Hunyuan-GameCraft-2: Instruction-following Interactive Game World Model

Tencent Hunyuan*

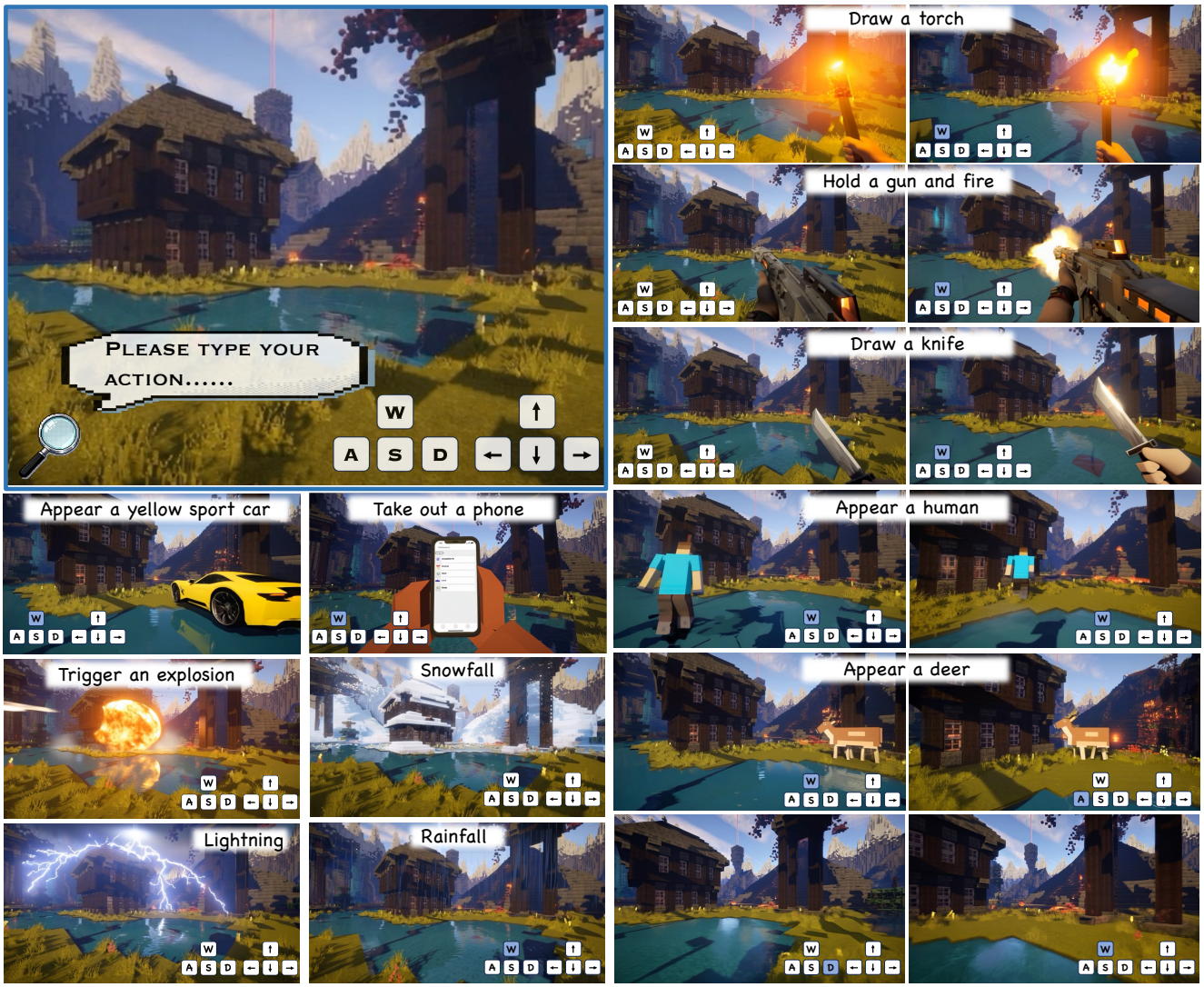

Figure 1. ***Hunyuan-GameCraft-2*** advances generative game world models from static game scene video synthesis to open-ended, instruction-following interactive simulation. We simulate a series of action signals from a single image. The left and right frames depict key moments from game video sequences generated in response to different action inputs. Our model can accurately produce content aligned with each interaction, supports high-fidelity game video generation with temporal and 3D consistency. As shown, instructions such as *"draw a torch"* or *"hold a gun and fire"*, and key/mouse action dynamically guide camera motion and dynamic video content editing, producing temporally coherent, causally grounded interactive videos with realistic scene continuity and a consistent style.

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 ***Hunyuan-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 long-horizon, temporally coherent, and causally grounded interactive game videos that faithfully respond to diverse user instructions such as "open the door," "draw a torch," or "trigger an explosion".*

1. Introduction

The rapid advancement of diffusion models [20, 30, 34, 44, 47, 56] has significantly advanced dynamic game content creation [31, 67]. Beyond static image or short video synthesis, recent cutting-edge achievements, from RTFM [60] and to Genie series [2], mark world model can serve as the foundation for immersive, controllable virtual experiences, marking a crucial step toward AI-driven **"playable worlds"** that can both simulate and respond to user intent.

Existing world models can be categorized into ***3D-based*** and ***video-based*** approaches. 3D-based world models [25, 33, 40, 50, 59, 60] 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 [4, 30, 56, 58], the video-based technical pathway [2, 13, 31, 43, 61, 64, 67] has shown remarkable potential. These works learn world dynamics directly from large-scale video data [7, 13, 35, 69] through implicit end-to-end representation learning. Notably, the Genie series [2, 43] introduces latent action modeling to simulate player-driven physical interactions, while Matrix-Game [67] and Hunyuan-GameCraft [31] integrate discrete gameplay 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*". Consequently, compelling us to rigorously define the concept of "***interaction***" within the context of 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 [13, 31, 61, 64, 67] to embodied motion sensing [41]. Grounded in this perspective, two key challenges hinder this progress: (1) the lack of a formal definition and 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 **Hunyuan-GameCraft-2**, an interactive game world model focusing on 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 [62] 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, enabling real-time interactive video generation.

To comprehensively evaluate interactive performance

across different models, we introduce InterBench, a new benchmark that systematically measures key dimensions of interactive behavior — including interaction completeness, action effectiveness, causal coherence, and physical plausibility. Extensive experiments on InterBench and general video-quality metrics demonstrate the effectiveness of our framework, achieving state-of-the-art performance in generating interactive videos that faithfully respond to user instructions while maintaining high visual fidelity and temporal coherence.

In general, our main contributions are as follows:

- We propose a unified controllable video generation framework integrating text, keyboard, and mouse signals for semantically grounded interactions.
- We leverage autoregressive distillation and randomized long-video tuning to ensure efficient and stable long-horizon generation, with KV-recache for multi-turn inference and real-time 16 FPS performance through engineering optimizations.
- Through extensive quantitative and qualitative experiments, we comprehensively validate the effectiveness of our proposed framework, demonstrating superior performance in generating interactive videos that faithfully respond to user instructions while maintaining visual quality and temporal coherence.

2. Related Works

2.1. Long Video Extension

Maintaining temporal coherence in long video generation is a principal challenge, primarily stemming from the "train-short-test-long" discrepancy in diffusion models, which often causes semantic drift and accumulating artifacts. To surmount this, one major line of work seeks to better align the training process with inference conditions. Methods such as Self-Forcing [26] condition the model on its own predictions to simulate error accumulation, while Rolling-Forcing [38] incrementally updates context through rolling windows. A complementary strategy integrates explicit memory structures, as seen in Memory-Forcing [23] and StreamingT2V [19], to preserve long-range dependencies and global dynamics. Beyond adapting the existing training loop, other research explores more fundamental shifts in the generative paradigm. These include alternative formulations like next-frame prediction models [14, 15], hybrid diffusion-autoregressive frameworks such as DiffusionForcing [8], and test-time adaptation for inference refinement [11]. The research frontier is also advancing toward interactive and structured synthesis. For example, LongLive [62] introduces a KV-recache mechanism for responsive semantic control, and MAGI-1 [1] autoregressively generates temporal blocks to mitigate error propagation through explicit partitioning.

2.2. Interactive Video-based World Model

Unlike traditional video generation models which produce predetermined sequences, interactive world models dynamically respond to user inputs, enabling the creation of explorable and playable game environments(Tab. 1).

Early explorations in this domain often utilized game environments like Minecraft as a testbed. Models such as MineWorld [17], Matrix [13], and GameFactory [64] demonstrated the ability to generate video conditioned on discrete user actions, typically keyboard and mouse inputs. Similarly, Yume [42] generates interactive video from a single image prompt controlled by discrete keyboard commands. While pioneering, these models were often limited to specific games and simple action spaces. Nevertheless, the scope of interaction these models support remains highly limited.

Subsequent research advanced generalization and long-term consistency. Genie2 [43] introduced a foundation model capable of generating diverse, action-controllable 2D worlds from single images. To tackle the challenge of consistency in extended simulations, WorldMem [61] introduced a memory bank framework to address long-term consistency issues, while recent works like PAN [51] also focus on achieving interactive long-range world simulation.

Building upon these foundations, researchers began to explore more flexible interaction. GameGen-X [7] integrates multi-modal control signals for open-world games. Critically, Genie3 [2] and Hunyuan-GameCraft [31] advance this paradigm by unifying discrete keyboard and mouse signals into a shared, continuous action space. This emerging fusion of direct controls and language prompts shows immense potential. However, prompts in these latest works are predominantly used for world setup and high-level guidance, rather than as a direct, interactive control mechanism. Consequently, the richness of interaction is still fundamentally constrained by the discrete nature of physical input devices.

2.3. Text-guided Video Generation and Editing

Text-based control over video synthesis has advanced significantly through two main paradigms: enhancing semantic understanding and executing structured plans.

The first paradigm focuses on *enriching the initial prompt*. This is achieved by fusing representations from Large Language Models (LLMs) for more nuanced inputs [39, 49], using LLMs to rephrase or expand simple queries [66], or employing lightweight adapters to bridge domain gaps [66, 68]. The second, more sophisticated paradigm treats text as a script or plan to be executed. Here, LLMs act as *"directors"*, decomposing high-level prompts into a sequence of frame-by-frame descriptions for temporally evolving scenes [21, 22]. This concept extends to orchestrating complex, multi-scene videos with explicit spatial layouts and consistency constraints [36, 37]. A related approach is found in video editing, where textual instructions guide discrete

Table 1. Comparison of recent interactive game world models.

Model	Resolusions	Training Data	Action type	Action space	Scene Generalizable	Scene dynamic	Scene memory	Real time
GameNGen [54]	240×192	Gameplay	Keyboard	Discrete	Closed	✗	✔	✗
Oasis [12]	640×360	Gameplay video	Key+Mouse	Discrete	Closed	✗	✗	✗
GameGen-X [7]	1280×720	Gameplay video	Instruction	Discrete	Closed	✗	✔	✗
Matrix [13]	1280×720	Gameplay + Rendered	Key	Discrete	Closed	✔	✔	✗
Matrix-Game [67]	1280×720	Gameplay + Rendered	Key+Mouse	Discrete	Closed	✔	✗	✗
Genie 2 [43]	1280×720	Unknown	Key+Mouse	Discrete	Closed	✔	✗	✗
Genie 3 [2]	1280×720	Unknown	Key+Mouse	Discrete	Closed	✔	✔	✔
GameFactory [64]	640×360	Gameplay video	7 Keys+Mouse	Discrete	Closed	✔	✔	✗
GameCraft [31]	1280×720	Gameplay + Rendered	Key+Mouse	Continuous	Closed	✔	✔	✗
GameCraft-2	864×480	Gameplay + Rendered+ **Interaction**	Key+Mouse+**Prompt-based Instrution**	Continuous	**Open-ended**	✔	✗	✔

tasks like style transfer or object manipulation, often within video-to-video frameworks that enable zero-shot or end-to-end control [9, 28, 32, 45, 46].

Despite their power, these methods are fundamentally non-interactive. Whether enhancing a prompt or executing a script, they perform one-off transformations based on a static, predefined set of commands. They lack the core concepts of state transition and continuous feedback, where an action perpetually redefines future possibilities. In stark contrast, GameCraft-2 introduces true interaction, where user prompts continuously drive the evolution of a dynamic world state, addressing a fundamentally different objective than either planned generation or scripted editing.

3. Interactive Video Data Construction

3.1. The Scarcity of Interactive Video Data

In world models, data fundamentally determines the depth and breadth of how the model understands and reproduces the real world. However, not all videos containing dynamic content are suitable for training world models that emphasize ***interaction***. Current training data for such models mainly come from the following sources:

Real-world capture: Videos are directly recorded from real or game environments. Such data exhibit high realism, but their collection is expensive and time-consuming. The diversity of scenes and behaviors is limited by physical constraints, making large-scale expansion difficult.

Simulation-based generation: Videos are rendered using game engines like *Unreal Engine*. This approach provides excellent controllability, allowing precise control of camera perspectives, object interactions, and data quality. However, the high cost of modeling and rendering severely limits the diversity and scalability of generated scenes.

Internet video resources: Massive amounts of data can be obtained by crawling public video platforms such as *YouTube* offering unmatched volume and diversity, but the data quality is highly inconsistent and often noisy, with excessive camera shake or irrelevant content. Complex and costly multi-stage cleaning pipelines are required to extract usable samples.

Public academic datasets: These datasets are typically well-annotated and of high quality, but they are limited in scale and domain coverage, far from sufficient for training general-purpose interactive world models.

3.2. 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**

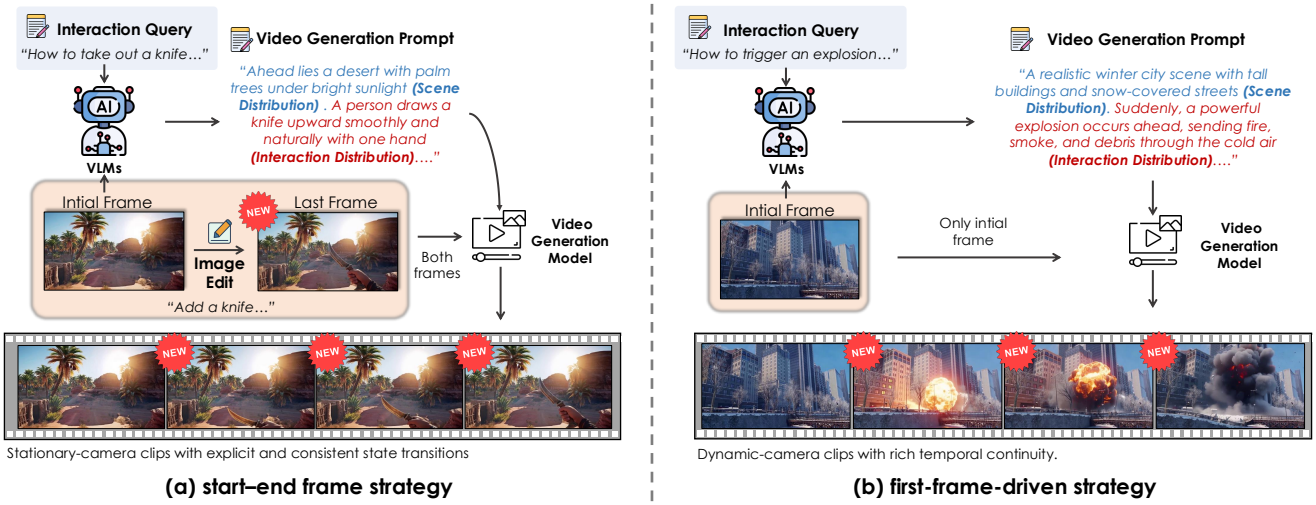

Figure 2. **Showcase of our Synthetic Interaction Video Pipeline.** ***(a) The start–end frame strategy*** uses a VLM and an image-editing model to construct both initial and edited target frames, enabling controlled state transitions for stationary-camera scenarios. ***(b) The first-frame-driven strategy*** relies solely on the initial frame and VLM-generated prompts, allowing the video generator to create dynamic, motion-rich interactions with flexible camera movement.

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 A.3.

3.3. 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 B (Figs. 16-18).

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 [5]), to synthesize these requisite frames on demand, providing a scalable source of high-quality inputs for our video generation pipeline.

3.4. 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 B Figs. 14 and 15.

Scene and Action-aware Data Partition. We employ a two-stage partitioning strategy to process the raw videos. First, PySceneDetect [6] segments long videos into visually coherent 6-second clips. Subsequently, we use RAFT-based optical flow [52] 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 [29]. Next, luminance filtering eliminates scenes that are poorly lit [3]. Finally, a VLM-based semantic check verifies content consistency across frames, retaining only clips with clean visual structure and accurate motion alignment [57].

Camera Annotation. We reconstruct 6-DoF camera trajectories for each clip using VIPE [24]. 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 su-

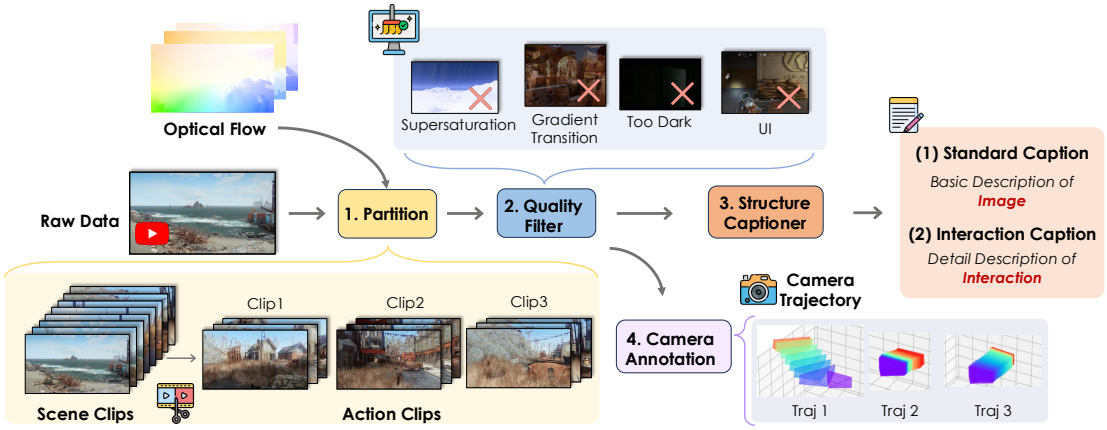

Figure 3. **Pipeline of the Data Curation System.** Our pipeline consists of four stages: ***(1) Partition,*** which segments long gameplay videos into scene- and action-level clips using scene detection and optical-flow cues; ***(2) Quality Filtering***, which removes low-quality frames via visual assessment, luminance checks, and VLM-based semantic filtering; ***(3) Structured Captioning***, which produces both standard and interaction-centric captions for each clip; ***(4) Camera Annotation***, which reconstructs 6-DoF trajectories to capture viewpoint motion. These steps convert raw gameplay footage into clean, structured, and interaction-aware training data.

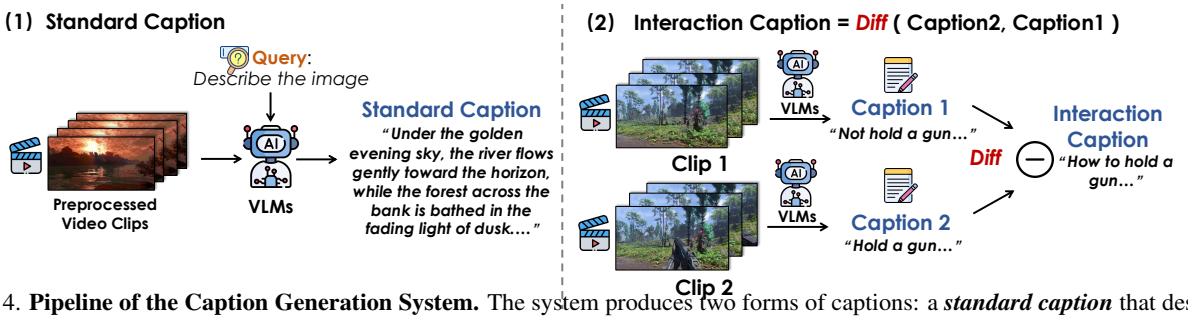

Figure 4. **Pipeline of the Caption Generation System.** The system produces two forms of captions: a ***standard caption*** that describes the visual content of each clip, and an ***interaction caption*** derived by computing the semantic difference between consecutive clips. This enables both scene-level descriptions and explicit interaction-oriented annotations for supervision.

pervision, 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 $I_{t\to t+1}$).

4. Method

We present **GameCraft-2**, an interactive game video model focusing on free-form instruction-based control. The overall framework is illustrated in Fig. 5. In particular, GameCraft-2 unifies a natural action-injected causal architecture, image-conditioned autoregressive long video generation, and diverse multi-prompt interaction into a cohesive framework. This section will introduce GameCraft-2's model architecture, training, and inference procedures.

4.1. Model Architecture

The main architechture of **GameCraft-2** is based on a 14B image-to-video mixture-of-experts (MoE) foundation video generation model [56]. Our objective is to extend this image-to-video diffusion model into an action-controllable generator. As discussed in Sec. 1, its action space includes both keyboard inputs and free-form text prompts.

For keyboard and mouse signal injection (W, A, S, D, ↑, ←, ↓, →, Space, etc.), we adopt the methodology from GameCraft-1 [31], mapping these discrete action signals to continuous camera control parameters. During training, annotated camera parameters are encoded as Plücker embeddings [18] and integrated into the model through token addition. At inference, user inputs are converted into camera trajectories to derive these parameters.

As for prompt-based interaction injection, we observe that the base model struggles to express certain interactive verbs, largely due to the higher semantic and spatial com-

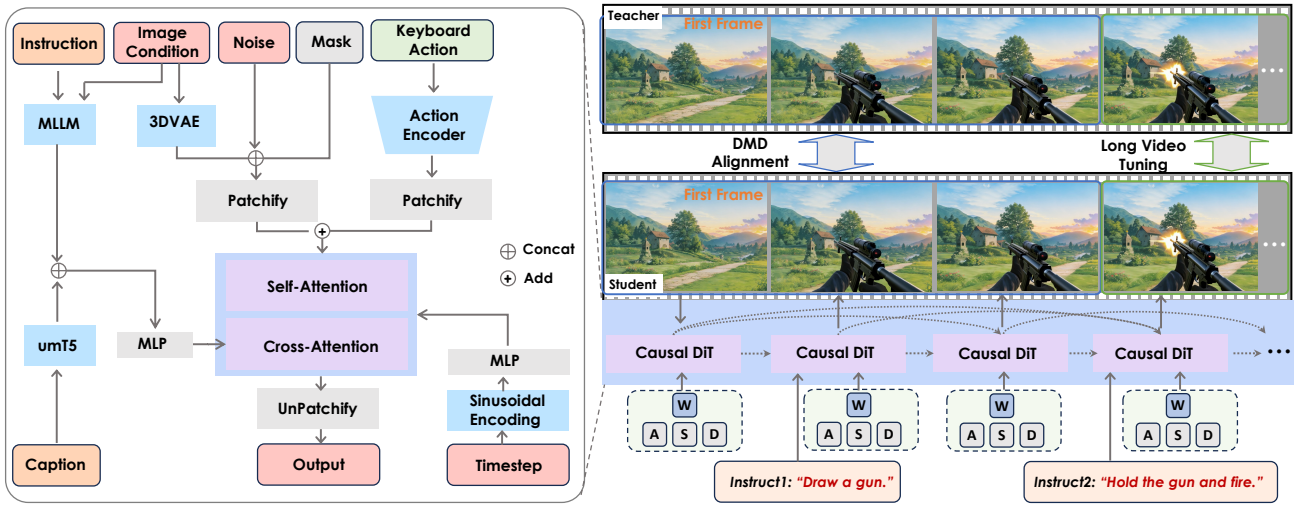

Figure 5. Model architecture of GameCraft-2. Given a reference image and the corresponding action, the keyboard/mouse signal, and prompt-based instruction, we inject these options to the main architecture (See Sec. 4.1). During training and inference, we leverage self-forcing post-training for long-video extension(See Sec. 4.2), and KV-cache/recache for multi-action switching(See Sec. 4.3). To maintain the long-term video quality, we design a randomized long video tuning scheme(See Sec. 4.2).

plexity of interaction texts compared to scene descriptions. Such texts are often tightly coupled with specific visual regions or object instances. To mitigate this, we leverage a multimodal large language model (MLLM) [57] to extract, reason and inject interaction information to the main model, which can enrich interaction-related textual guidance, improving the model's ability to differentiate between general text instructions and fine-grained interactive behaviors during training. This camera-conditioned control, when combined with text-based scene and interaction inputs, forms a unified mechanism that enables GameCraft-2 to navigate and interact seamlessly within its environment.

4.2. Training Procedure

To achieve long-term and real-time interactive video generation, it's necessary to distill the foundational bidirectional model into a few-step causal generator. In this work, we scale the comprehensive autoregressive distillation technique, Self-Forcing [26], to a 14B Mixture-of-Experts (MoE) image-to-video model. This scheme is specifically tailored to enhance generation quality and efficiency for long video generation, which often features large and rapid scene variations. We introduce random extension tuning to mitigate error accumulation. The training process is organized into four major stages: (1) Action-Injected Training, (2) Instruction-Oriented Supervised Fine-Tuning, (3) Autoregressive Generator 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 the model's interactive capabilities, we constructed a dataset of 150K samples by augmenting real-world footage with procedurally generated synthetic videos (details in Sec. 3). These synthetic sequences are pivotal, as they provide high-fidelity supervision across diverse interaction types (e.g., state transitions, subject interactions), thereby establishing a tight correspondence between actions and their visual outcomes. In the subsequent stage, we freeze the camera encoder's parameters and exclusively fine-tune the MoE experts. This process is designed to refine the model's 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 [26, 48, 55, 62]. In prac-

Table 2. **Detailed training configurations across different stages. CP denotes context parallelism.**

Training Stage	Dataset	Data type	CP	#iters
Action-Injected Training	1M	Game-play & Render Video	1	100k
Instruction-Oriented SFT	150K	Game-play & Synthetic Video	1	20k
Autoregressive Generator Distillation	200K	Game-play & Synthetic Video	4	10K
Randomized Long-Video Extension Tuning	100K	Game-play Long video	4	3K

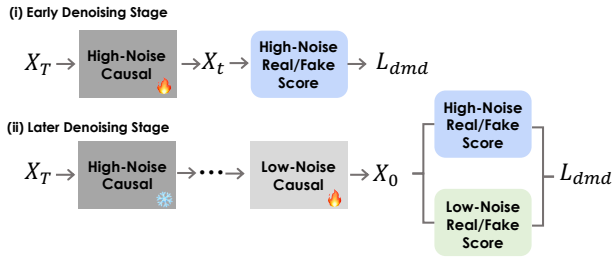

Figure 6. **Distillation Schedule for Self-Forcing post training on the MoE Model.**

tice, we scale the Self-Forcing [26] distillation scheme to a 14B Mixture-of-Experts (MoE) image-to-video model [56]. Building upon its 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: Previous arts [26, 63] updates the KV cache for causal attention using a direct sliding-window approach. However, this can lead to a degradation in generation quality over time, as later steps cannot reference the initial conditioning frame, causing drift. Therefore, inspired by prior work [48, 55, 62], we designate the initial frame as a sink token, which is always retained in the KV cache. This modification serves two critical functions: firstly, it improves and stabilizes generation quality. Secondly, in our specific task, the sink token provides information about the coordinate system origin. This ensures that camera parameters injected during the autoregressive process remain consistently aligned with the initial frame, thereby avoiding the need for a recache at each autoregressive step due to shifts in the coordinate origin. Additionally, we employ Block Sparse Attention [16] for local attention, better suited for our autoregressive, block-wise generation process. Specifically, the target block being generated can attend to a set of preceding blocks. This local attention, combined with the aforementioned sink attention, constitutes the full KV cache, enhancing generation quality while also accelerating the generation speed.

Distillation Schedule: Due to the unique nature of the MoE architecture, the high-noise expert presents greater challenges in training and convergence than the low-noise expert [56], particularly during SFT or distillation. To address this, we assign distinct learning rates to each expert. Concurrently, we re-define the target list of denoising timesteps for distillation based on the noise level boundary that separates the two experts. This ensures that the teacher and student models maintain consistency in their selection of the high- or low-noise expert during the distillation process.

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 [10, 62], roll out long video sequences from a causal generator and apply distributional moment distance (DMD) 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 (either the ground truth or teacher priors). 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. To maintain consistency with the training strategy, our inference process employs a fixed-length self-attention KV cache with a rolling update mech-

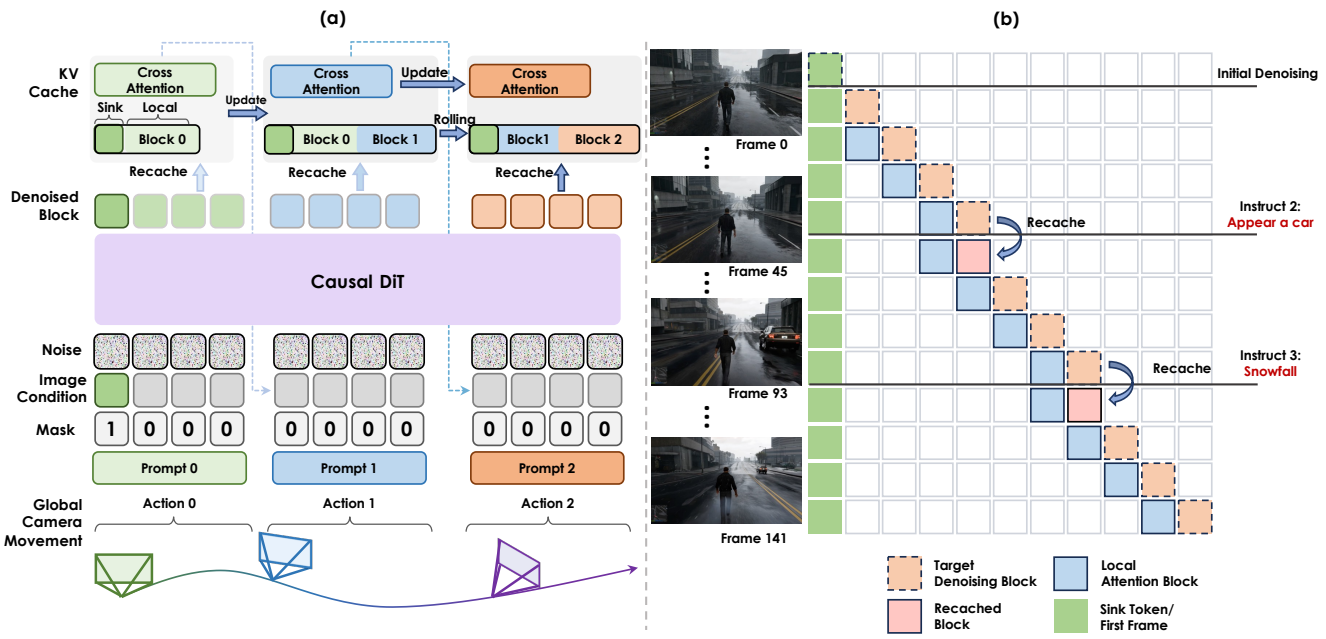

Figure 7. **Multi-turn Interactive Inference.** Figure 7.(a) illustrates our block-wise autoregressive inference pipeline for long video generation, along with the corresponding KV cache updating mechanism. The initial frame is retained as sink tokens at the start of the KV cache window, while local attention is derived from the recently generated blocks; The prompt recaching mechanism is depicted in Figure 7.(b), this strategy effectively enhances both the responsiveness and accuracy of the interaction when processing new prompts.

anism to facilitate efficient autoregressive generation, as depicted in Fig. 7. Specifically, sink tokens is permanently retained at the beginning of the cache window. The subsequent segment functions as a local attention window, maintaining the N frames preceding the target denoising block throughout multi-turn interactions. The complete KV cache is composed of these sink tokens and the local attention component, which is implemented using block sparse attention. This design not only enhances autoregressive efficiency but also effectively prevents 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.

4.4. Real-time Interaction Acceleration

Adopting the Distribution Matching Distillation (DMD) paradigm, we compress the model's denoising process into 4 steps, eliminating the need for Classifier-Free Guidance (CFG) to significantly reduce computational overhead. To further accelerate inference and ensure low latency, we implement a suite of advanced system optimizations:

- **FP8 Quantization:** We utilize 8-bit floating-point precision for model weights and activations. This reduces memory bandwidth consumption and leverages the hardware acceleration capabilities of modern GPUs, maintaining generation quality while maximizing throughput.
- **VAE Parallel Inference:** To address the bottleneck of decoding long video sequences, we parallelize the VAE decoding process. By processing multiple latent frame chunks simultaneously rather than sequentially, we drastically reduce the final image reconstruction time.
- **SageAttention:** We replace FlashAttension with SageAttention [65], a highly optimized attention kernel. This method quantizes the Q/K matrices and optimizes memory access patterns, significantly speeding up the core attention mechanism within the transformer backbone.
- **Sequence Parallelism:** To handle long-context video generation efficiently, we adopt sequence parallelism. This splits the video token sequence across multiple GPUs, allowing for the parallel computation of attention and feed-forward layers, thereby scaling our inference capability to longer horizons.

Collectively, these techniques boost the inference speed from 2 FPS to 16 FPS, successfully achieving the performance threshold required for real-time interaction.

Algorithm 1 Randomized Extended Long-Video Tuning

Require: Student G_θ, Real Score T_{real}, Fake Score T_{fake}, Dataset $\mathcal{D}$, Cache size L, Window K, Max length $N_{\max}$, Timesteps $\{t_1, \ldots, t_T\}$

loop
  $V_{\text{gt}} \sim \text{Sample}(\mathcal{D})$ *# Sample a ground truth video*
  $N \sim \text{Sample}(\mathcal{U}(K, N_{\max}))$ *# Randomize rollout length*
  $V_{\text{pred}} \leftarrow [V_{\text{gt}}[0]]$, KV $\leftarrow \emptyset$ *# Initialize with the first frame*
  Step 1: Autoregressive Rollout
  for $j = 1$ to N/K **do**
    $V_{\text{prev}} \leftarrow \text{LastKFrames}(V_{\text{pred}}, K)$
    # Extend sequence autoregressively
    V_{chunk}, KV $\leftarrow G_\theta(V_{\text{prev}}, \text{KV}, \text{sink\_token})$
    $\text{Append}(V_{\text{pred}}, V_{\text{chunk}})$
  end for
  Step 2: Randomized Window Sampling
  $i \sim \text{Sample}(\mathcal{U}\{1, \ldots, N - K + 1\})$
    # Uniformly sample a predicted window
  $W \leftarrow V_{\text{pred}}[i : i + K - 1]$
  Step 3: Interleaved Forcing Logic
    # Self-Forcing: Condition on predicted history
  $c_{\text{student}} \leftarrow V_{\text{pred}}[i - 1]$
    # Teacher-Forcing: Condition on ground truth
  $c_{\text{teacher}} \leftarrow V_{\text{gt}}[i - 1]$
  Step 4: Distillation with Distinct Conditions
  $t \sim \text{Sample}(\{t_1, \ldots, t_T\}), \epsilon \sim \mathcal{N}(0, \mathbf{I})$
    # Apply forward diffusion noise
  $x_t(W) \leftarrow \text{NoiseSchedule}(W, t, \epsilon)$
    # Calculate DMD loss
  $\mathcal{L} \leftarrow \text{DMD}\big(T_{\text{fake}}(x_t(W), t, c_{\text{student}}), T_{\text{real}}(x_t(W), t, c_{\text{teacher}})\big)$
    # Update generator parameters
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
end loop

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**. For fairness, all baselines are evaluated under their recommended or commonly adopted inference configurations, detailed as follows:

- **HunyuanVideo.** We use the official configuration with the following settings: `FLOW_SHIFT=7.0`, `EMBEDDED_CFG_SCALE=6.0`, 50 denoising steps, and with `flow_reverse` and `i2v_stability` enabled for enhanced temporal robustness.
- **Wan2.2 A14B.** We use the UniPC sampler, setting `sample_shift=5.0`, `sample_steps=40`, `boundary=0.900`, and using a dual-stage CFG with scales of (3.5, 3.5) for both noise regimes.
- **LongCatVideo.** We use the default high-quality inference setup with a `guidance_scale` of 4, 50 denoising steps, and enabled compilation optimizations for efficiency.

Resolution and Dataset. To comprehensively evaluate controllable video generation, we constructed a test suite organized around three core interaction dimensions: **(1) Environmental Interactions**, **(2) Actor Actions**, and **(3) Entity and Object Appearances**. To support this framework, we curated a custom test set of **100 images**, covering a wide diversity of scenes (indoor/outdoor, natural/urban), lighting conditions, and visual styles (realistic, game-like, cartoon). Furthermore, we built specialized subsets for specific actions, such as an additional **20 images** of closed doors for evaluating the *open door* action. For all evaluations, models are required to generate videos at a unified resolution of **832×448** and a fixed length of **93 frames**.

5.2. Evaluation Metrics.

To comprehensively evaluate the performance of our model in video generation, we employ two complementary families of metrics: general video-quality metrics and our interaction-focused evaluation suite, **InterBench**. The general metrics assess foundational aspects such as visual fidelity, temporal consistency, and motion realism, providing a baseline measurement of overall video quality. However, such metrics alone are insufficient for capturing the causal structure, action execution, and state transitions that are essential to interactive video generation. To bridge this gap, InterBench introduces **six interaction-centric dimensions**, each specifically designed to assess core properties of interactive behavior—including interaction completeness, action effectiveness, causal coherence, and physical plausibility. Together, these two metric families form a holistic and rigorous evaluation framework for interactive video models.

5.2.1. General Metrics.

To provide a comprehensive assessment of our model, we adopt a diverse set of evaluation metrics. For **video realism**, we use the Fréchet Video Distance (FVD) [53], which jointly captures spatial fidelity and temporal dynamics. **Visual quality** is quantified using Image Quality and Aesthetic scores, reflecting both low-level perceptual clarity and higher-level visual appeal. We further measure temporal consistency to evaluate cross-frame coherence and detect artifacts such as flickering or structural instability. For **dynamic performance**, we adapt the Dynamic Degree metric from VBench [27]. Instead of the original binary motion classification, we directly report absolute optical flow magnitudes, referred to as Dynamic Average. This continuous formulation provides a more nuanced characterization of motion intensity and naturalness.

For **interactive camera control performance**, we employ a multi-faceted evaluation protocol. We use the **Relative Pose Error** (**RPE trans** and **RPE rot**) to measure trajectory control accuracy, computed after applying a Sim3 Umeyama alignment between the predicted reconstructed trajectory and the ground truth. This alignment removes

Table 3. Composition of our curated evaluation test set. The data is hierarchically organized by high-level category, sub-category, and complexity level (Basic and Extended), with the number of prompts specified for each fine-grained subset.

Category	Sub-category	Level	Subset	Num
Environmental Interactions	Weather	Basic	Snow	**100**
			Rain	**100**
			Lightning	**100**
	Physical event	Extended	Explosion	**100**
Actor Actions	Primitive actions	Basic	Draw gun	**100**
			Draw knife	**100**
			Take out torch	**100**
	Composite actions	Extended	Draw and fire gun	**100**
			Take out and operate phone	**100**
			Open door	**20**
Entity & Object Appearances	Animals	Basic	Cat	**25**
		Basic	Dog	**25**
		Basic	Wolf	**25**
		Basic	Deer	**25**
		Extended	Dragon	**100**
	Vehicles	Basic	Red SUV	**25**
			Blue truck	**25**
			Yellow sports car	**25**
			Black off-road car	**25**
	Humans	Extended	Human appearances	**100**

Table 4. Quantitative comparison with recent related works.

Model	Visual Quality				Temporal	RPE		FPS↑
	FVD↓	Image Quality↑	Dynamic Average↑	Aesthetic↑	Temporal Consistency↑	Trans↓	Rot↓	
GameCraft	1554.2	0.69	67.2	0.67	0.95	0.08	0.20	0.25
GameCraft-PCM	1883.3	0.67	43.8	0.65	0.93	0.08	0.20	6.6
Matrix-Game	2260.7	0.72	31.7	0.65	0.94	0.18	0.35	0.06
Matrix-Game-2.0	1920.6	0.62	20.5	0.49	0.84	0.08	0.25	16
GameCraft-2								16

scale and global pose discrepancies, allowing RPE to specifically reflect local motion fidelity and frame-to-frame control precision. By examining both translational and rotational components, the metric provides a clearer view of how accurately the model responds to interactive inputs and how reliably it maintains the intended motion trajectory.

5.2.2. InterBench: Benchmarking Action-Level Interaction in Video Generation

To rigorously assess *action-level interaction* in generated videos, we propose **InterBench**, a six-dimensional evaluation protocol tailored to interactive video generation. Instantiated using a vision-language model (VLM) as an automatic evaluator, InterBench is designed to measure not only whether an interaction is triggered, but also its fidelity, smoothness, and physical plausibility over time. The six core dimensions are defined below. For a comprehensive discussion of the protocol, please refer to Appendix C.

1. **Interaction Trigger Rate.** A fundamental binary metric that assesses whether the requested interaction was successfully initiated. This serves as a gateway check, separating cases where the model completely ignored the prompt from those where it attempted the action.
2. **Prompt–Video Alignment.** Evaluates the semantic fidelity between the video and the full prompt. This

Figure 8. **Inference results by Hunyuan-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.**

dimension has two facets: *static alignment* (maintaining the scene's context and objects) and *dynamic alignment* (executing the correct action as specified).

3. **Interaction Fluency.** Measures the temporal naturalness and visual coherence of the interaction process. It penalizes temporal artifacts such as sudden jumps, flickering, or object teleportation that break the illusion of continuous motion and a stable timeline.
4. **Interaction Scope Accuracy.** Examines whether the spatial extent of an interaction's effects is appropriate. It ensures that global events (like *weather changes*) affect the entire scene, while local actions (like *"lighting a torch"*) have a contained but realistic area of influence.
5. **End-State Consistency.** Assesses whether the interaction converges to a stable and correct final state that persists until the end of the video. This distinguishes successful actions from those that are only partially completed or whose effects vanish prematurely.
6. **Object Physics Correctness.** Evaluates the physical plausibility of interacting agents and objects. This includes maintaining the structural integrity of rigid bodies (no unnatural deformation), ensuring realistic motion kinematics, and preserving correct contact relationships (e.g., no penetration between hands and objects).

Scoring Protocol. Each video is evaluated against the InterBench dimensions using a discrete, ordinal scoring system. Specifically, **Interaction Trigger Rate** is assessed with a binary value (success/failure), while the remaining five dimensions receive multi-level ordinal scores to capture varying degrees of interaction quality. These per-video scores are then averaged to produce a score for each interaction category. A final, global InterBench score is obtained by aggregating these category-level results. This hierarchical scoring protocol enables both fine-grained analysis of specific failure modes and high-level comparison of interactive capabilities across different models.

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.

5.3. Interaction Evaluation

Quantitative Results on Interaction Evaluation We present the quantitative results for the three interaction categories in Table 5. The evaluation follows our proposed

Figure 9. **Qualitative examples illustrating the six dimensions of our InterBench protocol.** Each row showcases a distinct evaluation dimension: **(1) Interaction Trigger Rate, (2) Prompt–Video Alignment, (3) Interaction Fluency, (4) Interaction Scope Accuracy, (5) End-State Consistency**, and **(6) Object Physics Correctness**. For each dimension, we present a high-scoring example from our model (right) against a low-scoring failure case from a baseline (left), demonstrating the rating scale used in our evaluation. GameCraft-2 consistently achieves higher ratings across all dimensions.

InterBench protocol (Sec. 5.2.2), structured around its six core dimensions: **Trigger**, **Align**, **Fluency**, **Scope**, **End-State**, and **Physics**. To provide a single aggregated metric for comparison, we also compute a weighted *Overall* score:

$$\text{Overall} = \frac{1}{6}\Big(5 \times \text{Trigger} + \text{Align} + \text{Fluency} + \text{Scope} + \text{EndState} + \text{Physics}\Big).$$

A category-wise quantitative analysis (Table 5) reveals that GameCraft-2's superiority begins with its exceptionally high success rate in initiating interactions. The model achieves `Trigger` scores of 0.962 for **Environmental Interactions** and a near-perfect 0.983 for **Actor Actions**, far surpassing all baselines. Beyond successful initiation, GameCraft-2 excels in modeling the fidelity of these interactions. This is particularly evident in its physical realism, where it outperforms the next-best model by margins of 0.683 in `Physics` for Environmental Interactions and over 0.52 in **Entity & Object Appearances**. Furthermore, it demonstrates substantial gains in temporal coherence and final state stability, with `Fluency` and `EndState` scores improving by +0.70 and +0.63, respectively, for Actor Actions. Collectively, these results underscore GameCraft-2's advanced capability not only to trigger interactions reliably but also to render them with high fidelity in semantics, dynamics, and physical consistency.

Qualitative Analysis. For an intuitive demonstration of performance differences, we present a qualitative comparison in Figs. 19 -21. 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. In Environmental Interactions, its generated effects, such as snowfall, achieve global coverage and dynamic accumulation, rendering them more physically plausible. For Actor Actions, **GameCraft-2** produces more coherent action sequences, enabling characters to stably grasp and precisely manipulate objects while ensuring stable final states. In Entity & Object Appearances, the model consistently maintains the structural integrity and identity of objects, seamlessly integrating them into the scene's lighting and perspective. Crucially, this robustness extends to concepts outside our specific training categories; for instance, the model adeptly handles interactions involving a "Phone" or the appearance of a "Dragon", showcasing strong generalization capabilities. Collectively, these qualitative ex-

Table 5. **Quantitative performance evaluation of our model**, GameCraft-2, 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**

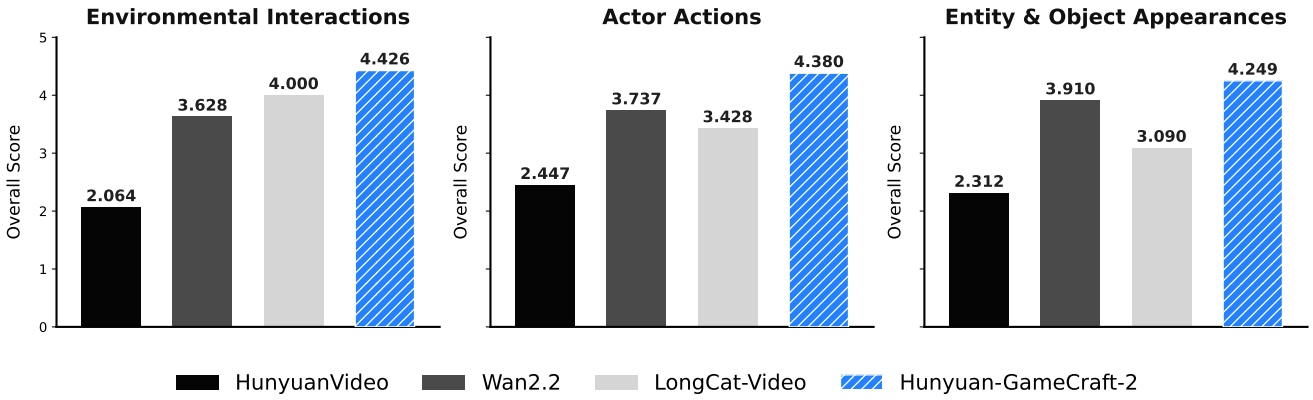

Figure 10. **Comparison of Environmental Interactions with Baseline Models.** Qualitative results showing the fidelity and consistency of environment-level effects. Our approach better preserves global influence and temporal stability.

amples not only corroborate our quantitative findings but also concretely showcase **GameCraft-2**'s robust capability to generate semantically accurate, temporally coherent, and physically plausible videos of complex interactions.

6. 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 extremely long sequences. 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.

7. Conclusion

In this work, we introduced **GameCraft-2**, an interactive game world model capable of generating high-fidelity, controllable video in response to free-form text instructions and keyboard/mouse actions. We formally defined interactive video data and proposed automated pipelines for its curation and synthesis, effectively addressing the data bottleneck that has hindered progress in this domain. Our model unifies multimodal control signals within a robust training framework, leveraging a novel randomized long-video tuning scheme and efficient inference mechanisms like KV-

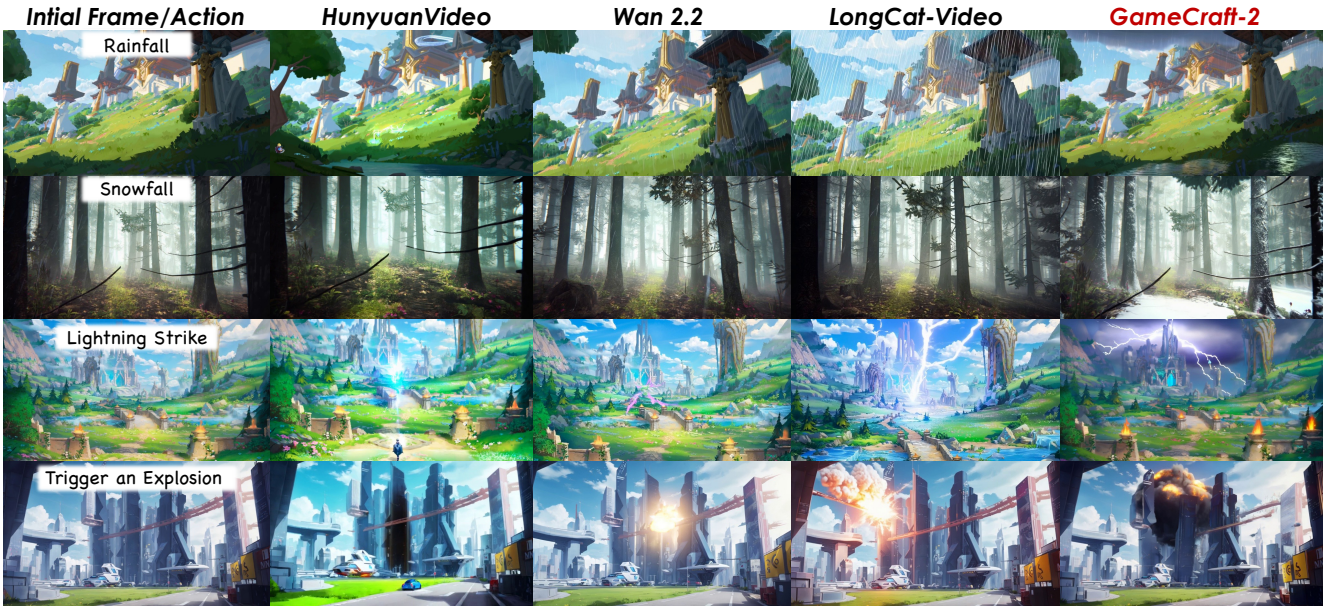

Figure 11. **Comparison of Environmental Interactions with Baseline Models.** Qualitative results showing the fidelity and consistency of environment-level effects. Our approach better preserves global influence and temporal stability.

recache to achieve stable, long-horizon, and real-time interactive generation. To rigorously evaluate our contributions, we introduced **InterBench**, a new benchmark specifically designed to assess action-level interaction quality. Extensive experiments demonstrate that GameCraft-2 significantly outperforms existing state-of-the-art models across all dimensions of interaction fidelity, visual quality, and temporal coherence. By pushing the frontier from passive video synthesis to active, user-driven world generation, our work marks a significant step toward creating truly playable and immersive AI-generated virtual experiences.

References

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

[2] 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.

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

[4] 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.

[5] 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.

[6] Brandon Castellano. PySceneDetect.

[7] Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video

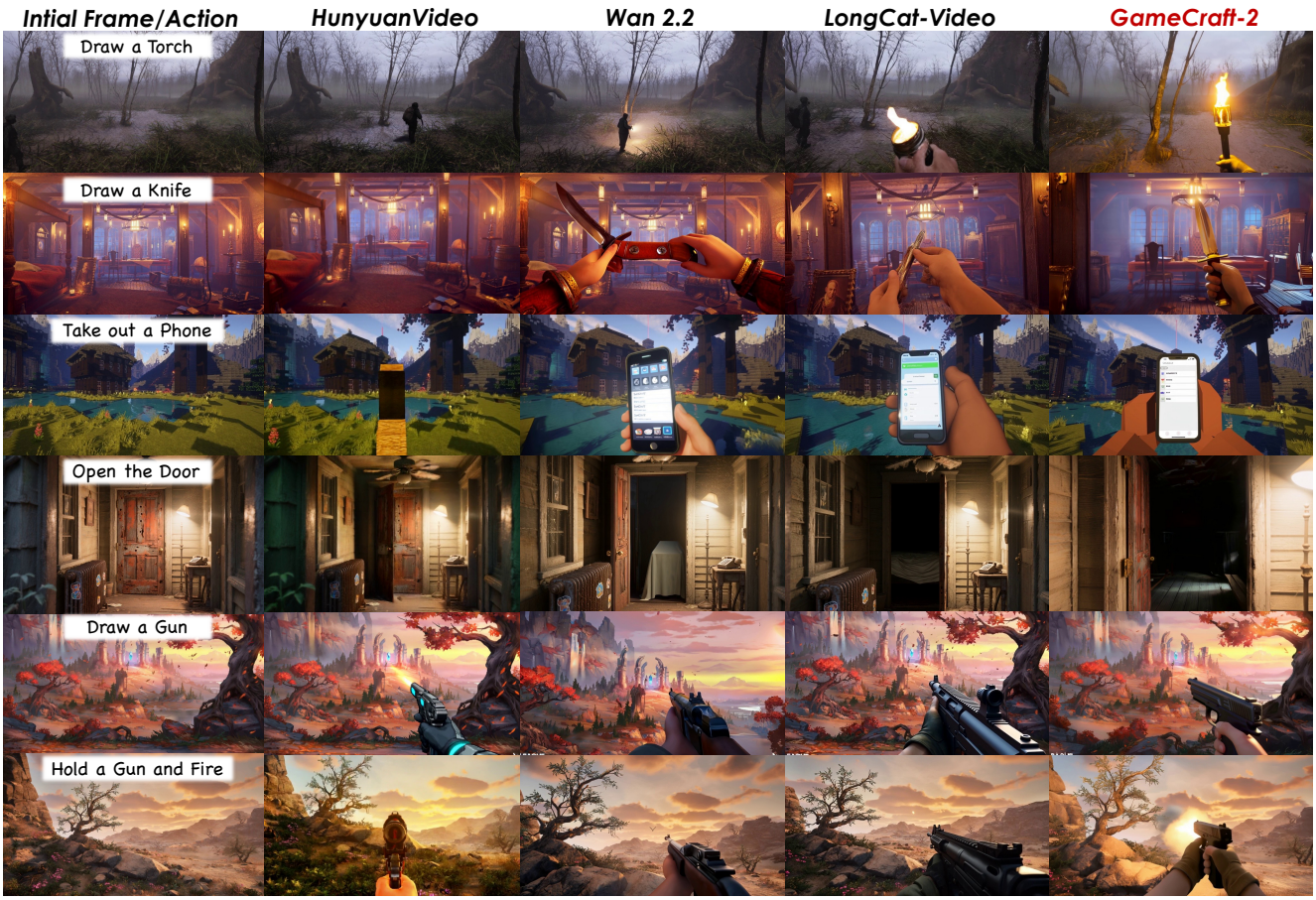

Figure 12. **Comparison of Actor-Action Interactions with Baseline Models.** Visual comparisons illustrating the quality of action-level interactions across representative prompts. Our method produces more coherent and physically consistent actions than all baselines.

generation. In *International Conference on Learning Representations*, 2025.

[8] 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.

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

[10] 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.

[11] 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.

[12] Decard. Oasis: A universe in a transformer. https://www.decart.ai/articles/oasis-interactive-ai-video-game-model, 2024.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] Roberto Henschel, Levon Khachatryan, Hayk Poghosyan,

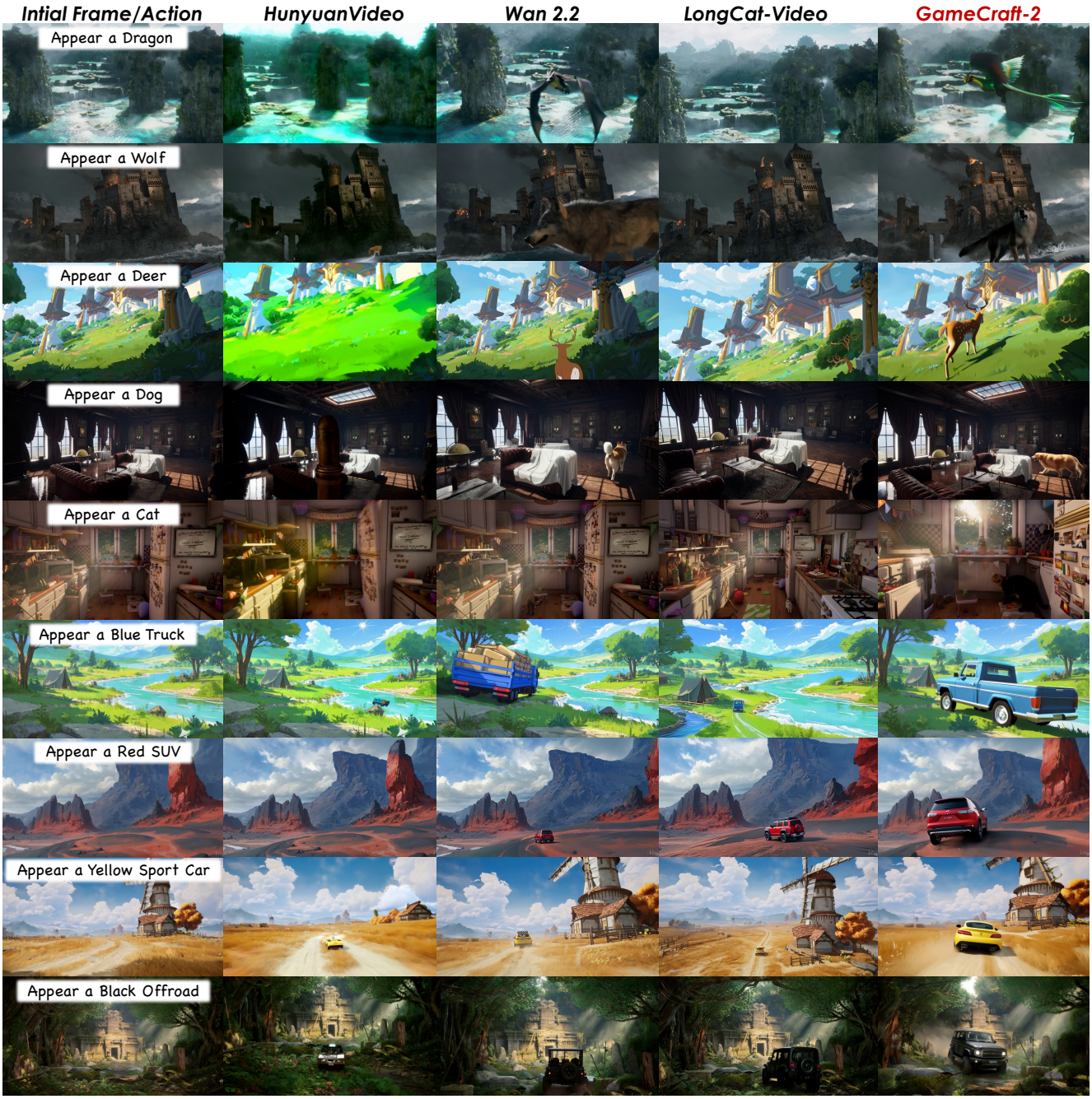

Figure 13. **Comparison of Entity and Object Appearance Interactions with Baseline Models.** Visual comparisons of object emergence and interaction correctness. Our method delivers more accurate, stable, and physically plausible object behaviors.

Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024.

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

[21] 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.

[22] 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.

[23] Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi,

Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft, 2025.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

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

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

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

[35] 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.

[36] 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.

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

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

[39] 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.

[40] 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.

[41] 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.

[42] Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model, 2025.

[43] 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.

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

[45] 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.

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

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

[48] 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.

[49] 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.

[50] 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.

[51] PAN Team, Jiannan Xiang, Yi Gu, Zihan Liu, Zeyu Feng, Qiyue Gao, Yiyan Hu, Benhao Huang, Guangyi Liu, Yichi Yang, Kun Zhou, Davit Abrahamyan, Arif Ahmad, Ganesh Bannur, Junrong Chen, Kimi Chen, Mingkai Deng, Ruobing Han, Xinqi Huang, Haoqiang Kang, Zheqi Liu, Enze Ma, Hector Ren, Yashowardhan Shinde, Rohan Shingre, Ramsundar Tanikella, Kaiming Tao, Dequan Yang, Xinle Yu, Cong Zeng, Binglin Zhou, Zhengzhong Liu, Zhiting Hu, and Eric P. Xing. Pan: A world model for general, interactable, and long-horizon world simulation, 2025.

[52] 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.

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

[54] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

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

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

[61] 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].

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

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

A. 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.

A.1. Positive Examples: Interactive Video Data

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

A.2. Negative Examples: Non-Interactive Video Data

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 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.

A.3. 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.

B. 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.

B.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. 14), as well as a comprehensive range of environments spanning natural and urban scenes under various lighting, weather, and terrain conditions (Fig. 15). This diversity is crucial for training robust and generalizable world models.

B.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. 16), **Actor Actions** involving complex body-object coordination (Fig. 17), and **Entity/Object Appearances** that introduce new subjects into the scene with high fidelity (Fig. 18).

Figure 14. **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.

Figure 15. **Comprehensive Environment Diversity in Our Dataset.** Examples of outdoor and indoor environments across natural and urban scenes, under diverse lighting, weather, and terrain conditions.

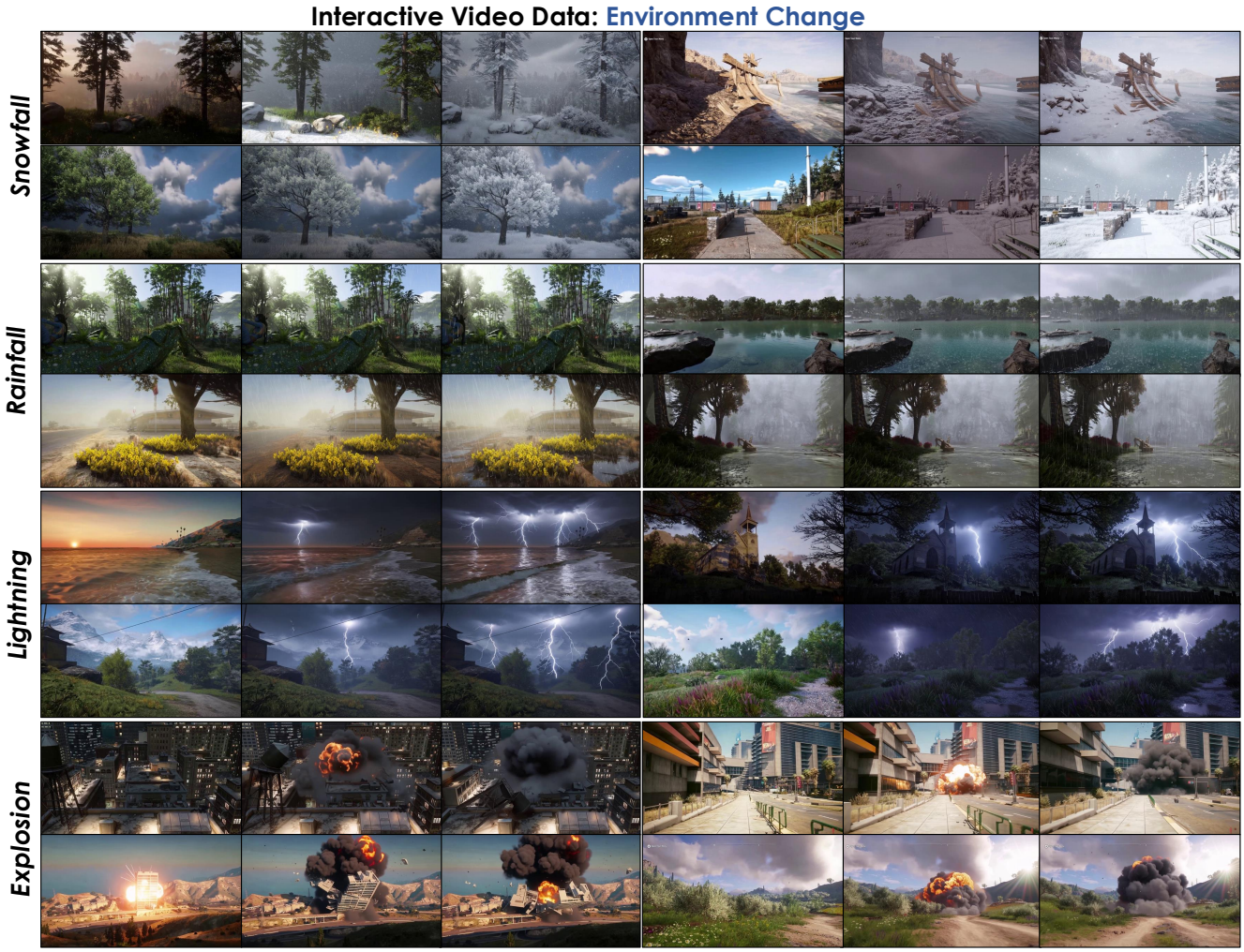

Figure 16. **Synthetic Examples of Environmental Interactions.** Examples of synthetic scene-change interactions generated by our pipeline, covering *snowfall*, *rainfall*, *lightning*, and *explosions*.

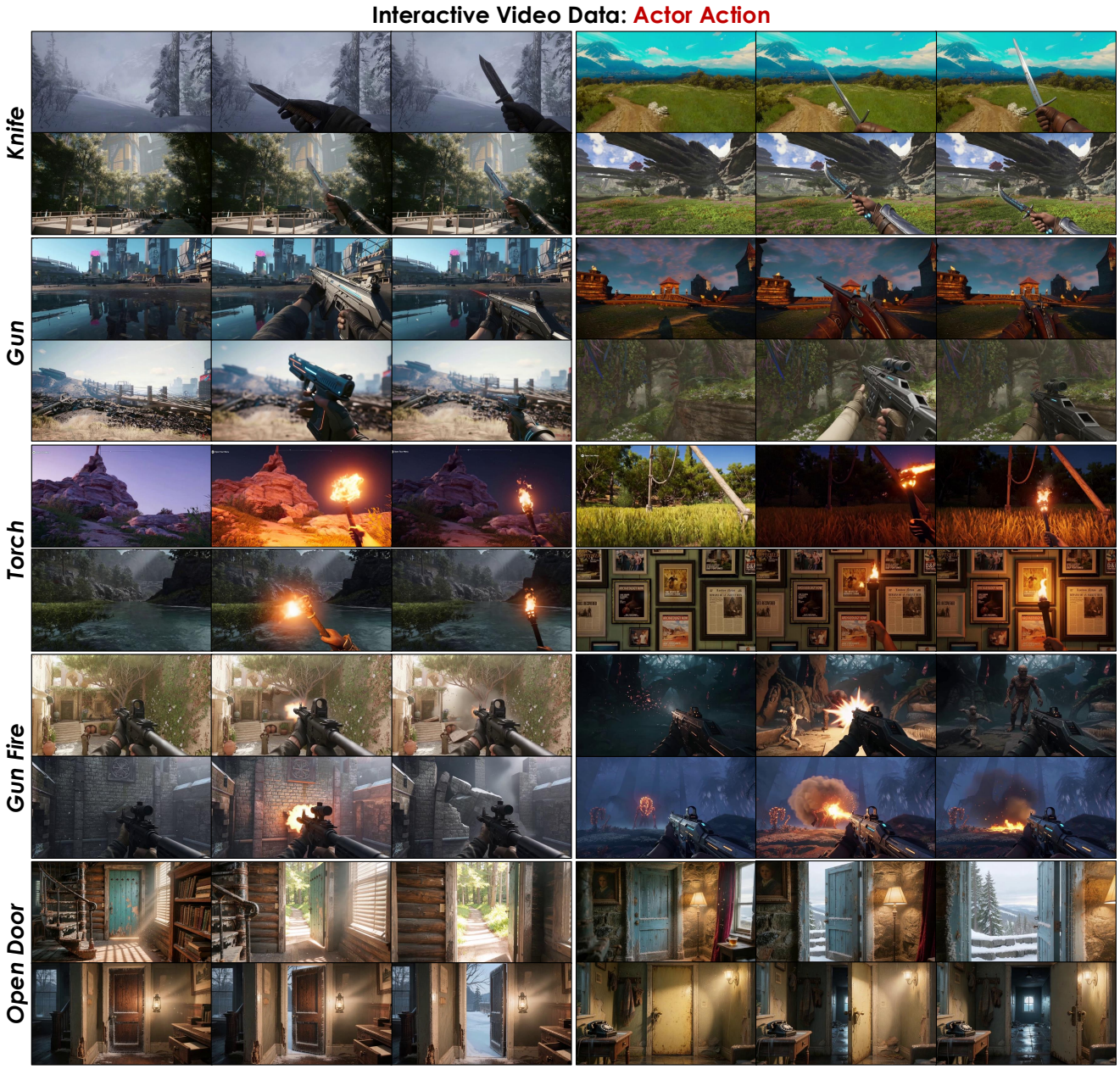

Figure 17. **Synthetic Examples of Actor Actions.** Examples illustrating the range of interactive behaviors synthesized by our pipeline. ***Actor-driven interactions*** include *drawing a knife*, *drawing a gun*, *drawing a torch*, *firing a weapon*, and *opening a door*, which require consistent body–object coordination and temporal coherence.

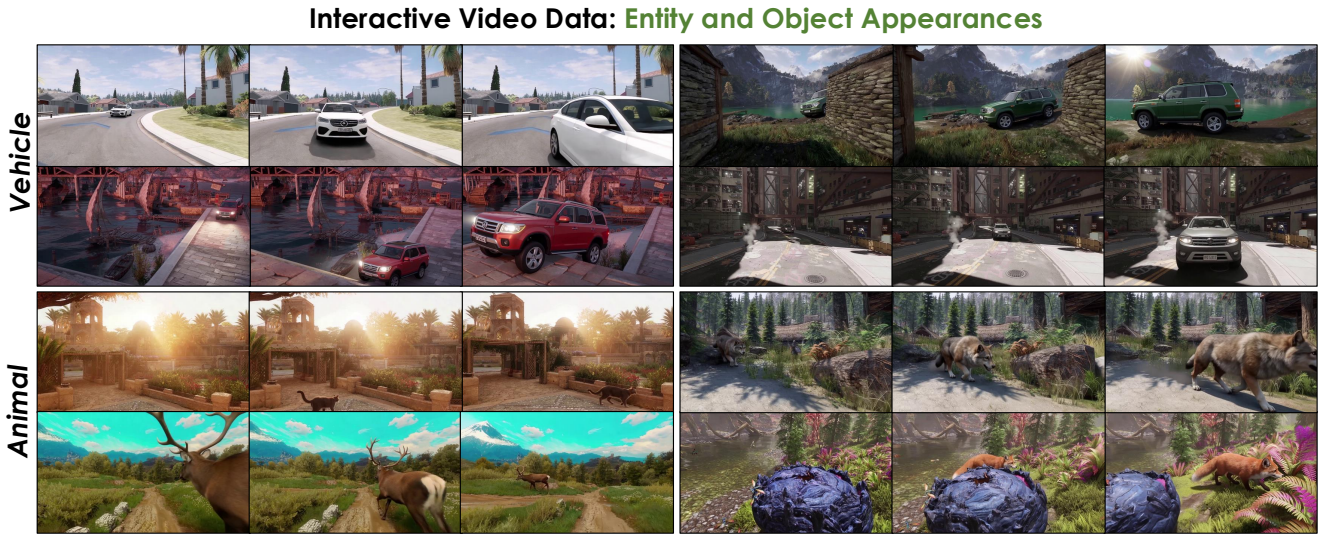

Figure 18. **Synthetic Examples of Entity/Object Appearances.** Examples illustrating the range of interactive behaviors synthesized by our pipeline. ***Entity- and object-level interactions*** include *animal intrusion* and *vehicle entry*, showing the pipeline's capability to introduce new entities with realistic geometry, scale consistency, and stable identity across frames.

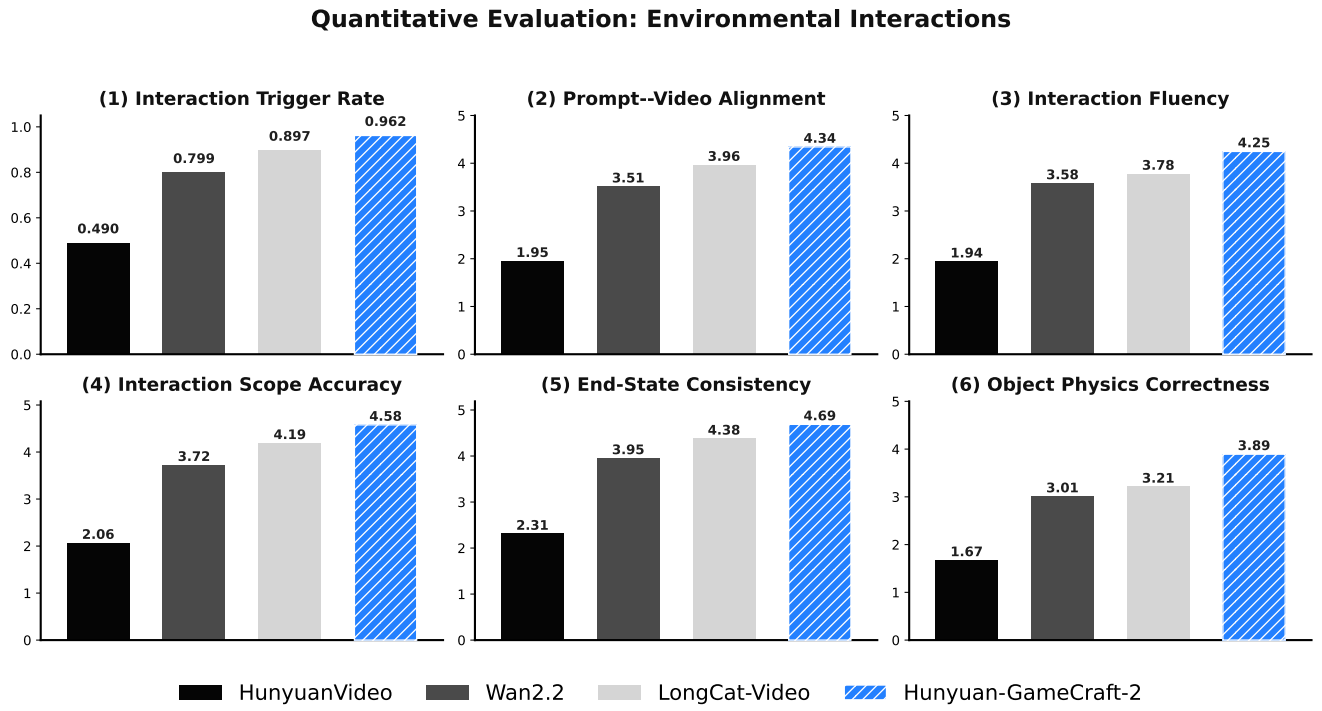

Figure 19. **Comparison of Environmental Interactions with Baseline Models.** Qualitative results showing the fidelity and consistency of environment-level effects. Our approach better preserves global influence and temporal stability.

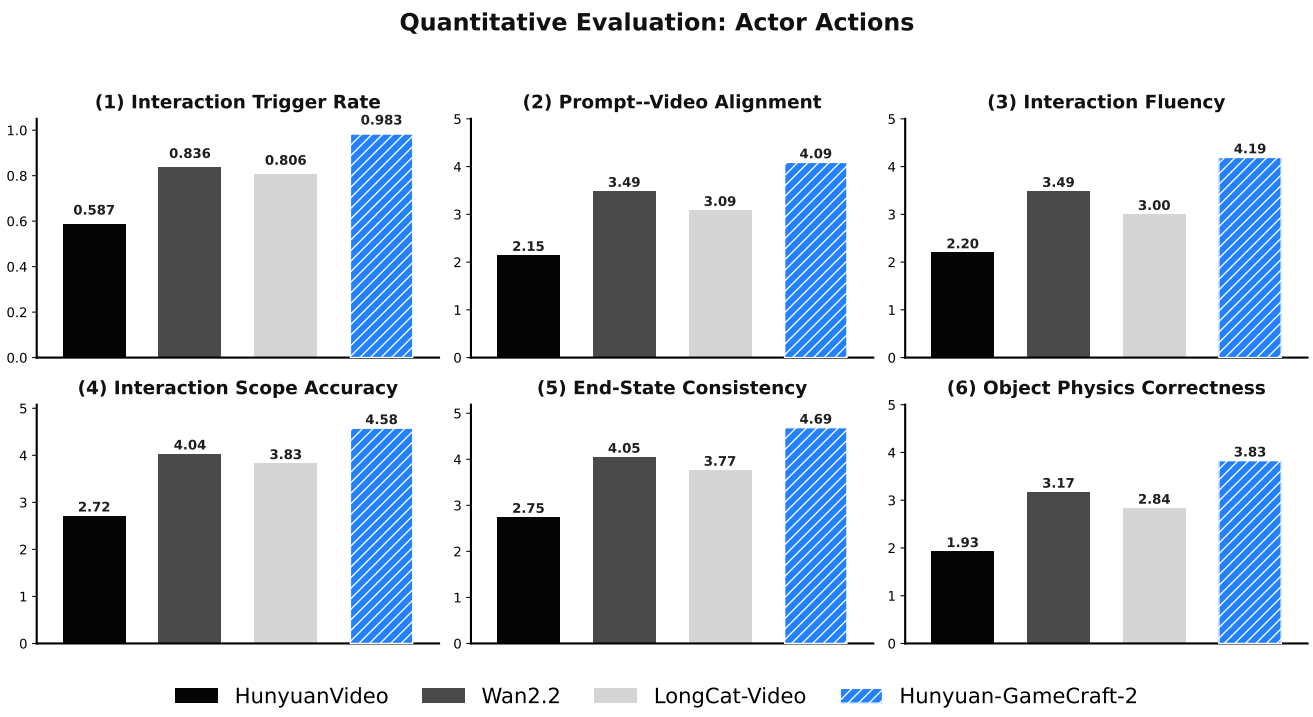

Figure 20. **Comparison of Actor-Action Interactions with Baseline Models.** Visual comparisons illustrating the quality of action-level interactions across representative prompts. Our method produces more coherent and physically consistent actions than all baselines.

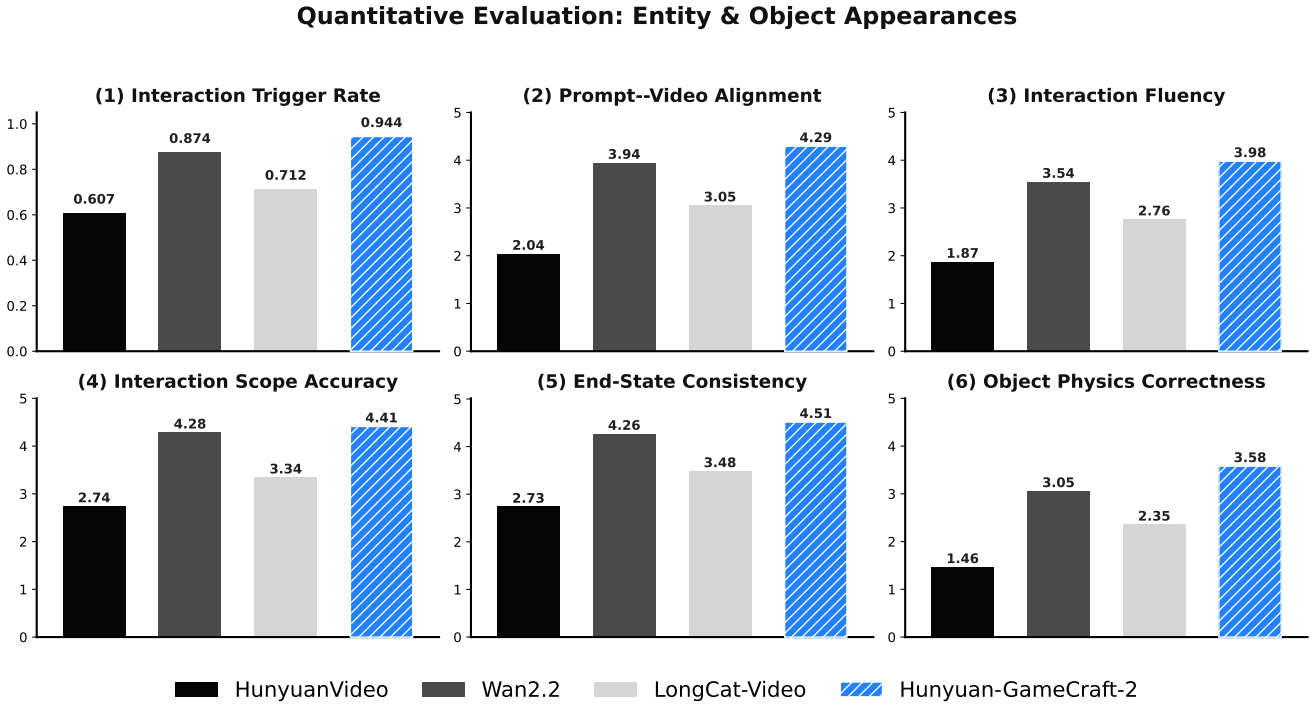

Figure 21. **Comparison of Entity and Object Appearance Interactions with Baseline Models.** Visual comparisons of object emergence and interaction correctness. Our method delivers more accurate, stable, and physically plausible object behaviors.

C. 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 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 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.