MemoBench: Benchmarking World Modeling in Dynamically Changing Environments

Haoyu Chen[1] Kaichen Zhou[1,2] Hang Hua[3] Kaile Zhang[4] Jingwen Qian[5]
Wufei Ma[6] Haonan Chen[1] Chunjiang Liu[7] Yizhou Zhao[7] Xiaoyuan Wang[7]
Weiyue Li[1] Alan Yuille[6] Paul Pu Liang[2] Yilun Du[1,8]

[1]Harvard University [2]MIT [3]MIT-IBM Watson AI Lab [4]Boston University
[5]Google [6]JHU [7]CMU [8]Kempner Institute

Abstract. Video generation models aspire to simulate dynamic environments, and several benchmarks now evaluate memory consistency across frames. However, most assess consistency only while the target remains in view, and the few that force objects out of view evaluate static scenes where nothing changes during occlusion. To bridge this gap, we introduce MemoBench, a diagnostic benchmark built around the disappear-and-reappear paradigm in dynamically changing environments: a target object undergoes a physical process, disappears from view, and must be correctly recovered in its updated state upon reappearance. We curate 360 ground-truth clips spanning synthetic and real-world scenes, and design an evaluation suite combining automated metrics with VQA-based assessment across four diagnostic pillars. Evaluation of ten state-of-the-art models reveals key insights and open challenges regarding memory consistency under the disappear-and-reappear paradigm. Our dataset, code, and leaderboard are available at https://github.com/MemoBench-Team.

Keywords: World Generation, Video Generation, Memory Consistency

1 Introduction

The real world is inherently dynamic, continuously evolving regardless of whether anyone is watching: ice melts, flames flicker, pedestrians walk, and traffic flows. Faithfully modeling such dynamically changing environments is crucial to applications ranging from autonomous driving and robotic manipulation to embodied tasks, where an agent must reason about how the world has changed beyond its field of view. Recent progress in video generation [51, 84] has shown that generative models can serve as *world generators*, capturing environment dynamics and enabling prediction under actions or interventions [13, 15, 57].

Despite this ambition, a fundamental challenge remains under-explored: *visual memory under partial observability*. In cognitive science, **object permanence**, the understanding that objects continue to exist when out of sight, is among the earliest cognitive milestones. An analogous capability is crucial for

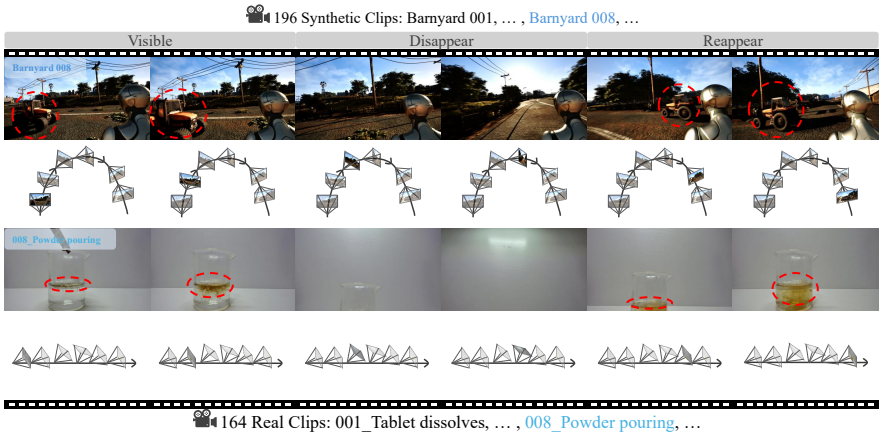


Fig. 1: Overview of MemoBench. Rows 1–2 show a synthetic Visible–Disappear–Reappear sequence and its camera trajectory; Rows 3–4 show a real-world state-change sequence (powder pouring). MemoBench contains 196 synthetic and 164 real-world clips, evaluated with automated metrics and LLM-judged VQA.

video generation: as the virtual camera moves, objects inevitably leave and re-enter the field of view, and the generative model must faithfully reproduce their appearance, position, and any ongoing state changes upon return [35]. This **disappear-and-reappear** pattern is ubiquitous in everyday experience. Yet current video generation benchmarks seldom treat this as an explicit evaluation target, leaving it unclear whether generative models truly *remember* or merely *regenerate* scene content.

Existing benchmarks have advanced the evaluation of world generation along multiple axes, including visual quality, temporal coherence, physical adherence, and scene consistency [1,9,26,31], but they predominantly evaluate what is *continuously visible* across frames. To our knowledge, none directly tests whether a generative model can maintain and update the state of objects that have temporarily left the field of view, under simultaneous camera and scene dynamics, leaving it unclear whether models can preserve identity, geometry, and evolving physical state across periods of occlusion.

To fill this gap, we introduce MemoBench, a simple yet comprehensive diagnostic benchmark for *world modeling in dynamically changing environments*. Each example follows a disappear-and-reappear structure: (i) the target object is *visible* and undergoing a physical process; (ii) the camera pans away and the target *disappears* from view while the process continues naturally; and (iii) the camera returns and the target *reappears*, and the generative model must recover its updated state.As illustrated in Fig. 1, MemoBench includes both synthetic and real-world examples following this Visible–Disappear–Reappear paradigm, together with camera trajectories and a comprehensive evaluation setup. MemoBench provides camera trajectories and depth maps, enabling evaluation grounded in both geometry and physical state evolution.

Our contributions are as follows:

- We introduce MemoBench, the first benchmark that evaluates memory consistency in world generation through the disappear-and-reappear paradigm, comprising 360 high-quality ground-truth videos at 1920×1080 resolution spanning diverse scenes and physical-state changes.
- We design a comprehensive evaluation suite combining automated metrics (video quality, Object Reappearance Score, pixel-level fidelity, and camera controllability) with LLM-judged VQA across four diagnostic dimensions.
- We benchmark ten state-of-the-art world generation models, revealing that no current model reliably maintains object memory across occlusion, and identifying key open challenges for future work.

2 Related Work

Video Generation as World Simulation. Video generation has evolved from synthesizing short clips to world simulation, modeling physics, causal dynamics, and persistent state of environments [13,15,19]. A large body of work has pushed the boundary of realistic video synthesis [4–7,18,30,34,36,39,42,45,46,48,49,51,53,55,63,65,68,69,76,79,80,84], with notable models including CogVideoX [73], Open-SoRA [84], LTX-Video [14] and LingBot-World [57] which explicitly targets long-term memory. Despite these advances, it remains unclear how well today's models preserve a persistent world state, rather than generating visually convincing frames.

Camera-Controllable Video Generation. A critical step toward faithful world simulation is generating videos conditioned on explicit camera trajectories, enabling controlled traversal of 3D environments. Early methods [16, 67] introduced camera pose conditioning modules for diffusion models, with later methods [11, 37, 40, 41, 44, 59, 64, 70, 74, 81, 82] improving camera control, 3D consistency, and geometry-aware panoramic data construction through multi-view constraints, explicit 3D representations, geometric conditioning, and simulation. Recent camera-controllable image-to-video (CI2V) models [8,27,32,57,60] accept camera pose sequences, allowing precise viewpoint control essential for autonomous driving and embodied AI. This architectural distinction carries important evaluation implications: CI2V models can execute trajectories that move objects in and out of view, whereas I2V models may exhibit *inactivity*, trivially satisfying visual consistency checks by keeping the viewpoint mostly static.

Evaluation Benchmarks for Video Generation. Most evaluation benchmarks evaluate video quality through dimensional decomposition [26,43,56,75] or physical adherence [1,2,28], while recent work evaluates generators as world simulators [31, 52]. However, these all evaluate single-viewpoint clips and to our knowledge none tests whether models maintain world state when previously observed content reappears. Recent work has made progress. WorldScore [9] evaluates scene consistency across multi-view sequences constrained by camera trajectories, but does not test object permanence. World-in-World [77] evaluates

Table 1: Comparison with recent world generation benchmarks. Scene Trav. = spatial traversal within a generated sequence; Phys. Adh. = physical adherence; Obj. Perm. = object permanence via disappear-and-reappear. MemoBench is the only benchmark that explicitly evaluates memory consistency through the disappear-and-reappear paradigm.

Benchmark	# Eval.	Eval. Type	Scene Trav.	Long Seq.	Camera Ctrl.	Scene Consist.	Phys. Adh.	Obj. Perm.
VideoPhy-2 [2]	3,940	Text	✗	✗	✗	✗	✓	✗
WorldModelBench [31]	350	Text+Img	✗	✗	✗	✗	✓	✗
WorldSimBench [52]	2,831	Text+Img	✗	✗	✗	✗	✓	✗
World-in-World [77]	1,079	Episode	✓	✓	✓	✗	✗	✗
VBench 2.0 [26]	946	Text	✓	✗	✓	✓	✓	✗
WorldScore [9]	3,000	Img+Traj	✓	✓	✓	✓	✗	✗
MemoBench (Ours)	360	Img+Traj+GT Video	✓	✓	✓	✓	✓	✓

world models in closed-loop embodied settings, focusing on task-level success rather than fine-grained visual consistency of individual objects.

As summarized in Tab. 1, these benchmarks collectively advance the evaluation of scene traversal, camera control, and scene consistency, yet they do not address the joint challenge of *dynamic camera viewpoints* and *dynamic scene content*, which tests whether a model can maintain the evolving state of a target object after it disappears from view and correctly recover it upon reappearance. Our MemoBench fills this gap through the disappear-and-reappear paradigm, requiring models to maintain memory of objects that leave the field of view and recover their evolved state when they reappear, directly probing memory consistency under simultaneous camera and scene motion.

3 MemoBench

3.1 Data Curation Framework

Our data curation pipeline comprises two parallel workflows as illustrated in Fig. 2: a synthetic pipeline and a real-world pipeline.

Synthetic pipeline. We initialize diverse 3D scenes in Unreal Engine 5 and place animated target objects along predefined paths. A virtual camera is attached to a first-person observer who follows a scripted trajectory: the observer first faces the target (Visible), performs a head turn or U-turn that moves the target out of the field of view (Disappear), and continues along the trajectory until the target re-enters the frame (Reappear). Each clip is rendered at 1920×1080 (60 FPS) with per-frame RGB, metric depth, camera intrinsics, and camera-to-world poses exported automatically.

Real-world pipeline. We record diverse physical-state-change processes in controlled indoor settings using a fixed-position camera that pans away from the target object and then returns, creating the same three-phase structure. Camera intrinsics are obtained from manufacturer calibration, while extrinsic poses are

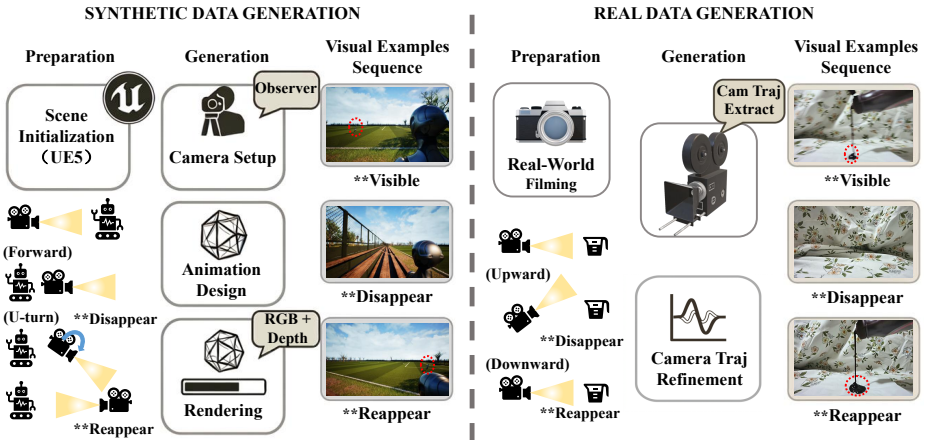


Fig. 2: Data curation pipeline for MemoBench. Left: synthetic data (196 clips, 14 scene subdomains across 5 environment categories) generated in Unreal Engine 5. Right: real-world data (164 clips, 30 physical-state-change processes across 7 categories) captured in controlled indoor settings.

estimated from the recorded RGB frames using MapAnything [29], followed by trajectory smoothing to obtain clean per-frame camera-to-world poses.

3.2 Dataset Overview

MemoBench comprises 360 ground-truth video clips organized into two complementary subsets. The synthetic subset (196 clips) focuses on *spatial diversity*, spanning 14 scene subdomains across five environment categories with rich ego-motion driving the disappear-and-reappear structure. The real-world subset (164 clips) focuses on *material diversity*, covering 30 physical-state-change processes across seven categories that depend on properties such as viscosity, elasticity, and thermal conductivity, which game engines cannot accurately model. Dataset statistics and breakdowns are provided in the supplementary material Fig. S1.

Each clip is human-annotated with two keyframe indices: d_{start} (the frame at which the target has completely disappeared from the FOV) and r_{start} (the frame at which the target has fully reappeared), which are linearly mapped to the generated-video length to set the disappear-and-reappear interval for evaluation.

3.3 Evaluation Setup

Given an input reference image (the first frame), a text prompt, and optionally a camera-control signal, a generative model produces a short video of T frames (see supplementary Tab. S1 for per-model configurations). Since the ground-truth (GT) and generated videos may differ in frame count and frame rate, GT frames are uniformly downsampled by linearly interpolating frame indices before computing per-frame metrics.

We report two complementary evaluations: **(1) Automated metrics** (Sec. 3.4) computed directly from the generated and GT videos; **(2) VQA-based evaluation** (Sec. 3.5) using Yes/No questions grouped into diagnostic dimensions.

3.4 Automated Metrics

Phase Structure. Using the two annotated keyframes defined in Sec. 3.3, each evaluation clip is divided into three non-overlapping phases: Visible (V): frames $[0, d_{\text{start}})$, where the target is fully in view; Disappeared (D): frames $[d_{\text{start}}, r_{\text{start}})$, where the target is completely out of view; Reappear (R): frames $[r_{\text{start}}, N-1]$, where the target has fully re-entered the field of view. Unless specified, we exclude the D phase from motion and geometry metrics by design.

Normalization to a 0–100 Scale. Each raw metric m is mapped to a percentage score via clipped min–max normalization:

$$\mathcal{N}(m; a, b) = 100 \cdot \text{clip}_{[0,1]}\left(\frac{m-a}{b-a}\right), \tag{1}$$

where $[a, b]$ is a predefined valid range for that metric and $\text{clip}_{[0,1]}(\cdot)$ denotes truncation to the interval $[0, 1]$. We use fixed ranges for all composite metrics (e.g., Aesthetic in $[1, 10]$; CLIP-IQA+ in $[0, 1]$).

General Video Quality. We report four metrics: Visual Quality, Motion Smoothness, Object Identity Consistency, and Geo3D Consistency.

Visual Quality. We average two no-reference quality signals over uniformly sampled frames from all phases: (i) *AestheticScore* from the LAION aesthetic predictor [54] (∼0–10); (ii) *ImageQuality* from CLIP-IQA+ [62] ($[0, 1]$). Both are mapped to $[0, 100]$ via Eq. 1 and averaged:

$$S_{\text{vq}} = \frac{1}{2}\big(\mathcal{N}(s_a; 1, 10) + \mathcal{N}(s_q; 0, 1)\big), \tag{2}$$

where s_a denote the mean AestheticScore over sampled frames and s_q denote the mean CLIP-IQA+ score over sampled frames.

Motion Smoothness. We follow VBench [26] and use RAFT-Large [58] optical flow to measure temporal smoothness. For consecutive sampled frame pairs within the V and R phases, RAFT predicts a dense flow from frame i to $i+1$, we warp frame i with bilinear sampling, and compute the mean L1 photometric error $\bar{e}$. We define:

$$S_{\text{ms}} = \mathcal{N}\left(\exp\left(-\frac{\bar{e}}{\tau}\right);\ 0,\ 1\right), \tag{3}$$

where $\tau = 0.15$ is a temperature parameter. Lower warp error implies smoother motion; the D phase is excluded by design.

Object Identity Consistency. We use DINOv2 ViT-B/14 [50] patch tokens to measure foreground object stability across the reappearance phase. For each sampled R-phase frame, we compute per-patch cosine similarity against the generated first frame I_0 patch tokens, and select the top-k% most similar patches

(k=40) to focus on the persistent foreground object. Let $\bar{c}_{\text{top}}^{(t)}$ and $c_{\text{top}}^{(t),\min}$ denote the mean and minimum similarity over the top-k% patches for frame t. We aggregate across all sampled R-phase frames:

$$S_{\text{oc}} = \alpha \cdot \overline{\bar{c}_{\text{top}}} + (1-\alpha) \cdot \min_t c_{\text{top}}^{(t),\min}, \tag{4}$$

where $\overline{\bar{c}_{\text{top}}}$ is the mean of per-frame top-k% means, $\min_t c_{\text{top}}^{(t),\min}$ is the global minimum across all sampled frames, and $\alpha = 0.7$.

Geo3D Consistency. Motion smoothness relies on optical flow, which captures pixel-level displacement but is sensitive to large camera motions and occlusions. To assess whether the underlying scene structure remains consistent, we compare per-frame depth maps estimated by Depth Anything V2 [72]. High cosine similarity between consecutive depth maps indicates stable 3D geometry, while low similarity reveals artifacts such as depth collapse or scene drift. Each depth map is min–max normalized to $[0, 1]$, flattened, and L2-normalized. We compute cosine similarity between consecutive depth maps within the V and R phases separately, obtaining per-phase mean ($\bar{d}$) and minimum ($d^{\min}$) similarities:

$$S_{\text{gc}} = \alpha \cdot \frac{\bar{d}_V + \bar{d}_R}{2} + (1-\alpha) \cdot \min(\, d_V^{\min},\ d_R^{\min}\,), \tag{5}$$

where $\alpha = 0.7$. The D phase is excluded by design.

Memory-Specific Metrics. We report five metrics across three groups: Object Reappearance Score (ORS), Pixel-Level Fidelity including PSNR, SSIM, and LPIPS, and Camera Controllability.

Object Reappearance Score (ORS). A key requirement of our evaluation is verifying whether the target object reappears during the R phase. Because the camera viewpoint in the R phase generally differs from the V phase (especially in synthetic clips with free camera trajectories), spatial metrics such as mask IoU between phases are unreliable. We therefore adopt a detection-based approach using SAM-3 [3], a text-prompted segmentation model.

For each R-phase frame, we query SAM-3 with the target object's text description and apply coverage filtering (0.05%–50% of image area, with a 0.05%–70% fallback) to reject spurious large-area masks (e.g., robot body) and noise. A frame is considered a detection if at least one valid mask is returned, and we record the highest confidence score among valid masks. ORS is defined as:

$$S_{\text{ors}} = \frac{n_d}{n_R} \cdot \frac{1}{n_d} \sum_{i=1}^{n_d} p_i, \tag{6}$$

where n_R is the total number of R-phase frames, n_d is the number of frames with a valid detection, and p_i is the confidence score of the i-th detected frame. A high ORS indicates the model reliably regenerates a recognizable target object when it reappears; a low ORS suggests the object is absent, unrecognizable, or blended into the background.

Pixel-Level Fidelity. For clips where a ground-truth reference video is available, we compute per-frame pixel-level fidelity between the generated and GT frames.

We report three complementary metrics per phase: PSNR [20] (↑) measuring signal fidelity; SSIM [66] (↑) measuring structural similarity; and LPIPS [78] (↓) measuring perceptual distance using a VGG backbone. Scores are computed separately for the V, D, and R phases as well as the full video (V+D+R), allowing phase-level analysis of where fidelity degrades. In our main results we report whole-video averages.

Camera Controllability. We estimate per-frame camera-to-world poses from generated frames using MapAnything [29], a feed-forward pose estimator that scales to large evaluation without multi-view optimization [86], and align the estimated trajectory to the GT via the first frame. We evaluate rotation error only, as the disappear-and-reappear paradigm is driven by camera heading changes and monocular translation is scale-ambiguous. We define:

$$S_{\text{cc}} = \text{clip}_{[0,1]}\left(1 - \frac{E_{\text{rot}}}{\max(\Theta_{\text{gt}},\ \theta_0)}\right), \tag{7}$$

where E_{rot} is the ATE rotation RMSE (degrees) after first-frame alignment, Θ_{gt} is the end-to-end net GT rotation, and $\theta_0 = 10^\circ$ prevents instability when the camera returns close to its starting orientation.

Prompt Fidelity. We report one metric: ImageReward Score.

ImageReward Score. We compute ImageReward [71] on uniformly sampled frames paired with the prompt. Raw scores ($\sim$−2 to +2) are first mapped to $[0, 1]$ via sigmoid, then normalized to $[0, 100]$ via Eq. 1 with $[a, b] = [0, 1]$.

3.5 VQA-based Metrics

Pipeline. Automated metrics primarily capture pixel-level fidelity and low-level perceptual quality, but often fail to measure whether a generated video correctly follows the prompt, maintains object identity over time, or preserves physical plausibility. Our VQA-based metric is designed to complement these automated signals by evaluating higher-level semantic correctness and temporal reasoning.

Recent work [10, 21–24, 33, 38, 47, 61, 83] shows that VQA-based evaluation provides a reliable and scalable framework for assessing multimodal generation models [25]. Building on this line, we introduce a multi-stage VQA metric driven by an LLM evaluator (Gemini-3.1-Pro [12]); an overview is illustrated in Fig. 3.

Given a generated clip, an LLM question generator conditions on the prompt and first frame to produce 24 polarity-balanced Yes/No questions (six per dimension). To mitigate acquiescence bias, we adopt mixed polarity: positive questions verify expected behaviors (*Yes* → Pass), while negative questions probe failure modes (*Yes* → Fail).

The question bank is refined through three stages. **(1) Ground-truth filtering:** the evaluator answers each question using the ground-truth video and removes those answered incorrectly, ensuring self-consistency. **(2) Failure filtering:** the remaining questions are tested on curated failure clips from the same scene, and questions that fail to penalize known errors are removed. **(3) Human cross-validation:** Ph.D.-level researchers and experienced AI engineers review

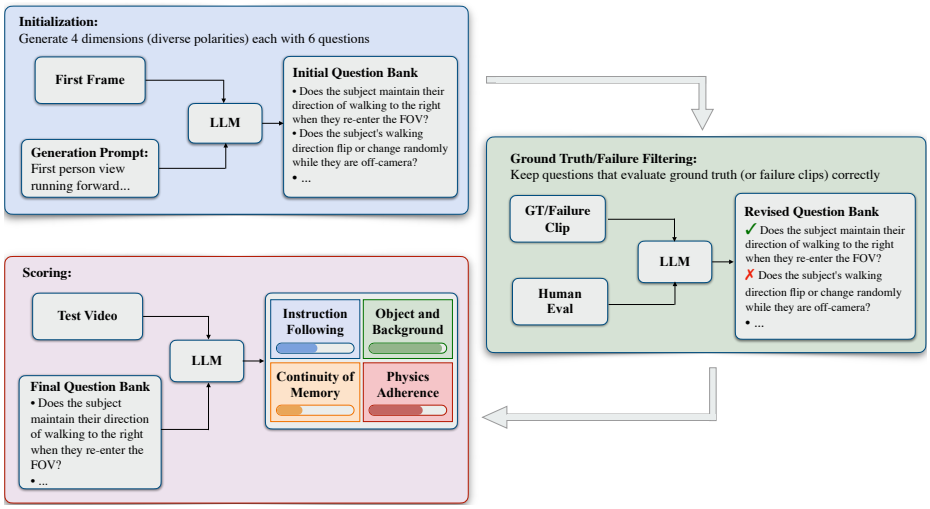


Fig. 3: VQA evaluation pipeline. An LLM generates 24 polarity-balanced Yes/No questions (6 per dimension) from the prompt and first frame. Questions are filtered through ground-truth and failure-clip evaluation, then validated by human reviewers. The final question bank is applied to each generated video, producing per-dimension pass rates across four diagnostic dimensions.

the refined question bank together with the failure cases to verify that each question is unambiguous, correctly polarized, and answerable from the video. The final validated question bank is then applied by an LLM scorer to each generated video, producing per-dimension pass rates.

Dimensions. The VQA-based evaluation covers four dimensions:
Instruction Following assesses whether the generated video faithfully executes the spatiotemporal instructions specified in the prompt, including camera motions, subject trajectories, and ordered events.
Object & Background Consistency probes the consistency of foreground objects and background elements across frames, detecting artifacts such as morphing, identity switches, or unexpected scene changes.
Continuity of Memory measures object permanence—whether the model maintains the identity, trajectory, and state of a subject after it disappears from the field of view and before it reappears. This dimension most directly aligns with the disappear-and-reappear paradigm of MemoBench.
Physics Adherence evaluates physical plausibility, including natural locomotion, consistent gravity, and coherent lighting and shadows as subjects move through the scene.

Human Validation. To validate the reliability of our VLM-generated questions and ground-truth answers, we conduct a human correlation study. We randomly sample 96 questions (8 per scene across 12 scenes), covering all four dimensions with mixed polarity, and distribute them across four interleaved survey versions.

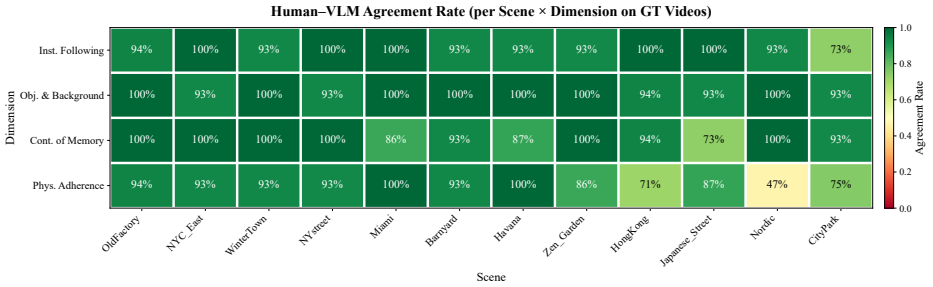


Fig. 4: Human–VLM agreement on ground-truth videos. Agreement rate per scene and dimension across 30 human responses. Overall agreement reaches 92.9%, indicating strong alignment between our VQA-based evaluation and human judgments.

In total, 30 responses are collected from Ph.D.-level researchers and experienced AI engineers, each answering Yes/No on the ground-truth videos. Fig. 4 shows the per-scene, per-dimension agreement between human majority answers and VLM-generated ground-truth answers. The results yield an overall agreement of 92.9% with Cohen's $\kappa = 0.85$, confirming that our VQA evaluation closely aligns with human judgment.

4 Evaluation Results

Testing Models. We evaluate ten world generation models on MemoBench across three categories. We assess five camera-instructed image-to-video (CI2V) models: LingBot-World [57], Wan2.2 [60], FantasyWorld [8], HunyuanWorld-Play [27], and HunyuanGameCraft [32]; two 3D-based models that synthesize novel views from explicit scene representations: Matrix-Game 2.0 [17] and Stable Virtual Camera [85]; and three open-source image-to-video (I2V) models without explicit camera conditioning: Open-SoRA [51], LTX-Video [14], and CogVideoX [73]. Implementation details and generation configurations are provided in the supplementary material (Sec. A.1 and Sec. A.2).

4.1 Analysis of Automated Metrics

Explicit 3D representations enable precise but not universal trajectory control. Pose-conditioned view synthesis pipelines, such as Stable Virtual Camera, render images directly from explicit camera poses, following the specified trajectory by construction. As shown in Table 2, this leads to strong Camera Controllability and the highest pixel-level fidelity, although HunyuanWorldPlay achieves the highest overall Camera Controllability. In contrast, even though Matrix-Game 2.0 also relies on an explicit 3D representation, it achieves controllability comparable to I2V models. The key difference lies in the conditioning interface: when geometry is accessed through action-conditioned dynamics

Table 2: Automated evaluation of 10 world generation models on MemoBench. Models are grouped into CI2V, 3D-based, and I2V categories. ↑: higher is better; ↓: lower is better. **Bold**: best; underline: second best.

	General Video Quality				Memory-Specific					Prompt
Model	VisQual↑	MotSmooth↑	ObjConsist↑	3DConsist↑	ORS↑	PSNR↑	SSIM↑	LPIPS↓	CamCtrl↑	ImgReward↑
CI2V Models										
LingBot-World [57]	47.4	57.6	59.0	88.2	0.381	14.41	0.490	0.482	37.4	36.7
Wan2.2 [60]	40.0	54.0	50.7	84.5	0.328	13.76	0.469	0.529	29.8	26.1
FantasyWorld [8]	51.0	55.2	47.6	88.7	0.276	13.23	0.427	0.571	27.2	30.7
HunyuanWorldPlay [27]	43.5	66.6	61.5	90.6	**0.582**	14.35	0.471	0.505	**69.9**	24.5
HunyuanGameCraft [32]	54.2	51.9	46.6	85.9	0.266	12.81	0.388	0.603	54.2	8.9
3D-based Models										
Matrix-Game 2.0 [17]	**61.2**	83.6	46.5	93.7	0.157	13.49	0.376	0.550	17.3	22.3
Stable Virtual Camera [85]	43.3	63.1	59.5	88.5	0.294	**15.36**	**0.523**	**0.455**	65.2	22.3
I2V Models										
Open-SoRA [51]	49.7	68.3	47.2	89.7	0.182	12.54	0.384	0.566	16.8	31.3
LTX-Video [14]	44.9	**84.4**	**81.6**	**94.1**	0.330	13.42	0.455	0.518	17.1	**37.1**
CogVideoX [73]	40.1	59.8	54.0	94.0	0.251	12.07	0.480	0.592	12.0	34.9

rather than explicit pose conditioning, the underlying 3D structure is not fully leveraged for precise trajectory control. What ultimately determines trajectory precision is whether the model's conditioning mechanism directly exposes geometric degrees of freedom, or instead relies on implicit, learned transitions. This is visually confirmed in Fig. 5: Stable Virtual Camera reproduces the GT pan-away-and-return trajectory with consistent viewpoint progression, whereas Matrix-Game 2.0 drifts to an entirely different scene despite also operating on an explicit 3D representation.

Camera inactivity inflates consistency metrics. LTX-Video tops three of four General Video Quality metrics (Table 2) and also obtains a relatively high ORS of 0.330, despite having Camera Controllability comparable to the other I2V baselines. This contradiction arises because LTX-Video barely moves the camera: when consecutive frames are nearly identical, flow-based smoothness, depth consistency, and identity similarity are trivially maximized. The same mechanism inflates its ORS, since the target object never leaves the frame and SAM-3 detects it throughout the R phase by default. This exposes a limitation of standard video quality metrics: they cannot distinguish a model that preserves appearance across genuine viewpoint changes from one that simply avoids moving. Fig. 6 illustrates this behavior, with LTX-Video retaining a nearly fixed viewpoint throughout the sequence.

A trade-off emerges between geometric fidelity and perceptual quality. Stable Virtual Camera leads all pixel-level metrics, yet its Visual Quality score remains relatively low in Table 2, whereas Matrix-Game 2.0 achieves the highest Visual Quality but ranks among the lowest in SSIM. Both 3D-based models also obtain relatively low ImageReward scores. This pattern suggests that geometric consistency and perceptual naturalness are not yet aligned in current methods. HunyuanGameCraft further demonstrates this metric mismatch: it achieves the second-highest Visual Quality score, while obtaining the lowest ImageReward

Fig. 5: Qualitative comparison of camera controllability on a real-world clip. SVC follows the prescribed trajectory closely, while Wan2.2 and Matrix-Game 2.0 fail to reproduce the intended viewpoint changes.

and the weakest GT-aligned pixel fidelity among the CI2V models. These results show that no-reference visual quality, prompt-image alignment, and GT-aligned geometric fidelity capture different aspects of generation quality and should not be interpreted interchangeably. As illustrated in Fig. 7, Matrix-Game 2.0 produces sharp frames but drifts from the GT viewpoint, whereas Stable Virtual Camera better preserves scene geometry while introducing rendering artifacts such as blurring, seams, and depth-inpainting errors.

Camera conditioning alone does not ensure object memory. All five CI2V models share explicit camera conditioning, yet their object memory performance varies notably (Table 2): HunyuanWorldPlay achieves the highest CI2V ORS, while LingBot-World leads all CI2V pixel-level fidelity metrics. In contrast, FantasyWorld achieves higher Visual Quality than LingBot-World but a substantially lower ORS. This gap within the same model category reveals that camera conditioning does not by itself encourage the model to maintain a representation of objects that have left the field of view. A model can produce aesthetically better frames while failing to recall what it previously observed, suggesting that object permanence must be explicitly targeted during training rather than as a byproduct of camera-conditioned generation. As shown in Fig. 6, both LingBot-World and FantasyWorld receive the same camera trajectory, yet LingBot-World produces a recognizable return to the target region while FantasyWorld generates frames that bear little resemblance to the ground-truth reappearance.

ORS reveals memory failures and reliable reappearance remains open. No model exceeds an ORS of 0.6 (Table 2), indicating that even the top performer does not reliably re-detect the target object throughout the R phase. However, ORS must be interpreted jointly with Camera Controllability: LTX-Video obtains an ORS of 0.330 despite its low Camera Controllability, indicating

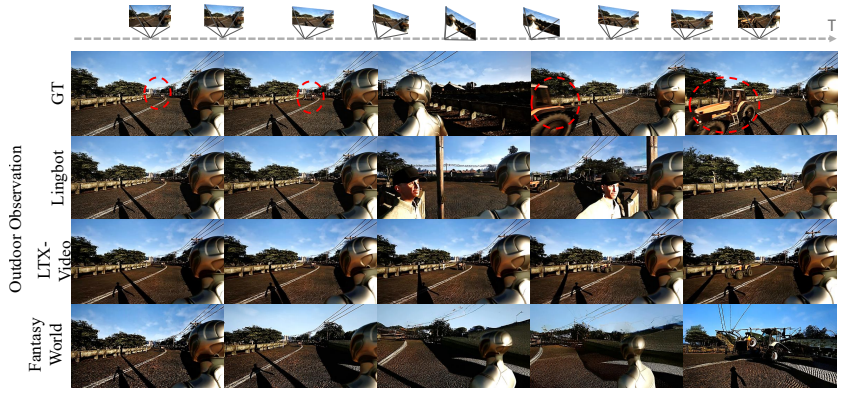


Fig. 6: Camera inactivity vs. active trajectory following. LTX-Video produces nearly static frames, while LingBot-World and FantasyWorld follow the trajectory but fail to recover the target object upon reappearance.

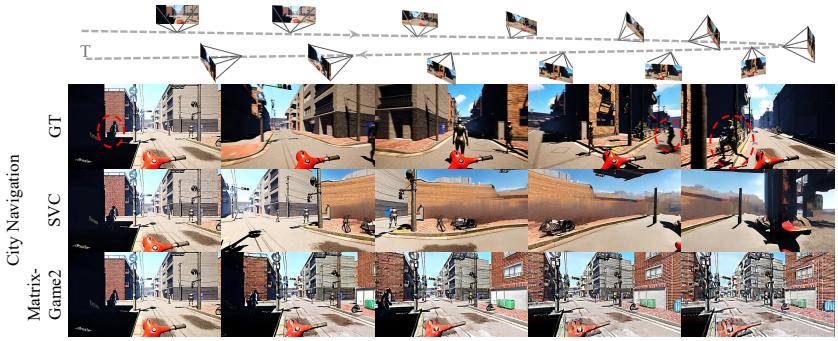


Fig. 7: Qualitative comparison of geometric fidelity and perceptual quality on a synthetic clip. SVC preserves scene geometry but introduces artifacts, while Matrix-Game 2.0 produces visually sharper frames but drifts from the GT viewpoint.

that part of its score can be attributed to camera inactivity rather than genuine disappearance and reappearance. Among models that actually execute the trajectory, HunyuanWorldPlay leads, followed by LingBot-World, Wan2.2, and Stable Virtual Camera. The low absolute values indicate that current models lack a persistent internal representation of disappeared objects. Once the target leaves the frame, the model's "memory" degrades rapidly, and the reappeared content is either absent, hallucinated, or unrecognizable. As shown in Fig. 6, even LingBot-World, one of the top-performing models on ORS, fails to recover the target object faithfully upon reappearance, and LTX-Video's apparent success reflects a static viewpoint rather than genuine object recall. Overall, no single model simultaneously achieves strong Camera Controllability, high ORS, and competitive Visual Quality. Closing this gap is a core challenge that MemoBench exposes for future world generation models.

4.2 Analysis of VQA-based Evaluation

Camera inactivity inflates VQA scores; Instruction Following exposes the gap. LTX-Video achieves the highest scores on two of the four VQA dimensions (Table 3) and ranks second on Physics Adherence, narrowly behind HunyuanWorldPlay. As an I2V model without camera conditioning, its strong consistency-oriented scores mirror the inflation pattern observed in automated metrics. However, Instruction Following reveals a different ranking: LingBot-World leads, closely followed by HunyuanWorldPlay, and CI2V models occupy the top four positions, while I2V models cluster at lower scores. This indicates that camera conditioning generally improves the execution of spatiotemporal instructions, whereas consistency-oriented VQA scores can be inflated when a model avoids the requested viewpoint change. Notably, LingBot-World's advantage in Instruction Following does not transfer to other dimensions: its Object & Background score remains substantially lower than that of LTX-Video, suggesting that actively following the trajectory introduces inconsistencies that static models avoid by not moving.

Table 3: VQA evaluation across four semantic dimensions on MemoBench. Each dimension is scored 0–100 (↑: higher is better). Models are grouped into CI2V, 3D-based, and I2V categories. **Bold**: best; underline: second best.

Model	Inst.Fol. ↑	Obj.&Bkg. ↑	Cont.Mem. ↑	Phys.Adh. ↑
CI2V Models				
LingBot-World [57]	**64.2**	44.4	42.1	53.6
Wan2.2 [60]	50.6	30.2	36.8	38.9
FantasyWorld [8]	50.5	25.6	37.1	33.6
HunyuanWorldPlay [27]	61.6	66.4	55.6	**63.6**
HunyuanGameCraft [32]	41.6	71.6	48.4	61.0
3D-based Models				
Matrix-Game 2.0 [17]	37.5	12.7	36.5	21.8
Stable Virtual Camera [85]	49.7	23.8	29.6	33.3
I2V Models				
Open-SoRA [51]	43.2	66.8	48.3	59.7
LTX-Video [14]	41.0	**76.6**	**57.0**	63.5
CogVideoX [73]	40.5	52.4	42.7	42.8

Semantic evaluation reveals artifacts missed by automated metrics. Matrix-Game 2.0 records the lowest Object & Background and Physics Adherence scores across all models (Tab. 3), despite achieving the highest Visual Quality and second-highest Motion Smoothness in automated evaluation. Stable Virtual Camera shows a similar trend: strong pixel-level fidelity but below-average VQA scores. These results suggest that rendering artifacts—such as warping seams, depth inpainting errors, and texture flickering—are largely in-

visible to no-reference quality metrics but are penalized by VQA evaluation, which focuses on semantic correctness rather than perceptual sharpness.

Continuity of Memory remains a major bottleneck. The highest Continuity of Memory score is achieved by LTX-Video, whose score may be inflated by camera inactivity (Tab. 3). Among models that actively follow the trajectory, HunyuanWorldPlay achieves the highest score, although only slightly more than half of the memory-related questions are answered correctly. Together with the low ORS values observed in automated evaluation, this result confirms that current models fail to maintain a reliable representation of objects once they leave the field of view, both at the signal level and the semantic level.

5 Conclusion

This paper introduces MemoBench, a novel benchmark that evaluates memory consistency in world generation through the disappear-and-reappear paradigm. By combining automated metrics with a VQA pipeline across 360 clips, we found that current models struggle to maintain a persistent representation of objects that leave the field of view. No model exceeds an Object Reappearance Score of 0.4, and models without camera conditioning inflate consistency scores by generating near-static video rather than executing viewpoint changes. Even among camera-conditioned models, object permanence does not emerge as a byproduct of trajectory control, indicating that memory must be explicitly addressed in model design. These findings point to future work on persistent state representations, memory-aware training objectives, and evaluation protocols that account for camera inactivity. We hope MemoBench serves as a useful tool for tracking progress on these challenges.

Acknowledgements

YZ was supported in part by the SoftBank Group–ARM Fellowship. This work was supported in part by the Office of Naval Research (ONR) under Grant No. N000142412696 and also by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University.

References

1. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520 (2024) 2, 3
2. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. arXiv preprint arXiv:2503.06800 (2025) 3, 4
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025) 7
4. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023) 3
5. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: CVPR (2024) 3
6. Chen, T.S., Lin, C.H., Tseng, H.Y., Lin, T.Y., Yang, M.H.: Motion-conditioned diffusion model for controllable video synthesis. arXiv preprint arXiv:2304.14404 (2023) 3
7. Chen, X., Wang, Y., Zhang, L., Zhuang, S., Ma, X., Yu, J., Wang, Y., Lin, D., Qiao, Y., Liu, Z.: Seine: Short-to-long video diffusion model for generative transition and prediction. In: ICLR (2023) 3
8. Dai, Y., Jiang, F., Wang, C., Xu, M., Qi, Y.: Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. arXiv preprint arXiv:2509.21657 (2025) 3, 10, 11, 14, S2, S3
9. Duan, H., Yu, H.X., Chen, S., Fei-Fei, L., Wu, J.: Worldscore: A unified evaluation benchmark for world generation. arXiv preprint arXiv:2504.00983 (2025) 2, 3, 4
10. Feng, Y., Li, Y., Liu, C., Chen, Y., Jiang, F., Huang, Y., Hua, H., Yuan, Z., Zheng, K., Niu, L., et al.: Visual aesthetic benchmark: Can frontier models judge beauty? arXiv preprint arXiv:2605.12684 (2026) 8
11. Ge, X., Pan, Y., Zhang, Y., Li, X., Zhang, W., Zhang, D., Wan, Z., Lin, X., Zhang, X., Liang, J., et al.: Airsim360: A panoramic simulation platform within drone view. In: CVPR (2026) 3
12. Google: Gemini 3.1 pro (2026), https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, accessed: 2026-03-02 8
13. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018) 1, 3
14. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024) 3, 10, 11, 14, S2, S3
15. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104 (2023) 1, 3
16. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024) 3
17. He, X., Peng, C., Liu, Z., Wang, B., Zhang, Y., Cui, Q., Kang, F., Jiang, B., An, M., Ren, Y., et al.: Matrix-game 2.0: An open-source real-time and streaming interactive world model. arXiv preprint arXiv:2508.13009 (2025) 10, 11, 14, S2, S3

18. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022) 3
19. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Adv. Neural Inform. Process. Syst. (2022) 3
20. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: ICPR (2010) 8
21. Hu, Y., Hua, H., Yang, Z., Shi, W., Smith, N.A., Luo, J.: Promptcap: Prompt-guided image captioning for vqa with gpt-3. In: CVPR (2023) 8
22. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: CVPR (2023) 8
23. Hua, H., Shi, J., Kafle, K., Jenni, S., Zhang, D., Collomosse, J., Cohen, S., Luo, J.: Finematch: Aspect-based fine-grained image and text mismatch detection and correction. In: ECCV (2024) 8
24. Hua, H., Tang, Y., Zeng, Z., Cao, L., Yang, Z., He, H., Xu, C., Luo, J.: Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. arXiv preprint arXiv:2410.09733 (2024) 8
25. Hua, H., Zeng, Z., Song, Y., Tang, Y., He, L., Aliaga, D., Xiong, W., Luo, J.: Mmig-bench: Towards comprehensive and explainable evaluation of multi-modal image generation models. arXiv preprint arXiv:2505.19415 (2025) 8
26. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR (2024) 2, 3, 4, 6
27. HunyuanWorld, T.: Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. arXiv preprint (2025) 3, 10, 11, 14, S2, S3
28. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective. arXiv preprint arXiv:2411.02385 (2024) 3
29. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025) 5, 8
30. Kuaishou: Kling (2024), https://kling.kuaishou.com/en, accessed: 2026-03-01 3
31. Li, D., Fang, Y., Chen, Y., Yang, S., Cao, S., Wong, J., Luo, M., Wang, X., Yin, H., Gonzalez, J.E., et al.: Worldmodelbench: Judging video generation models as world models. arXiv preprint arXiv:2502.20694 (2025) 2, 3, 4
32. Li, J., Tang, J., Xu, Z., Wu, L., Zhou, Y., Shao, S., Yu, T., Cao, Z., Lu, Q.: Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. arXiv preprint arXiv:2506.17201 (2025) 3, 10, 11, 14, S2, S3
33. Li, W., Zhao, M., Dong, W., Cai, J., Wei, Y., Pocress, M., Li, Y., Yuan, W., Wang, X., Hou, R., et al.: Grading scale impact on llm-as-a-judge: Human-llm alignment is highest on 0-5 grading scale. arXiv preprint arXiv:2601.03444 (2026) 8
34. Li, Y., Meng, S., Yang, C., Feng, W., Liu, J., An, Z., Wang, Y., Tian, Y.: A comprehensive survey of interaction techniques in 3d scene generation. Authorea Preprints (2026) 3
35. Lillemark, H.J., Huang, B., Zhan, F., Du, Y., Keller, T.A.: Flow equivariant world models: Memory for partially observed dynamic environments. arXiv preprint arXiv:2601.01075 (2026) 2

36. Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., et al.: Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131 (2024) 3
37. Lin, X., Song, M., Zhang, D., Lu, W., Li, H., Du, B., Yang, M.H., Nguyen, T., Qi, L.: Depth any panoramas: A foundation model for panoramic depth estimation. In: CVPR (2026) 3
38. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. arXiv preprint arXiv:2404.01291 (2024) 8
39. Liu, C., Wang, X., Lin, Q., Xiao, A., Chen, H., Wen, S., Zhang, H., Qi, L., Yang, M.H., Jeni, L.A., et al.: Mosiv: Multi-object system identification from videos. arXiv preprint arXiv:2603.06022 (2026) 3
40. Liu, M., Liu, J., Zhang, Y., Li, J., Yang, M.Y., Nex, F., Cheng, H.: 4dstr: Advancing generative 4d gaussians with spatial-temporal rectification for high-quality and consistent 4d generation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026) 3
41. Liu, M., Zhang, D., Liu, J., Cui, J., Xie, H., Chen, G., Ye, H., Yang, M.Y., Nex, F., Cheng, H.: Driveva: Video action models are zero-shot drivers. arXiv preprint arXiv:2604.04198 (2026) 3
42. Liu, P., Song, L., Zhang, D., Hua, H., Tang, Y., Tu, H., Luo, J., Xu, C.: Emo-avatar: Efficient monocular video style avatar through texture rendering. arXiv preprint arXiv:2402.00827 **1** (2024) 3
43. Liu, Y., Li, L., Ren, S., Gao, R., Li, S., Chen, S., Sun, X., Hou, L.: Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. Adv. Neural Inform. Process. Syst. (2023) 3
44. Liu, Y., Lin, X., Li, X., Yang, B., Wang, C., Sunkavalli, K., Hold-Geoffroy, Y., Tan, H., Zhang, K., Xie, X., et al.: Omniroam: World wandering via long-horizon panoramic video generation. arXiv preprint arXiv:2603.30045 (2026) 3
45. Luma AI: Luma dream machine (2024), https://lumalabs.ai/dream-machine, accessed: 2026-03-01 3
46. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation. arXiv preprint arXiv:2303.08320 (2023) 3
47. Ma, W., Chen, H., Zhang, G., Chou, Y.C., Chen, J., de Melo, C., Yuille, A.: 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In: ICCV (2025) 8
48. OpenAI: Sora (2024), https://openai.com/index/sora/, accessed: 2026-03-01 3
49. OpenAI: Sora2 (2025), https://openai.com/index/sora-2/, accessed: 2026-03-01 3
50. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 6
51. Peng, X., Zheng, Z., Shen, C., Young, T., Guo, X., Wang, B., Xu, H., Liu, H., Jiang, M., Li, W., et al.: Open-sora 2.0: Training a commercial-level video generation model in 200 k. arXiv preprint arXiv:2503.09642 (2025) 1, 3, 10, 11, 14, S2, S3
52. Qin, Y., Shi, Z., Yu, J., Wang, X., Zhou, E., Li, L., Yin, Z., Liu, X., Sheng, L., Shao, J., et al.: Worldsimbench: Towards video generation models as world simulators. arXiv preprint arXiv:2410.18072 (2024) 3, 4
53. Runway ML: Gen-3 alpha (2024), https://runwayml.com/research/introducing-gen-3-alpha, accessed: 2026-03-01 3

54. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Adv. Neural Inform. Process. Syst. (2022) 6
55. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022) 3
56. Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., Liu, X.: T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In: CVPR (2025) 3
57. Team, R., Gao, Z., Wang, Q., Zeng, Y., Zhu, J., Cheng, K.L., Li, Y., Wang, H., Xu, Y., Ma, S., Chen, Y., Liu, J., Cheng, Y., Yao, Y., Zhu, J., Meng, Y., Zheng, K., Bai, Q., Chen, J., Shen, Z., Yu, Y., Zhu, X., Shen, Y., Ouyang, H.: Advancing open-source world models. arXiv preprint arXiv:2601.20540 (2026) 1, 3, 10, 11, 14, S2
58. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV (2020) 6
59. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In: ECCV (2024) 3
60. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 3, 10, 11, 14, S2
61. Wang, C., Lin, X., Liu, J., Liu, Y., Wang, Z., Qi, D., Yan, Y., Chen, X.: PanoWorld: Towards spatial supersensing in 360-degree panorama world. arXiv preprint arXiv:2605.13169 (2026) 8
62. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023) 6
63. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023) 3
64. Wang, X., Zhao, Y., Ye, B., Shan, X., Lyu, W., Qi, L., Chan, K.C., Li, Y., Yang, M.H.: Holigs: Holistic gaussian splatting for embodied view synthesis. arXiv preprint arXiv:2506.19291 (2025) 3
65. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. IJCV (2025) 3
66. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP (2004) 8
67. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: SIGGRAPH (2024) 3
68. Xiang, J., Liu, G., Gu, Y., Gao, Q., Ning, Y., Zha, Y., Feng, Z., Tao, T., Hao, S., Shi, Y., et al.: Pandora: Towards general world model with natural language actions and video states. arXiv preprint arXiv:2406.09455 (2024) 3
69. Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., Wong, T.T.: Dynamicrafter: Animating open-domain images with video diffusion priors. In: ECCV (2024) 3
70. Xu, D., Nie, W., Liu, C., Liu, S., Kautz, J., Wang, Z., Vahdat, A.: Camco: Camera-controllable 3d-consistent image-to-video generation. arXiv preprint arXiv:2406.02509 (2024) 3

71. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Adv. Neural Inform. Process. Syst. (2023) 8
72. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Adv. Neural Inform. Process. Syst. (2024) 7
73. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024) 3, 10, 11, 14, S2, S3
74. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024) 3
75. Yuan, S., Huang, J., Xu, Y., Liu, Y., Zhang, S., Shi, Y., Zhu, R.J., Cheng, X., Luo, J., Yuan, L.: Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. Adv. Neural Inform. Process. Syst. (2024) 3
76. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. IJCV (2025) 3
77. Zhang, J., Jiang, M., Dai, N., Lu, T., Uzunoglu, A., Zhang, S., Wei, Y., Wang, J., Patel, V.M., Liang, P.P., et al.: World-in-world: World models in a closed-loop world. arXiv preprint arXiv:2510.18135 (2025) 3, 4
78. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 8
79. Zhao, Y., Chen, H., Liu, C., Li, Z., Herrmann, C., Hur, J., Li, Y., Yang, M.H., Raj, B., Xu, M.: Masiv: Toward material-agnostic system identification from videos. arXiv preprint arXiv:2508.01112 (2025) 3
80. Zhao, Y., Liu, C., Chen, H., Raj, B., Xu, M., Baltrusaitis, T., Rundle, M., Wu, H., Ghasedi, K.: Total-editing: Head avatar with editable appearance, motion, and lighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025) 3
81. Zhao, Y., Wang, Y., Wang, X., Wu, Y., Zhang, H., Haji-Ali, M., Abdal, R., Mirzaei, A., Li, Y., Menapace, W., et al.: Geostream: Toward precise camera controlled streaming video generation. arXiv preprint arXiv:2606.15162 (2026) 3
82. Zheng, G., Li, T., Jiang, R., Lu, Y., Wu, T., Li, X.: Cami2v: Camera-controlled image-to-video diffusion model. arXiv preprint arXiv:2410.15957 (2024) 3
83. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems **36**, 46595–46623 (2023) 8
84. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024) 1, 3
85. Zhou, J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. In: CVPR (2025) 10, 11, 14, S2, S3
86. Zhou, K., Wang, Y., Chen, G., Beaudouin, G., Zhan, F., Liang, P.P., Wang, M.: Page-4d: Disentangled pose and geometry estimation for 4d perception. In: ICLR (2025) 8

A Supplementary Materials

Table of Contents

A.1 Model Configurations

We evaluate ten open-source models spanning three categories: camera-conditioned image-to-video generation (CI2V), standard image-to-video generation (I2V), and novel view synthesis (NVS). Table S1 summarizes the key specifications of each model, including output resolution, frame rate, generated video length, and whether the model supports explicit camera pose conditioning.

The CI2V models (LingBot-World, Wan2.2, FantasyWorld HunyuanWorld-Play, and HunyuanGameCraft) all accept camera trajectories as input, allowing direct control over viewpoint changes. The NVS models (Matrix-Game 2.0 and Stable Virtual Camera) approach the problem from a 3D reconstruction perspective, synthesizing novel views given target camera poses. The I2V models (Open-SoRA, LTX-Video, and CogVideoX) do not support camera conditioning and instead rely on text prompts or learned priors to determine camera motion. Including these non-camera-conditioned baselines allows us to assess how much explicit camera control contributes to generation quality and geometric consistency.

Table S1: Model configurations used in our evaluation. We summarize the output resolution, frame rate, video length, and camera-conditioning support for each baseline. All models are open-source. CI2V and I2V denote camera-conditioned and standard image-to-video generation, respectively; NVS denotes novel view synthesis.

Model	Type	Ability	Res.	FPS	Length (s / frames)	Open Source	Camera
LingBot-World [57]	Video	CI2V	464 × 832	16	5.1 / 81	✓	✓
Wan2.2 [60]	Video	CI2V	480 × 832	16	5.1 / 81	✓	✓
FantasyWorld [8]	Video	CI2V	480 × 832	16	5.1 / 81	✓	✓
HunyuanWorldPlay [27]	Video	CI2V	480 × 848	24	5.2 / 125	✓	✓
HunyuanGameCraft [32]	Video	CI2V	720 × 1280	24	4.1 / 99	✓	✓
Matrix-Game 2.0 [17]	3D	NVS	352 × 640	60	4.8 / 285	✓	✓
Stable Virtual Camera [85]	3D	NVS	576 × 1024	16	5.0 / 80	✓	✓
Open-SoRA [51]	Video	I2V	192 × 336	24	5.4 / 129	✓	×
LTX-Video [14]	Video	I2V	480 × 832	25	5.2 / 129	✓	×
CogVideoX [73]	Video	I2V	480 × 832	16	5.1 / 81	✓	×

A.2 Implementation Details

All experiments are conducted on a server equipped with four NVIDIA RTX Pro 6000 GPUs. We reproduce each baseline using its official codebase and publicly available checkpoints. Below we summarize the inference configuration for each model.

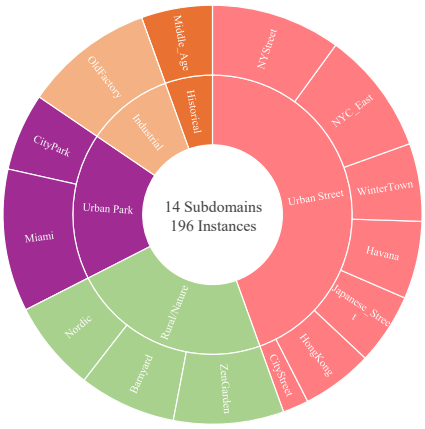
LingBot [57]. We use the official `lingbot-world-base-cam` checkpoint and generate 81 frames at 464×832 resolution with the UniPC solver, 70 sampling steps, a guidance scale of 5.0, and 16 fps output.

FantasyWorld [8]. We use the Wan2.1-I2V-14B-480P variant and generate 81 frames at 336×592 resolution with 50 flow-matching steps and a guidance scale of 5.0.
HunyuanWorldPlay [27]. We use the official `HY-WorldPlay` distilled checkpoint (`ar_distilled_action_model`) and generate 125 frames at 480×848 resolution with 4 flow-matching steps, a guidance scale of 1.0, and 24 fps output.
HunyuanGameCraft [32]. We use the official `Hunyuan-GameCraft-1.0` distilled checkpoint (`mp_rank_00_model_states_distill.pt`) and generate 99 frames at 720×1280 resolution with 8 flow-matching steps, a guidance scale of 1.0, and 24 fps output.
Stable Virtual Camera [85]. We use the v1.1 checkpoint and run two-pass Euler EDM sampling with 50 steps each, guidance scales of 3.0 and 2.0, outputting 80 frames at 576×1024 and 16 fps.
Matrix-Game [17]. We use the Universal distilled checkpoint with 3 denoising steps, generating videos at 352×640 and 60 fps with duration matched to the ground-truth sequence.
VideoX-Fun. We use the Wan2.2-A14B camera-control checkpoint and generate 81 frames at 480×832 with the Flow-UniPC sampler, 50 steps, and a guidance scale of 6.0.
Open-Sora [51]. We use the v2.0 checkpoint and generate 129 frames at 256px (16:9) resolution with 50 rectified-flow steps, text guidance scale 7.5, image guidance scale 3.0, and 24 fps output.
LTX-Video [14]. We use the 13B (v0.9.8-dev) checkpoint and generate 129 frames at 480×832 with a two-pass multi-scale pipeline ($30 + 13$ steps) and dynamic guidance scales, at 25 fps.
CogVideoX [73]. We use the CogVideoX1.5-5B-I2V checkpoint and generate 81 frames at 480×832 with a DPM scheduler, 50 steps, a guidance scale of 6.0, and 16 fps.

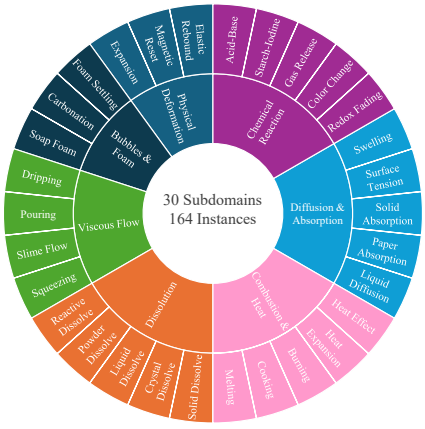
A.3 Dataset Statistics

The synthetic subset contains 196 clips spanning 14 scene subdomains (Fig. S1a) across five environment categories, featuring diverse target objects and action types. Sequences are typically 260–300 frames long, rendered at 1920×1080 (60 FPS).

The real-world subset contains 164 clips captured at 1920×1080, emphasizing diversity of physical-state changes: 30 common state-change processes grouped into seven major categories (Fig. S1b). Sequences range from 103–349 frames.
Camera trajectory statistics. Table S2 summarizes the camera-trajectory distribution of the two subsets. The real-world subset primarily contains controlled horizontal pans and vertical tilts, whereas the synthetic subset contains more diverse viewpoint changes, including U-turns, forward motion, head turns, and vertical motion. For each trajectory type, we report the number of clips, the mean total camera rotation, and the mean temporal gap between departure and reappearance. The trajectory counts sum to 164 real-world clips and 196 synthetic clips, matching the sizes of the two subsets.



(a) Synthetic: 14 scene subdomains across five environment categories.



(b) Real-world: 30 state-change processes across seven major categories.

Fig. S1: Dataset overview of MemoBench.

Table S2: Camera trajectory statistics of MemoBench. Rotation denotes the mean total camera rotation, and gap denotes the mean number of frames between departure and reappearance.

Subset	Trajectory	# Clips	Rotation (°)	Gap (frames)
Real-world	Pan L→R	56	72	59
Real-world	Pan R→L	51	63	65
Real-world	Tilt U→D	57	38	70
Synthetic	U-turn	68	178	113
Synthetic	Forward	58	92	81
Synthetic	Head turn	41	70	76
Synthetic	Vertical	29	27	77

The synthetic trajectories generally involve larger viewpoint changes and longer temporal gaps. In particular, U-turn sequences exhibit the largest mean camera rotation (178°) and the longest mean departure-to-reappearance gap (113 frames), providing a challenging setting for evaluating memory across substantial viewpoint changes.

A.4 More Dataset Examples

We provide additional dataset examples for both the synthetic and real-world subsets of MemoBench.

Synthetic Data. Figs. S2 and S3 show representative sequences from the synthetic scenes. Each row displays five uniformly sampled frames from a single

Fig. S2: Synthetic dataset examples (1/2). Representative scene from the synthetic subset of MemoBench, showing sampled frames across the V-D-R phases.

clip, covering the V-phase (target object visible), D-phase (camera departs), and R-phase (camera returns).

Real-World Data. Figs. S4 to S10 present examples from the real-world subset, organized by the seven state-change categories. During the D-phase, the camera pans away while the physical transformation occurs off-screen.

A.5 More Qualitative Results

Figs. S11 to S16 show additional qualitative comparisons. Each figure visualizes the camera trajectory alongside sampled frames from V, D, and R phases for a subset of models.

A.6 Additional Radar Visualizations

We provide two radar-plot visualizations to summarize VQA performance from complementary perspectives. Fig. S17a compares overall performance across the key evaluation dimensions, while Fig. S17b presents a fine-grained VQA-focused breakdown of model behavior.

Fig. S3: Synthetic dataset examples (2/2). Representative scene from the synthetic subset of MemoBench, showing sampled frames across the V-D-R phases.

Fig. S4: Real-world examples: Dissolution. Sampled V-D-R frames capturing dissolution processes such as salt dissolving or sugar melting, where the target object gradually loses its solid form during the camera's absence.

Fig. S5: Real-world examples: Combustion & Heat. Sampled V-D-R frames showing heat-driven state changes such as candle burning or paper burning, where the object's shape and material properties transform irreversibly.

Fig. S6: Real-world examples: Diffusion & Absorption. Sampled V-D-R frames depicting diffusion and absorption processes such as ink spreading in water or liquid soaking into fabric.

Fig. S7: Real-world examples: Chemical Reaction. Sampled V-D-R frames showing chemical reactions such as oxidation or effervescence, where the object undergoes compositional changes during the D-phase.

Fig. S8: Real-world examples: Viscous Flow. Sampled V-D-R frames capturing viscous flow processes such as pouring, dripping, and slime deformation, where fluid dynamics govern the state change.

Fig. S9: Real-world examples: Bubble & Foam. Sampled V-D-R frames showing foam settling, soap bubble evolution, and carbonation reactions, where transient structures form and collapse over time.

Fig. S10: Real-world examples: Physical Deformation. Sampled V-D-R frames depicting mechanical deformations such as crushing, tearing, or bending, where the object's geometry changes through applied force.

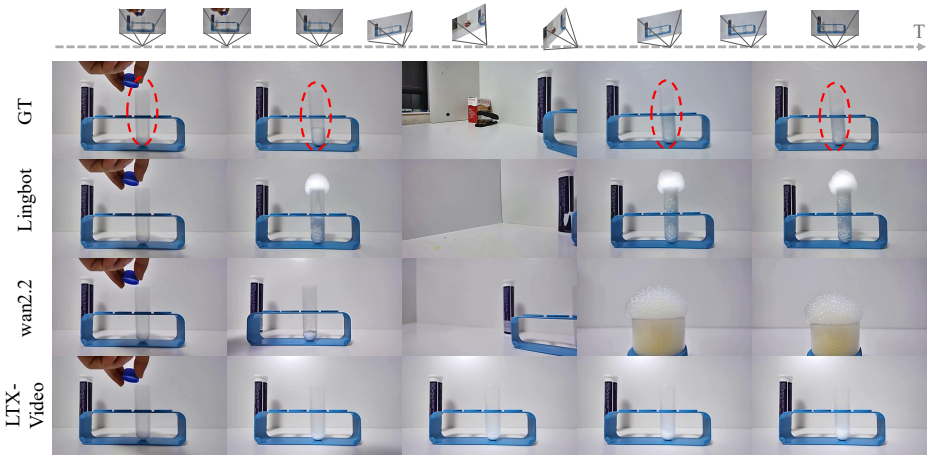


Fig. S11: Qualitative comparison on a real-world clip. The camera trajectory is shown above, with sampled frames from LingBot, Wan2.2, and LTX-Video.

Fig. S12: Qualitative comparison on a synthetic clip. The camera pans away from and returns to the scene, with sampled frames from LingBot, Wan2.2, and LTX-Video.

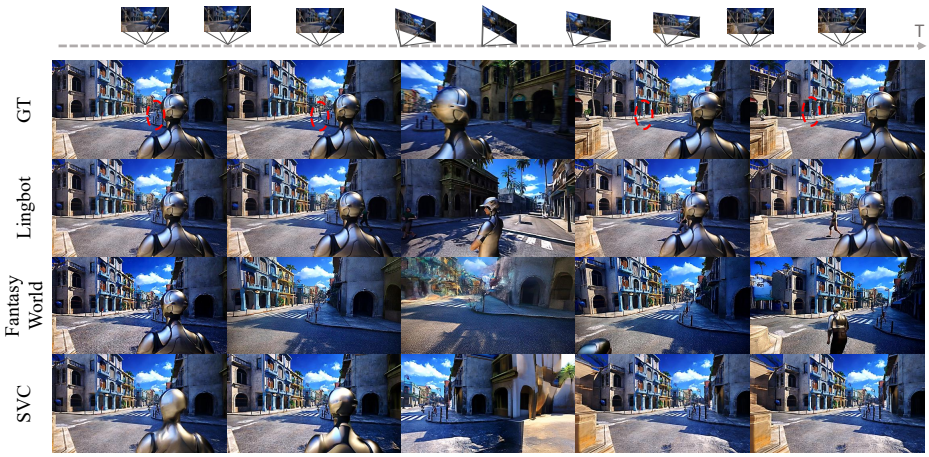


Fig. S13: Qualitative comparison on a synthetic clip. The camera trajectory is shown above, with sampled frames from LingBot, FantasyWorld, and SVC.

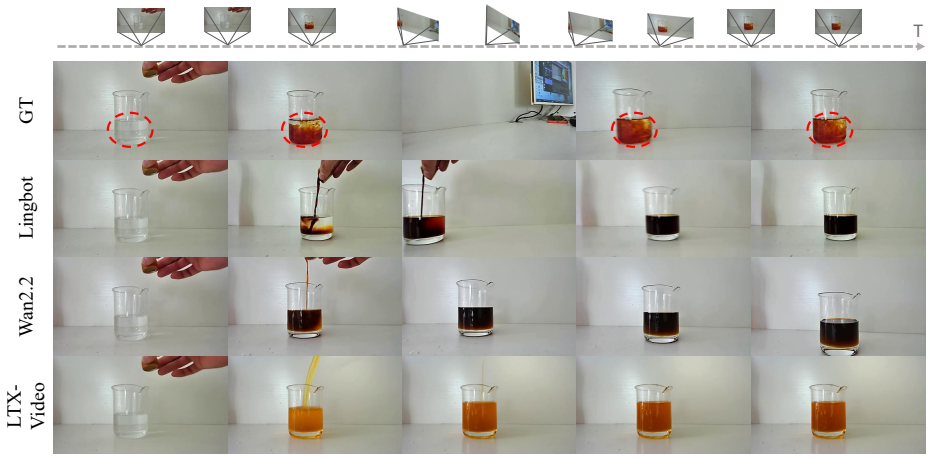


Fig. S14: Qualitative comparison on a real-world clip. The camera departs and returns while the physical state change progresses, with sampled frames from LingBot, Wan2.2, and LTX-Video.

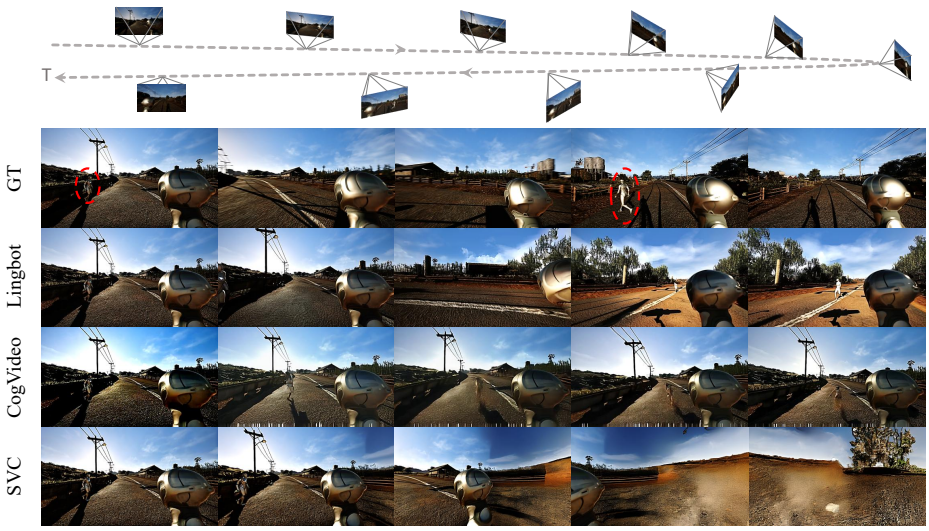
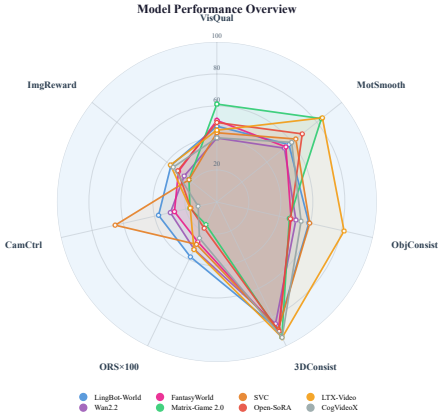
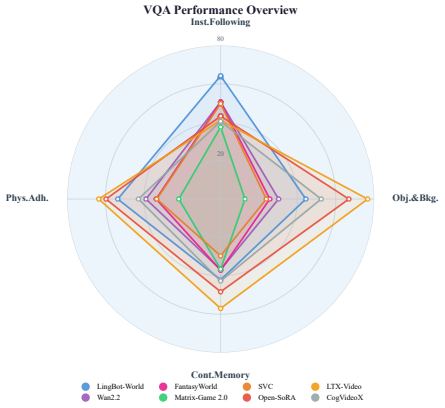


Fig. S15: Qualitative comparison on a synthetic clip. The camera trajectory is shown above, with sampled frames from LingBot, CogVideoX, and SVC below.

Fig. S16: Qualitative comparison on a real-world clip. The camera departs and returns while the physical state change progresses, with sampled frames from LingBot, FantasyWorld, and SVC.



(a) Overall radar comparison. Per-model scores across the main evaluation dimensions.



(b) Fine-grained VQA radar analysis. Per-model breakdown across four VQA dimensions.

Fig. S17: Radar visualizations for detailed VQA evaluation. The two panels provide complementary summaries: a high-level cross-dimension view and a fine-grained VQA-focused breakdown.

A.7 Ablation Studies

We conduct six ablation and diagnostic studies to validate the robustness of our evaluation metrics and the necessity of our design choices. These studies use the models and subsets specified in each subsection and are intended as controlled diagnostics rather than a reproduction of the complete ten-model leaderboard.

ORS Robustness Analysis The Object Revisit Score (ORS) relies on SAM-3 text-prompted segmentation to detect whether a target object reappears in the R-phase. We verify that ORS is stable under perturbations to (a) the coverage-filtering thresholds used to discard spurious masks, and (b) the text prompt formulation.

Coverage threshold sweep. ORS filters SAM-3 masks by image-area coverage $[\text{cov}_{\min}, \text{cov}_{\max}]$ to exclude noise (tiny masks) and background (overly large masks). We sweep eight threshold configurations on 30 clips for two representative models. Table S3 reports the results: for LingBot-World the mean ORS varies by only 0.013 across all configurations (0.456–0.469), and for StableVirtualCamera by 0.020 (0.325–0.345).

Prompt variation sweep. We test five prompt formulations: the original subject phrase extracted from the scene description, and four rephrasings. Table S4 shows that semantically equivalent prompts (*original* vs. *the* ⟨*subject*⟩) yield nearly identical ORS, while truncated prompts (*short*: first word only) degrade substantially. This gap indicates that ORS captures semantic object identity rather than low-level pattern matching.

Stratification by object size and reappearance angle. We further stratify ORS by GT mask area (small $< 2\%$, medium 2–10%, large $> 10\%$) and by camera rotation at the reappearance frame. Larger objects yield higher ORS (LingBot-World: large 0.67, medium 0.45, small 0.31), consistent with the intuition that larger targets are easier for the model to regenerate faithfully. Objects reappearing after extreme rotations ($> 120°$) yield near-zero ORS for StableVirtualCamera (0.005), which we attribute to generation failures at large viewpoint changes rather than detector limitations, since SAM-3 achieves a 100% detection rate on all clips.

Motion-Gated Evaluation A potential confound in camera-controllable generation benchmarks is *camera inactivity*: a model that ignores the camera trajectory and produces a near-static video may receive artificially high pixel-fidelity scores. To disentangle generation quality from camera compliance, we re-evaluate all metrics on subsets filtered by the total GT camera rotation magnitude.

Table S5 reports results at the $\geq 90°$ threshold (80 clips per model). The most notable finding is the trade-off between camera tracking and object permanence: StableVirtualCamera reaches 92.43 Camera Controllability yet drops to 0.012

Table S3: ORS sensitivity to coverage thresholds. Mean ORS (± std) across 30 clips under different mask-coverage filter ranges $[\text{cov}_{\min}, \text{cov}_{\max}]$. The default setting is highlighted.

$\text{cov}_{\min}$ (%)	$\text{cov}_{\max}$ (%)	LingBot-World	SVC
0.01	30	0.469 ± 0.391	0.345 ± 0.404
0.03	40	0.462 ± 0.396	0.345 ± 0.404
0.05	50	0.462 ± 0.396	0.345 ± 0.404
0.05	60	0.462 ± 0.396	0.345 ± 0.404
0.05	70	0.462 ± 0.396	0.345 ± 0.404
0.10	50	0.462 ± 0.396	0.326 ± 0.411
0.10	70	0.462 ± 0.396	0.326 ± 0.411
0.20	50	0.456 ± 0.401	0.325 ± 0.412

Table S4: ORS sensitivity to text prompt formulation. Mean ORS (± std) across 30 clips under different prompt templates for SAM-3.

Prompt style	LingBot-World	SVC
Original subject phrase	0.462 ± 0.396	0.345 ± 0.404
"the ⟨subject⟩"	0.459 ± 0.366	0.328 ± 0.406
"detect ⟨subject⟩ in the scene"	0.284 ± 0.347	0.241 ± 0.313
"a photo of ⟨subject⟩"	0.168 ± 0.304	0.122 ± 0.282
First word only	0.030 ± 0.029	0.015 ± 0.011

ORS, while LingBot-World balances both axes (CamCtrl 75.04, ORS 0.281). This trade-off is invisible in the aggregate evaluation and can only be surfaced by jointly reporting both metrics under controlled camera motion, underscoring the need for MemoBench's multi-dimensional protocol. I2V models (CogVideoX, LTX-Video) show stable scores across thresholds, as expected for models that do not condition on camera poses.

Per-Phase Fidelity Breakdown The V-D-R paradigm enables phase-aware evaluation. We separately compute pixel-fidelity metrics (PSNR, SSIM, LPIPS) for the V-phase (Visible) and R-phase (Reappear) and report the fidelity drop Δ upon object reappearance.

Table S6 reveals consistent R-phase degradation across all eight models. LingBot-World suffers the largest PSNR drop ($\Delta = 5.24\,\text{dB}$) despite having the highest V-phase fidelity among CI2V models, suggesting that its generative prior does not maintain coherence across the occlusion gap. Matrix-Game2 exhibits the steepest perceptual degradation ($\Delta\text{SSIM} = 0.239$, $\Delta\text{LPIPS} = 0.302$). These phase-level differences would be masked in an aggregate fidelity score, motivating the per-phase breakdown in our protocol.

Table S5: Motion-gated evaluation on clips with $\geq 90^\circ$ total GT camera rotation.

Model	n	CamCtrl ↑	ORS ↑	PSNR ↑	SSIM ↑	ObjConsist ↑	3DConsist ↑
SVC	80	**92.43**	0.012	12.56	0.34	38.33	91.42
LingBot-World	80	75.04	**0.281**	12.12	0.33	50.49	88.60
Wan2.2	80	65.02	0.057	11.05	0.27	22.29	83.78
LTX-Video	80	51.08	0.248	12.13	0.33	**84.50**	91.79
Open-SoRA	80	50.49	0.093	12.81	0.31	65.56	92.31
Matrix-Game2	80	46.81	0.008	12.49	0.27	39.99	**95.26**
CogVideoX	45	83.32	0.211	11.05	0.31	67.62	95.48
FantasyWorld	80	70.81	0.062	11.43	0.27	30.92	85.98

Table S6: Per-phase pixel fidelity breakdown. V = Visible phase, R = Reappear phase. Δ denotes the fidelity drop upon reappearance (positive = degradation).

	PSNR ↑			SSIM ↑			LPIPS ↓		
Model	V	R	Δ	V	R	Δ	V	R	Δ
SVC	18.68	14.56	+4.12	0.632	0.483	+0.148	0.295	0.473	+0.178
LingBot-World	17.96	12.72	+5.24	0.611	0.420	+0.191	0.309	0.545	+0.236
Matrix-Game2	17.02	11.90	+5.12	0.527	0.289	+0.239	0.335	0.637	+0.302
Wan2.2	16.20	12.81	+3.38	0.552	0.421	+0.131	0.374	0.581	+0.207
FantasyWorld	16.09	11.93	+4.16	0.525	0.376	+0.149	0.406	0.628	+0.222
LTX-Video	15.91	12.74	+3.16	0.543	0.420	+0.123	0.363	0.543	+0.180
Open-SoRA	15.23	11.00	+4.23	0.467	0.339	+0.128	0.401	0.658	+0.257
CogVideoX	14.18	11.17	+3.01	0.555	0.441	+0.114	0.408	0.672	+0.264

Metric Sensitivity Analysis Each metric in our pipeline depends on hyperparameters (*e.g.*, the fraction of DINOv2 patch tokens retained, the optical-flow outlier threshold, or the depth-sampling density). We sweep these parameters and measure Kendall's τ rank correlation against the default configuration over 200 clip–model pairs to verify that model rankings are not artifacts of a particular parameter choice.

Table S7 reports the results. RAFT Motion Smoothness is perfectly rank-preserving ($\tau = 1.000$) across the entire threshold range $[0.05, 0.30]$, indicating that the outlier threshold affects absolute scores but not relative ordering. DINOv2 Object Identity Consistency maintains $\tau \geq 0.910$ even when the top-k fraction varies from 0.2 to 0.8. Depth Anything V2 Geo3D Consistency is the most sensitive of the three, yet still achieves $\tau \geq 0.860$ across all sampling densities. All correlations are statistically significant ($p < 10^{-4}$).

Camera Pose Estimation Validation Camera Controllability is derived from the Absolute Trajectory Error (ATE) between MapAnything-estimated and GT

Table S7: Metric sensitivity analysis. Kendall's τ rank correlation between default and variant hyperparameter settings.

Metric	Parameter variant	Kendall's τ
DINOv2 ObjConsist	top_k = 0.2	0.947
	top_k = 0.3	0.976
	top_k = 0.4 (default)	1.000
	top_k = 0.5	0.978
	top_k = 0.6	0.958
	top_k = 0.8	0.910
RAFT MotSmooth	τ = 0.05	1.000
	τ = 0.10	1.000
	τ = 0.15 (default)	1.000
	τ = 0.20	1.000
	τ = 0.30	1.000
DepthV2 Geo3D	n_sample = 2	0.873
	n_sample = 3 (default)	1.000
	n_sample = 5	0.900
	n_sample = 7	0.860

camera poses. We perform two sanity checks: (1) verifying GT pose integrity, and (2) examining whether the ATE distribution across models is consistent with their architectural priors. For (1), we recompute the total rotation from raw GT `poses.npy` files and compare against the values stored in the evaluation CSVs. Across all 159 synthetic clips the Pearson correlation is $r = 1.0000$ with a maximum absolute difference below 0.01°. For (2), Table S8 reports the per-model ATE distribution on synthetic data. StableVirtualCamera, which directly conditions on target camera extrinsics, achieves the lowest ATE ($8.3° \pm 6.6°$). LingBot-World ranks second (20.7°). I2V models that receive no camera input (CogVideoX, LTX-Video, Open-SoRA) cluster near 63–65°, consistent with near-random camera trajectories. The Spearman correlation between ATE and Camera Controllability is moderate ($\rho = -0.355$); the non-linear mapping from ATE to the coverage-based CamCtrl score accounts for this gap, since CamCtrl saturates once the ATE falls below the coverage radius.

Initial-State Conditioning vs. Backbone Capacity To disentangle the effect of initial-state conditioning from backbone capacity, we conduct a controlled Wan2.2 ablation on 50 clips. The *with V-frame* setting provides the first frame of the V phase as an image condition together with the text prompt and camera trajectory. The *without V-frame* setting removes this image condition and uses only text and camera conditioning.

Table S8: Camera pose estimation validation. Per-model ATE rotation RMSE (degrees) on synthetic data, sorted by ATE. CamCtrl is reported on a 0–100 scale.

Model	n	ATE (°) ↓	CamCtrl ↑	GT Rot (°)
SVC	159	8.3 ± 6.6	86.10	93.6
LingBot-World	159	20.7 ± 24.1	93.06	93.6
FantasyWorld	159	34.4 ± 24.5	52.76	93.6
Wan2.2	159	44.9 ± 36.0	66.15	93.6
Matrix-Game2	159	58.1 ± 30.6	32.52	93.6
Open-SoRA	159	63.0 ± 38.9	32.77	93.6
LTX-Video	159	63.6 ± 39.1	32.55	93.6
CogVideoX	86	64.5 ± 38.3	31.15	95.7

Table S9: Wan2.2 initial-state-conditioning and backbone-capacity ablation on 50 clips. "V-frame" denotes the first frame of the visible phase provided as an image condition.

Variant	PSNR↑	LPIPS↓	ORS↑	ObjConsist↑	MotSmooth↑	3DConsist↑
5B w/o V-frame	8.70	0.85	0.27	47.1	61.7	90.7
5B w/ V-frame	12.9	0.65	0.36	49.2	65.5	88.7
14B w/o V-frame	8.90	0.74	**0.53**	**80.7**	**79.7**	**97.0**
14B w/ V-frame	**13.6**	**0.58**	0.21	45.6	57.3	83.7

As shown in Table S9, providing the V-phase frame improves GT-aligned fidelity more substantially than scaling the backbone from 5B to 14B. Adding the V-frame improves PSNR by 4.2 dB for the 5B model and 4.7 dB for the 14B model, while also reducing LPIPS by 0.20 and 0.16, respectively. In comparison, scaling the backbone from 5B to 14B produces substantially smaller improvements under matched conditioning.

Interestingly, the 14B model without the V-frame achieves the highest ORS, Object Consistency, Motion Smoothness, and Geo3D Consistency, despite its substantially lower GT-aligned fidelity. This result shows that internally self-consistent generation does not necessarily recover the correct post-occlusion state, motivating the joint use of GT-aligned fidelity and self-consistency metrics.

A.8 Detailed VQA Pipeline

We present the three-stage VQA pipeline used in our evaluation. Each stage is illustrated with its prompt template, followed by concrete filtering examples.

Stage 1: LLM Question Generation. Given the start frame and generation prompt, we ask an LLM to generate candidate Yes/No questions (six per dimension).

`Question Generation Prompt`

System Role:
You are an expert LLM judger, specializing in "World Model" evaluation. Your task is to generate questions used to evaluate AI-generated videos against text instructions.
Input Data:
Start Frame (Ground Truth) as attached image

Generation Prompt: `{generation_prompt}`

Task:
Generate 24 Yes/No questions.
Evaluation Dimensions & Constraints:
Generate six questions for each of the four dimensions below. We use **mixed polarity**, meaning no dimension has a fixed yes/no preference.
- **Instruction Following**: Did the video follow the requested movements and events?
- **Object and Background**: Is there inconsistency in subject identity or background details?
- **Continuity of Memory**: Does the model preserve object location/trajectory while out of frame?
- **Physics Adherence**: Are lighting, shadows, and motion physically plausible?

Output Format:
Output JSON with columns: [ID, Dimension, Question].

Stage 2: Question Filtering. Candidate questions are filtered using both ground-truth and failure-case references.

`GT & Failure Filtering Prompt`

System Role:
You are an expert LLM judger, specializing in "World Model" evaluation. Your task is to audit AI-generated videos against specific text instructions.
Input Data:
Test Video as attached video
Questions: `{questions}`
Task:
Watch the test video and answer each Yes/No question.
Output Format:
Output JSON with columns: [ID, Dimension, Question, Answer (Yes/No), Verdict (Pass/Fail), Reasoning].

`Revised Failure Filtering Prompt (with Hint)`

System Role:
You are an expert LLM judger, specializing in "World Model" evaluation. Your task is to audit AI-generated videos against specific text instructions and known failure hints.
Input Data:
Test Video as attached video
Questions: `{questions}`
Hint: `{hint}`
Tasks:
1. Answer the Yes/No questions for the test video.
2. Audit these answers against known failures and remove unstable questions.
Output Format:
Output JSON with columns: [ID, Dimension, Question, Answer (Yes/No), Verdict (Pass/Fail), Reasoning].

Example Legend In the examples below, ✓ denotes a question retained after filtering; ✗ Failure denotes a question removed due to instability under failure-case checking; and ✗ GT denotes a question removed because it is inconsistent with the GT reference.
Example: Nordic #001 (5/8 questions remain after filtering).

– *Instruction Following* — (1) Does the observer re-encounter the subject after completing the U-turn? ✓ (2) Does the observer execute a U-turn after the subject has exited the field of view? ✗ Failure
– *Object & Background* — (1) Does the subject maintain its silver robotic appearance throughout? ✓ (2) Does the Nordic architecture maintain its structural details during camera rotation? ✓
– *Continuity of Memory* — (1) Does the street layout change its configuration after the observer turns around? ✓ (2) Is the subject in a logically consistent position when the observer turns back? ✗ Failure
– *Physics Adherence* — (1) Do shadows move realistically as the observer changes perspective? ✓ (2) Does the subject's walking speed remain consistent and natural? ✗ Failure

Example: Zen Garden #005 (4/8 questions remain after filtering).
– *Instruction Following* — (1) Does the observer successfully perform a U-turn and see the person again? ✓ (2) Does the person continue moving consistently after the fox completes the turn? ✗ Failure
– *Object & Background* — (1) Is the Zen Garden aesthetic and landscape preserved throughout? ✓ (2) Does the ground texture flicker or disappear as the fox runs? ✗ GT
– *Continuity of Memory* — (1) Is the character's walking animation continuous even off-focus? ✓ (2) Does the person reappear at a location consistent with their original trajectory? ✗ Failure
– *Physics Adherence* — (1) Do shadows cast by trees remain in a fixed orientation relative to the sun? ✓ (2) Does the foliage jitter unnaturally as the camera moves past it? ✗ GT

Stage 3: Final Evaluation. The filtered question bank is applied to each test video. A VLM answers the remaining questions, and per-dimension scores are aggregated from binary verdicts.

Judge Prompt

System Role:
You are an expert judge specialized in evaluating answer correctness for world-model outputs.
Input Data:
Test Video as attached video (generated output)
Evaluation Questions: `{filtered_questions}`
Task:
Watch the test video and answer each question with a final Yes/No decision and reasoning.
Output Format:
Output JSON with columns: [ID, Dimension, Question, Answer (Yes/No), Verdict (Pass/Fail), Reasoning].

Table S10: Failure taxonomy for LingBot-World. The reported counts are non-exclusive, since one generated sequence may exhibit multiple failure types.

Split	Obj. vanish	Id. drift	State reset	Teleport	Bkg. halluc.	Cam. drift
Synthetic	85	49	5	54	131	72
Real	15	83	2	13	57	35

A.9 Failure Analysis

We further analyze the failure modes of LingBot-World on the synthetic and real-world subsets of MemoBench. We categorize the observed failures into six types: object disappearance, identity drift, state reset, teleportation, background hallucination, and camera drift. The categories are non-exclusive, and a single generated sequence may exhibit multiple failure types. This analysis is intended as a case study of recurring model failures rather than a comparison across different models.

The two subsets exhibit different dominant failure patterns. Background hallucination is the most frequent failure on the synthetic subset, followed by object disappearance and camera drift. In contrast, identity drift is the most frequent failure on the real-world subset, where the target object may undergo a physical-state change while outside the field of view. These results indicate that memory failures extend beyond object disappearance and also affect object identity, scene layout, physical state, and camera execution.

The failure categories are reflected by complementary components of our evaluation protocol. Object disappearance is primarily captured by ORS; identity drift by Object Consistency and Object & Background VQA; state reset by R-phase GT-aligned fidelity and clip-specific VQA; teleportation by Motion Smoothness and VQA; background hallucination by whole-frame fidelity, 3D Consistency, and Object & Background VQA; and camera drift by Camera Controllability.