GameCraft-2: Instruction-following Interactive Game World Model

Anonymous CVPR submission

Paper ID 2

Figure 1. ***GameCraft-2*** advances generative game world models from static scene synthesis to open-ended, instruction-following simulation. From a single image, we simulate action signals; left/right frames show key moments under different inputs. GameCraft-2 generates interaction-aligned videos with temporal/3D consistency. Instructions (e.g., *"draw a torch"*, *"hold a gun and fire"*) and key/mouse actions control camera motion and content updates, producing coherent videos with consistent style. W/A/S/D indicate movement, ↑, ←, ↓, → adjust view angles, and white denotes idle.

Abstract

*Recent advances in generative world models have enabled remarkable progress in creating open-ended game environments, evolving from static scene synthesis toward dynamic, interactive simulation. However, current approaches remain limited by rigid action schemas and high annotation costs, restricting their ability to model diverse in-game interactions and player-driven dynamics. To address these challenges, we introduce **GameCraft-2**, a new paradigm of instruction-driven interaction for generative game world modeling. Instead of relying on fixed keyboard inputs, our model allows users to control game video contents through natural language prompts, keyboard, or mouse signals, enabling flexible and semantically rich interaction within generated worlds. We formally define the concept of Interactive Video Data and develop an automated pipeline that converts large-scale, unstructured text–video pairs into causally aligned interactive datasets. Built upon a 14B image-to-video Mixture-of-Experts (MoE) foundation model, our model incorporates a text-driven interaction injection mechanism for fine-grained control over camera motion, character behavior, and environment dynamics. We introduce an interaction-focused benchmark, InterBench to evaluate interaction performance comprehensively. Extensive experiments demonstrate that our model generates temporally coherent, and causally grounded interactive game videos that faithfully respond to diverse and free-form user instructions such as "open the door", "draw a torch", or "trigger an explosion".*

1. Introduction

The rapid progress of diffusion models [18, 27, 31, 41, 44, 51] has accelerated dynamic game content creation [17, 28, 64]. Beyond static images and short videos, recent systems such as RTFM [55] and the Genie series [1] demonstrate that world models can underpin immersive, controllable virtual experiences, advancing AI-driven **"playable worlds"** that simulate and respond to user intent.

Existing world models can be categorized into ***3D-based*** and ***video-based*** approaches. 3D-based world models [22, 30, 37, 47, 54, 55] emphasize geometric consistency and physical accuracy, enabling detailed world reconstruction and memory persistence. However, they are often limited to scripted or static interactions, lacking the creative flexibility and open-ended gameplay dynamics essential for interactive game environments. With recent improvements

in video foundation models [3, 27, 51, 53], the video-based technical pathway [1, 11, 17, 28, 40, 56, 61, 64] has shown remarkable potential. These works learn world dynamics directly from large-scale video data [6, 11, 32, 66] through implicit end-to-end representation learning. Notably, the Genie series [1, 40] introduces latent action modeling to simulate player-driven interactions, while Matrix-Game [64] and Hunyuan-GameCraft [28] integrate discrete actions (e.g., W/A/S/D, mouse movements) into a unified representation space, achieving continuous, high-fidelity video generation that responds to user inputs.

These frontier works mark a fundamental shift in focus from the world's static appearance ("what the world looks like") to its interactive dynamics ("how we interact with it"), necessitating a rigorous definition of "interaction" within world models, especially in game scenarios.

We formally define interaction in world models as "*actions executed by an explicit agent that trigger state transitions in the environment with clear causal relationships and physical or logical validity.*" This definition encompasses diverse input modalities, from mouse and keyboard operations [11, 28, 56, 61, 64] to embodied motion sensing [38]. Grounded in this perspective, two key challenges hinder this progress: (1) the lack of a formal definition and a scalable construction pipeline for interactive video data, and (2) multi-turn interactions in long video generation while maintaining video quality and interaction accuracy.

To address these challenges, we present **GameCraft-2**, an interactive game world model for free-form instruction-following control. We begin by formally defining interaction within the context of generative world models, and develop two automated pipelines for interactive video data construction and refinement. These pipelines, for the first time, enable the efficient transformation of large-scale, unstructured text–video pairs into open-domain interactive datasets enriched with implicit causal labels.

For model training, our model integrates text-based instructions and keyboard/mouse action signals into a unified controllable video generator, enabling flexible, semantically grounded, and causally consistent interaction within dynamic game environments. To support efficient long-horizon video generation, we employ a comprehensive autoregressive distillation strategy that transfers the bidirectional video generator into a causal autoregressive model. Subsequently, a randomized image-to-long-video extension tuning scheme is introduced to alleviate error accumulation during extended rollouts, ensuring stable and coherent long-form generation. For multi-turn interactive inference, we following LongLive [57] to employ a KV-recache mechanism to enhance the accuracy and stability of multi-turn interactions in autoregressive long video generation. In addition, we incorporate several engineering acceleration optimizations, boosting the model's inference speed to 16 FPS on 8 NVIDIA H20 GPUs, enabling real-time interactive video generation.

To comprehensively evaluate interactive performance across models, we introduce InterBench, a benchmark that measures key dimensions of interactive behavior: interaction completeness, action effectiveness, causal coherence, and physical plausibility. Extensive experiments on InterBench demonstrate the effectiveness of our framework, achieving SOTA performance in generating interactive videos that respond faithfully to user instructions while maintaining high visual fidelity and temporal coherence.

In general, our main contributions are as follows:

- We propose a **Unified Interaction Framework** and scalable data engine that unifies keyboard/mouse signals with real-time **text instructions**. We then post-train **Wan2.2 14B MoE** on automatically generated/annotated/cleaned data for strong controllability and **emergent generalization** to unseen action combinations.
- We leverage **autoregressive distillation** and **randomized long-video tuning** for efficient and stable long-horizon generation, and introduce **KV-recache** for multi-turn inference with real-time **16 FPS** performance.
- We establish **InterBench**, a benchmark that combines VLM-based evaluation with classical video metrics to quantify both visual fidelity and **instruction-following accuracy** in dynamic interactions.

2. Related Works

2.1. Long Video Extension

Long video generation primarily struggles with "train-short-test-long" discrepancies, leading to semantic drift and artifacts. To mitigate this, Self-Forcing [23] and its follower [9, 35, 39, 57] align training with inference via self-conditioning and rolling windows. Alternative paradigms explore next-frame prediction [12, 13], hybrid diffusion-autoregressive models like DiffusionForcing [7], or test-time adaptation [10]. Recently, LongLive [57] introduced KV-recache to improve interactive semantic control responsiveness.

2.2. Interactive Video-based World Model

Unlike static video models, interactive world models respond dynamically to inputs for explorable environments. Early game-specific models like MineWorld [15], Matrix [11], and GameFactory [61] focused on discrete actions (e.g., keyboard/mouse). To enhance generalization and consistency, Genie2 [40] proposed a 2D foundation model, while WorldMem [56] utilized memory banks for long-term simulation. Recent works such as GameGen-X [6], Genie3 [1], and Hunyuan-GameCraft [28] unify control signals into shared continuous spaces. However, these models mainly use prompts for initialization rather than real-time interaction, which is still limited by physical inputs.

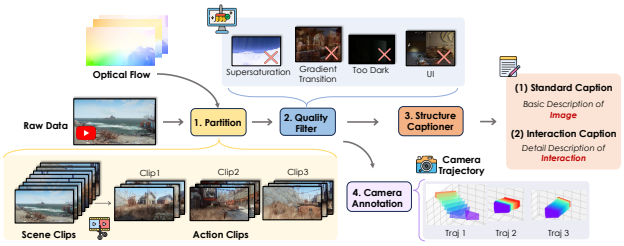

Figure 2. **Data Curation Pipeline.** (1) **Partition:** segment raw videos into scene/action clips via scene detection and optical flow; (2) **Quality Filtering:** discard low-quality frames using visual, luminance checks; (3) **Structured Captioning:** generate standard and interaction captions per clip; (4) **Camera Annotation:** recover 6-DoF camera trajectories to model viewpoint motion.

2.3. Text-guided Video Generation and Editing

Text-driven synthesis follows two paradigms: semantic enrichment and structured execution. The former enhances prompts via LLM representations or adapters [36, 46, 63, 65]. The latter treats text as a script, using LLMs to decompose prompts into frame-level descriptions or layouts [19, 20, 33, 34]. Related editing works leverage video-to-video frameworks for zero-shot or end-to-end control [8, 25, 29, 42, 43]. Unlike these one-off, non-interactive methods, GameCraft-2 enables continuous interaction where prompts iteratively evolve a dynamic world state through feedback loops.

3. Interactive Video Data Construction

3.1. Definition of Interactive Video Data

Unlike traditional video, **Interactive Video Data** is defined as a temporal sequence recording a *causally driven state-transition process*. It captures how agents or environments move from a defined **initial state** to a distinct **final state**. We categorize these interactions into: **(1) Environmental Interactions** (scene changes), **(2) Actor Actions** (embodied agent-driven), and **(3) Entity Appearances** (subject introduction). Each includes varying complexity levels.

3.2. Controllable Synthetic Data Production

To address data scarcity, we develop a synthetic pipeline leveraging foundation models to generate high-quality interactive transitions. Given an initial frame F_t (which can be synthesized via T2I models to ensure diversity), we employ two strategies based on the interaction type:

- **Start-End Frame Strategy:** For stationary scenes (e.g., weather changes), a VLM-guided editing model generates a target end-frame F'_t, ensuring strong controllability over the final state.
- **First-Frame-Driven Strategy:** For dynamic actions (e.g., opening a door), the model generates sequences *from* F_t to ensure temporal continuity.

3.3. Game Scene Curation and Annotation

For raw videos, we implement a robust curation pipeline:

- **Action-aware Partition & Filtering:** We split videos into action-aware clips and perform a three-stage quality screen (fidelity, luminance, semantic consistency) to discard noisy or inconsistent clips.
- **Camera Annotation:** We reconstruct 6-DoF camera trajectories for each clip to provide precise translational and rotational metadata for viewpoint modeling.
- **Structured Dual-Captioning:** To supervise both states and transitions, we generate a *Standard Caption* (C_t) for static visual layout and an *Interaction Caption* ($I_{t\to t+1}$) capturing the semantic delta between clips. This enables the model to jointly learn appearance perception and action-level reasoning.

4. Method

We present **GameCraft-2**, an interactive game video model for free-form instruction-based control (Fig. 3). It unifies an action-injected causal architecture, image-conditioned autoregressive long-video generation, and multi-prompt interaction within a single framework. We then describe the architecture, training, and inference procedures.

4.1. Model Architecture

Our model builds on a 14B image-to-video MoE diffusion foundation model [51], which we extend into an action-controllable generator. As in Sec. 1, actions comprise both keyboard inputs and free-form text prompts.

For keyboard/mouse control (e.g., W/A/S/D, arrows), we map discrete inputs to continuous camera parameters following Hunyuan-GameCraft [28]. We encode annotated parameters as Plücker embeddings [16] and inject them via token addition during training; at inference, user inputs are converted into camera trajectories to derive the parameters.

For prompt-based interaction injection, we find the base model underperforms on interactive verbs due to their higher semantic/spatial complexity and tight grounding to specific regions or instances. We therefore use an MLLM [52] to extract, reason about, and inject interaction cues, strengthening fine-grained guidance and helping distinguish scene-level instructions from interactive behaviors during training. Combined with camera-conditioned control, this yields a unified mechanism for coherent navigation and interaction.

4.2. Training Procedure

To enable long-horizon, real-time interactive generation, we distill a bidirectional foundation model into a few-step causal generator. Specifically, we scale Self-Forcing [23] to a 14B MoE image-to-video model [51], improving quality and efficiency under rapid scene changes. To reduce error accumulation, we propose random extension tuning. Training

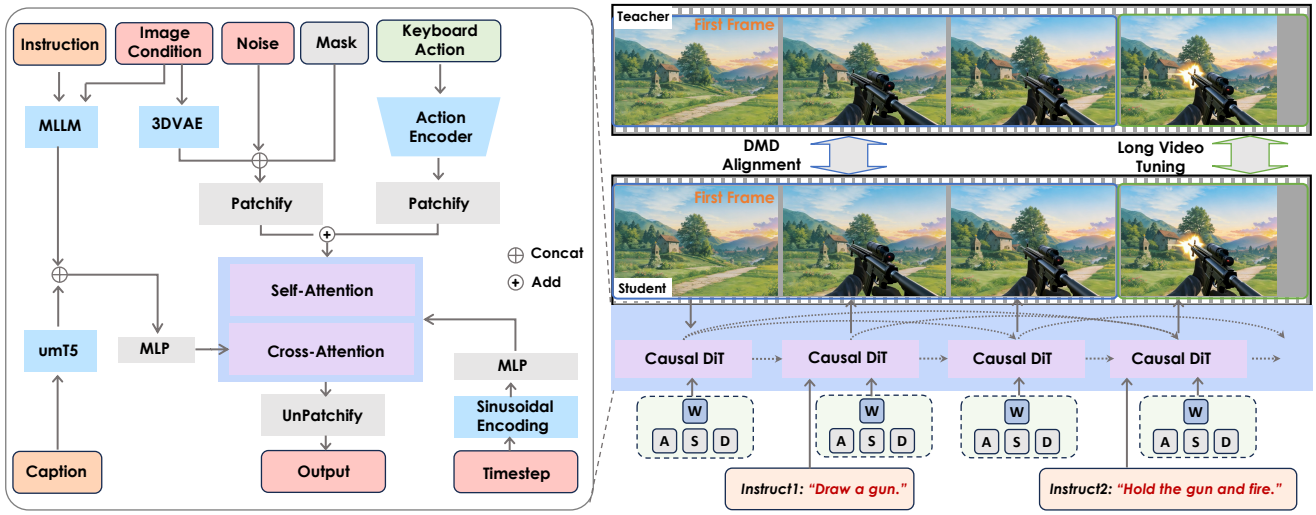

Figure 3. Model architecture of GameCraft-2. Given a reference image and the corresponding action, the keyboard/mouse signal, and prompt-based instruction, we inject these options to the main architecture. During training and inference, we leverage self-forcing post-training for long-video extension, and KV-cache/recache for multi-action switching. To maintain the long-term video quality, we design a randomized long video tuning scheme.

proceeds in four stages: (1) Action-Injected Training, (2) Instruction-Oriented SFT, (3) Autoregressive Distillation, and (4) Randomized Long-Video Extension Tuning.

4.2.1. Action-Injected Training

The primary objective of this stage is to establish a fundamental understanding of 3D scene dynamics, lighting, and physics. We load the pre-trained weights and finetune the model with the flow-matching objective for architectural adaptation. In order to improve the long-term consistency, we adopted a curriculum learning strategy. Specifically, we organized the training into three phases, exposing the model to video data of 45, 81, and 149 frames in 480p in sequence. This stepped approach allows the model to first solidify its understanding of short-term motion dynamics before gradually adapting its attention mechanisms to handle the complex dependencies required for longer-duration coherence. Besides, we randomly choose long and short captions during training, and concatenate interactive captions for interaction learning. This option will help the model to have an initial perception of the injection of interactive information.

4.2.2. Instruction-Oriented Supervised Fine-Tuning

To enhance interactivity, we built a 150K-sample dataset by augmenting real-world footage with procedurally generated synthetic videos. These sequences provide high-fidelity supervision across diverse interactions, establishing a tight correspondence between actions and visual outcomes. Next, we freeze the camera encoder and fine-tune only the MoE experts to improve alignment with semantic control cues.

4.2.3. Autoregressive Generator Distillation

For interactive world models, extending fixed-length video generators to high-quality autoregressive long-video generation is essential. Prior works have made preliminary attempts on long video generation [23, 45, 50, 57]. Building upon the high- and low-noise MoE architecture and camera parameter injection, we introduce targeted adaptations to the attention mechanism and the distillation protocol. These modifications are specifically tailored to optimize performance within the autoregressive distillation process.

Sink Token and Block Sparse Attention: Sliding-window KV updates in prior causal attention [23, 60] can cause drift as later tokens lose access to the initial conditioning frame. Following [45, 50, 57], we keep the first frame as a *sink token* in the KV cache, improving long-horizon stability and preserving the coordinate origin so injected camera parameters remain aligned, avoiding per-step recaching. We further use Block Sparse Attention [14] for local attention, where each target block attends to selected preceding blocks. Sink + block-sparse local attention together constitute the KV cache, boosting quality and speed for block-wise generation.

Distillation Schedule: Since the high-noise MoE expert is harder to train [51], especially under SFT/distillation, we use expert-specific learning rates. We also set timestep targets by the inter-expert noise boundary, ensuring teacher and student choose the same expert throughout distillation.

4.2.4. Randomized Extended Long-Video Tuning

Our approach to enabling long-form video generation is motivated by the observation that the foundation model, despite being pre-trained on short clips, implicitly captures the global visual data distribution. Previous methods [9, 57], roll

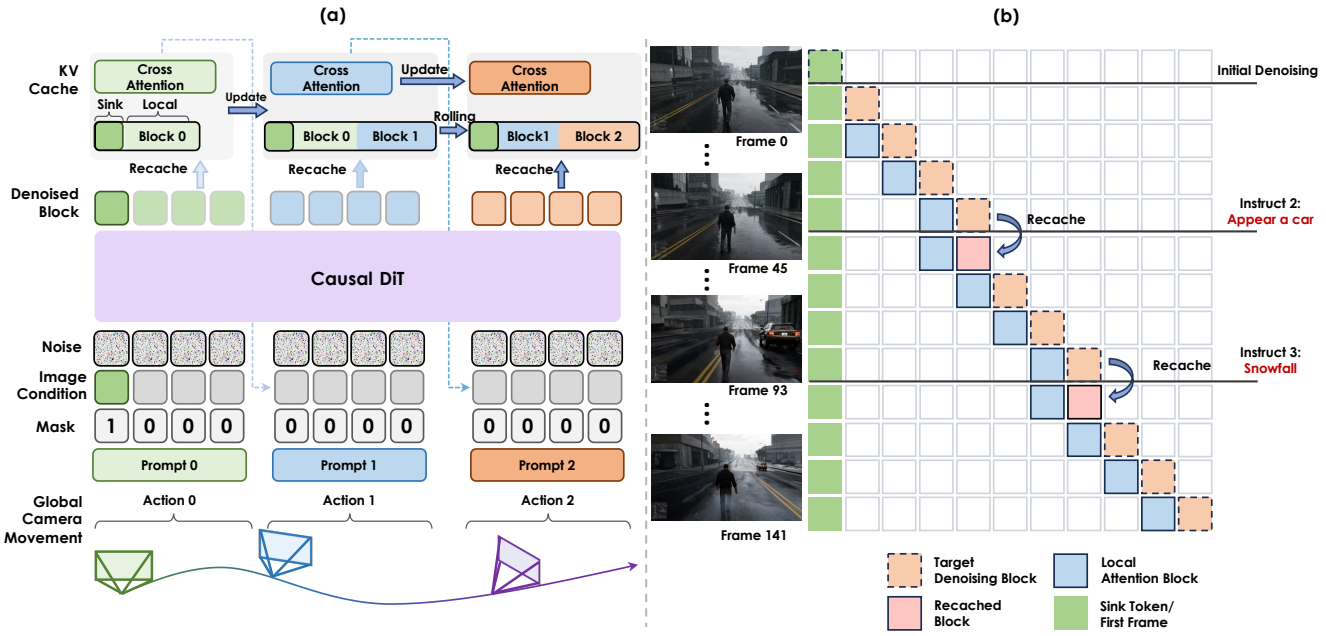

Figure 4. **Multi-turn Interactive Inference.** (a) shows block-wise autoregressive long-video inference with KV-cache updates: the initial frame is kept as sink tokens, while local attention uses recent blocks. (b) illustrates prompt re-caching, which improves interaction responsiveness and accuracy for new prompts.

out long video sequences from a causal generator and apply distributional moment distance (DMD) [58, 59] alignment on the extended frames. This strategy effectively mitigates error accumulation during autoregressive generation.

Building upon this insight, we adopt a randomized extension tuning strategy using a dataset of long-form gameplay videos exceeding 10 seconds. In this stage, the model autoregressively rolls out N frames, and contiguous T-frame windows are uniformly sampled to align the predicted and target distributions. Furthermore, we randomly extend the predicted videos from the causal generator to varying lengths, promoting robustness across different temporal horizons. In practice, while rolling out at window $W = V[i : i + K - 1]$, the student generator uses sink token and KV cache and autogressively extend long video, and the fake score teacher model uses the last frame in the previous clean predicted chunk $V[i - 1]$ as image condition; while the real score uses the ground truth frame in the original video.

To mitigate the potential erosion of interactive capabilities inherent in few-step distillation, we adopt a training paradigm that interleaves self-forcing with teacher-forcing. The rationale for this approach is to compel the model to master state recovery and maintain temporal stability. Crucially, this is achieved by exposing it to diverse states at arbitrary points along the generation trajectory, rather than limiting such corrective training solely to the initial phase.

4.3. Multi-turn Interactive Inference

Self-attention KV Cache. We use a fixed-length KV cache with rolling updates for efficient autoregressive generation. Sink tokens are permanently retained at the cache start, while a local window keeps the N frames preceding the target denoising block across multi-turn interactions. The cache comprises sink tokens and a block-sparse local attention component, improving efficiency and preventing quality drift.

ReCache Mechanism. We employ a recache mechanism to enhance the accuracy and stability of multi-turn interactions in autoregressive long video generation. Upon receiving a new interaction prompt, the model extracts the corresponding interaction embeddings to recompute the last autoregressive block and update both the self-attention and cross-attention KV caches. This strategy provides precise historical context for the subsequent target block with minimal computational overhead, thereby ensuring accurate and responsive feedback to facilitate a smoother user experience.

5. Experiments

5.1. Model and Dataset Configurations.

We compare our method against several SOTA image-to-video generation foundation models, including **Hunyuan-Video**, **Wan2.2 A14B**, and **LongCatVideo**. To comprehensively evaluate controllable generation, we generate 93-frame videos at 832×448 resolution using a curated test set spanning diverse scenes and specific action scenarios.

5.2. InterBench: A Comprehensive Benchmark for Interactive Video Generation

To evaluate video generation, we adopt two metric families: general video-quality metrics and our interaction-focused

Figure 5. **Inference results by GameCraft-2 on multi-action control.** In our case, blue-lit keys indicate key presses. W, A, S, D represent transition movement and ↑, ←, ↓, → denote changes in view angles.

Figure 6. **Inference results by GameCraft-2 on instruction-based multi-action control.** Except W, A, S, D, ↑, ←, ↓, and →, the red notes denote action instructions.

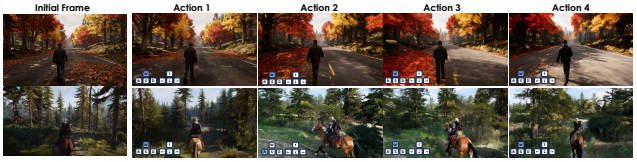

Figure 7. **Inference results by GameCraft-2 on the third-perspective long-term game video generation.**

suite, **InterBench**. While general metrics capture fidelity, temporal consistency, and motion realism, they miss key interactive properties such as causality, action execution, and state transitions. InterBench addresses this with six interaction-centric dimensions (e.g., interaction completeness, action effectiveness, causal coherence), yielding a holistic evaluation of interactive video models.

General Quality Metrcis To comprehensively assess our model, we adopt diverse metrics. For **video realism**, we use Fréchet Video Distance (FVD) [49] to capture spatial fidelity and temporal dynamics. **Visual quality** is measured by Image Quality and Aesthetic scores, reflecting perceptual clarity and visual appeal. We evaluate temporal consistency for cross-frame coherence and artifacts (e.g., flickering, structural instability). For **dynamic performance**, we adapt VBench's Dynamic Degree [24] by reporting absolute optical-flow magnitudes (*Dynamic Average*) instead of binary motion classes, yielding a finer measure of motion intensity and naturalness.

Geometric Camera Control. For interactive camera control, we report Relative Pose Error (RPE) in translation and rotation after Sim3 Umeyama alignment between predicted and ground-truth trajectories. This removes global pose and scale ambiguity, so RPE reflects local frame-to-frame motion

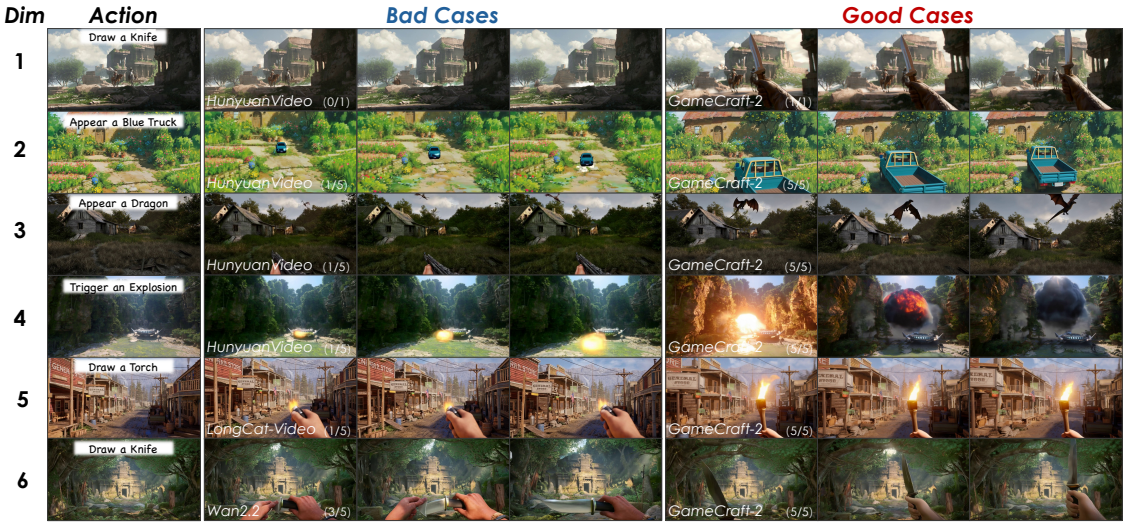

Figure 8. **Qualitative examples for the six InterBench dimensions.** Rows correspond to **(1) Trigger Rate, (2) Prompt–Video Alignment, (3) Fluency, (4) Scope Accuracy, (5) End-State Consistency**, and **(6) Object Physics**. For each, we compare a high-scoring result from GameCraft-2 (right) with a low-scoring baseline failure (left), illustrating the rating scale.

Table 1. **Quantitative comparison of recent controllable video generation models.**

Model	Visual Quality				Temporal Consistency ↑	RPE	
	FVD↓	Image Quality↑	Dynamic↑	Aesthetic↑		Trans↓	Rot↓
GameCraft	**1554.2**	0.69	**67.2**	0.67	0.95	**0.08**	0.20
GameCraft-PCM	1883.3	0.67	43.8	0.65	0.93	**0.08**	0.20
Matrix-Game	2260.7	**0.72**	31.7	0.65	0.94	0.18	0.35
Matrix-Game-2.0	1920.6	0.62	20.5	0.49	0.84	**0.08**	0.25
GameCraft-2	1856.3	0.70	45.2	**0.71**	**0.96**	**0.08**	**0.17**

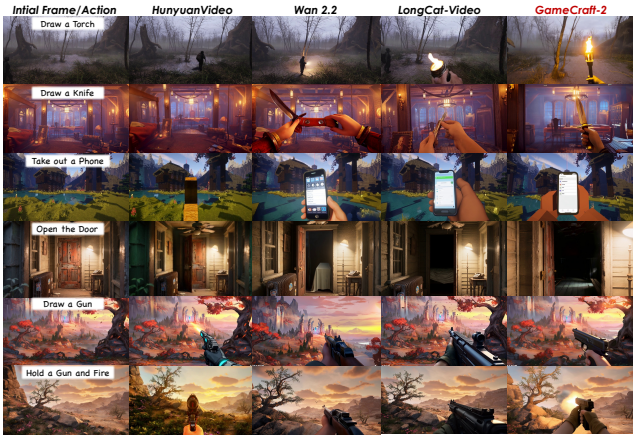

Figure 9. **Comparison of Actor-Action Interactions with Baseline Models.** Visual comparisons illustrating the quality of action-level interactions across representative prompts. Our method produces more coherent and physically consistent actions than all baselines.

fidelity and control precision under interactive inputs.

Action-Level Interaction Semantics. To evaluate *action-level interaction*, we introduce **InterBench**, a six-dimensional protocol for interactive video generation. Using a Vision-Language Model (VLM) as an automatic evaluator, it measures interaction fidelity, smoothness, and physical plausibility. The six dimensions are:

1. **Interaction Trigger Rate.** A binary check of whether the requested interaction is initiated, distinguishing ignored prompts from attempted actions.
2. **Prompt–Video Alignment.** Measures semantic fidelity to the full prompt, including *static alignment* (scene/context consistency) and *dynamic alignment* (correct action execution).
3. **Interaction Fluency.** Assesses temporal naturalness and visual coherence, penalizing artifacts such as jumps, flicker, and object teleportation.
4. **Interaction Scope Accuracy.** Evaluates whether interaction effects have appropriate spatial extent (global events affect the whole scene; local actions remain localized).
5. **End-State Consistency.** Checks whether the interaction reaches a correct, stable end state.
6. **Object Physics Correctness.** Evaluates physical plausibility, including rigid-body integrity, realistic kinematics,

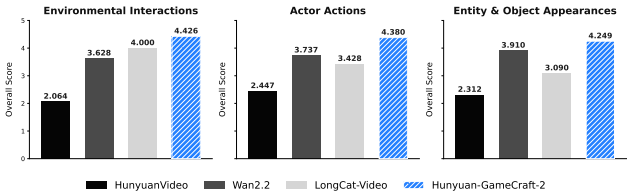

Figure 10. **Comparison of Environmental Interactions with Baseline Models.** Qualitative results showing the fidelity and consistency of environment-level effects. Our approach better preserves global influence and temporal stability.

and correct contact relationships (no penetration).

Hierarchical scoring combines a binary **Trigger** with five ordinal dimensions into a weighted *Overall* score for prioritization.

5.3. Interaction Evaluation

Quantitative Results on InterBench InterBench results show that GameCraft-2 best balances complex interaction logic with cinematic control. It achieves near-perfect semantic responsiveness with `Trigger` > 0.96 for both Environmental and Actor actions (Figure 10) outperforming Wan2.2 and HunyuanVideo, while also improving physical fidelity (`Physics` $+0.68$) and `EndState` stability. For detailed subclass scores, please see the appendix. Importantly, this does not sacrifice visual quality. As shown in Tab 1. GameCraft-2 attains the highest **Temporal Consistency** (0.96) and **Aesthetic** (0.71) among control-centric models. With mixed real and synthetic data, it further delivers precise camera control, yielding the lowest rotation error (**Rot** 0.17) and competitive translation accuracy. Overall, these metrics confirm that GameCraft-2 unifies high-level physical reasoning with high-fidelity generation.

Qualitative Analysis The results clearly highlight the superior performance of GameCraft-2 over baseline models. Baselines frequently exhibit noticeable deficiencies when handling complex interactions. For instance, environmental effects often lack dynamic evolution and realistic lighting interactions. Actor actions are commonly plagued by object deformation, motion incoherence, and inaccurate hand-object contact. Furthermore, newly generated entities tend to suffer from identity drift, unstable geometry, and poor integration with the scene. In contrast, GameCraft-2 demonstrates substantially higher fidelity and consistency across all interaction categories. For Actor Actions, GameCraft-2 produces more coherent action sequences, enabling characters to stably grasp and precisely manipulate objects while ensuring stable final states. These qualitative examples support the quantitative results and demonstrate GameCraft-2's robust generation of semantically accurate, temporally coherent, physically plausible interaction videos.

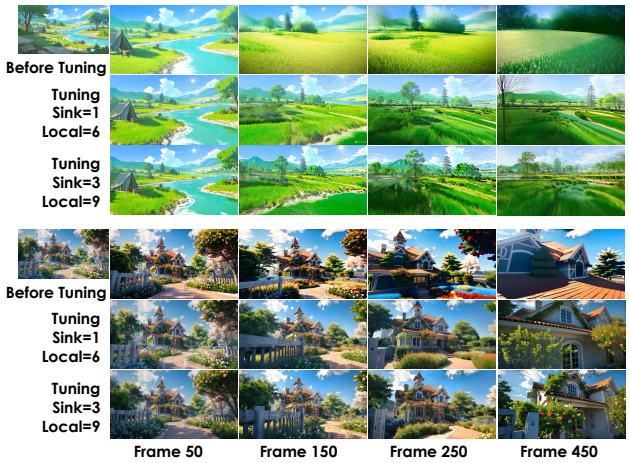

Figure 11. **Qualitative Analysis of Long-Video Tuning and Cache Settings. Row 1**: Baseline results without Long-Video Tuning (sink token size = 1, local attention size = 6). **Row 2**: Incorporates Long-Video Tuning upon the baseline. **Row 3**: Further modifies setting based on Row 2 by increasing the sink token size to 3 and local attention size to 9. Input prompts and camera parameters remain consistent across all samples.

Analysis of Long-Video Tuning and Cache Settings We qualitatively analyze the impact of long-video tuning and KV cache settings, specifically regarding the sink tokens and local attention. We compare generated frames at aligned time steps, the integration of randomized extended long-video tuning substantially improves video fidelity and motion consistency beyond the 450th frame. Moreover, expanding the sink tokens and local attention size can enrich the detail but increases the artifacts. These observations confirm both the efficacy of our tuning strategy and the importance of leveraging sink tokens and local attention to maintain robust context.

6. Conclusion

In this work, we present **GameCraft-2**, an interactive game world model that generates high-fidelity, controllable video from free-form text and keyboard/mouse actions. We formalize interactive video data and develop automated curation and synthesis pipelines to alleviate the data bottleneck. GameCraft-2 unifies multimodal controls under a robust training framework, combining randomized long-video tuning with efficient inference (e.g., KV-recache) for stable long-horizon and real-time generation. We further introduce **InterBench**, a benchmark for action-level interaction quality. Experiments show that GameCraft-2 consistently surpasses prior state-of-the-art methods in interaction fidelity, visual quality, and temporal coherence, advancing video synthesis from passive generation to user-driven, playable world creation.

References

[1] Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, et al. Genie 3: A new frontier for world models. 2025. 1, 2

[2] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000. 13

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[4] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025. 12

[5] Brandon Castellano. PySceneDetect. 12

[6] Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations*, 2025. 2

[7] 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. 2

[8] Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset, 2023. 3

[9] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025. 2, 4

[10] Karan Dalal, Daniel Koceja, Gashon Hussein, Jiarui Xu, Yue Zhao, Youjin Song, Shihao Han, Ka Chun Cheung, Jan Kautz, Carlos Guestrin, et al. One-minute video generation with test-time training. *arXiv preprint arXiv:2504.05298*, 2025. 2

[11] Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*, 2024. 2

[12] Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, and Jun Xiao. Vid-gpt: Introducing gpt-style autoregressive generation in video diffusion models. *arXiv preprint arXiv:2406.10981*, 2024. 2

[13] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025. 2

[14] Junxian Guo, Haotian Tang, Shang Yang, Zhekai Zhang, Zhijian Liu, and Song Han. Block Sparse Attention. https://github.com/mit-han-lab/Block-Sparse-Attention, 2024. 4

[15] Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft, 2025. 2

[16] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 3

[17] Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025. 1, 2

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models, 2020. arXiv:2006.11239 [cs]. 1

[19] Susung Hong, Junyoung Seo, Heeseong Shin, Sunghwan Hong, and Seungryong Kim. Direct2v: Large language models are frame-level directors for zero-shot text-to-video generation, 2024. 3

[20] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator, 2023. 3

[21] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers arXiv:2508.10934*, 2025. 13

[22] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025. 1

[23] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion, 2025. 2, 3, 4

[24] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6

[25] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators, 2023. 3

[26] KolorsTeam. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024. 13

[27] 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. 1, 2

[28] Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition, 2025. 1, 2, 3

[29] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing, 2023. 3

[30] Xinyang Li, Tengfei Wang, Zixiao Gu, Shengchuan Zhang, Chunchao Guo, and Liujuan Cao. FlashWorld: High-quality 3D Scene Generation within Seconds, 2025. 1

[31] Zhimin Li, Jianwei Zhang, and and others Lin. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. 1

[32] Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, Zizhen Li, Fanrui Zhang, Jiaxin Ai, Zhixiang Wang, Yuwei Wu, Tong He, Jiangmiao Pang, Yu Qiao, Yunde Jia, and Kaipeng Zhang. Sekai: A video dataset towards world exploration, 2025. 2

[33] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models, 2024. 3

[34] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning, 2024. 3

[35] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time, 2025. 2

[36] Mushui Liu, Yuhang Ma, Yang Zhen, Jun Dan, Yunlong Yu, Zeng Zhao, Zhipeng Hu, Bai Liu, and Changjie Fan. Llm4gen: Leveraging semantic representation of llms for text-to-image generation, 2024. 3

[37] Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025. 1

[38] Xiaoxiao Long, Qingrui Zhao, Kaiwen Zhang, Zihao Zhang, Dingrui Wang, Yumeng Liu, Zhengjie Shu, Yi Lu, Shouzheng Wang, Xinzhe Wei, Wei Li, Wei Yin, Yao Yao, Jia Pan, Qiu Shen, Ruigang Yang, Xun Cao, and Qionghai Dai. A survey: Learning embodied intelligence from physical simulators and world models, 2025. 2

[39] Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025. 2

[40] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. 2024. 2

[41] William Peebles and Saining Xie. Scalable Diffusion Models with Transformers, 2023. arXiv:2212.09748 [cs]. 1

[42] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing, 2023. 3

[43] Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. Instructvid2vid: Controllable video editing with natural language instructions, 2024. 3

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, 2022. arXiv:2112.10752 [cs]. 1

[45] Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Schechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025. 4

[46] Shuai Tan, Biao Gong, Yutong Feng, Kecheng Zheng, Dandan Zheng, Shuwei Shi, Yujun Shen, Jingdong Chen, and Ming Yang. Mimir: Improving video diffusion models for precise text understanding, 2024. 3

[47] HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, Yihang Lian, Yulin Tsai, and Wangand others. HunyuanWorld 1.0: Generating Immersive, Explorable, and Interactive 3D Worlds from Words or Pixels, 2025. 1

[48] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 13

[49] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 6

[50] Florentina Voboril, Vaidyanathan Peruvemba Ramaswamy, and Stefan Szeider. Streamllm: Enhancing constraint programming with large language model-generated streamliners. In *2025 IEEE/ACM 1st International Workshop on Neuro-Symbolic Software Engineering (NSE)*, pages 17–22. IEEE Computer Society, 2025. 4

[51] 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. 1, 2, 3, 4

[52] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 13

[53] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. 2

[54] WorldLabs. Generating worlds. https://www.worldlabs.ai/blog, 2024. 1

[55] WorldLabs. Rtfm: A real-time frame model. https://www.worldlabs.ai/blog/rtfm, 2025. 1

[56] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WORLDMEM: Long-term Consistent World Simulation with Memory, 2025. arXiv:2504.12369 [cs]. 2

[57] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation, 2025. 2, 4

[58] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024. 5

[59] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024. 5

[60] Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models, 2025. 4

[61] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025. 2

[62] Jintao Zhang, Jia Wei, Haofeng Huang, Pengle Zhang, Jun Zhu, and Jianfei Chen. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration, 2025. 14

[63] Xiangjun Zhang, Litong Gong, Yinglin Zheng, Yansong Liu, Wentao Jiang, Mingyi Xu, Biao Wang, Tiezheng Ge, and Ming Zeng. Rise-t2v: Rephrasing and injecting semantics with llm for expansive text-to-video generation, 2025. 3

[64] Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Zedong Gao, Eric Li, Yang Liu, and Yahui Zhou. Matrix-game: Interactive world foundation model. *arXiv*, 2025. 1, 2

[65] Shihao Zhao, Shaozhe Hao, Bojia Zi, Huaizhe Xu, and Kwan-Yee K. Wong. Bridging different language models and generative vision models for text-to-image generation, 2024. 3

[66] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images, 2018. 2

A. Interactive Video Data Construction

Current video data suitable for training interactive world models remain scarce. Real-world captured videos offer high realism but are costly, time-consuming to collect, and difficult to scale. Simulation-based generation using engines such as *Unreal Engine* provides strong controllability over viewpoint and interactions, yet the heavy modeling and rendering costs restrict scene diversity. Internet videos from platforms like *YouTube* offer massive volume and variety, but their highly inconsistent quality and abundant noise demand complex cleaning pipelines. Public academic datasets, while well-annotated and reliable, are limited in scale and domain coverage. As a result, none of these sources simultaneously satisfy the requirements of interactivity, large scale, and broad diversity, leaving high-quality interactive video data fundamentally insufficient. This scarcity highlights the need for a clearer understanding of what truly qualifies as *interactive* data, which we formalize in the following analysis.

A.1. Definition of Interactive Video Data

> Interactive Video Data refers to a temporal sequence that explicitly records a **causally driven state-transition process**, in which agents or the environment transition from a clearly defined ***initial state*** to a significantly different ***final state***. The importance of such data lies in its ability to faithfully capture *how an event evolves over time*, rather than in visual complexity.

A video segment is considered *interactive* if it satisfies any of the following properties:

- **Significant State Transition.** The video must contain a recognizable and non-trivial macroscopic change of state. It should present clearly distinguishable *pre-condition* and *post-condition* states, with the temporal content between them forming the *transition process*.
- **Subject Emergence or Interaction.** The main content involves explicit subjects, including:
  1. *Emergence*: a new subject appears in a previously empty context.
  2. *Action-driven*: a subject performs an action that changes its own state or affects the environment.
- **Scene Shift or Evolution.** The video records a fundamental shift or evolution of the scene or background, rather than minor or random perturbations.

Interactive videos thus possess **explicit causal structure**, **clear state transitions**, and **perceivable action agents**, enabling world models to learn interpretable action–outcome mappings. Following this definition, we systematically organize interactive data into three principal categories to structure our analysis: **(1) Environmental Interactions**, which encompass global or local scene changes; **(2) Actor Actions**, which are driven by an embodied agent; and **(3) Entity and Object Appearances**, which involve the introduction of new subjects. To facilitate a nuanced evaluation, each category is further divided into *simple* and *complex* settings, reflecting varying degrees of difficulty. Specific examples for each category are provided in Appendix D.2.

A.2. Synthetic Data Construction

To address the scarcity and high annotation cost of interactive video data, we propose a controllable Synthetic Interaction Video Pipeline for large-scale automated production. While generating synthetic data for training video models has been underexplored, we argue it is now feasible by leveraging the advanced world knowledge and visual representation capabilities of recent foundation models. The effectiveness of our pipeline in producing diverse, high-quality data is showcased in Appendix E (Figs. 18-20).

We generate interactive videos starting from an initial frame F_t. To handle diverse visual contexts, we first employ a Vision-Language Model (VLM) to analyze F_t and, guided by a high-level instruction (e.g., *"taking out a torch"*), generate a customized, scene-specific prompt. Based on the interaction type, we then apply one of two distinct strategies:

1. **Start-End Frame Strategy:** For stationary scenes requiring explicit state transitions (e.g., environmental changes like *"making it snow"*), a VLM guides an image editing model to generate a target end-frame F'_t. This provides strong controllability over the final state.
2. **First-Frame-Driven Strategy:** For dynamic actions involving significant camera motion (e.g., *"opening a door"*), the model generates freely from only the initial frame. This approach avoids distortions and yields smoother camera movement and temporal continuity.

Sourcing specific initial frames for certain interactions, such as *"opening a door"*, is a significant bottleneck, as manual curation is both costly and inefficient. To address this, we leverage an advanced text-to-image model (e.g., HunyuanImage-3.0 [4]), to synthesize these requisite frames on demand, providing a scalable source of high-quality inputs for our video generation pipeline.

A.3. Game Scene Data Curation

We build our dataset from over 150 AAA games (e.g., *Assassin's Creed*, *Cyberpunk 2077*), which provides extensive diversity in environments, lighting, artistic styles, and camera viewpoints is showcased in Appendix E Figs. 17.

Scene and Action-aware Data Partition. We employ a two-stage partitioning strategy to process the raw videos. First, PySceneDetect [5] segments long videos into visually coherent 6-second clips. Subsequently, we use RAFT-based

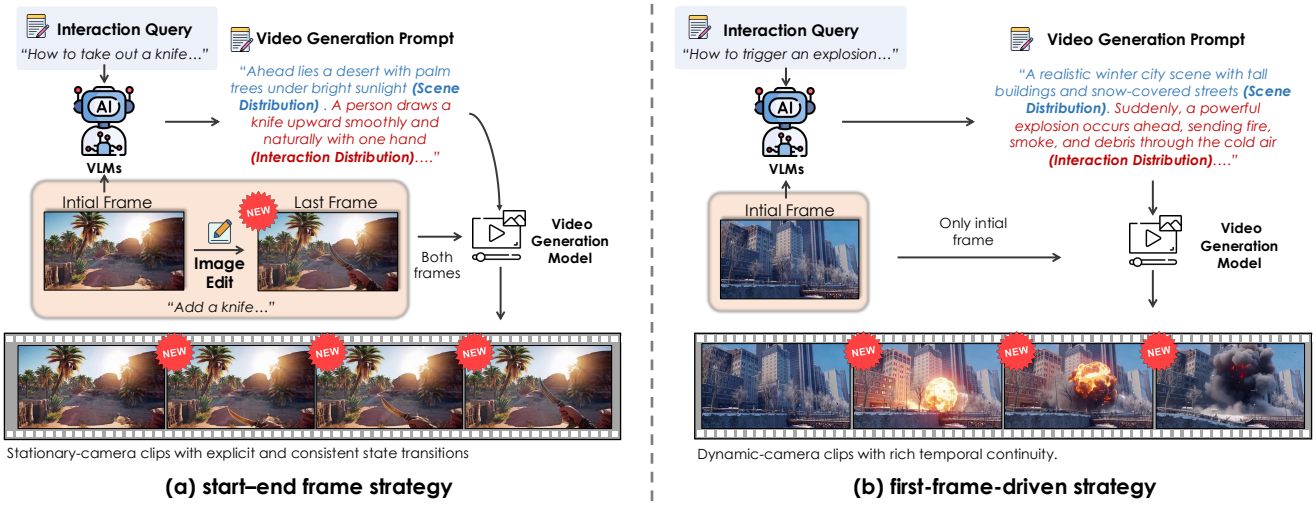

Figure 12. **Showcase of our Synthetic Interaction Video Pipeline.** ***(a) The start–end frame strategy*** uses a VLM and an image-editing model to construct both initial and edited target frames, enabling controlled state transitions for stationary-camera scenarios. ***(b) The first-frame-driven strategy*** relies solely on the initial frame and VLM-generated prompts, allowing the video generator to create dynamic, motion-rich interactions with flexible camera movement.

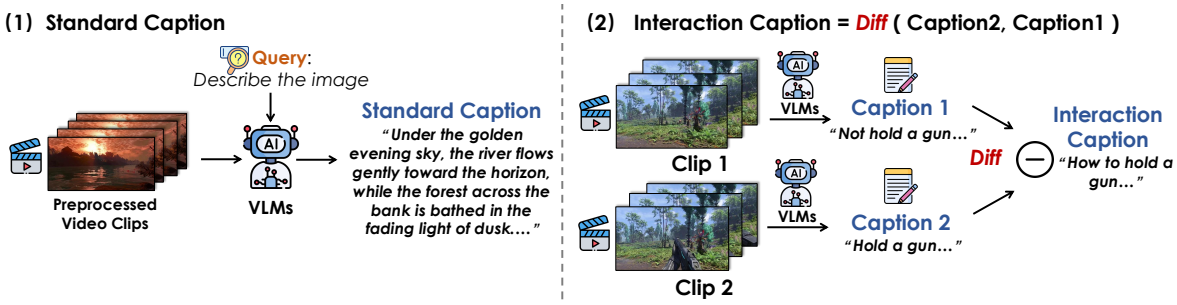

Figure 13. **Pipeline of the Caption Generation System.** The system produces two forms of captions: a ***standard caption*** that describes the visual content of each clip, and an ***interaction caption*** derived by computing the semantic difference between consecutive clips. This enables both scene-level descriptions and explicit interaction-oriented annotations for supervision.

optical flow [48] to localize fine-grained action boundaries, ensuring each clip preserves temporal integrity for training.

Data Filtering. To ensure data quality, we perform a three-stage filtering process. A learning-based model first removes low-fidelity or artifact-heavy frames [26]. Next, luminance filtering eliminates scenes that are poorly lit [2]. Finally, a VLM-based semantic check verifies content consistency across frames, retaining only clips with clean visual structure and accurate motion alignment [52].

Camera Annotation. We reconstruct 6-DoF camera trajectories for each clip using VIPE [21]. This process yields frame-by-frame translational and rotational motion estimates, providing precise metadata for training camera-aware models and enforcing spatio-temporal consistency.

Structured Captioning. To provide interaction-aware supervision, we devise a structured captioning scheme with two components. First, a **Standard Caption** (C_t), generated by a VLM for each clip, describes the static visual content. Second, an **Interaction Caption** ($I_{t\to t+1}$) captures the state transition between adjacent clips. This interaction is computed as the semantic difference between their respective standard captions:

$$I_{t\to t+1} = \Delta(\Phi(C_{t+1}), \Phi(C_t)),$$

where Φ is a semantic encoder and Δ is a difference operator. This dual-component approach enables the model to jointly learn **appearance-level perception** (from C_t) and **action-level reasoning** (from $I_{t\to t+1}$).

B. Detailed Experimental Setup

B.1. Model Configurations

- **HunyuanVideo.** We use the official inference configuration to ensure reproducibility, with FLOW_SHIFT=7.0 and EMBEDDED_CFG_SCALE=6.0 controlling motion dynamics and conditional guidance, respectively. All experiments are conducted with 50 denoising steps. In addition, both flow_reverse and i2v_stability are enabled to improve temporal robustness across frames.

- **Wan2.2 A14B.** We employ the UniPC sampler with sample_shift=5.0, sample_steps=40, and boundary=0.900. A dual-stage classifier-free guidance (CFG) strategy is used, with guidance scales of (3.5, 3.5) applied to the two noise regimes to maintain consistent conditioning.
- **LongCatVideo.** We adopt the default high-quality inference setup, using a guidance_scale of 4 and 50 denoising steps. Compilation optimizations are enabled to improve inference efficiency while preserving generation quality.

B.2. Prompt Design.

To ensure fair and controlled evaluations, we designed a standardized, two-part prompt strategy. This approach constructs two complementary components for each test image: an ***interaction prompt*** to specify the dynamic target action or event, and a ***base prompt*** to describe static scene attributes and anchor the generation process to the input image's appearance. During inference, these two prompts are concatenated into a single conditioning sentence and fed directly to each model. This decoupled design not only ensures that all models receive identical instructions for fair comparison but, critically, it also enables controlled evaluations by allowing us to systematically vary interaction instructions while keeping the visual context constant.

B.3. Test Datasets.

To comprehensively evaluate interaction-aware video generation, we construct two complementary test datasets, each targeting different aspects of interaction capability.

Viewpoint-Oriented Interaction Test Set. The first test set is designed to evaluate interaction robustness under diverse viewpoints and camera motions. It consists of **75 base images**, where each image is associated with multiple viewpoint and camera control instructions, including translations (front, back, left, right, up, down), diagonal viewpoints (upper-left, upper-right, lower-left, lower-right), and rotational motions (turn left and turn right). For each image–viewpoint combination, a corresponding interaction prompt is applied, resulting in a total of **900 generated videos**. This dataset is used for the quantitative evaluation reported in Table 1, enabling a systematic assessment of model performance across varied camera perspectives.

Fine-Grained Interaction Test Set. In addition to the viewpoint-oriented dataset, we construct a second test set focusing on detailed and fine-grained interaction behaviors. This dataset is organized around three core interaction dimensions: **(1) Environmental Interactions**, **(2) Actor Actions**, and **(3) Entity and Object Appearances**. Each dimension is further divided into fine-grained categories and interaction types to support controlled and diagnostic evaluation.

For **Environmental Interactions**, the dataset includes weather-related effects (e.g., snow, rain, and lightning) as well as large-scale physical events (e.g., explosions), assessing whether models can produce globally consistent and temporally stable scene-level changes. The **Actor Actions** dimension covers both primitive actions, such as drawing a weapon or taking out a tool, and composite, multi-step actions, such as drawing and firing a gun or taking out and operating a phone. To further probe action-specific generalization, we additionally include a specialized subset of **20 closed-door images** for evaluating the *open door* action, which requires consistent state transitions across frames. The **Entity and Object Appearances** dimension spans a diverse set of entities, including animals, vehicles, and human subjects, as well as extended unseen entities such as dragons, enabling evaluation of object emergence, appearance consistency, and generalization to novel concepts.

Overall, this fine-grained interaction test set consists of **100 images**, covering a wide range of scenes (indoor/outdoor, natural/urban), lighting conditions, and visual styles (realistic, game-like, and cartoon). For all evaluations, models are required to generate videos at a unified resolution of **832×448** with a fixed length of **93 frames**, ensuring consistent and fair comparison across methods.

C. Real-time Acceleration Details

To further accelerate inference and minimize latency, we incorporate several system-level optimizations:

- **FP8 Quantization** reduces memory bandwidth and leverages GPU acceleration while preserving visual quality;
- **Parallelized VAE decoding** enables simultaneous latent-frame reconstruction, mitigating bottlenecks in long-sequence decoding;
- **SageAttention** [62] replaces FlashAttention with an optimized quantized attention kernel for faster transformer computation; and
- **Sequence parallelism** distributes video tokens across multiple GPUs, supporting efficient long-context generation.

Together, these techniques boost inference speed to 16 FPS on 8 NVIDIA H20 GPUs, achieving real-time interactive video generation with stable quality and low latency.

D. Illustrative Examples of Interactive Video Data

To provide a comprehensive understanding of our definition, we present representative examples that clarify the boundary between interactive and non-interactive video data.

D.1. Positive Examples: Interactive Video Data

The following examples satisfy one or more properties of interactive video data, exhibiting clear causal structures and

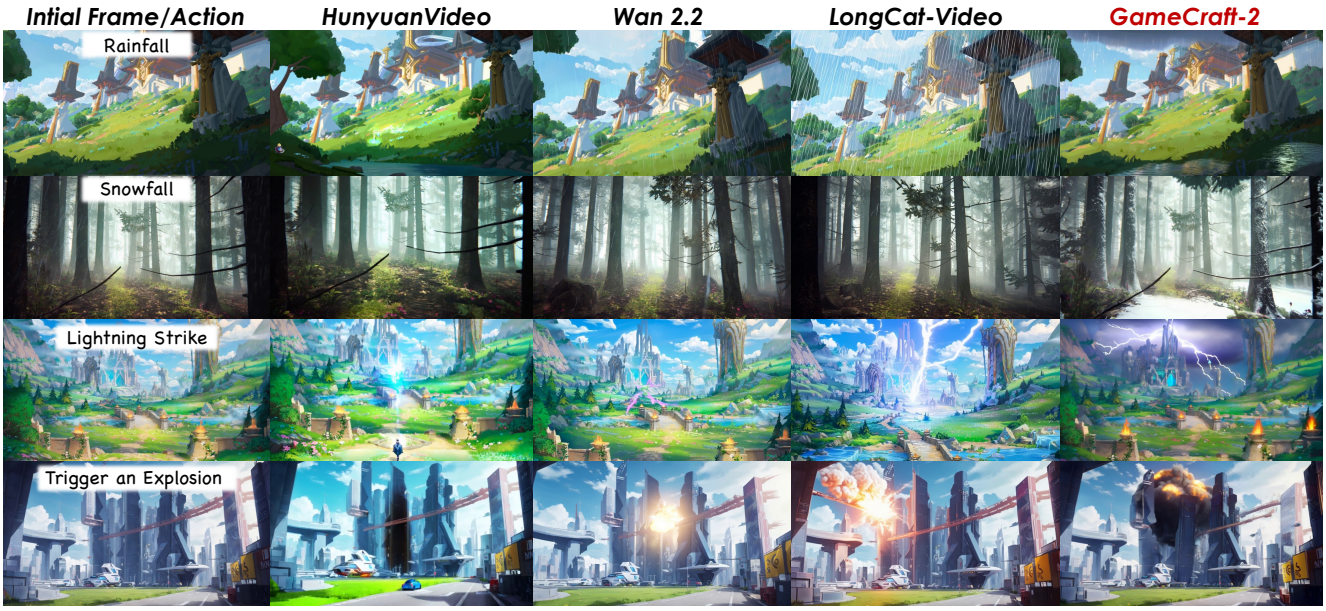

Figure 14. **Comparison of Environmental Interactions with Baseline Models.** Qualitative results showing the fidelity and consistency of environment-level effects. Our approach better preserves global influence and temporal stability.

perceivable state transitions.

Subject Emergence.

- *Example 1: Vehicle Appearance.* An empty street (initial state) transitions as a car enters from off-screen and parks at the roadside (transition process), culminating in a scene depicting *"a car parked on the street"* (final state). The automobile constitutes the emergent core subject, transforming the scene from vacant to occupied.
- *Example 2: Object Retrieval.* From a first-person perspective, the frame initially contains only a pair of hands (initial state). The hands retrieve a key from a pocket and hold it prominently (transition process), resulting in a final state of *"hands holding a key"* (final state). The key represents the emergent core subject.

Action-Driven Interaction.

- *Example 3: Door Opening.* The scene begins with a subject standing before a closed door (initial state). The subject pushes the door open (transition process), leading to a fully open door (final state). This exemplifies direct interaction where the subject acts upon an object, inducing a clear state change.
- *Example 4: Weapon Discharge.* A character aims a firearm at a target (initial state), pulls the trigger (transition process), resulting in projectile impact and target destruction (final state). This demonstrates action-consequence causality with observable physical effects.

Environmental State Evolution.

- *Example 5: Weather Transition.* A clear sky (initial state) undergoes gradual cloud accumulation followed by snowfall of increasing intensity (transition process), ultimately blanketing the entire scene in heavy snow (final state). This represents a fundamental transformation of the environmental weather attribute.
- *Example 6: Spatial Transition.* Upon opening a door, the camera view shifts from an interior room (initial scene) to an exterior courtyard (final scene). This exemplifies a discrete scene transition driven by subject action, fundamentally altering the observational context.

These examples, though visually dynamic, lack the defining characteristics of interactive video data.

Continuous Static Process.

- *Example 1: Sustained Blizzard.* A 10-second video segment depicting continuous heavy snowfall. Although visually dynamic, the macroscopic state remains constant as *"actively snowing throughout,"* lacking a transition from *"no snow"* to *"snow present."* The absence of state evolution disqualifies this as interactive data.

Stochastic Background Activity.

- *Example 2: Busy Intersection.* A scene featuring continuous pedestrian and vehicular traffic at a crowded intersection. While abundant motion exists, there is no singular

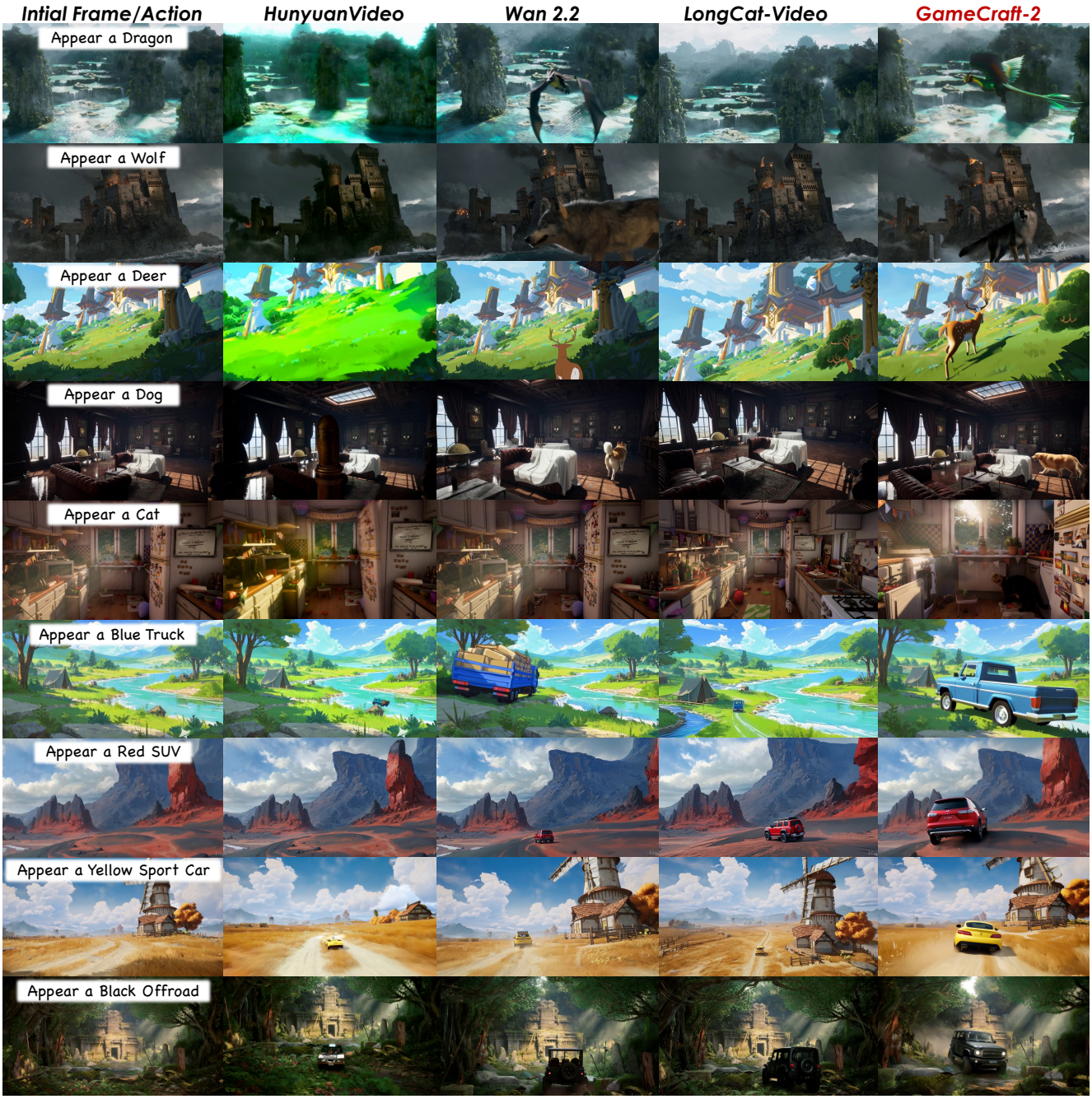

Figure 15. **Comparison of Entity and Object Appearance Interactions with Baseline Models.** Visual comparisons of object emergence and interaction correctness. Our method delivers more accurate, stable, and physically plausible object behaviors.

event-driven macroscopic state change with definitive beginning and end points. The scene's overarching state persistently remains *"busy intersection,"* lacking a coherent causal narrative.

Generalized Motion without Core Subject.

- *Example 3: Ambient Environmental Fluctuations.* Ripples propagating across a water surface or leaves swaying in wind. These phenomena typically constitute random environmental perturbations rather than state transitions driven by specific subjects or events with explicit causal chains. They lack the purposeful, agent-driven transformation characteristic of interactive data.

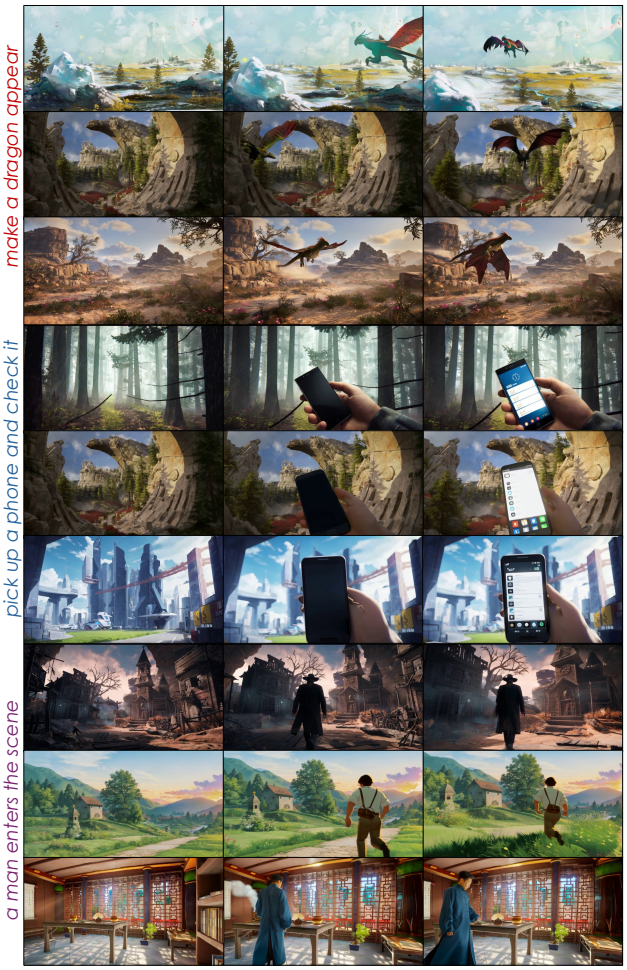

Figure 16. **Generalization Beyond Training Distribution.** GameCraft-2 generates coherent, physically plausible interactions for unseen entities/actions (e.g., a man suddenly appearing, a dragon emerging, taking out a phone), despite such cases not occurring in training, indicating it learns transferable interaction dynamics rather than memorizing visuals.

D.2. Interaction Categories

Following the definition of interactive data in the main text, we provide here a detailed breakdown of the three principal interaction categories used to structure our dataset and analysis. Each category includes both simple and complex settings to reflect different levels of difficulty and to facilitate a fine-grained evaluation of model capabilities.

(1) Environmental Interactions. These interactions reflect global or local scene changes. *Simple cases* include atmospheric effects such as *snowfall* and *rainfall*. *Complex cases* involve more substantial causal transformations, such as *lightning strikes* or *triggering an explosion*, which require coherent illumination changes, particle dynamics, and physically plausible propagation.

(2) Actor Actions. These interactions are driven by an embodied or first-person actor. *Simple cases* include basic manipulation actions such as *drawing a gun* or *drawing a knife*. *Complex cases* require multi-step or environment-affecting interactions, such as *drawing a torch to illuminate the surroundings*, *firing a gun*, *taking out a phone and operating it*, or *opening a door*. These demand consistent body–object coordination and temporal stability.

(3) Entity and Object Appearances. These interactions introduce new entities into the scene. *Simple cases* include the appearance of a single human or common object. *Complex cases* involve entities with more distinct geometry or motion priors, such as animals (cat, dog, deer, wolf, dragon) or vehicles (red SUV, yellow sports car, blue truck, black off-road vehicle), which require accurate spatial placement, scale consistency, and stable identity preservation.

E. Dataset Showcase

This appendix provides visual examples from our constructed dataset, which is composed of two primary sources: curated real-world gameplay footage and synthetically generated interactive videos. The following sections showcase the diversity and quality of each data type.

E.1. Curated Gameplay Data

The following figures illustrate the rich diversity of our curated gameplay data, collected from over 150 AAA games. As shown, the dataset covers a wide array of interaction contexts, including both first-person and third-person viewpoints (Fig. 17), as well as a comprehensive range of environments spanning natural and urban scenes under various lighting, weather. This diversity is crucial for training robust and generalizable world models.

E.2. Synthetic Interaction Data

Generated by our synthetic data pipeline, the following examples demonstrate the pipeline's capability to create controlled and high-quality interactive videos. These examples cover the three main interaction categories defined in our work: **Environmental Interactions** such as weather changes and explosions (Fig. 18), **Actor Actions** involving complex body-object coordination (Fig. 19), and **Entity/Object Appearances** that introduce new subjects into the scene with high fidelity (Fig. 20).

F. Detailed Comparison Across Interaction Dimensions

In this section, we present a concise comparison of our approach against baseline models across three key dimensions

Figure 17. **Examples of First-person and Third-person Interactive Gameplay Videos.** Samples showing diverse actor actions under different viewpoints, illustrating rich interactive semantics captured from our gameplay collection.

of interactive video generation: environmental interactions, actor–action dynamics, and entity or object appearance behaviors. As shown in Tab 2, our method achieves higher temporal stability, more coherent action execution, and more accurate object emergence, consistently outperforming existing models across diverse interaction scenarios.

G. InterBench: A Detailed Protocol for Benchmarking Action-Level Interaction

Motivation and Design Philosophy. Existing video generation benchmarks, such as Fréchet Video Distance or CLIP Score, primarily assess perceptual quality, temporal consistency, and static text-video alignment. While valuable, they are ill-suited for evaluating *interactive* video generation, where the primary task is to render a causal change in response to a specific action command. These metrics cannot distinguish between a correctly executed action and a visually plausible but semantically incorrect video. To fill this critical gap, we designed **InterBench**, an evaluation protocol specifically tailored to measure the fidelity of action-level interactions. Its philosophy is to deconstruct the complex concept of a "good interaction" into a set of distinct, measurable, and interpretable dimensions, enabling a fine-grained analysis of model capabilities and failure modes.

Interaction Trigger Rate. This dimension serves as the most fundamental, gateway assessment. It asks the question: *Did the requested interaction happen at all?"* This metric is designed to isolate the model's basic ability to acknowledge and act upon an instruction, separating cases where the model successfully initiated the action from those where it completely failed to respond. This is a binary metric:

- **1 (Success):** The requested interaction is initiated in the video. For instance, for the prompt *draw a gun,"* this score is given if a gun becomes visible. If this score is given, the subsequent dimensions are evaluated on their respective scales.
- **0 (Failure):** The requested interaction does not occur at all. The model ignores or completely misunderstands the interaction prompt. If this score is given, all subsequent

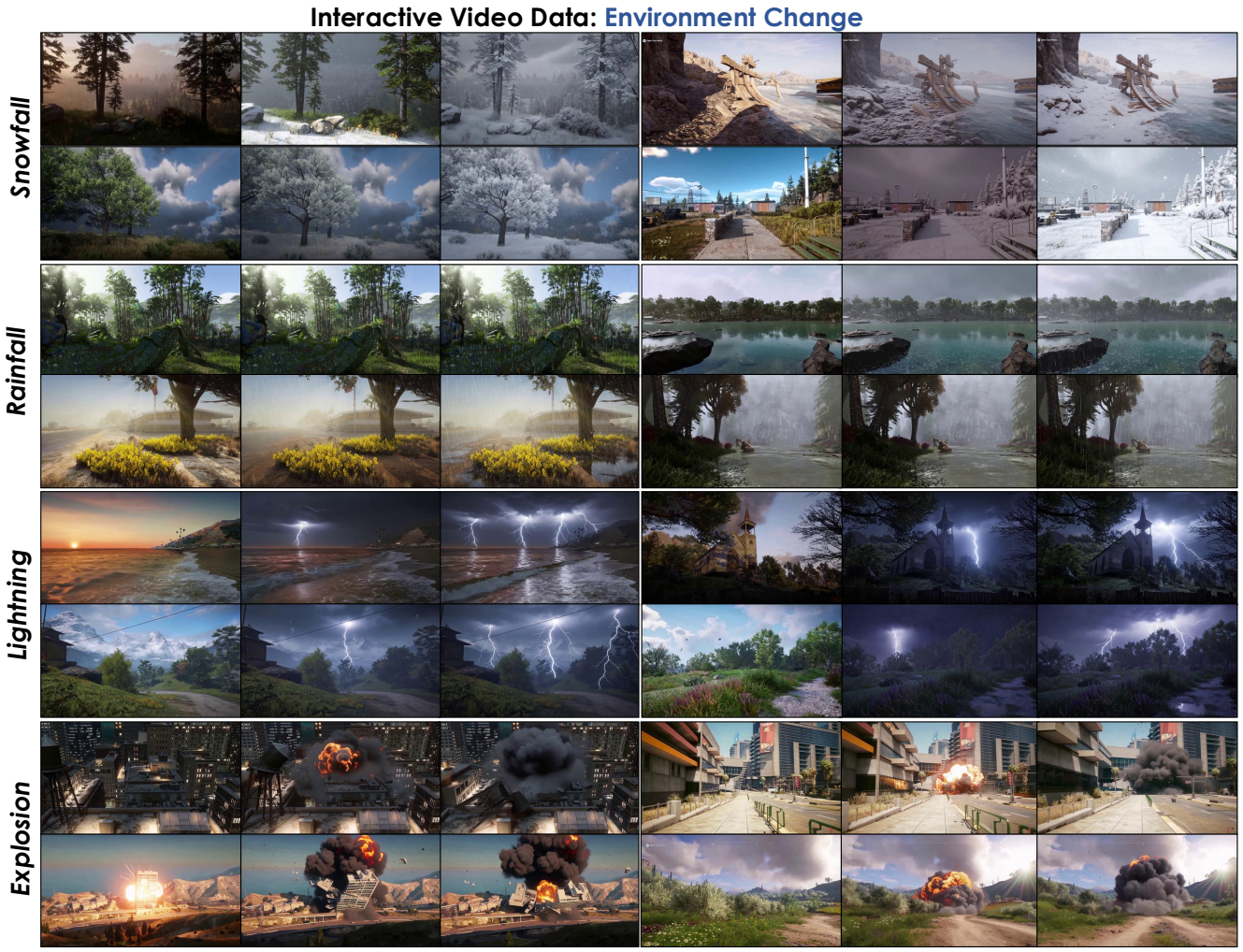

Figure 18. **Synthetic Examples of Environmental Interactions.** Examples of synthetic scene-change interactions generated by our pipeline, covering *snowfall*, *rainfall*, *lightning*, and *explosions*.

Table 2. **Quantitative performance evaluation**, against state-of-the-art competitors on the InterBench protocol. Scores are presented across six key interaction dimensions, with our model's superior results highlighted.

Category	Method	Trigger	Align	Fluency	Scope	EndState	Physics
Environmental Interactions	**Wan2.2 A14B**	0.799	3.511	3.579	3.722	3.951	3.008
	LongCat-Video	0.897	3.963	3.777	4.188	4.377	3.210
	HunyuanVideo	0.490	1.950	1.940	2.065	2.308	1.670
	GameCraft-2	**0.962**	**4.342**	**4.247**	**4.578**	**4.688**	**3.893**
Actor Actions	**Wan2.2 A14B**	0.836	3.490	3.488	4.036	4.054	3.175
	LongCat-Video	0.806	3.089	3.005	3.832	3.771	2.839
	HunyuanVideo	0.587	2.147	2.202	2.717	2.748	1.931
	GameCraft-2	**0.983**	**4.087**	**4.191**	**4.576**	**4.686**	**3.828**
Entity & Object Appearances	**Wan2.2 A14B**	0.874	3.943	3.545	4.281	4.265	3.054
	LongCat-Video	0.712	3.050	2.758	3.340	3.482	2.352
	HunyuanVideo	0.607	2.037	1.870	2.736	2.734	1.462
	GameCraft-2	**0.944**	**4.292**	**3.978**	**4.410**	**4.514**	**3.578**

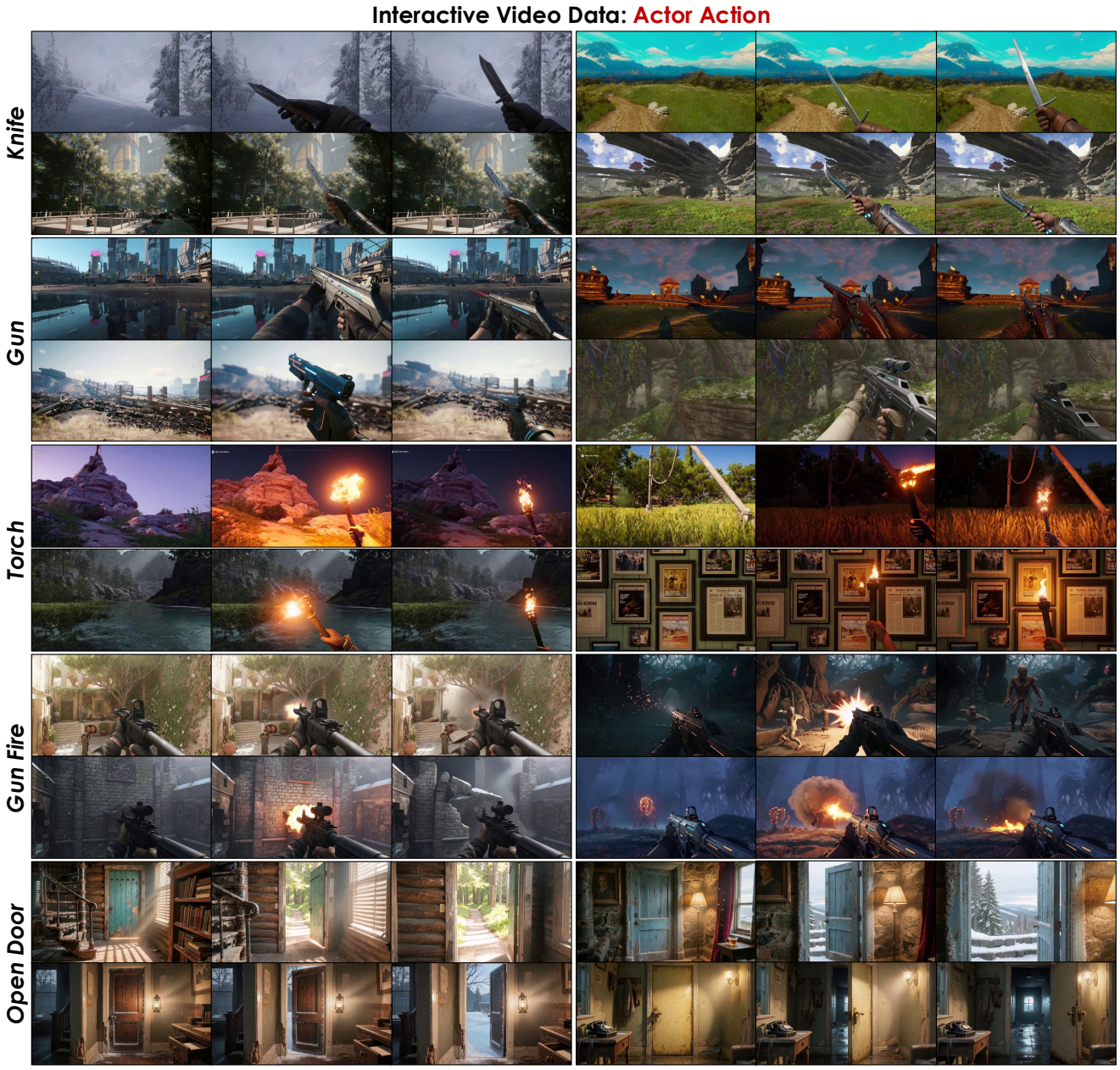

Figure 19. **Synthetic Examples of Actor Actions.** Examples illustrating the range of interactive behaviors synthesized by our pipeline. ***Actor-driven interactions*** include *drawing a knife*, *drawing a gun*, *drawing a torch*, *firing a weapon*, and *opening a door*, which require consistent body–object coordination and temporal coherence.

dimensions are automatically scored 0.

Prompt–Video Alignment. Beyond simply triggering an action, this dimension evaluates the semantic fidelity of the generated video with respect to the *entire* prompt (both the base scene description and the interaction command). It ensures the interaction happens in the *right way* and the *right context*, encompassing both static and dynamic alignment. This metric is scored on a 0-1-3-5 ordinal scale, contingent on the interaction being triggered:

- **5 (Excellent):** Both the static context (scene, style) and the dynamic action perfectly match the prompt's description.
- **3 (Moderate):** The primary action is correct, but there are minor semantic deviations in the scene's context or the specifics of the action's execution.
- **1 (Poor):** A recognizable interaction occurs, but it involves

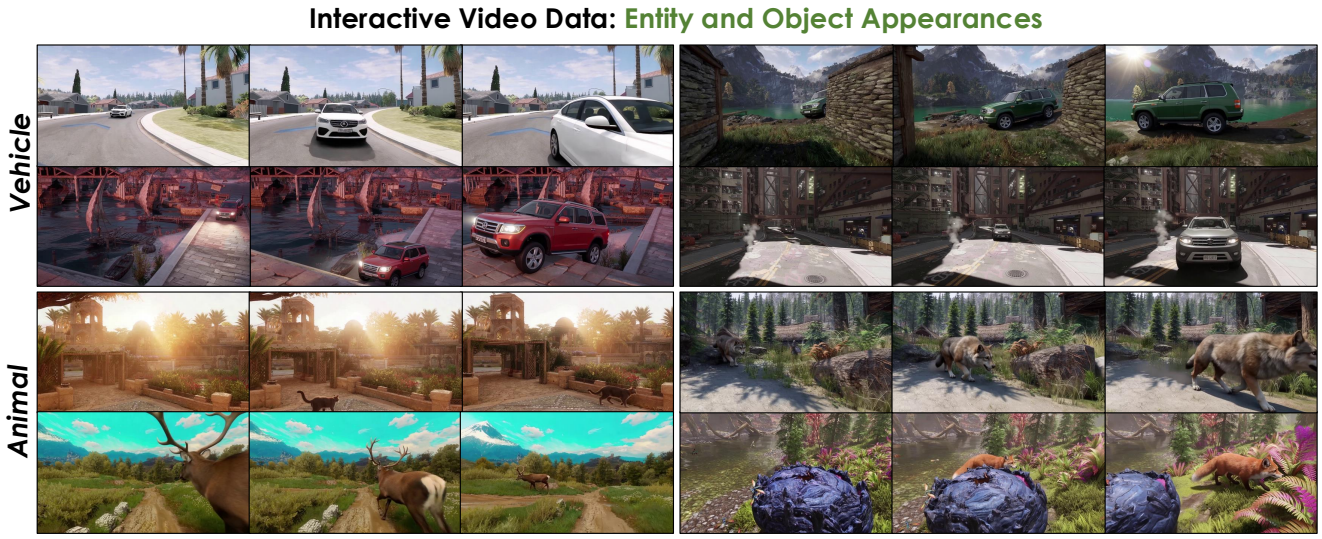

Figure 20. **Synthetic Examples of Entity/Object Appearances.** Examples illustrating the range of interactive behaviors synthesized by our pipeline. ***Entity- and object-level interactions*** include *animal intrusion* and *vehicle entry*, showing the pipeline's capability to introduce new entities with realistic geometry, scale consistency, and stable identity across frames.

a major semantic error, such as performing the wrong action (e.g., closing instead of opening a door) or generating a scene that bears no resemblance to the base prompt.
- **0 (Failure):** The triggered video content shows no meaningful semantic alignment with either the prompt's context or its specified action.

Interaction Fluency. This dimension measures the temporal naturalness and continuity of the interaction process. It specifically penalizes temporal discontinuities such as abrupt teleportation of objects, noticeable frame jumps, unrealistic motion jitter, and structural tearing of geometry, particularly around the interacting regions. This metric is scored on a 0-1-3-5 ordinal scale:

- **5 (Excellent):** The motion is perfectly smooth, continuous, and natural, with no temporal artifacts present.
- **3 (Moderate):** The motion is generally continuous but contains minor, non-disruptive artifacts like slight jitter or a single inconspicuous jump-cut.
- **1 (Poor):** The interaction is plagued by severe temporal artifacts (e.g., constant flickering, object teleportation) that significantly disrupt the viewing experience.

Interaction Scope Accuracy. This metric assesses a model's spatial reasoning by examining whether the spatial extent and environmental influence of an interaction are plausible and consistent with its expected scope (global or local). This metric is scored on a 0-1-3-5 ordinal scale:

- **5 (Excellent):** The spatial influence of the interaction is physically and semantically correct (e.g., global effects are global, local effects are local and propagate realistically).
- **3 (Moderate):** The scope is generally correct but with minor inaccuracies, such as a global effect not covering the entire scene or a local effect having a slightly incorrect area of influence.
- **1 (Poor):** The scope is fundamentally wrong. For example, a global event is rendered as a tiny local patch, or a local effect implausibly affects the entire scene.

End-State Consistency. A successful interaction must not only be initiated correctly but also *converge* to a stable and correct outcome. This dimension evaluates the final state of the video to ensure the result of the action persists as expected. This metric is scored on a 0-1-3-5 ordinal scale:

- **5 (Excellent):** The interaction converges to the correct final state, which remains stable until the end of the video.
- **3 (Moderate):** The final state is mostly correct but exhibits minor instability, such as slight flickering, object drift, or subtle geometric inconsistencies.
- **1 (Poor):** The interaction fails to converge correctly. The final state is incorrect, highly unstable (e.g., oscillating), or the effects of the action vanish prematurely.

Object Physics Correctness. This dimension focuses on the physical plausibility and structural integrity of the objects and agents involved in the interaction, evaluating whether their behavior adheres to basic physical principles like object permanence, rigidity, and kinematics. This metric is scored on a 0-1-3-5 ordinal scale:

- **5 (Excellent):** All objects and agents maintain structural integrity and interact in a physically plausible manner. There is no unnatural deformation, interpenetration, or kinematic errors.
- **3 (Moderate):** Minor physical inaccuracies are present, such as slight object warping during movement or brief, non-critical interpenetration between an agent and an object.
- **1 (Poor):** Severe physical violations occur. Objects unnaturally deform, agents pass through solid objects, or motion is kinematically impossible.

H. Limitation and Future Work

Despite its advancements, our framework has several limitations that highlight avenues for future research. First, while our randomized long-video tuning strategy alleviates error accumulation in autoregressive generation, it does not entirely eliminate it, and semantic drift may still manifest in long sequences (More than 500 frames). This is partly attributable to our model's lack of an explicit long-term memory mechanism, a crucial component for advanced world models, as it relies instead on the finite capacity of its KV cache. Furthermore, the scope of supported interactions is currently centered on single-step, immediate-effect actions. Enabling multi-stage tasks that require logical reasoning or planning remains a significant future challenge. Finally, although we achieve real-time performance at 16 FPS, further optimization is required to reduce latency for highly reactive gameplay and to enable deployment on more accessible hardware.