, t) - (\mathbf{x}_1^m - \mathbf{x}_0^m)\|_2^2\right]. \tag{3}$$

The joint training objective is

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{traj}}\mathcal{L}_{\text{traj}}. \tag{4}$$

Positional alignment. RGB and track-map tokens corresponding to the same frame share the same temporal position. To distinguish the two modalities, we apply a fixed spatial RoPE shift to the track-map tokens. Let $\mathbf{p}_{t,x,y}$ denote the spatiotemporal position of an RGB token. The corresponding track-map token uses

$$\mathbf{p}_{t,x,y}^{\text{traj}} = \mathbf{p}_{t,x+\Delta,y}, \tag{5}$$

where Δ is a fixed spatial offset. This preserves temporal correspondence while allowing the transformer to separate visual and trajectory modalities.

3.3 Decoupled Scene and Dynamic Memory

MotionMem maintains both a global *2D memory* of historical observations and a coordinate-aligned *patch memory*. The patch memory contains scene and dynamic components because they follow different transition mechanisms: static scene content can be propagated according to camera motion, while dynamic objects additionally require predicted object motion to align their historical appearance with the target frame.

3.3.1 SCENE MEMORY

Following MosaicMem (Yu et al., 2026b), we use depth-based patch warping for static scene recall. For target view t, we select K historical frames with maximum visible-scene coverage:

$$\mathcal{H}_t = \arg \max_{\mathcal{H},\, |\mathcal{H}|=K} \text{Coverage}(\mathcal{H}, t). \tag{6}$$

Their latent patches are projected into the target view using camera poses and median patch depth:

$$P_t = \mathcal{T}_{h \to t} \left(P_h,\, \text{Median}_{\mathbf{q} \in P_h}\, d_h(\mathbf{q}) \right). \tag{7}$$

Each latent patch is assigned to the nearest latent grid and we keep the closest patch at overlaps. The resulting static scene memory $\mathbf{M}_t^s$ is provided to the transformer as $\mathbf{X}_{\text{scene}}$.

3.3.2 TRAJECTORY-GUIDED DYNAMIC MEMORY

Dynamic objects cannot be aligned by camera motion alone. Given historical and target track maps, we first match each instance by its persistent identity code and select the historical frame where the instance occupies the largest image area. Its appearance is cropped directly from the RGB frame,

$$h_i^* = \arg\max_h \left| \mathbf{T}_h^i \right|, \qquad \mathbf{P}_i^{\text{img}} = \mathcal{C} \left(\mathbf{I}_{h_i^*}, \mathbf{T}_{h_i^*}^i \right), \tag{8}$$

where $\mathcal{C}$ denotes instance-aware image cropping. The image patch is resized according to the spatial extent of the target track map, encoded by the VAE, and placed at the target instance location,

$$\mathbf{P}_{t,i} = \mathcal{E} \left(\text{Resize} \left(\mathbf{P}_i^{\text{img}}, \mathbf{T}_t^i \right) \right), \tag{9}$$

where $\mathcal{E}$ denotes VAE encoding. The resulting dynamic patches are directly composited with the scene memory according to the target track maps,

$$\mathbf{M}_t = \text{Composite} \left(\mathbf{M}_t^s, \left\{ \mathbf{P}_{t,i}, \mathbf{T}_t^i \right\}_i \right). \tag{10}$$

Thus, historical observations provide object appearance, while the target track maps determine its future location and scale. During training, dynamic patches are injected with probability 0.5 to improve robustness to imperfect trajectory prediction.

3.4 PLAN–RETRIEVE–GENERATE INFERENCE

At inference, we switch the same diffusion transformer between *joint* RGB–trajectory modeling and RGB-only generation, forming a *plan–retrieve–generate* procedure.

Plan. For the first denoising step, we enable the joint mode introduced in Sec. 3.2. The retrieved 2D memory, anchor, and target therefore contain paired RGB–trajectory tokens, while the scene memory provides camera-aligned static appearance:

$$\left(\tilde{\mathbf{I}}_{\text{target}}, \hat{\mathbf{T}}_{\text{target}} \right) = \mathcal{F}_\theta \left(\left[\mathbf{X}_{\text{ret2D}}^{\text{joint}}; \mathbf{X}_{\text{scene}}; \mathbf{X}_{\text{anchor}}^{\text{joint}}; \mathbf{X}_{\text{target}}^{\text{joint}} \right], t \right), \tag{11}$$

The model jointly predicts provisional RGB and target trajectories. We discard $\tilde{\mathbf{I}}_{\text{target}}$ and retain only $\hat{\mathbf{T}}_{\text{target}}$ as the motion plan.

Retrieve. The predicted trajectories are then used to retrieve and place historical dynamic appearance following Sec. 3.3. These dynamic patches are composited with the camera-aligned scene memory, yielding the updated patch memory $\hat{\mathbf{X}}_{\text{scene+dyn}}$.

Generate. After memory update, trajectory tokens are removed and the remaining denoising steps operate on RGB tokens only:

$$\hat{\mathbf{I}}_{\text{target}} = \mathcal{F}_\theta \left(\left[\mathbf{X}_{\text{ret2D}}^{\text{rgb}}; \hat{\mathbf{X}}_{\text{scene+dyn}}; \mathbf{X}_{\text{anchor}}^{\text{rgb}}; \mathbf{X}_{\text{target}}^{\text{rgb}} \right], t \right). \tag{12}$$

Thus, trajectory prediction is performed only once for planning, while the remaining denoising steps focus on RGB generation conditioned on the trajectory-updated memory. The generated RGB frames and predicted track maps are then stored as history for the next chunk.

Table 1: Quantitative comparison on MBench and WBench Ying et al. (2026) (higher is better). MBench is evaluated with VQA-based metrics, while WBench reports its standard dimensions. AR and BI denote streaming and bidirectional models, respectively. Best results are in bold.

	MBench (VQA)					WBench (Ying et al., 2026)					
Method	Avg.	Instruction	Identity	Background	Number	Avg.	Quality	Setting	Interact.	Consist.	Physical
Autoregressive / Streaming (AR)											
MatrixGame-3.5 [5B, AR]	74.35	**87.50**	91.25	52.31	66.32	75.56	78.65	67.72	83.39	79.71	68.32
WorldPlay [8B, AR]	79.72	78.12	88.75	84.62	67.37	78.11	78.70	65.39	87.19	87.45	**71.80**
Ours [5B, AR]	83.59	78.91	91.25	80.00	**84.21**	76.30	79.64	64.01	**88.34**	87.70	61.80
Bidirectional (BI)											
HyDRA [1.3B, BI]	17.22	42.19	1.25	13.85	11.58	54.28	61.34	55.09	58.52	73.17	23.27
LiveWorld [14B, BI]	73.37	75.00	76.25	85.38	56.84	75.99	80.30	74.22	86.61	81.86	56.98
WorldDirector [14B, BI]	86.91	86.88	90.00	90.77	80.00	76.09	75.46	74.00	85.16	81.75	64.08
Ours [5B, BI]	**87.30**	79.69	**92.50**	**93.85**	83.16	**80.67**	**80.86**	**74.31**	87.34	**89.31**	71.55

3.5 Distillation to Real-Time Generation

To enable interactive world generation, we distill the multi-step MotionMem model into a causal few-step streaming generator. We first initialize causal autoregressive generation with teacher-forcing (Zhu et al., 2026), and then adopt self-gradient forcing (Zhuang et al., 2026) with distribution matching distillation (DMD) (Yin et al., 2024), conditioning the model on its own generated history while reducing the number of denoising steps. During causal distillation and inference, we maintain a 20-latents rolling KV cache over recent video states, enabling bounded per-chunk computation.

4 Experiments

4.1 Experimental Setup

Training Data. We train MotionMem on a mixture of synthetic and real-world videos with diverse camera motion and dynamic subjects. Our training set consists of three sources: (1) 4K Mixamo-based character animation videos, including 2K single-character and 2K multi-character sequences; (2) 25K videos from the HyDRA (Chen et al., 2026c) training set; and (3) a 3P dataset containing 20K videos game footage from EgoCS-400K (Guo et al., 2026). Together, these datasets provide diverse examples of subject motion, out-of-view dynamics, and camera-induced scene changes.

Evaluation Benchmarks. We evaluate MotionMem on two complementary benchmarks. First, we use the standard WBench (Ying et al., 2026) benchmark, including both first-person (1P) and third-person (3P) settings, to evaluate general world-generation quality, camera controllability, and long-horizon consistency. Second, we construct MBench, a dedicated benchmark for dynamic memory. Each sequence starts from an image containing single or multiple subjects, while the camera rotates in place such that subjects naturally leave and later re-enter the field of view.

Baselines. We compare against two types of world models: general interactive models that use streaming generation, including WorldPlay (Sun et al., 2025) and Matrix-Game 3.5 (Qian et al., 2026), and dynamic-memory models built on bidirectional architectures, including WorldDirector (Wang et al., 2026a), LiveWorld (Duan et al., 2026), and HyDRA (Chen et al., 2026c).

Metrics. We report the standard WBench metrics across five dimensions: *Quality*, *Setting*, *Interaction*, *Consistency*, and *Physical*, together with the overall average score. For MBench, we use only the VQA-based evaluation protocol (Chen et al., 2026b) and report four dimensions: *Instruction Following*, *Identity*, *Background*, and *Number*. These assess whether the generated sequence follows the prompt, preserves subject identity and background appearance, and maintains the correct number of subjects after they leave and re-enter the view. We focus MBench on these memory-related properties; physical plausibility and interaction quality are not its primary evaluation targets and are instead covered by the standard WBench metrics.

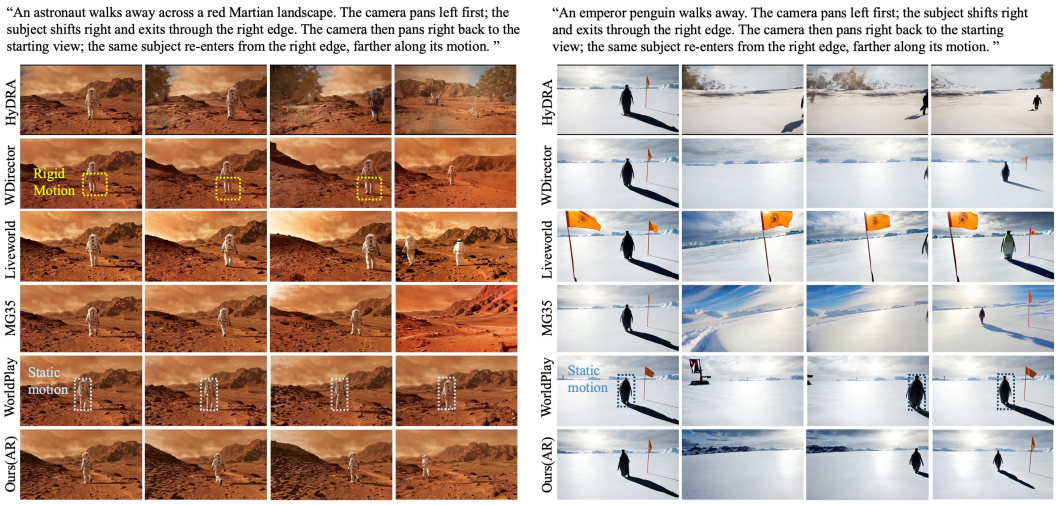

Figure 3: Qualitative comparison. HyDRA (Chen et al., 2026c) suffers from visible artifacts, WorldDirector (Wang et al., 2026a) produces rigid motion, LiveWorld (Duan et al., 2026) introduce fusion artifacts, and streaming baselines (Qian et al., 2026; Sun et al., 2025) exhibit scene drift or conservative dynamics. MotionMem better preserves dynamic subjects as they leave and re-enter.

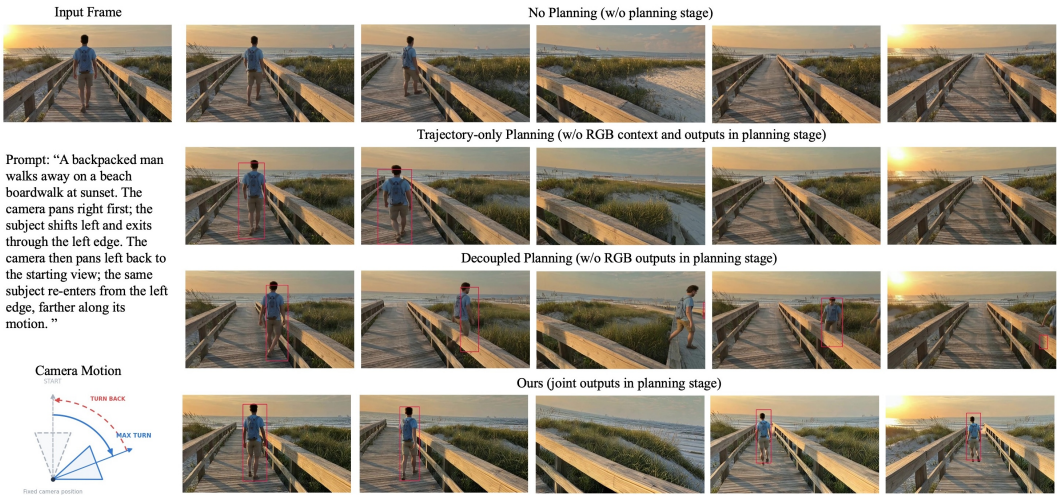

Figure 4: Ablation study on future trajectory planning. Track-only planning lacks visual cues for object orientation and motion, trajectory-only prediction provides insufficient visual constraints, and removing planning entirely weakens dynamic-state propagation across chunks. Joint RGB–trajectory planning produces more coherent motion and more reliable re-entry.

4.2 Qualitative Comparison

Fig. 3 further compares the models on the cases where subjects leave the current view and must continue evolving before reappearing. HyDRA (Chen et al., 2026c), which is designed to condition on a long historical context, degrades noticeably in our image-to-video initialization and produces visible artifacts. WorldDirector (Wang et al., 2026a) can plan object motion with an external LLM, but the resulting trajectories are often overly rigid and resemble simple translation rather than articulated motion, as illustrated by the astronaut's feet. Its external API-based planning also introduces more than 10 s of latency before video generation. LiveWorld (Duan et al., 2026) separately evolves videos under dynamic and static camera settings to preserve both scene structure and object dynamics. This multi-branch generation substantially increases inference cost, while inconsistencies between the independently generated views can introduce visible artifacts, e.g. another astronaut.

Table 2: MBench ablation using the same VQA-based dimensions as Table 1. Each variant changes one component of the 5B streaming model (AR); Ours [5B, AR] is the full model. Higher is better.

Method	Avg.	Instruction	Identity	Background	Number
No Planning	81.25	78.12	**93.75**	81.54	71.58
Trajectory-only Planning	73.47	71.88	87.50	89.23	45.26
Decoupled Planning	75.13	**84.38**	81.25	**93.85**	41.05
Latent-level DynPat Injection	80.06	78.94	87.50	81.54	72.26
Ours [5B, AR]	**83.59**	78.91	91.25	80.00	**84.21**

Table 3: WBench (Ying et al., 2026) ablation using the same dimensions as Table 1; differences are small because WBench mostly contains static (1P) or center-subject scenes (3P). Higher is better.

Method	Avg.	Quality	Setting	Interact.	Consist.	Physical
No Planning	75.80	79.26	63.32	87.73	86.73	**61.98**
Latent-level DynPat Injection	75.56	**79.67**	63.88	87.32	87.14	59.79
No DynPat Injection	75.24	79.42	63.61	87.08	86.52	59.57
Ours [5B, AR]	**76.30**	79.64	**64.01**	**88.34**	**87.70**	61.80

Streaming baselines exhibit a different failure mode. Matrix-Game 3.5 (Qian et al., 2026) shows pronounced scene drift over long horizons and can fail to maintain dynamic entities, e.g., the astronaut disappears instead of continuing its motion. WorldPlay (Sun et al., 2025), while providing stable streaming generation, tends to produce overly conservative dynamics, with subjects such as the astronaut and penguin remaining nearly static. In contrast, MotionMem plans future trajectories and uses them to retrieve dynamic appearance before RGB generation, enabling subjects to evolve while unobserved and re-enter the view with more consistent appearance, identity, and motion.

4.3 Quantitative Comparison

Table 1 compares the models on MBench and WBench. On MBench, the bidirectional MotionMem achieves the highest overall VQA score (87.30), exceeding WorldDirector by 0.39 points, and leads in identity (92.50) and background consistency (93.85). The streaming model scores 83.59 overall, above both streaming baselines, and obtains the highest number-preservation score (84.21). On WBench, bidirectional MotionMem achieves the highest overall score (80.67), while the 5B streaming model scores 76.30, slightly below the 8B WorldPlay (78.11). The gap mainly comes from the physical score, which may be affected by the smaller model size and differences in training data.

4.4 Ablation Study

Future Trajectory Planning. We compare different trajectory-planning designs in Fig. 4 and Table 2. Without RGB context, track-only planning lacks sufficient cues for object orientation and motion, leading to implausible trajectories and incorrect re-entry. Adding RGB observations improves visual grounding, but predicting trajectories separately from RGB still produces less natural motion. In contrast, joint RGB–trajectory planning yields more coherent dynamics and reliable re-entry, as shown in Fig. 4. Removing planning entirely further weakens dynamic-state propagation, causing subjects to disappear after leaving the view. Quantitatively, the full model achieves the highest MBench average (83.59) and substantially improves number preservation (84.21), supporting the importance of explicit joint planning.

Feature Injection. We further compare image- and latent-level dynamic memory injection in Fig. 5. Directly resizing retrieved latent patches introduces visible structural distortions, especially around fine-grained regions such as legs and feet. Our image-level design instead resizes the historical RGB patch before VAE encoding, better preserving appearance and geometry. This advantage is also reflected on MBench, where the full model improves the average score from 80.06 to 83.59 over latent-level injection. On WBench, the differences are smaller (Table 3), consistent with its mostly static or center-subject scenes, where dynamic memory is less frequently stressed.

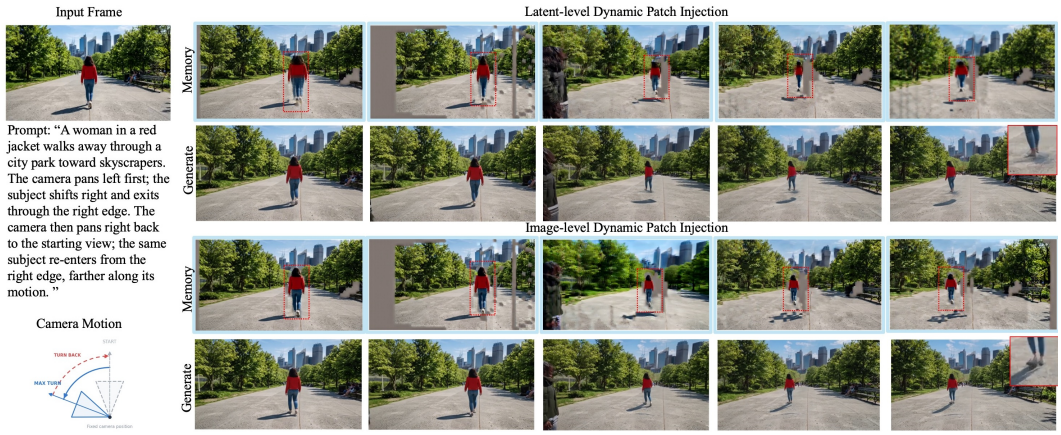

Figure 5: Ablation study on dynamic memory injection. Directly resizing latent patches introduces visible structural distortions, especially around fine-grained regions, whereas ours image-space transformation followed by VAE encoding better preserves object appearance and geometry.



Figure 6: Limitations of MotionMem. **Top:** trajectories model only spatial evolution, without an notion of time. **Bottom:** trajectories lack explicit physical constraints (e.g., a non-rolling ball).

5 Conclusion and Limitations

We presented **MotionMem**, a plan–retrieve–generate framework for long-horizon streaming world modeling with explicit dynamic memory. MotionMem jointly models RGB videos and instance-level trajectories, first forecasting future object motion and then using the predicted trajectories to retrieve and align historical appearance before RGB generation. By decoupling static scene memory from trajectory-guided dynamic memory, the model can preserve previously observed content while allowing dynamic entities to continue evolving outside the current view. We further distill MotionMem into a causal few-step generator that updates memory and produces each video chunk online from generated history. This enables real-time interaction with dynamic memory, whereas the compared dynamic-memory methods rely on bidirectional generation. Experiments demonstrate improved consistency in scene revisitation and out-of-view dynamic evolutions.

MotionMem still has two main limitations, as illustrated in Fig. 6. First, our trajectory representation primarily models *spatial motion*, but does not explicitly capture temporal evolution beyond object displacement. For example, after the camera turns away and later returns, the digital clock may retain its previous reading rather than advancing with elapsed time. Second, the predicted trajectories are not explicitly constrained by physical dynamics. For objects whose future states depend on velocity, inertia, gravity, or interactions, the model may preserve their appearance and continuity but fail to evolve them according to physically plausible dynamics, as shown by the basketball example. Future work could extend dynamic memory from spatial trajectories toward richer time-aware state representations and physics-grounded world evolution.

AI USE STATEMENT

Generative AI was used to assist with language editing of the manuscript. GPT-5.6-luna was also used as the video-question-answering judge for MBench, with the same fixed questions and scoring protocol applied to all methods, and as the object-trajectory planner when adapting the WorldDirector (Wang et al., 2026a) baseline to our evaluation setting. These model-assisted evaluation and baseline procedures are described in Sec. A.3 and Sec. A.4, respectively. The authors reviewed the manuscript and take responsibility for the study design, experiments, and conclusions.

ETHICS STATEMENT

This work studies video generation and does not involve recruited human participants or animal experiments. Training uses Mixamo-based synthetic animations, videos from the HyDRA (Chen et al., 2026c) training set, and third-person game footage. As with other generative video systems, MotionMem could be used to produce misleading synthetic content, and its outputs may reflect biases or limitations in the source data. We therefore present generated videos as synthetic research outputs and describe the data sources and evaluation setting so that these limitations can be assessed.

REPRODUCIBILITY STATEMENT

The method is specified in Secs. 3.2–3.5, including joint RGB–trajectory modeling, scene and trajectory-guided dynamic memory, plan–retrieve–generate inference, and distillation for streaming generation. The appendix documents training-data and MBench construction (Sec. A.1), model initialization and training settings (Sec. A.2), the VQA scoring protocol (Sec. A.3), and baseline adaptations (Sec. A.4). We report quantitative comparisons on MBench and WBench (Ying et al., 2026) together with ablations of trajectory planning and memory injection. Reproducing the reported results also requires access to the cited third-party datasets, pretrained models, and evaluation model.

REFERENCES

Hanlin Chen, Jiaxin Wei, Xibin Song, Yifu Wang, Steve Wang, Hongdong Li, Pan Ji, and Gim Hee Lee. Robust dreamer: Deviation-aware latent gaussian memory for action-controlled ar video generation. *arXiv preprint arXiv:2605.30855*, 2026a.

Haoyu Chen, Kaichen Zhou, Hang Hua, Kaile Zhang, Jingwen Qian, Wufei Ma, Haonan Chen, Chunjiang Liu, Yizhou Zhao, Xiaoyuan Wang, et al. Memobench: Benchmarking world modeling in dynamically changing environments. In *European Conference on Computer Vision*, pp. 658–676. Springer, 2026b.

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

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

Rongjin Guo, Dong Liang, Yuhao Liu, Fang Liu, Tianyu Huang, Gerhard P. Hancke, and Rynson W. H. Lau. Egocs-400k: An egocentric gameplay dataset for world models, 2026. Project page: https://EgoCS-400K.github.io.

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

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198*, 2025.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolo. *Zenodo*, 2023.

Wayne King, Zeyue Xue, Yuxuan Bian, Jie Huang, Haoran Li, Yaowei Li, Yaofeng Su, Yuming Li, Haoyu Wang, Shiyi Zhang, et al. Echo-memory: A controlled study of memory in action world models. *arXiv preprint arXiv:2606.09803*, 2026.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25690–25699, 2025.

Hansen Jin Lillemark, Benhao Huang, Fangneng Zhan, Yilun Du, and Thomas Anderson Keller. Flow equivariant world models: Memory for partially observed dynamic environments. *arXiv preprint arXiv:2601.01075*, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

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

Yu Meng, Xiangyang Luo, Letian Li, Wenyuan Jiang, Chen Gao, Xinlei Chen, Yong Li, and Xiao-Ping Zhang. Tethercache: Stabilizing autoregressive long-form video generation with gated recall and trusted alignment. *arXiv preprint arXiv:2606.13035*, 2026.

Jisu Nam, Yicong Hong, Chun-Hao Paul Huang, Feng Liu, JoungBin Lee, Jiyoung Kim, Siyoon Jin, Yunsung Lee, Jaeyoon Jung, Suhwan Choi, et al. Worldcam: Interactive autoregressive 3d gaming worlds with camera pose as a unifying geometric representation. *arXiv preprint arXiv:2603.16871*, 2026.

Yuta Oshima, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Worldpack: Compressed memory improves spatial consistency in video world modeling. *arXiv preprint arXiv:2512.02473*, 2025.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, et al. Matrix-game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

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.

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

Yijing Tu, Shaojin Wu, Mengqi Huang, Wenchuan Wang, Yuxin Wang, Chunxiao Liu, and Zhendong Mao. Stream-t1: Test-time scaling for streaming video generation. *arXiv preprint arXiv:2605.04461*, 2026.

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

Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Qingyan Bai, Ka Leong Cheng, Yue Yu, Yixuan Li, Yihao Meng, Zichen Liu, et al. Worlddirector: Building controllable world simulators with persistent dynamic memory. *arXiv preprint arXiv:2607.02517*, 2026a.

Jiahao Wang, Luoxin Ye, TaiMing Lu, Junfei Xiao, Jiahan Zhang, Yuxiang Guo, Xijun Liu, Rama Chellappa, Cheng Peng, Alan Yuille, et al. Evoworld: Evolving panoramic world generation with explicit 3d memory. *arXiv preprint arXiv:2510.01183*, 2025.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models. *arXiv preprint arXiv:2606.09828*, 2026b.

Zun Wang, Han Lin, Jaehong Yoon, Jaemin Cho, Yue Zhang, and Mohit Bansal. Anchorweave: World-consistent video generation with retrieved local spatial memories. *arXiv preprint arXiv:2602.14941*, 2026c.

Zhengxuan Wei, Xu Guo, Xinghui Li, Xunzhi Xiang, Min Wei, Yiran Zhu, Qiulin Wang, Xintao Wang, Pengfei Wan, Xiangwang Hou, et al. Geometry-aware implicit memory for video world models. *arXiv preprint arXiv:2606.02436*, 2026.

Xiaofei Wu, Guozhen Zhang, Zhiyong Xu, Yuan Zhou, Qinglin Lu, and Xuming He. Pack and force your memory: Long-form and consistent video generation. *arXiv preprint arXiv:2510.01784*, 2025.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Jung Yi, Minjae Kim, Paul Hyunbin Cho, Wooseok Jang, Sangdoo Yun, and Seungryong Kim. Worldkv: Efficient world memory with world retrieval and compression. *arXiv preprint arXiv:2605.22718*, 2026.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024.

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 *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Jiwen Yu, Jianxiong Gao, Jianhong Bai, Yiran Qin, Kaiyi Huang, Quande Liu, Xintao Wang, Pengfei Wan, Kun Gai, and Xihui Liu. Memlearner: Learning to query context memory for video world models. *arXiv preprint arXiv:2606.31734*, 2026a.

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, Animesh Garg, et al. Mosaicmem: Hybrid spatial memory for controllable video world models. *arXiv preprint arXiv:2603.17117*, 2026b.

Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. Tinyhistory: Lightweight video history embeddings via two-stage context learning. *arXiv preprint arXiv:2512.23851*, 2025.

Yisu Zhang, Chenjie Cao, Tengfei Wang, Xuhui Zuo, Junta Wu, Jianke Zhu, and Chunchao Guo. Worldstereo: Bridging camera-guided video generation and scene reconstruction via 3d geometric memories. *arXiv preprint arXiv:2603.02049*, 2026.

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, 2026.

Junhao Zhuang, Shiyi Zhang, Yuxuan Bian, Yaowei Li, Yawen Luo, Yijun Liu, Weiyang Jin, Songchun Zhang, Xianglong He, Xuying Zhang, et al. Self gradient forcing: Native long video extrapolation. *arXiv preprint arXiv:2607.20368*, 2026.

A APPENDIX

A.1 DATASET CONSTRUCTION

Real-World Videos. For real-world videos, we estimate camera poses and depth using DA3 (Lin et al., 2025). Dynamic instances are first detected using YOLO (Jocher et al., 2023) and subsequently segmented and tracked with SAM2 (Ravi et al., 2024). These annotations provide the camera, depth, and consistent instance masks required for constructing scene and dynamic memories.

Synthetic Videos. Apart from dataset from HyDRA (Chen et al., 2026c), we also construct synthetic training sequences to provide controlled examples of long-horizon object motion and out-of-view dynamics. We extend short character walking animations from Mixamo into longer sequences by repeating the motion cycle, and render the subjects against randomly sampled HDR backgrounds. For each sequence, we randomly sample the initial position and orientation of the subject. The camera trajectory is then designed to first keep the subject visible, subsequently move away such that the subject leaves the field of view, and optionally revisit the previously observed region. This construction provides explicit supervision for both continuous subject motion and long-term re-entry.

MBench Construction. MBench contains 115 cases covering humans, vehicles, and animals. Each sequence contains 192 frames. The camera initially observes the target subject, rotates approximately 70° away from it, and then returns to the original viewing direction. This controlled leave-and-return trajectory evaluates whether a model can maintain subject identity and appearance while producing plausible dynamics during the unobserved interval.

A.2 TRAINING DETAILS

MotionMem is built upon Wan2.2-TI2V-5B (Wan et al., 2025) and trained in three stages: the bidirectional model for 800 steps, the autoregressive model by teacher-forcing Zhu et al. (2026) for 1,000 steps, and the self-gradient forcing model for few steps model (Zhuang et al., 2026) 1,800 steps, all at a spatial resolution of 480×832. We support up to 16 instance slots, discard instances with a bounding-box area ratio below 0.01, randomly permute instance colors during training, and independently drop text conditioning with probability 0.1. All stages use AdamW with $(\beta_1, \beta_2) = (0, 0.999)$ and weight decay 0.01, with gradient norms clipped to 1.0, and adopt a 1,000-step diffusion noise schedule with timestep shift 5.0. Scene memory is constructed offline using DA3 (Lin et al., 2025) camera poses and depth estimates together with SAM2 (Ravi et al., 2024) instance masks; reprojection is resolved using a z-buffer that prioritizes the nearest surface with a filling threshold of 0.95, while unfilled regions are represented by zero-valued latents. The autoregressive model is initialized from the bidirectional checkpoint. For DMD (Yin et al., 2024), the generator is initialized from the autoregressive checkpoint, while the real and fake score networks are initialized from the bidirectional checkpoint, with the real score network kept frozen. The generator and fake score network use learning rates of 2×10^{-6} and 4×10^{-7}, respectively, with one generator update for every five fake-score updates and gradient clipping at 10.0; no adversarial loss is used. Distillation uses four base denoising step IDs, $\{1000, 750, 500, 250\}$, mapped through the shifted noise schedule, with random early exits during training and the complete four-step schedule at inference without student classifier-free guidance. We maintain an exponential moving average of the generator parameters with decay 0.99 and use the EMA model for validation.

A.3 VQA-BASED EVALUATION

We evaluate long-term dynamic memory in MBench using a VQA-based protocol with GPT-5.6-luna as the video judge. Each benchmark instance is associated with a fixed set of scene-specific yes/no questions, which are shared across all evaluated methods. The questions cover four dimensions: *Instruction Following*, *Object & Background Consistency*, *Continuity of Memory*, and *Number*. These dimensions respectively evaluate instruction compliance, visual consistency across viewpoints, state preservation after objects leave and re-enter the field of view, and consistency of object counts.

Each question is assigned either positive or negative polarity. For positive questions, a "Yes" answer indicates a pass, while for negative questions describing failure cases, a "No" answer indicates a pass. The evaluator returns a structured response containing the question ID, dimension, polarity,

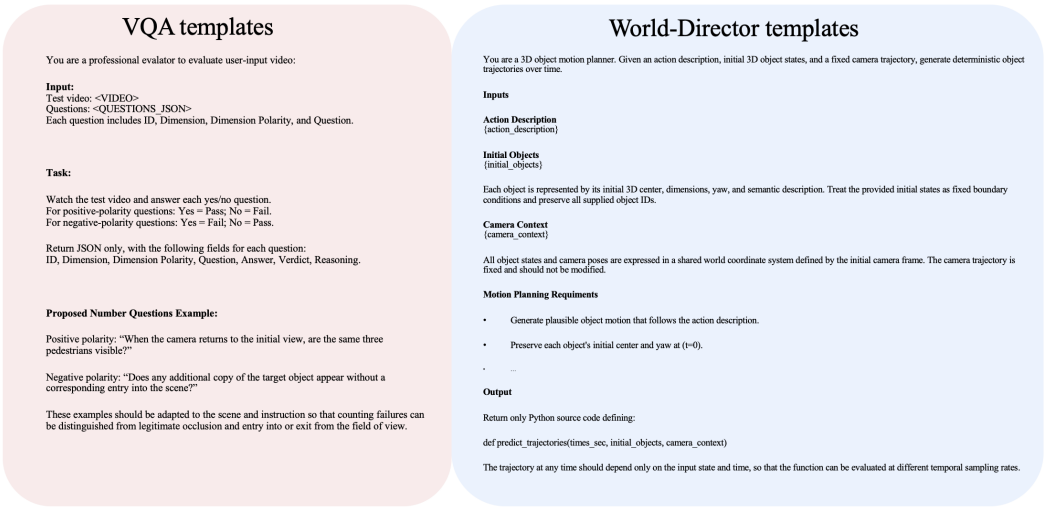

Figure 7: Persistent memory across repeated revisits.MotionMem accumulates observations from earlier exploration into historical memory, enabling later revisits to reuse previously recovered content and maintain a persistent world state across repeated exploration.

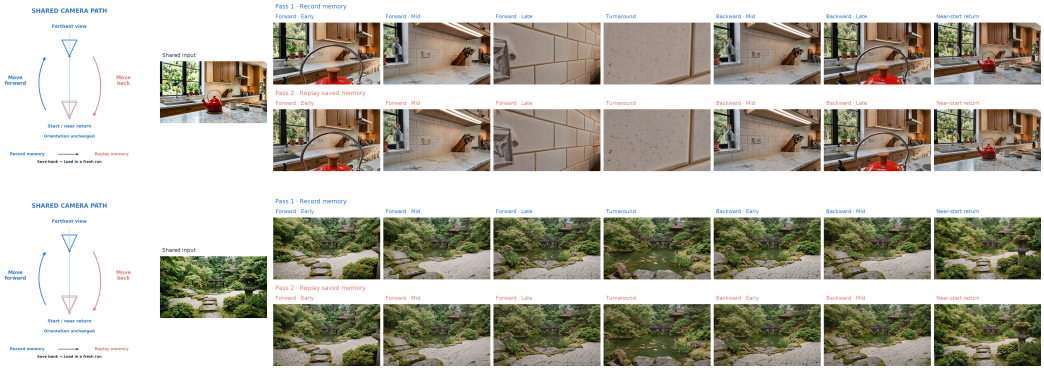

Figure 8: Persistent memory across repeated revisits.MotionMem accumulates observations from earlier exploration into historical memory, enabling later revisits to reuse previously recovered content and maintain a persistent world state across repeated exploration.

answer, verdict, and a brief justification. For each video, the score of each dimension is computed as the proportion of questions receiving a passing verdict.

For Wbench (Ying et al., 2026), We use Doubao-Seed-2.0-lite instead of Qwen3-VL-30B-A3B.

A.4 Baseline Adaptation

We adapt two baselines to the input and evaluation settings of MBench and WBench while preserving their original generation pipelines as much as possible. Unless otherwise specified, we use the official implementations and retain their original inference procedures.

WorldDirector. WorldDirector originally employs an external language model to jointly plan camera and object trajectories before video generation. Since WBench/MBench provides a predefined camera trajectory shared by all methods, we remove the camera-planning component and directly provide the ground-truth camera trajectory to the planner. The planner therefore predicts only object trajectories conditioned on the given camera motion, using the prompt shown in the right panel of Fig. 7. We use GPT-5.6 Luna as the trajectory planner and otherwise retain the original WorldDirector generation pipeline.

HyDRA. HyDRA requires a 77-frame historical context, whereas WBench/MBench provides a single input image. To adapt it to image-to-video setting, we repeat the input image to construct the required 77-frame initialization context. All subsequent generation follows the original procedure.

A.5 Additional Results and Applications

We provide additional qualitative results to further examine MotionMem beyond the settings reported in the main paper.

Additional Re-entry Results. We provide additional re-entry examples involving diverse dynamic entities, including humans, animals, and vehicles, as shown in Fig. 9. These results further demonstrate that trajectory-guided dynamic memory preserves subject identity and appearance while allowing objects to continue evolving during out-of-view intervals and re-enter the scene consistently.

Closed-Loop Camera Motion. We further evaluate MotionMem under an in-place 360° camera rotation, as illustrated in Fig. 10. The camera gradually moves away from the initial view and eventually returns to the same region, forming a closed spatial loop that challenges long-term scene memory. When the loop is closed, MotionMem retrieves previously observed scene content rather than reconstructing the revisited region solely from recent autoregressive context.

Persistent Memory across Repeated Revisits. We additionally evaluate repeated visits to the same region, as shown in Fig. 8. After each revisit, the recovered observation is incorporated into the historical memory and can be reused during subsequent revisits. This allows MotionMem to progressively accumulate and reuse previously recovered content, demonstrating persistent memory across long-term scene exploration.

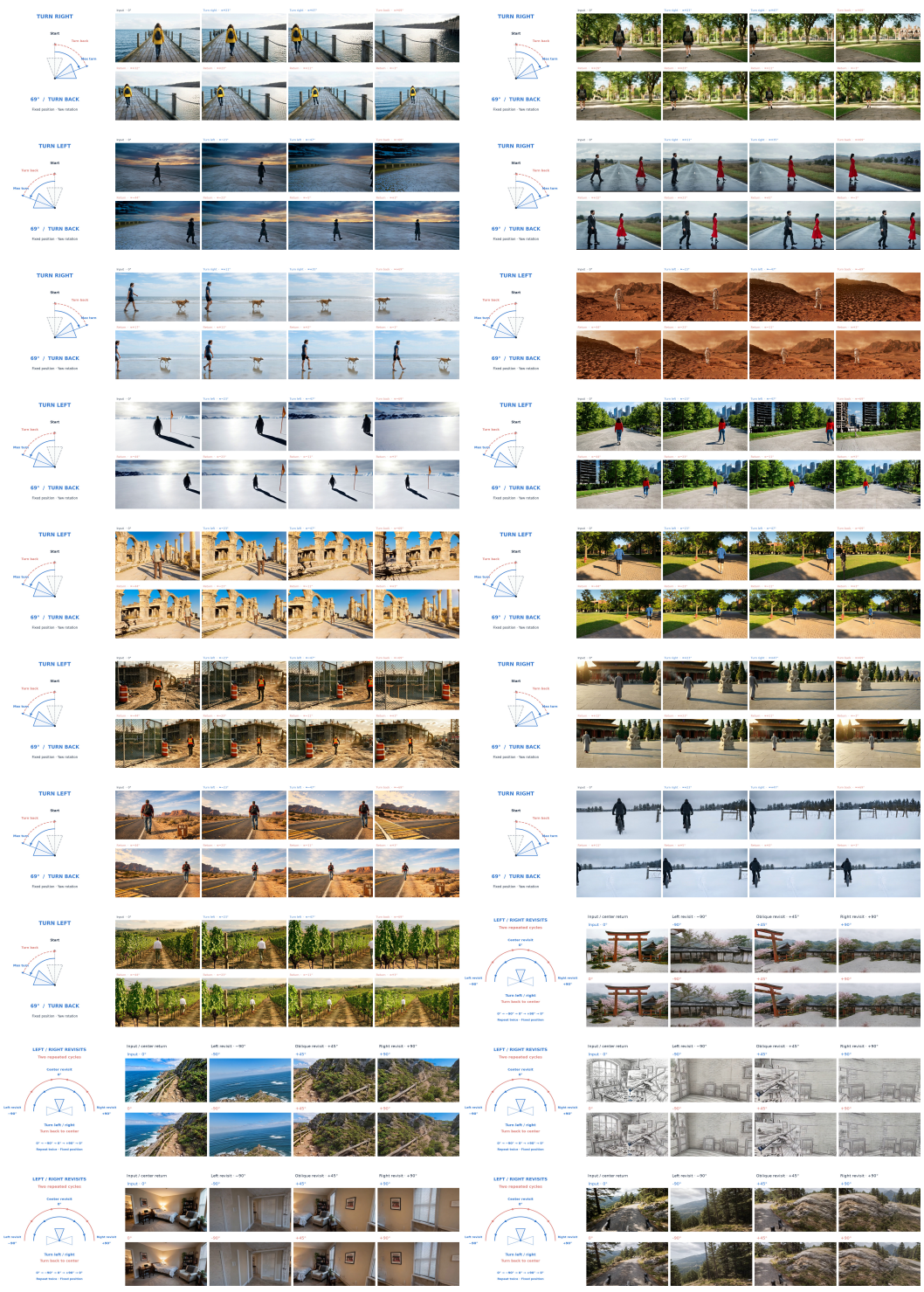

Figure 9: Dynamic objects leaving and re-entering the view. MotionMem maintains object identity and appearance while allowing subjects to continue evolving during out-of-view intervals, and recovers consistent dynamic states when they re-enter the camera view.



Figure 10: 360° closed-loop camera rotation. The camera performs a full in-place rotation and eventually returns to the initial region. MotionMem retrieves previously observed scene content when closing the loop, demonstrating long-term scene consistency under large camera motion.