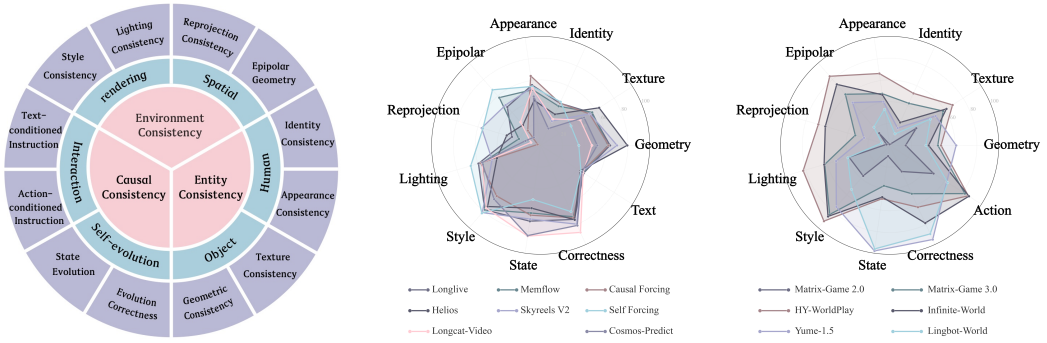
MBENCH: A COMPREHENSIVE BENCHMARK ON MEMORY CAPABILITY FOR VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in video-based world models have demonstrated an unprecedented ability to synthesize high-fidelity visual sequences. However, a fundamental gap persists between visually plausible video generation and the functional requirements of a world model, particularly in maintaining a stable and reasonable internal state over extended temporal horizons. While existing benchmarks primarily emphasize visual quality, motion coherence, and text-video alignment, they largely overlook memory, the core capability of a world model to preserve consistency across long-term horizons and complex interactions. To address this gap, we present **MBench**, a comprehensive benchmark dedicated to quantifying and evaluating the memory capability of video world models. We systematically decompose the memory capability of video world models into three hierarchical and complementary core dimensions: entity consistency, environment consistency, and causal consistency, which are further refined into 12 quantifiable sub-dimensions for comprehensive characterization of long-term memory. Our benchmark is built upon rigorously curated real-captured long videos, and evaluated by rule-based quantitative matrices and VLM to enable objective and comprehensive consistency assessment. Extensive evaluations of mainstream state-of-the-art video world models reveal critical systemic limitations of existing methods in long-term state retention, providing a standardized benchmark and clear research direction to advance the field.



(a) **MBench Dimension** (b) **Text-Conditioned Models** (c) **Action-Conditioned Models**

Figure 1: **Overview of MBench.** (a) MBench consists of a three-level hierarchical taxonomy for comprehensive evaluation. (b) We visualize the evaluation results of 8 video world models with text-conditioned interaction. (c) We visualize the evaluation results of 6 video world models with action-conditioned interaction.

1 INTRODUCTION

In recent years, video generation models (OpenAI, 2025; Yang et al., 2025b; Seedance et al., 2026; Wan et al., 2025; Google DeepMind, 2025) have achieved remarkable breakthroughs in synthesizing high-fidelity and temporally coherent realistic videos, establishing a solid foundation for the

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.

Table 1: **Comparison of Video Generation and World Model Evaluation Benchmarks.**

Method	Entity Consistency	Environment Consistency	Causal Consistency	Action Condition	Long Video Support
VBench (Huang et al., 2024)	✓	✓	×	×	×
ChronoMagicBench (Yuan et al., 2024)	×	×	×	×	×
T2V–CompBench (Sun et al., 2025)	✓	×	×	×	×
VBench++ (Huang et al., 2025b)	✓	✓	×	×	×
OpenS2V–Eval (Yuan et al., 2025)	✓	×	×	×	×
VBench 2 (Zheng et al., 2025a)	✓	✓	×	×	×
LoCoT2V–Bench (Zheng et al., 2025b)	×	×	×	×	✓
WorldModelBench (Li et al., 2025)	×	×	×	×	×
WorldScore (Duan et al., 2025)	×	✓	×	×	×
WorldBench (Upadhyay et al., 2025)	×	×	×	×	×
MIND (Ye et al., 2026)	×	✓	×	✓	×
STEVO-Bench (Ma et al., 2026)	×	×	✓	×	×
Omni-WorldBench (Wu et al., 2026a)	×	✓	×	✓	×
WorldMark (Xu et al., 2026)	×	✓	×	✓	✓
MBench (Ours)	✓	✓	✓	✓	✓

development of powerful video-based world models (Bruce et al., 2024; HunyuanWorld, 2025; Mao et al., 2025b). These models have accelerated advancements across diverse high-impact domains, including robotics (Liao et al., 2025; Unitree, 2025; Ali et al., 2025), autonomous driving (Hu et al., 2023; Wang et al., 2023b), and interactive immersive gaming (Wang et al., 2026; Ye et al., 2025). Rather than merely producing visually plausible standalone video clips, a functional world model must faithfully maintain the inherent dynamics, physical laws, and semantic rules of the real world, and reliably support long-horizon prediction, reasoning, and interaction. While recent video generation models can produce minute-long visual sequences (Huang et al., 2025a; ai et al., 2025; Team et al., 2025; Yang et al., 2025a), a fundamental gap persists between visual synthesis and world modeling, which requires long-term memory to maintain stable, consistent internal representations of world states across extended temporal horizons and interaction.

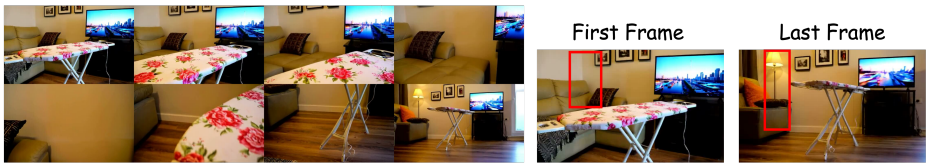
This critical gap is further exacerbated by the inadequacy of existing evaluation benchmarks for video world models. Currently, dominant evaluation protocols (Huang et al., 2024; Zheng et al., 2025a; Duan et al., 2025; Upadhyay et al., 2025) predominantly center on visual quality, motion coherence, and text-video alignment. While these metrics effectively quantify how realistic or well-aligned a generated video appears, they entirely overlook the long-term memory of world models. Unlike one-off short video generation, a functional world model must reliably retain the persistent properties of the world, where the identity and attributes of entities should not arbitrarily change, the layout and structure of the environment should remain stable across camera motions and scene dynamics, and the causal consequences of prior events should consistently govern future scene evolution. However, existing benchmarks fail to systematically challenge or measure these core memory capabilities. As shown in Table 1, most benchmarks are limited to short-horizon generation tasks, and lack structured decomposition of memory-related consistency requirements. As a result, the field currently lacks a standardized way to distinguish between models that merely generate visually appealing frames, and those that can truly simulate a predictable real world.

To this end, we present MBench, a comprehensive benchmark dedicated to systematically quantifying and evaluating the long-term memory capability of generative video world models. As illustrated in Figure 1a, our benchmark is built upon a three-level hierarchical taxonomy that decomposes world model memory into three foundational, complementary, and functionally critical dimensions:

1. Entity Consistency focuses on the model's ability to retain the persistent identity and attributes of individual participants throughout the generated sequence, which is bifurcated into Object Consistency and Human Consistency, ensuring that both inanimate items and human agents maintain their unique characteristics over time. As shown in Figure 2a, a model with robust entity memory must preserve the specific texture and geometric integrity of an object (e.g., the intricate patterns on a bowl) even during complex camera movements or occlusions.
2. Environment Consistency measures the stability of the "stage" upon which the simulation unfolds. It consists of two primary second-level modules: Spatial Consistency, which evaluates



(a) **Entity Consistency**



(b) **Environment Consistency**



(c) **Causal Consistency**

Figure 2: **Cases of different evaluation dimensions.**

the stability of 3D scene layouts and relative positioning, and Rendering Consistency, which monitors global scene properties like lighting and stylistic coherence. A typical failure in this dimension is illustrated in the center of Figure 2b, where the spatial arrangement of furniture or background elements may drift or collapse as the perspective shifts.

3. Causal Consistency assesses whether the model can uphold the logical relationships between prior states and future dynamics, ensuring the world follows established physical and semantic rules. This dimension is divided into Self-evolution, which tests the model's internal adherence to inherent physical laws, and Interaction, which measures the model's ability to remember and propagate the consequences of external prompts or actions. A typical failure in this dimension is illustrated in Figure 2c, where after an explosion, the camera passes through the resulting cloud of dust and smoke, only to reveal a perfectly intact street lined with orderly houses instead of the ruins and debris that the explosion should have caused. This exposes a fundamental breakdown in causal reasoning: the model fails to carry forward the logical consequences of a destructive event into the subsequent visual state.

This systematic taxonomy enables a multi-faceted evaluation that spans from individual object-level details to holistic scene-level layouts, while simultaneously encompassing both the persistence of static properties and the logical validity of dynamic temporal evolutions. This hierarchical framework enables MBench to provide a granular diagnostic of a world model's memory, moving beyond simple visual quality to probe the underlying logical and structural stability of the simulated world.

Our benchmark is built upon a rigorously curated dataset of real-captured long videos, filtered to select clips that pose meaningful, graded challenges to long-term temporal stability, where we strategically introduce contextual cues designed to provoke and expose latent memory failures during the generative process. To ensure a granular assessment, we employ a hybrid evaluation strategy that integrates normalized heuristic metrics with VLM-augmented Visual Question Answering (VQA). Specifically, MBench comprises over 1040 standardized evaluation cases, through which we have conducted an extensive audit of 12 state-of-the-art video world models. As illustrated in Figure 1b and 1c, our experimental results uncover prevalent systemic vulnerabilities in existing methods. Most advanced models exhibit significant temporal drift and forgetfulness as the simulation horizon extends. By providing a rigorous, multi-faceted diagnostic tool, MBench reveals that current generative paradigms struggle to maintain a persistent world state under interaction or long-range evolution. Ultimately, our work establishes a standardized evaluation protocol and delineates a clear research trajectory for the development of truly consistent and reliable video world models.

2 BENCHMARK CONSTRUCTION

In this section, we present the detailed design of MBench, a comprehensive benchmark dedicated to evaluating the long-term memory capability of video world models.

2.1 DATA COLLECTION

We aggregate data from five diverse real-world video datasets: DL3DV (Ling et al., 2024), Tanks and Temples (Knapitsch et al., 2017), OpenHumanVID (Li et al., 2024), SpatialVID (Wang et al., 2025), and Physics-aware-video (BAAI-DatacCube, 2025). These datasets collectively cover a wide range of scenarios including indoor and outdoor environments, human-object interactions, dynamic camera motions, and physical state transitions. Video durations range from 5 seconds to 15 minutes, providing sufficient temporal context for evaluating long-term memory retention.

Furthermore, we use a Vision Language Model (VLM) to systematically select videos that pose meaningful challenges to the three dimensions of memory consistency. The model is specifically prompted to evaluate the video's complexity across three core challenges: entity consistency, by identifying entities undergoing occlusion, attribute changes, or complex motion trajectories; environment consistency, by detecting significant camera motions, scene layout changes, or dynamic background elements; and causal consistency, by recognizing physical state transitions, or causal event chains.

2.2 PROMPT GENERATION

For each curated video clip, we generate a detailed description of the overall scene and main activity, which covers precise attributes, positions of entities, and chronological sequence of key events.

Multi-Granularity Caption Generation. For video continuation models, we further segment the complete structured description into five semantically coherent consecutive paragraphs along the temporal axis, and prepend designed camera control instructions to each paragraph. These instructions introduce memory-triggering camera motions including translation, rotation, zooming and dynamic occlusion, which force the model to maintain consistent world state representations under changing viewing conditions, thus effectively measuring its long-term memory capability. These multi-granularity captions also serve as the foundation for generating comprehensive ground truth question-answering pairs, ensuring our assessment covers all aspects of memory consistency.

Action Sequence Generation. For action-conditioned world models, we adopt an Exit-and-Reenter paradigm, which is specifically designed to rigorously test the model's ability to retain the complete state of entities when they are temporarily outside the camera's field of view. For a video with initial frame I_0 and a predefined target entity E, we construct a standardized action sequence $A = [a_1, a_2, \ldots, a_T]$ consisting of three consecutive phases: first, the exit phase, where the camera moves along a predefined trajectory to make the target entity E completely exit the field of view; second, the wait phase, where the camera remains stationary for a fixed duration while E stays entirely invisible; and third, the reenter phase, where the camera moves along the exact reverse trajectory to bring E back into the field of view and return to the initial camera position and orientation.

3 BENCHMARK EVALUATION

To quantify the long-term memory capability of video world models, we design a multi-dimensional, automated evaluation pipeline across three core dimensions: **entity consistency**, **environment consistency**, and **causal consistency**.

3.1 TRIGGER-CONDITIONED SCORING MECHANISM

A critical challenge in evaluating generative world models is accounting for generative stochasticity: models may generate static or overly conservative content to avoid memory challenges entirely, leading to artificially inflated consistency scores. To address this issue, we propose a *Trigger-Conditioned Scoring* mechanism that decouples performance into instruction responsiveness and memory reliability.

Our evaluation proceeds in two sequential stages:

1. **Trigger Verification:** For each sample, we first employ a Vision-Language Model (VLM) to verify whether the generated sequence successfully executes the memory-triggering event specified in the prompt (e.g., an object exiting the camera's field of view). We record a binary trigger label $y \in \{0, 1\}$.
2. **Consistency Evaluation:** For samples that successfully trigger the challenge ($y = 1$), we compute the corresponding metric. For failed samples, no consistency score is calculated.

Based on this process, we define the trigger coverage metric C_k^{trig} to measure the model's responsiveness to instructions, and the memory reliability metric S_k^{rel} to measure the average consistency on successfully triggered samples:

$$C_k^{\text{trig}} = \frac{N_k^{\text{trig}}}{N_{\text{total}}}, \quad S_k^{\text{rel}} = \frac{1}{N_k^{\text{trig}}} \sum_{i=1}^{N_k^{\text{trig}}} S_k(i) \tag{1}$$

where N_{total} is the total number of test samples, N_k^{trig} is the number of successfully triggered samples for the k-th metric, and $S_k(i)$ is the raw evaluation score. To prevent models from achieving high rankings by simply avoiding challenges, we compute the final **M-Score** as the harmonic mean of reliability and coverage:

$$\text{M-Score}_k = 2 \times \frac{S_k^{\text{rel}} \times C_k^{\text{trig}}}{S_k^{\text{rel}} + C_k^{\text{trig}}} \tag{2}$$

This formulation penalizes overly conservative models while rewarding those that maintain high consistency under complex state transitions.

3.2 Entity Consistency

Entity consistency evaluates the model's ability to maintain the persistent identity, attributes, and trajectories of individual entities throughout the generation process. We evaluate this across two granular levels:

Object Consistency: We assess whether non-human entities preserve their geometric structures and visual textures when they undergo long-term occlusions or leave and re-enter the camera's field of view (FoV). We compute a *Geometry Consistency Score* by warping corresponding frames from departure-return trajectories and calculating local structural similarity. Additionally, we compute a *Texture Consistency Score* using deep semantic features extracted from segmented object tracks to measure visual invariance. Detailed formulations are provided in Appendix C.1.

Human Consistency: Since human characters require stricter identity preservation than general objects, we design two specialized metrics. *Identity Consistency* tracks detected facial embeddings using a rolling-average bipartite matching strategy and measures embedding dispersion against a global track centroid. *Appearance Consistency* evaluates holistic attributes (clothing, hairstyle) using semantic feature alignment on full-body masks. Formal definitions are deferred to Appendix C.1.

3.3 Environment Consistency

Environment consistency measures the stability of the scene's spatial structure and overall rendering characteristics.

Spatial Consistency: We evaluate the 3D geometric constraints of the generated scene under dynamic camera motions. Utilizing estimated camera poses, we compute two complementary metrics: *Epipolar Geometry*, which measures the alignment of non-adjacent frame pairs with epipolar constraints, and *Reprojection Consistency*, which quantifies the 3D-to-2D projection error of triangulated spatial points. See Appendix C.2 for mathematical details.

Rendering Consistency: We isolate geometric structures and evaluate visual rendering stability. The *Lighting Consistency Score* measures global brightness, color temperature, and color distribution shifts in CIELAB color space across spatially paired video segments. The *Style Consistency Score* employs the Gram Matrix distance of deep visual features to assess the preservation of stylistic and artistic rendering. Detailed derivations are detailed in Appendix C.2.

3.4 Causal Consistency

Causal consistency evaluates whether the model maintains the logical causal relationships between preceding events and future scene dynamics.

Self-Evolution: This dimension assesses the model's ability to autonomously simulate physical dynamics in temporarily unseen regions. We introduce a soft-gating mechanism that decouples this capability into *State Evolution* (whether the off-screen event is initiated) and *Evolution Correctness* (whether the progression conforms to physical laws), evaluated via a VLM. The integrated score is formulated in Appendix C.3.

Interaction: We evaluate how reliably the model responds to external control signals. For text-conditioned models, the *Text Interaction Score* measures cross-modal semantic alignment using joint vision-language embeddings across sequential segments. For action-conditioned models, the *Action Interaction Score* evaluates the alignment between the estimated camera trajectories (6 DoF twists) and the ground-truth control signals. Detailed definitions are provided in Appendix C.3.

4 Experiments

4.1 Experimental Settings

To comprehensively evaluate the capabilities of different generation paradigms on MBench, we select two primary categories of state-of-the-art baselines: text-conditioned world models and action-conditioned world models. We adopt an iterative extension setting for the text-conditioned models, including Memflow (Ji et al., 2025), Self Forcing (Huang et al., 2025a), Skyreels V2 (Chen et al., 2025a), Longlive (Yang et al., 2025a), Longcat-Video (Team et al., 2025), Cosmos-Predict 2.5 (Ali et al., 2025), Causal Forcing (Zhu et al., 2026), and Helios (Yuan et al., 2026).[1] Specifically, these models are tasked with autoregressively predicting the future content for 5 consecutive segments, with each segment consisting of approximately 5 seconds. For the action-conditioned world models, consisting of Matrix-Game 2.0 (He et al., 2025), HY-WorldPlay (HunyuanWorld, 2025), Yume-1.5 (Mao et al., 2025a), Matrix-Game 3.0 (Wang et al., 2026), Lingbot-World (Team et al., 2026), and Infinite-World (Wu et al., 2026b), we evaluate their performance by generating 25-second action-conditioned continuation rollouts following each model's native control interface. For trigger-conditioned scoring, we apply Seed 2.0 (ByteDance Seed Team, 2026) verify whether each generated video enters the intended memory challenge, such as an entity leaving and reappearing, a viewpoint departing and returning, or a causal event continuing through occlusion or viewpoint transfer. This trigger is an evaluability gate rather than a judgment of whether the model has already failed. We evaluate self-evolution metrics with Qwen3-VL-235B-A22B (Bai et al., 2025). All metric scores are mapped and normalized to a range of $[0, 100]$, where higher scores represent superior consistency.

4.2 MBench Evaluation

Table 2 reports the quantitative results on the current MBench evaluation suite. Overall, the results show that memory remains a substantial bottleneck for both text-conditioned video continuation models and action-conditioned world models. Models that produce visually plausible videos often fail to maintain a persistent world state after the target entity or camera view leaves the visible region. The failure is especially clear in spatial and causal metrics, where a model must recover a previously observed viewpoint or continue an off-screen physical process rather than merely preserve local frame quality.

4.2.1 Text-conditioned video continuation models

For text-conditioned continuation models, Self Forcing (Huang et al., 2025a) and LongLive (Yang et al., 2025a) are the most competitive baselines across several memory dimensions. Helios (Yuan et al., 2026) performs best on object geometry and object texture, while Self Forcing (Huang et al.,

[1] Self Forcing and Causal Forcing are not evaluated as standalone released long-video generators. We follow the Infinity-RoPE (Yesiltepe et al., 2025) implementation and use their corresponding base video generators under the same continuation protocol.

Table 2: **MBench quantitative results.** We report the M-Score for each dimension, which verify whether the model has successfully executed the memory-triggering event specified in the prompt.

Methods	Object		Human		Spatial		Rendering		Self-evolution		Interaction	
	Geometry	Texture	Identity	Appearance	Epipolar	Reprojection	Lighting	Style	State	Correctness	Text	Action
Text-conditioned continuation												
Memflow (Ji et al., 2025)	61.72	56.06	39.48	51.60	57.95	20.89	55.06	30.08	62.75	72.04	46.31	-
Self Forcing (Huang et al., 2025a)	34.97	33.02	43.92	54.58	67.44	55.19	66.83	30.15	50.19	66.84	43.91	-
Skyreels V2 (Chen et al., 2025a)	70.03	53.70	24.57	53.76	49.01	56.39	46.44	20.33	68.35	79.52	44.68	-
Longlive (Yang et al., 2025a)	63.57	55.41	42.51	55.89	46.68	27.51	59.44	25.26	70.32	74.69	46.97	-
Longcat-Video (Team et al., 2025)	46.96	43.13	26.56	52.98	28.28	9.26	56.10	27.46	84.17	87.83	46.25	-
Cosmos-Predict 2.5 (Ali et al., 2025)	51.90	47.31	16.95	45.42	9.73	14.68	55.95	22.66	83.67	80.81	45.08	-
Causal Forcing (Zhu et al., 2026)	62.23	53.36	42.53	64.37	18.10	2.88	57.44	27.48	64.79	73.10	44.90	-
Helios (Yuan et al., 2026)	79.43	63.70	31.33	41.64	24.79	32.46	41.79	25.26	58.27	75.08	43.17	-
Action-conditioned rollouts												
Matrix-Game 2.0 (He et al., 2025)	14.62	28.99	1.22	0.94	14.78	3.08	38.79	73.78	10.00	26.40	-	47.86
Matrix-Game 3.0 (Wang et al., 2026)	44.15	58.22	42.38	47.91	61.99	32.86	62.06	95.17	37.50	48.80	-	81.93
HY-WorldPlay (HunyuanWorld, 2025)	47.12	68.54	52.46	66.58	83.86	68.17	82.67	98.23	49.50	62.40	-	85.69
Yume-1.5 (Mao et al., 2025a)	60.96	49.99	17.41	40.57	51.86	24.55	51.21	92.05	97.90	95.00	-	62.20
Lingbot-World (Team et al., 2026)	33.20	44.54	11.57	33.53	22.12	7.57	40.06	85.87	96.00	89.40	-	63.32
Infinite-World (Wu et al., 2026b)	35.70	61.88	23.08	46.85	74.04	61.51	62.63	96.87	48.00	78.40	-	86.37

2025a) excels on human identity and most environment metrics, suggesting that training with self-generated histories improves robustness to long-horizon drift. Longlive (Yang et al., 2025a) achieves the best prompt-interaction score, and Causal Forcing (Zhu et al., 2026) achieves the strongest human appearance consistency, reflecting its advantage in maintaining semantic content over multiple generated segments. MemFlow (Ji et al., 2025) is also competitive on object texture and spatial epipolar consistency, which is consistent with the motivation of explicitly retrieving or reusing historical visual information.

Despite these relative differences, the absolute style scores remain low for all continuation models. Even the strongest models show much weaker scores on style consistency than on epipolar or lighting. This indicates that current long video generators can often keep a stable visual tone while still forgetting the underlying 3D layout. The causal prompt-interaction scores are relatively close across models, with LongLive (Yang et al., 2025a), Cosmos-Predict 2.5 (Ali et al., 2025), MemFlow (Ji et al., 2025), and Longcat-Video (Team et al., 2025) clustered near the top. However, such prompt alignment does not imply physical memory: causal scores still expose large gaps in whether models continue hidden events with plausible state evolution. These results support the central design of MBench: long-term memory must be evaluated through entity, environment, and causal consistency together, rather than through visual quality or instruction following alone.

4.2.2 Action-conditioned video world models

Among action-conditioned world models, HY-WorldPlay (HunyuanWorld, 2025) obtains the strongest overall performance on entity and rendering consistency. It achieves the highest human identity and appearance consistency scores, and also leads on object texture, lighting, and style consistency. This suggests that its rollout mechanism is comparatively effective at preserving both foreground entities and global rendering statistics when the camera exits and later returns to a scene. Matrix-Game 3.0 (Wang et al., 2026) also performs competitively on rendering style, closely following HY-WorldPlay and Infinite-World.

The spatial metrics expose a sharper gap. HY-WorldPlay (HunyuanWorld, 2025) has the highest epipolar and reprojection scores among the evaluated action-conditioned models, followed by Infinite-World (Wu et al., 2026b) and Matrix-Game 3.0(Wang et al., 2026). In contrast, Matrix-Game 2.0 (He et al., 2025) and Lingbot-World (Team et al., 2026) show substantially larger geometric errors, implying that the scene does not return to the same 3D configuration even when the camera motion semantically suggests a departure-return trajectory. For interface following, Infinite-World (Wu et al., 2026b) and HY-WorldPlay (HunyuanWorld, 2025) are strongest on action-motion alignment.

However, a deeper analysis of the causal metrics reveals a critical trade-off. Although HY-WorldPlay (HunyuanWorld, 2025) and Infinite-World (Wu et al., 2026b) excel at environment ren-

Table 3: **Metric-human correlation on pairwise preferences.** Each row is one of the twelve MBench consistency metrics (Section 3). We aggregate three-annotator majority votes on pairwise tasks, convert human preferences and automatic scores into model-level win ratios over eight text-conditioned continuation models or six action-conditioned world models, and report Spearman's ρ and Kendall's τ. Entries marked "-" indicate no valid automatic scores for that setting.

			Text-conditioned continuation		Action-conditioned rollouts	
Memory consistency	Sub-dimension	Metric	ρ	τ	ρ	τ
Entity	Object	Geometry Consistency	+0.97	+0.87	+0.97	+0.87
		Texture Consistency	+0.89	+0.73	+0.93	+0.80
	Human	Identity Consistency	+0.86	+0.71	+0.71	+0.60
		Appearance Consistency	+0.69	+0.50	+0.81	+0.67
Environment	Spatial	Epipolar Geometry	+0.16	+0.11	+0.94	+0.87
		Reprojection Consistency	+0.45	+0.36	+0.94	+0.87
	Rendering	Lighting Consistency	-0.49	-0.36	+0.71	+0.60
		Style Consistency	+0.56	+0.46	+0.71	+0.47
Causal	Self-evolution	State Progress	+0.83	+0.71	+0.54	+0.47
		Physical Plausibility	+0.81	+0.64	+0.43	+0.20
	Interaction	Text Interaction	+0.19	+0.11	–	–
		Action Interaction	–	–	+0.71	+0.60

dering and spatial consistency, their performance on self-evolution drops significantly. Both models achieve remarkably low state evolution scores compared to Yume-1.5 (Mao et al., 2025a) and Lingbot-World (Team et al., 2026). This contrast highlights a common failure mode in current action-conditioned models: the tendency to generate overly static scenes. By minimizing scene dynamics and object movements, HY-WorldPlay (HunyuanWorld, 2025) and Infinite-World (Wu et al., 2026b) successfully preserve 3D geometry and global lighting, yielding high spatial and rendering scores, but they fundamentally fail to actuate the intended physical processes or simulate meaningful state evolution. This discrepancy underscores the necessity of MBench's multi-dimensional evaluation: spatial memory and causal self-evolution must be assessed jointly to penalize models that merely "freeze" the scene to avoid temporal drift.

4.3 Human Alignment

In this section, we conduct a human study on the MBench annotation platform and compare automatic scores with human pairwise preferences over memory consistency. This main text focuses on the core correlation result, while more detailed trigger calibration and ablation analyses are deferred to the appendix.

4.3.1 Human annotation protocol

In total, the current annotation set contains 2,362 records from 22 unique annotators, including 400 binary judgments on text-conditioned videos, 1,020 pairwise judgments on text-conditioned video rollouts, and 942 pairwise judgments on action-conditioned video rollouts. The annotation pool covers all 14 video world models and all 12 dimensions. The binary interface asks whether a single generated continuation video contains a memory challenge, such as transition, exit-reentry, occlusion, or departure-return camera motion that requires the model to remember temporarily invisible content. The pairwise interface then asks annotators to compare two model outputs on dimension-specific memory consistency questions.

4.3.2 Metric-human correlation

Table 3 reports Spearman's ρ and Kendall's τ after converting both human votes and automatic scores into model-level win ratios. Entity consistency metrics show the strongest agreement on text-conditioned continuation, especially geometry consistency ($\rho = 0.97$). Rendering metrics are also aligned on action-conditioned rollouts ($\rho \approx 0.71$ for lighting and style), while spatial epipolar geometry is much more predictive for action-conditioned models ($\rho = 0.94$) than for continuation models ($\rho = 0.16$). Regarding the causal dimensions, self-evolution metrics show strong human correla-

tion for text-conditioned models ($\rho > 0.80$), but relatively lower alignment for action-conditioned ones ($\rho \approx 0.50$), since action-driven models currently struggle to generate complex causal state changes. Finally, interaction correlations remain setting-specific, with action interaction showing strong alignment ($\rho = 0.71$) while text interaction correlates poorly ($\rho = 0.19$), further validating that explicit control interfaces provide a more reliable anchor for human-metric agreement in memory evaluation.

5 DISCUSSION

Our experiments suggest that current long-video generation systems are still far from reliable video world models. Although recent models can synthesize visually plausible and temporally extended clips, their apparent coherence often breaks when the evaluation requires a persistent world state. Entity appearance, background layout, and causal states may remain locally convincing, yet drift after occlusion, camera motion, or long-horizon continuation.

Spatial understanding remains limited. A recurring bottleneck is spatial reasoning. On the generation side, many models struggle to follow camera-motion instructions specified in the prompt: the camera may move in the wrong direction, stop too early, or produce a visually plausible motion that does not match the requested trajectory. On the evaluation side, current vision models also have limited ability to infer camera motion from a sparse set of extracted frames. This makes spatial memory a particularly difficult axis: a model must both execute a controlled viewpoint change and recover the same 3D configuration when the view returns, while the evaluator must distinguish true geometric consistency from superficial visual similarity.

Action-conditioned models do not yet provide stable control. For action-conditioned world models, the response to actions is uneven across methods. Some models can produce approximate motion under common actions, but others respond with a clear delay, ignore part of the control signal, or collapse to frequent motion patterns seen during training. In these cases, the generated video may look reasonable as a generic rollout while failing as an interactive world simulation. This suggests that action conditioning should be evaluated not only by visual quality, but also by whether the model can execute novel or composed actions faithfully over time.

Causal memory is the central unresolved challenge. The causal dimension exposes an even deeper limitation. For many action-conditioned models, the generated scene remains mostly static: objects and agents may preserve their appearance, but the world itself does not evolve in a meaningful way. Such videos can obtain acceptable scores on appearance or rendering consistency while failing to simulate dynamic state transitions. This limitation becomes more severe when the relevant evolution occurs out-of-view. A reliable world model should continue hidden physical processes, remember their intermediate consequences, and reveal a plausible later state when the camera returns. Current models often lack this capability: hidden events may be ignored, reset, or replaced by visually plausible but causally unrelated content. This is why MBench separates trigger coverage from memory reliability. A model should not be rewarded for avoiding the event, and it should also not be rewarded for triggering an event without preserving its causal consequences. Future video world models therefore need stronger mechanisms for state tracking, action grounding, and latent causal simulation, rather than only better frame-level synthesis.

6 CONCLUSION

We introduced MBench, a benchmark for evaluating the memory capability of video world models across entity, environment, and causal consistency. Unlike standard video-generation evaluation, MBench explicitly tests whether a model can retain a persistent world state under long-horizon continuation, camera motion, occlusion, action conditioning, and hidden causal evolution. Through evaluations on both text-conditioned long-video generators and action-conditioned world models, we find that current systems still struggle to preserve stable entities, recover consistent spatial layouts, follow controlled actions, and simulate causal state changes that happen outside the visible field of view. These results suggest that memory should be treated as a key capability for future world models, rather than as a by-product of visual fidelity or prompt alignment. We hope MBench provides a concrete diagnostic tool and a common evaluation protocol for building video world models that are not only visually plausible, but also persistent, controllable, and causally coherent.

REFERENCES

Sand. ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. Magi-1: Autoregressive video generation at scale, 2025. URL https://arxiv.org/abs/2505.13211.

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai, 2025. URL https://arxiv.org/abs/2511.00062.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

BAAI-DatacCube. Physics-aware-videos. https://huggingface.co/datasets/BAAI-DataCube/Physics-aware-videos, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

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

ByteDance Seed Team. Seed2.0 model card: Towards intelligence frontier for real-world complexity. 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. Skyreels-v2: Infinite-length film generative model, 2025a. URL https://arxiv.org/abs/2504.13074.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. *arXiv preprint arXiv:2505.21996*, 2025b.

Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17702–17711, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024.

Google DeepMind. Veo-3 technical report. Technical report, Google DeepMind, May 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

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

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, et al. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18858–18868, 2025.

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

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2568–2577, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022b.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving, 2023. URL https://arxiv.org/abs/2309.17080.

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*, 2025a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b. doi: 10.1109/TPAMI.2025.3633890.

Team HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. *arXiv preprint*, 2025.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

Pengliang Ji, Chuyang Xiao, Huilin Tai, and Mingxiao Huo. T2vbench: Benchmarking temporal dynamics for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5325–5335, 2024.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *Advances in Neural Information Processing Systems*, 37: 89834–89868, 2024.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

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.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models. 2025.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. *arXiv preprint arXiv:2412.00115*, 2024.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng Yan, Maoqing Yao, and Guanghui Ren. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

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.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22139–22149, 2024a.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024b.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36:62352–62387, 2023.

Ziqi Ma, Mengzhan Liufu, and Georgia Gkioxari. Out of sight, out of mind? evaluating state evolution in video world models, 2026. URL https://arxiv.org/abs/2603.13215.

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*, 2025a.

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

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *arXiv preprint arXiv:2209.00588*, 2022.

OpenAI. Sora2. https://openai.com/zh-Hans-CN/index/sora-2/, 2025. URL https://openai.com/zh-Hans-CN/index/sora-2/.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing Video Generation for World Complexity, April 2026. URL http://arxiv.org/abs/2604.14148.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Skywork AI Matrix-Game Team. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. Technical report, 2026. URL https://github.com/SkyworkAI/Matrix-Game/blob/main/Matrix-Game-3/assets/pdf/report.pdf.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8406–8416, 2025.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, and Tong Zhang. Longcat-video technical report, 2025. URL https://arxiv.org/abs/2510.22200.

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.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1526–1535, 2018.

Unitree. UniFOLM-WMA-0: A world-model-action (wma) framework under UniFOLM family, 2025.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Rishi Upadhyay, Howard Zhang, Zhirong Lu, Lakshman Sundaram, Ayush Agrawal, Yilin Wu, Yunhao Ba, Alex Wong, Celso M de Melo, and Achuta Kadambi. How close are world models to the physical world? *arXiv preprint arXiv:2501.xxxxx*, 2025.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023a.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023b.

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

Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021.

Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, et al. Omni-worldbench: Towards a comprehensive interaction-centric evaluation for world models. *arXiv preprint arXiv:2603.22212*, 2026a.

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

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.

Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Kaipeng Zhang, and Yongtao Ge. Worldmark: A unified benchmark suite for interactive video world models. *arXiv preprint arXiv:2604.21686*, 2026.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

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

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, and Song Hanand Yukang Chen. Longlive: Real-time interactive long video generation. 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer, March 2025b. URL http://arxiv.org/abs/2408.06072. arXiv:2408.06072 [cs].

Deheng Ye, Fangyun Zhou, Jiacheng Lv, Jianqi Ma, Jun Zhang, Junyan Lv, Junyou Li, Minwen Deng, Mingyu Yang, Qiang Fu, Wei Yang, Wenkai Lv, Yangbin Yu, Yewen Wang, Yonghang Guan, Zhihao Hu, Zhongbin Fang, and Zhongqian Sun. Yan: Foundational interactive video generation, 2025. URL https://arxiv.org/abs/2508.08601.

Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. Mind: Benchmarking memory consistency and action control in world models, 2026. URL https://arxiv.org/abs/2602.08025.

Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. *arXiv preprint arXiv:2511.20649*, 2025.

Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1309–1320, 2023.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023.

Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *Advances in Neural Information Processing Systems*, 37:21236–21270, 2024.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*, 2025.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv e-prints*, pp. arXiv–2504, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025a.

Xiangqing Zheng, Chengyue Wu, Kehai Chen, and Min Zhang. Locot2v-bench: Benchmarking long-form and complex text-to-video generation. *arXiv preprint arXiv:2510.26412*, 2025b.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

Tianrui Zhu, Shiyi Zhang, Zhirui Sun, Jingqi Tian, and Yansong Tang. Memorize-and-generate: Towards long-term consistency in real-time video generation. *arXiv preprint arXiv:2512.18741*, 2025.

APPENDIX

In this supplementary material, we provide further details on the MBench benchmark construction and evaluation protocol. We clarify the AI assistence disclosure in Section A. We review preivous works on video generation and world models in Section B. Then, we report the detialed metric formulations of 12 dimensions in Section C. We describe the dataset composition for both evaluation settings in Section D. Section E elaborates on the evaluation and trigger mechanisms, including a comparison of the differences in computational strategies between Text-Conditioned Video Generation Models (MBench-T) and Action-Conditioned Video World Models (MBench-A). Section F details the human annotation protocol and validates the VLM trigger judge against human binary labels. Additional qualitative visualizations and case studies are provided in Section G.

A AI ASSISTANCE DISCLOSURE

In this work, we used generative AI tools for assisting with translation, implementing partial evaluation code, reformatting dataset metadata, supporting qualitative analysis of model failure cases, and interpreting experimental results. We have not used generative AI tools for generating synthetic datasets, developing core theoretical or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proofs, assisting with proof writing, nor proposing or refining the main research hypothesis. Formulating mathematical proofs is not applicable to this work.

Additionally, we used generative AI tools for drafting partial manuscript paragraphs, improving manuscript readability, and suggesting figure layout options.

We have reviewed all AI-assisted work. All AI-generated code modules were independently inspected, debugged, and experimentally validated by two authors to guarantee functional correctness. AI-translated and AI-drafted text content was manually revised for technical accuracy, logical consistency, and academic tone. Qualitative observations suggested by generative AI were cross-checked against raw experimental outputs and human annotation statistics before being included in the manuscript. We take responsibility for the final content of this work, including all text, claims, metrics, and artifacts produced with the aid of generative AI.

B RELATED WORKS

B.1 VIDEO GENERATION AND WORLD MODELS

Video Generation Models. Early video generation built on GANs (Vondrick et al., 2016; Tulyakov et al., 2018; Clark et al., 2019) and VAE-based autoregressive transformers (Yan et al., 2021; Villegas et al., 2022; Yu et al., 2023), producing short, low-resolution clips. Video Diffusion Models (Ho et al., 2022b) extended image diffusion to the temporal axis, and subsequent works scaled this paradigm through cascaded super-resolution (Ho et al., 2022a; Singer et al., 2022), latent diffusion (Blattmann et al., 2023; Guo et al., 2023; Chen et al., 2023), and the diffusion transformer (Peebles & Xie, 2023). Sora (Liu et al., 2024b) pioneered large-scale DiT training and demonstrated minute-level photorealistic synthesis, catalyzing a wave of foundation video models including CogVideoX (Yang et al., 2025b), HunyuanVideo (Kong et al., 2024), Wan (Wan et al., 2025), Veo-3 (Google DeepMind, 2025), Seedance 2.0 (Seedance et al., 2026), and MAGI-1 (ai et al., 2025). To break the short-clip ceiling, recent work couples diffusion with autoregressive rollout for streaming generation, exemplified by Self Forcing (Huang et al., 2025a), LongLive (Yang et al., 2025a), LongCat-Video (Team et al., 2025), and SkyReels-V2 (Chen et al., 2025a). Despite this progress, mainstream video generators are optimized for perceptual quality and prompt alignment, and produce standalone clips rather than coherent world dynamics.

World Models. A world model learns environment dynamics and predicts future states under given conditions, originating in reinforcement learning (Ha & Schmidhuber, 2018) and extended through recurrent (Hafner et al., 2023), transformer (Micheli et al., 2022), and diffusion-based (Alonso et al., 2024) dynamics. Large-scale video pretraining has since lifted this paradigm to open-domain settings, yielding three application-driven lines: interactive game-style worlds with discrete actions (Bruce et al., 2024; He et al., 2025; Skywork AI Matrix-Game Team, 2026; HunyuanWorld, 2025; Mao et al., 2025a; Ye et al., 2025; Yuan et al., 2026; Team et al., 2026; Wu et al., 2026b),

driving simulators conditioned on ego-trajectories (Hu et al., 2023; Wang et al., 2023b; Gao et al., 2024; Ali et al., 2025), and robotic foundation models conditioned on manipulation policies (Yang et al., 2023; Liao et al., 2025; Unitree, 2025). Across all three, the bottleneck is no longer per-frame fidelity but long-horizon state consistency: as rollouts extend, models forget previously generated content and drift away from grounded dynamics. This bottleneck motivates the systematic study of memory in video world models.

B.2 Memory Modeling in Video World Models

Long-horizon consistency is the defining challenge of video world models. Early long-video synthesis approached this through causal masking (Villegas et al., 2022), hierarchical diffusion (Yin et al., 2023), and temporal co-denoising (Wang et al., 2023a), but treated all history uniformly and left consistency to emerge implicitly from data. Recent approaches make memory a first-class design target and can be organized by where world-state information is stored.

In-context memory. In-context memory models retain history as attention activations within an extended or restructured context window. Compression-based designs allocate token budget asymmetrically across time, encoding recent frames densely while summarizing distant ones (Gu et al., 2025; Zhang & Agrawala, 2025; Jin et al., 2024). Streaming designs instead extend rollout horizons via sliding-window autoregression (Chen et al., 2025a; Team et al., 2025; Henschel et al., 2025; Qiu et al., 2023; Kim et al., 2024).

External memory. External memory architectures offload history into an explicit bank that is queried on demand, with retrieval keys ranging from camera pose (Xiao et al., 2025) to semantic relevance (Ji et al., 2025; Chen et al., 2025b) and pose-free hierarchical descriptors (Skywork AI Matrix-Game Team, 2026; Wu et al., 2026b; Zhu et al., 2025). This architecture decouples context length from compute and is naturally suited to scenarios where the model must revisit previously generated regions.

Parametric memory. Parametric memory approaches absorb history directly into model weights at inference time. Test-time training video models (Dalal et al., 2025) treat long-context tokens as an online dataset and adapt internal MLPs accordingly. A separate line of work tackles autoregressive train–test mismatch through auxiliary training schemes that expose models to noised or self-generated history (Chen et al., 2024; Song et al., 2025; Huang et al., 2025a; Yang et al., 2025a; Zhu et al., 2026), which improve robustness to imperfect history but are orthogonal to the memory architecture itself.

B.3 Evaluation for Video Generation and World Models

Early video generation evaluation relied on distribution-level metrics such as Inception Score (Salimans et al., 2016), Fréchet Inception Distance (Heusel et al., 2017), Fréchet Video Distance (Unterthiner et al., 2018), LPIPS (Zhang et al., 2018), and CLIPSIM (Wu et al., 2021), which capture frame realism and text–video alignment but reveal little about temporal or semantic consistency. Comprehensive benchmarks decompose quality into fine-grained perceptual dimensions (Huang et al., 2024; 2025b; Zheng et al., 2025a; Liu et al., 2024a; 2023; Sun et al., 2025; Han et al., 2025), with specialized variants probing physics (Bansal et al., 2024; Meng et al., 2024; Guo et al., 2025), time-lapse coherence (Yuan et al., 2024), and long-form generation (Ji et al., 2024; Zheng et al., 2025b). These protocols target content-creation quality rather than persistent world-state retention.

A recent wave of benchmarks targets world models directly, evaluating physical fidelity and instruction following (Li et al., 2025; Duan et al., 2025; Upadhyay et al., 2025), with two concurrent efforts probing memory under specific conditions (Ye et al., 2026; Ma et al., 2026). As summarized in Table 1, however, no existing benchmark jointly covers entity, environment, and causal consistency, supports both text- and action-conditioning, and probes long videos with explicit memory triggers. MBench fills this gap with a three-axis memory taxonomy and Trigger-Conditioned Scoring, which prevents conservative models from inflating scores by avoiding the challenges the benchmark is designed to expose.

C Detailed Metric Formulations

In this section, we provide the complete mathematical formulations and implementation details for the 12 sub-dimension evaluation metrics in MBench.

C.1 Entity Consistency Formulations

C.1.1 Object Consistency

Geometry Consistency. To evaluate the geometric self-consistency of generated videos in departure-return camera trajectories, we establish fold pairs by identifying the most viewpoint-aligned corresponding frames. Let frame i denote a frame from the forward pass (*departure*) and frame j be its paired counterpart from the return pass (*return*). For each target object, we extract its spatial mask M_i in the reference frame i using SAM 2 (Ravi et al., 2024). We project the return frame I_j to the viewpoint of the forward frame I_i via a warping operation, denoted as $\mathcal{W}_{j\to i}(\cdot)$. The metric is formulated as:

$$\text{Geometry Consistency Score} = \text{SSIM}\Big(I_i \odot M_i,\ \mathcal{W}_{j\to i}(I_j) \odot M_i\Big) \tag{3}$$

where $\odot$ represents the Hadamard product. This isolates the target object from the background to evaluate geometric details after off-screen periods.

Texture Consistency. Using SAM 2 (Ravi et al., 2024) to propagate precise segmentation masks, we extract semantic features $\{f_i\}$ from the target regions using the DINOv2 (Oquab et al., 2023) backbone. The score is computed as the average cosine similarity between each frame's feature and the global track centroid:

$$\text{Texture Consistency Score} = \frac{1}{N}\sum_{i=1}^{N}\frac{f_i \cdot \bar{f}}{\|f_i\|\|\bar{f}\|} \tag{4}$$

where $\bar{f} = \frac{1}{N}\sum_{j=1}^{N} f_j$ represents the global mean embedding of the object throughout the sequence of N frames.

C.1.2 Human Consistency

Identity Consistency. We detect and track faces across 32 uniformly sampled frames using a pre-trained ArcFace model (Deng et al., 2019). We maintain a feature track $\mathcal{T}_k$ storing embeddings $\{\phi_{k,i}\}$ for each subject k. A rolling-average matching strategy with greedy bipartite matching (threshold 0.15) is employed. The score is computed as:

$$\text{Identity Consistency Score} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{N_k}\sum_{i=1}^{N_k}\frac{\phi_{k,i}\cdot\bar{\phi}_k}{\|\phi_{k,i}\|\|\bar{\phi}_k\|}\right) \tag{5}$$

where K is the number of human tracks, N_k is the number of detections in track k, and $\bar{\phi}_k = \frac{1}{N_k}\sum_{j=1}^{N_k}\phi_{k,j}$ represents the centroid embedding.

Appearance Consistency. Building on the human tracks established above, we employ SAM 2 (Ravi et al., 2024) to extract full-body masks and generate semantic embeddings using DINOv2 (Oquab et al., 2023). The score is computed using the same centroid-based similarity formulation as the Identity Consistency metric to capture holistic appearance preservation.

C.2 Environment Consistency Formulations

C.2.1 Spatial Consistency

Epipolar Geometry. We estimate camera poses for all frames using DA3 (Lin et al., 2025). We establish a set of non-consecutive frame pairs $\mathcal{S}$ exhibiting similar camera poses. For each pair, we identify N matching point pairs $\{(p_i, p_i')\}$. The Epipolar Geometry Error E_{Epipolar} is:

$$E_{\text{Epipolar}} = \frac{1}{2N}\sum_{i=1}^{N}\left(d(p_i, F^T p_i') + d(p_i', F p_i)\right) \tag{6}$$

where F is the fundamental matrix derived from the estimated poses, and $d(p, l)$ is the pixel distance from point p to epipolar line l. The consistency score is computed as $\exp(-E_{\text{Epipolar}})$.

Reprojection Consistency. Using the same set $\mathcal{S}$, we triangulate 3D spatial points $\{X_i\}$. The Reprojection Error $E_{\text{Reproject}}$ is:

$$E_{\text{Reproject}} = \frac{1}{N} \sum_{i=1}^{N} \|p_i - K[R|t]X_i\|_2 \tag{7}$$

where K is the camera intrinsic matrix and $[R|t]$ represents the extrinsic parameters from DA3. The final score is $\exp(-E_{\text{Reproject}})$.

C.2.2 Rendering Consistency

Lighting Consistency. For a video V, we define paired segments $\mathcal{P} = \{(i, j)\}$ matching the departure-return passes. The overall lighting deviation $D_{\text{light}}(V)$ is a weighted combination of illumination and color shift in CIELAB space:

$$D_{\text{light}}(V) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \Big[\alpha|A_i - A_j| + (1-\alpha)\|c_i - c_j\|_2\Big] \tag{8}$$

where α is a balancing hyperparameter (set to 0.5). The illumination component A_i and color shift vector c_i for segment s_i are:

$$A_i = \frac{1}{|s_i|} \sum_{t\in s_i} \mu(L_t), \quad c_i = \frac{1}{|s_i|} \sum_{t\in s_i} [\mu(a_t), \mu(b_t)] \tag{9}$$

where $\mu(\cdot)$ represents the spatial mean of the channels. The normalized score is defined using an exponential decay function with a temperature parameter τ:

$$\text{Lighting Consistency Score} = \exp\left(-\frac{D_{\text{light}}(V)}{\tau}\right) \tag{10}$$

Style Consistency. We measure visual style drift using the Gram Matrix Distance of features extracted from a pre-trained VGG network:

$$\text{Style Consistency Score} = 1 - \frac{1}{T-1} \sum_{t=1}^{T-1} \|G_t - G_{t-1}\|_F \tag{11}$$

where G_t denotes the Gram matrix of frame t, and $\|\cdot\|_F$ denotes the Frobenius norm.

C.3 Causal Consistency Formulations

C.3.1 Self-Evolution

To evaluate out-of-view physical evolution during departure-return transitions, we employ a Vision-Language Model (VLM) as an automated evaluator. The VLM scores *State Evolution* (whether the event was actuated, mapping to a normalized gating weight $g_i \in [0, 1]$) and *Evolution Correctness* (physical plausibility, mapping to a normalized correctness score $s_i \in [0, 1]$). The final integrated score is defined via a soft-gating mechanism:

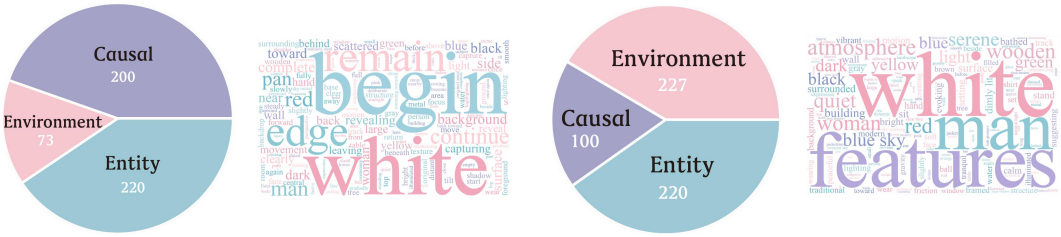
$$\text{Evolution Correctness Score} = g_i \cdot s_i \tag{12}$$

This formulation penalizes generations that fail to initiate the required action or freeze the scene to preserve appearance.

C.3.2 Interaction

Text-conditioned Instruction. For a video partitioned into M sequential segments with corresponding text prompts, we uniformly sample N frames from each segment. We employ OpenCLIP (Ilharco et al., 2021) to compute the cross-modal similarity score:

$$\text{Text Interaction Score} = \frac{1}{M \cdot N} \sum_{k=1}^{M} \sum_{i=1}^{N} \cos(f^{\text{clip}}_{\text{img}_{k,i}}, f^{\text{clip}}_{\text{text}_k}) \tag{13}$$



(a) **Prompts for Text-Conditioned Models** (b) **Prompts for Action-Conditioned Models**

Figure 3: **Prompt Statistics.**

where $f^{\text{clip}}_{\text{img}_{k,i}}$ and $f^{\text{clip}}_{\text{text}_k}$ are the image and text embedding vectors, respectively.

Action-conditioned Instruction. Given the estimated camera poses, the 6 DoF frame-wise twist $\xi_i \in \mathbb{R}^6$ at transition i is defined as:

$$\xi_i = [v_i; \omega_i] \tag{14}$$

where $\omega_i \in \mathbb{R}^3$ is the relative rotation (axis-angle) and $v_i \in \mathbb{R}^3$ is the relative translation. To handle scaling discrepancies across models, we evaluate the directional alignment with the ground-truth twist ξ_{gt}:

$$\text{Action Instruction Score} = \cos(\xi_i, \xi_{\text{gt}}) = \frac{\xi_i \cdot \xi_{\text{gt}}}{\|\xi_i\| \|\xi_{\text{gt}}\|} \tag{15}$$

D Dataset and Prompt Details

D.1 Data Composition

MBench is built upon five real-world video datasets covering diverse scenarios. Table 4 summarizes the number of evaluation samples per subset across the two benchmark settings. Here *Human* subset and *Object* subset are used to evaluate Entity-Human and Entity-Object dimensions respectively. While Environment and Causal subset is used for Causal and Environment Consistency dimensions.

Table 4: **MBench data composition.** Number of evaluation samples per subset.

	Human	**Object**	**Causal**	**Environment**	**Total**
MBench-A	120	100	100	227	547
MBench-T	120	100	200	73	493
Total	240	200	300	300	1,040

The source datasets cover complementary domains to construct our evaluation subsets: **DL3DV** (Ling et al., 2024), **Tanks and Temples** (Knapitsch et al., 2017), and **SpatialVID** (Wang et al., 2025) are utilized to build the environment and object subsets. **OpenHumanVID** (Li et al., 2024) is employed to construct the human subset, while **Physics-aware-video** (BAAI-DatacCube, 2025) contributes to the causal subset. These subsets are respectively deployed to evaluate three core memory dimensions: environment consistency, entity consistency, which comprises human and object, and causal consistency.

E Evaluation and Trigger Details

For each of the 12 sub-dimensions, the core metric formula is shared between text-conditioned video continuation models and action-conditioned world models, but the preprocessing steps differ in three key aspects: (a) how frame pairs or segments are selected, (b) how the memory trigger is determined, and (c) how the target entity is detected.

E.1 Pair Selection Strategies

MBench-T: Caption-segment-based pairs. For text-conditioned continuation videos partitioned into M caption segments, we sample frames from early segments (segment 0-1) and late segments (segment M–2 to M–1). Candidate pairs (i, j) are scored by a combination of pose-return similarity and temporal span: the pose-return score $\exp(-\|\mathbf{t}_i - \mathbf{t}_j\|/0.1 - \|\mathbf{R}_i - \mathbf{R}_j\|_{\text{deg}}/10.0)$ is computed from DA3 (Lin et al., 2025) extrinsics, and pairs are ranked primarily by this score and secondarily by temporal span. The top 5 pairs are retained.

MBench-A: Pose-based fold pairs. For action-conditioned videos featuring departure-return camera trajectories, we use DA3 (Lin et al., 2025) to estimate camera extrinsics for every frame. We detect the turnaround point where the camera is farthest from its initial pose, dividing frames into outbound (departure) and inbound (return) phases. Each outbound frame i is matched to the inbound frame j with the smallest combined translation and rotation distance:

$$j^* = \arg\min_j \left[\|\mathbf{t}_i - \mathbf{t}_j\| + \|\mathbf{R}_i - \mathbf{R}_j\|_{\text{deg}}/180 \right], \tag{16}$$

where $\mathbf{t}$ denotes the camera translation vector and $\mathbf{R}$ the rotation matrix. Up to 20 pairs with the largest temporal spans are retained for evaluation. When no clear turnaround is detected (e.g., insufficient frames), fixed symmetric pairs are used as a fallback.

E.2 Trigger Mechanism

Each MBench-T metric for text-conditioned video continuation models uses a VLM to verify whether the generated video has actually entered the intended memory challenge. The trigger judge receives 8 uniformly sampled frames from the video together with the caption segments, and returns a binary decision with a confidence score. The trigger prompt is specific to each subset:

- **Human**: "Does the video actually create a human memory challenge, such as a person appearing, leaving/being occluded across middle segments, and later reappearing?"
- **Object**: "Does the video actually create an object memory challenge for the target object, such as appearing, leaving/being occluded/changing view, and later reappearing?"
- **Environment**: "Does the video actually create an environment memory challenge, such as moving away from a scene/view/layout and later revisiting the same place or layout?"
- **Causal**: "Does the video actually create a causal memory challenge, such as an event or physical process that should continue while partially hidden or across later segments?"

A sample is considered evaluable for a given metric only if the trigger judge returns trigger. Models with low trigger coverage are penalized through the M-Score harmonic mean.

For MBench-A, we omit the VLM-based semantic trigger. Unlike text-conditioned generation where memory challenges (e.g., leaving and re-entering) can be easily ignored by the model due to the unconstrained nature of abstract prompts, action-conditioned models are mainly driven by explicit spatial trajectories. These predetermined camera or action movements deterministically enforce the physical occlusion and reappearance of scenes and objects, inherently activating the memory challenge without the need for semantic verification.

E.3 Entity Detection Details

MBench-T: Detection-based pipeline. Each object evaluation item is equipped with a semantic profile comprising a descriptive name and core visual attributes. Grounding DINO detects the target object in the initial frame using a text query formed by concatenating these attributes and the name. The resulting bounding box serves as a point prompt for SAM 2 to generate an initial mask, which is subsequently propagated across the video via SAM 2 tracking. For human entities, the detection process follows the same ArcFace pipeline as MBench-A, but utilizes per-caption-segment frame sampling rather than uniform sampling.

MBench-A: Anchor-based pipeline. Each evaluation item provides a ground-truth spatial anchor, which is parameterized by its center coordinates, width, and height, to explicitly specify the target object's location in the first frame. SAM 2 (Ravi et al., 2024) utilizes this bounding box anchor

to generate a precise initial mask, which is then propagated across all sampled frames via video tracking. For human entities, InsightFace (Buffalo_L) is employed to detect all faces within 32 uniformly sampled frames. The facial tracks are then constructed through greedy bipartite matching (with a distance threshold of 0.15) using rolling-average ArcFace (Deng et al., 2019) embeddings.

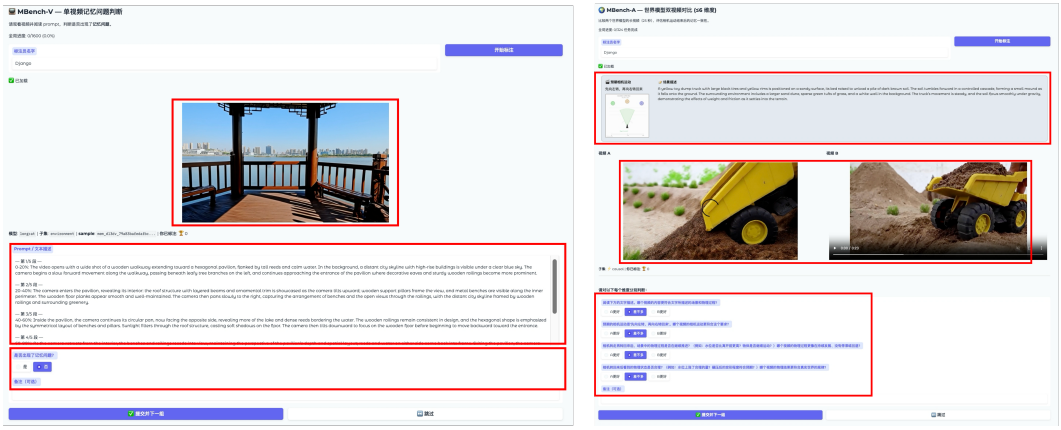
F Human Annotation Details

F.1 Annotation Protocol

Human annotations were collected through a web-based platform featuring two specialized interfaces:

Trigger Annotation Interface. Annotators view a single generated video continuation and answer whether it visibly triggers a specific memory challenge. This interface is primarily designed as a user study to collect human ground-truth labels for validating the effectiveness and reliability of our VLM-based trigger judge.

Metric Preference Interface. Annotators view two model outputs side by side (generated from the same input prompt or action sequence) and select which one better maintains a specific memory consistency property, or declare a tie. Dimension-specific questions are presented for each of the evaluated sub-dimensions.



(a) **Trigger Annotation Interface** (b) **Metric Preference Interface**

Figure 4: **Human Annotation Platform. (a) Trigger Annotation**: The top section displays the generated video, the middle shows the 5-segment continuation captions used for generation, and the bottom provides options to determine whether the intended memory challenge is successfully triggered. **(b) Metric Preference**: The top section presents contextual information (including captions, potential camera trajectories, and object masks/descriptions). The middle displays a side-by-side comparison of two generated videos, while the bottom lists specific questions corresponding to the memory dimensions evaluated in that particular subset.

F.2 Annotation Statistics

Table 5 summarizes the annotation pool used in this work. The pairwise comparisons evaluate the generated models across 12 fine-grained memory dimensions. To ensure the reliability of the collected data, annotator agreement is determined via a majority vote mechanism. Each video pair is independently evaluated by three annotators, and the majority consensus serves as the ground-truth human preference. Instances resulting in a complete tie (e.g., all three annotators provide divergent responses) are explicitly excluded from the final win-ratio calculations.

Table 5: **Human annotation statistics.**

	MBench-A Preference	MBench-T Preference	Trigger Validation
Total records	942	1,020	400
Unique annotators	17	11	4
Models covered	6	8	8

F.3 Validation of VLM-based Triggers

To validate the reliability of the VLM-based trigger judge, we conducted a user study collecting 400 human annotations across all four subsets (causal, object, environment, human). Annotators were asked whether a generated video visibly enters the intended memory challenge. Table 6 reports the accuracy, precision, recall, and F1-score of the VLM trigger decisions against the human majority vote.

Table 6: **Validation of VLM trigger against human labels.** In our evaluation settings, the VLM achieves high recall ($\geq$ 0.73) across all subsets, demonstrating its effectiveness in ensuring that genuine memory challenges are rarely missed. A false positive merely admits a sample into the downstream metric evaluation, whereas a false negative excludes genuinely informative data. With an overall recall of 0.90, the VLM serves as an effective gatekeeper for our benchmark.

Subset	Accuracy	Precision	Recall	F1
Causal	0.75	0.75	1.00	0.86
Object	0.86	0.83	1.00	0.91
Environment	0.45	0.44	0.89	0.59
Human	0.70	0.73	0.73	0.73
Overall	**0.66**	**0.65**	**0.90**	**0.76**

G More Visualizations and Case Studies

In this section, we provide qualitative visualizations to illustrate the effectiveness of the trigger mechanism and the discriminative power of MBench metrics across all three memory dimensions.

G.1 Trigger Mechanism Visualization

Figure 5 illustrates the trigger mechanism in action with two contrastive examples, both generated from the same five-stage caption prompt sequence describing a woman stepping away from and later returning to an office scene. The first example (5a) successfully triggers the memory challenge: although the generated video exhibits relatively weak appearance preservation, it faithfully follows the prompt and produces the required disappearance–reappearance event. The second example (5b) fails to trigger: the generated character maintains better visual quality, but the required exit–reentry event never occurs. As a result, this sample is excluded from consistency evaluation and contributes to the trigger coverage penalty in the M-Score.

G.2 Metric Scores Visualization

To demonstrate the discriminative power of our proposed metrics, we provide qualitative visual comparisons of high-scoring (successful) and low-scoring (failed) generated videos across the three core memory dimensions. For a fair comparison, each good–bad pair is generated conditioned on the identical input prompt or action sequence.

Entity Consistency. Figure 6 illustrates the evaluation of human and object consistency. The metrics effectively distinguish between models that preserve fine-grained facial and geometric details across the departure–return trajectory, and those that suffer from severe identity drift or object deformation upon re-entry.



(a) **Triggered Example.** Appearance Score: 10.37. Although the generated video exhibits relatively weaker visual quality, it successfully follows the prompt and produces the required disappearance–reappearance event of the target character.

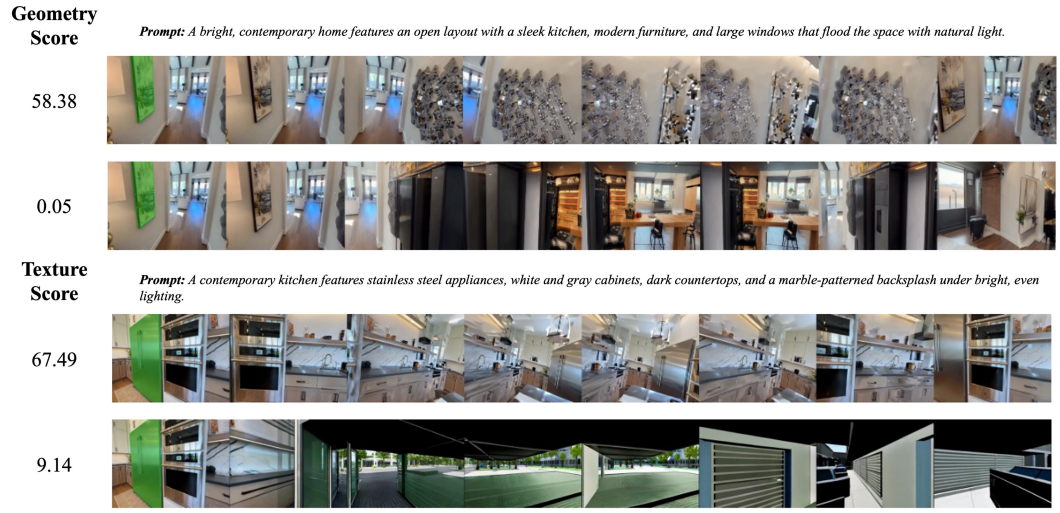


(b) **Non-triggered Example.** The generated character appearance is better preserved, but the required disappearance–reappearance event never occurs. The sample is considered non-triggered; its consistency scores are excluded from final score aggregation.

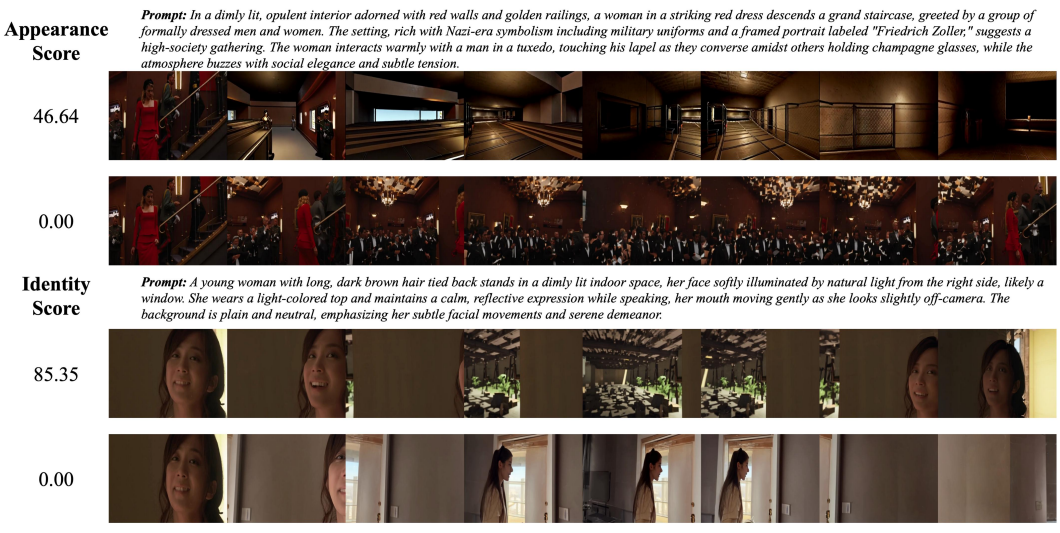
Figure 5: **Trigger mechanism for text-conditioned generation.** A triggered sample contributes its consistency scores to the reliability calculation; a non-triggered sample is excluded and contributes only to the trigger coverage penalty. The full prompt used for these examples is documented in the figure source. Both examples are generated using the same five-stage prompt sequence: *(1) A woman in her late 30s or early 40s stands closely embraced with a man in a softly lit office. She has shoulder-length dark brown hair, a few strands framing her face, wearing a white blouse with thin black vertical stripes under a dark blazer. Her left wrist displays a sleek black watch with a metallic band, and she holds the blazer tightly around him. (2) The woman begins to step away, moving steadily toward the right side of the frame. Her body turns slightly, revealing more of her profile as she glides forward. The blazer drapes over her arm, and the watch catches the light as her hand lifts gently. Only half of her remains visible as she edges out. (3) The office space is now empty—bookshelves lined with books and binders stand quietly, a framed abstract painting hangs on the wall, and soft natural light filters through unseen windows. The air feels still, the embrace gone, leaving only the calm, untouched environment behind. (4) From the left edge of the frame, the woman re-enters slowly, her silhouette emerging into view. Her dark blazer appears first, followed by the curve of her shoulder and the flash of her watch. Her hair falls gently across her face as she steps forward, partially visible again. (5) The woman is fully back in frame, standing centered once more. Her shoulder-length dark brown hair frames her face, the white striped blouse beneath the dark blazer, and the sleek black watch gleams on her left wrist. She is exactly as before—still, poised, and complete.*

Environment Consistency. Figure 7 visualizes spatial and rendering consistency during scene revisits. High scores accurately reflect mathematically preserved 3D layouts and global illumination, whereas low scores successfully capture severe geometric collapse and stylistic hallucinations.

Causal Consistency. We visualize causal self-evolution in Figure 8, highlighting the metrics' ability to heavily penalize physically implausible state transitions (e.g., hallucinatory morphing instead of physical destruction). Furthermore, Figures 9 and 10 demonstrate the evaluation of interaction adherence under action and text conditions, respectively. These metrics effectively penalize noisy motion control and semantic scene drift during long-term sequential execution.

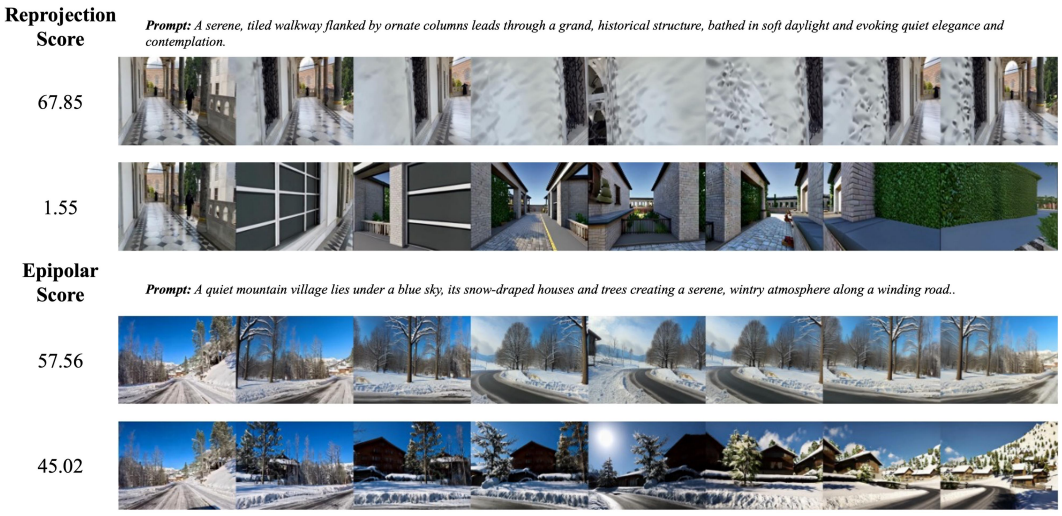


(a) **Object Consistency**

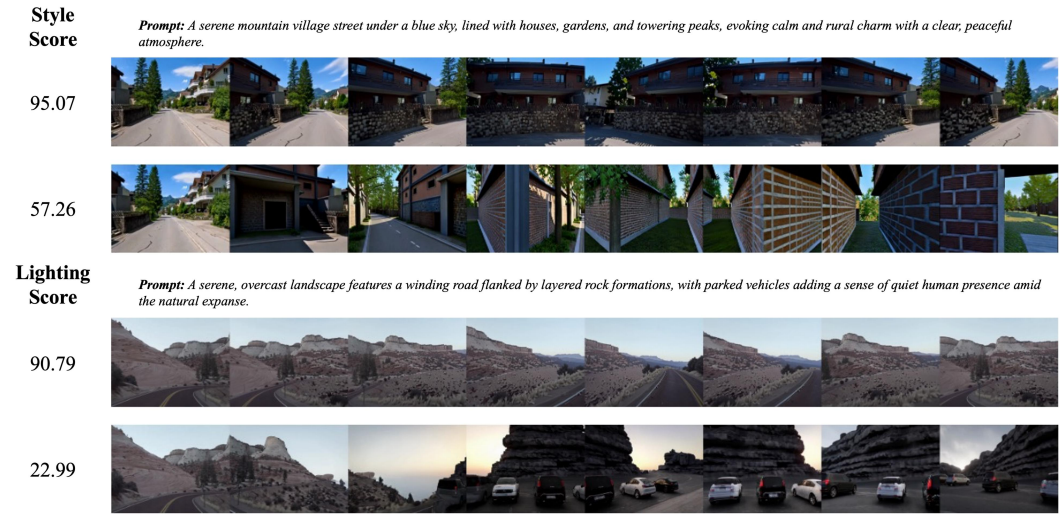


(b) **Human Consistency**

Figure 6: **Entity consistency visualization.** Qualitative comparison of human and object consistency under identical condition. **(a) Human Consistency**: Evaluated via Appearance and Identity scores. Low-scoring examples exhibit severe garment changes or facial distortion. **(b) Object Consistency**: Evaluated via Geometry and Texture scores.



(a) **Geometry Consistency**



(b) **Rendering Consistency**

Figure 7: **Environment consistency visualization.** Qualitative comparison of scene revisits. **(a) Spatial Consistency**: Measured by Reprojection and Epipolar scores. Poor models fail to maintain 3D multi-view geometry, resulting in misaligned layouts. **(b) Rendering Consistency**: Measured by Style and Lighting scores. Low-scoring examples show obvious shifts in global illumination and texture style upon returning to the initial viewport.

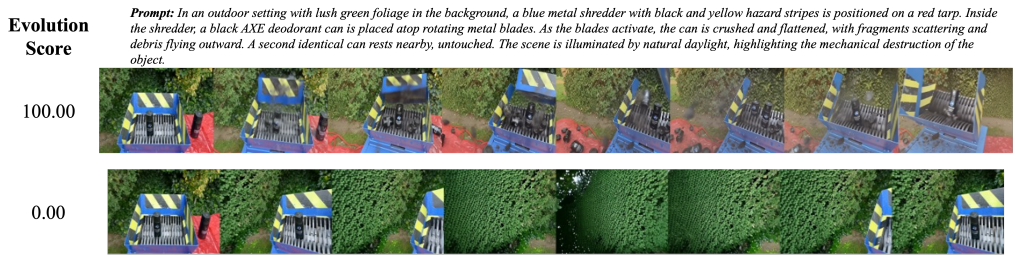
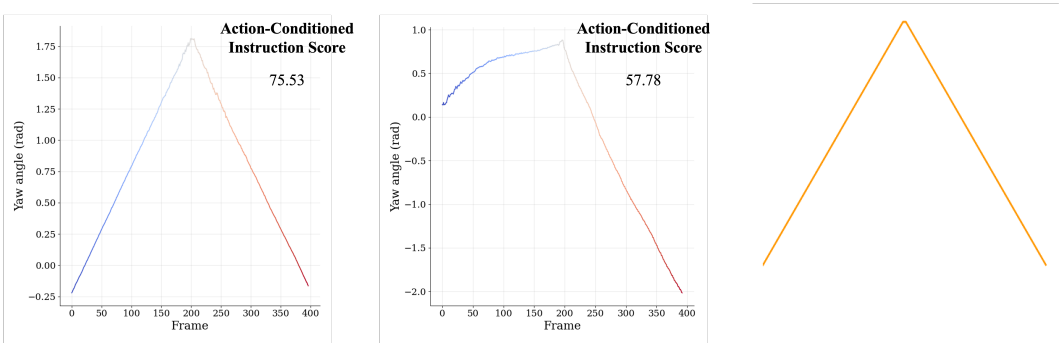


Figure 8: **Causal evolution visualization.** Qualitative comparison of state progression. The top low-scoring example (Score: 0.00) completely hallucinates the physical process, morphing the shredder into foliage. The bottom high-scoring example (Score: 100.00) correctly simulates the expected physical destruction of the deodorant can inside the shredder.



(a) **Extracted Camera Trajectories vs. Action Scores** (b) **Ideal Input**

Figure 9: **Action instruction adherence visualization.** **(a)** Extracted camera trajectories (yaw angle over frames) from generated videos. The left model tightly follows the outbound-and-return instruction, yielding a high score (75.53). The right model exhibits noisy and deviated movements, resulting in a lower score (57.78). **(b)** The ideal triangular input trajectory provided as the action condition.

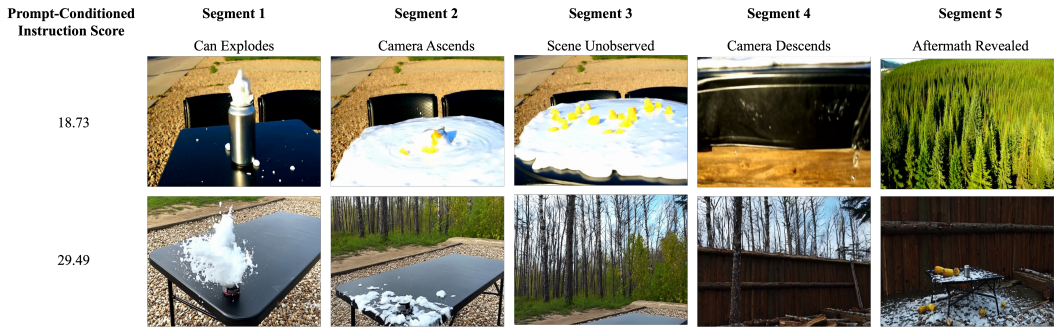


Figure 10: **Prompt instruction adherence visualization.** Qualitative comparison of text-conditioned sequential instruction following across 5 prompt segments. The top row (Score: 18.73) fails the challenge by completely hallucinating an unrelated drone-view forest in Segment 5 instead of revealing the table aftermath. The bottom row (Score: 29.49) successfully maintains scene consistency, returning to the initial black table and forest background in the final segment as instructed.