Position: Interactive Generative Video as Next-Generation Game Engine

Anonymous ICCV submission

Paper ID *****

Abstract

Modern game development faces significant challenges in creativity and cost due to predetermined content in traditional game engines. Recent breakthroughs in video generation models, capable of synthesizing realistic and virtual environments, present an opportunity to revolutionize game creation. In this position paper, we propose Interactive Generative Video (IGV) as the foundation for Generative Game Engines (GGE), enabling unlimited novel content generation in next-generation gaming. GGE leverages IGV's unique strengths in unlimited high-quality content synthesis, physics-aware world modeling, user-controlled interactivity, long-term memory capabilities, and causal reasoning. We present a comprehensive framework detailing GGE's core modules and a hierarchical maturity roadmap (L0-L4) to guide its evolution. Our work charts a new course for game development in the AI era, envisioning a future where AI-powered generative systems fundamentally reshape how games are created and experienced.

1. Introduction

Computer games have witnessed an ever-growing market demand, yet the gaming industry faces three critical challenges. First, current game engines rely heavily on premade assets and fixed logic scripts, leading to predetermined content that players will eventually exhaust, even in modern open-world games. Second, existing game engines cannot provide adaptive, personalized gaming content tailored to individual players' preferences, habits, and backgrounds. Third, developing high-quality games, especially AAA games, requires substantial human resources and extensive development time. How to rapidly create high-quality games with unlimited personalized content while minimizing costs remains a fundamental challenge for the entire gaming industry.

During the past year, video generation models have made remarkable progress [2, 6, 17, 38, 40, 56, 62, 63, 70, 74, 89], demonstrating unprecedented capabilities in large-scale motion dynamics, semantic understanding, concept composition, 3D consistency with physical laws, and long-term temporal coherence in both object structure and appearance. These advances show great potential for effectively simulating real-world physics [56, 85, 88], suggesting that these models could serve as capable world models for generating physically plausible videos.

Building upon these advances in video generation, we propose Interactive Generative Video (IGV), a new paradigm that extends video generation capabilities with interactive features. IGV centers around video generation while incorporating four key characteristics: user control over the generated content, memory of video context, understanding and simulation of physical rules, and causal reasoning intelligence. By combining these elements, IGV effectively constructs an interactive virtual world through video generation, functioning similarly to a simulator.

The virtual worlds created by IGV naturally align with video games as they provide interactive environments where players can explore and engage with dynamically generated content, representing a promising direction for next-generation gaming. Recent works [3, 7, 9, 12, 15, 16, 21, 23, 26, 35, 73, 81, 86, 92] have demonstrated this potential by training action-conditioned video generation models using action-video pairs collected from classic games like Atari [3], DOOM [73, 86], CS:GO [3], Minecraft [15, 26, 81, 92], and Super Mario Bros [86]. These models create interactive gaming experiences by iteratively generating predicted video frames in response to user action inputs.

However, as pointed out by some works [16, 21, 92], merely replicating existing games through IGV offers limited value over traditional game engines. The revolutionary potential of IGV lies in its ability to create infinite entirely new games through its powerful generative capabilities. Imagine a future where everyone can become a game designer, creating their own games by simply providing design instructions to video generation models, which then generate explorable virtual worlds. This will fundamentally transform both game development and gaming experiences.

In conclusion, this position paper argues that **Interactive Generative Video (IGV) serves as the core technology**

for Generative Game Engine (GGE). GGE will reduce barriers in game development while boosting productivity and creativity through AI-driven content generation.

In this position paper, we first introduce preliminary knowledge about video generation and AI-driven game applications in Sec. 2. Sec. 3 analyzes the core capabilities required for next-generation Generative Game Engines (GGE) and demonstrates why Interactive Generative Video (IGV) is uniquely positioned to fulfill these requirements. Sec. 4 presents our comprehensive framework for GGE, providing detailed definitions, analysis, and future prospects for each module within the framework. To guide future research and development, Sec. 5 proposes a hierarchical roadmap that outlines progressive milestones toward fully functional GGE systems. Finally, Sec.6, Sec.7 and Sec. 8 discuss alternative perspectives, address potential ethical issues and provide concluding remarks.

2. Preliminaries

Detailed preliminaries are in the Supplementary Material.

Video Generation Models. Video generation models have achieved significant breakthroughs with the rise of diffusion models [29, 45, 48, 66, 67], which have become the mainstream approach due to their superior generation quality [2, 6, 17, 38, 40, 56, 62, 63, 70, 74, 89]. The field has also made substantial progress in conditional video generation [27, 54, 83], particularly in camera control [4, 22, 28, 79, 87], where methods like MotionCtrl [79] and CameraCtrl [28] enable precise manipulation of camera movements. For autoregressive video generation, which is crucial for creating variable-length or infinite video sequences, two representative approaches have emerged: GPT-like next-token prediction methods [19, 39, 77] and Diffusion Forcing [10, 65].

AI-driven Game Applications. AI technologies have demonstrated diverse applications in game creation. In game video generation, recent works leveraging diffusion models [3, 15, 73] have achieved high-quality results, with open-domain methods [16, 21, 92] even enabling the creation of novel game content. AI-powered design assistants have enhanced the game development process by automating design completion [64] and generating multiple design suggestions [44, 52], thereby streamlining development and fostering creativity. Furthermore, intelligent game agents have evolved from traditional reinforcement learning approaches [8, 33] to more sophisticated LLM-based methods [75, 78], significantly improving performance in long-horizon tasks.

3. Why IGV for Generative Game Engine?

Computer games have witnessed an ever-growing market demand, yet developing high-quality games, especially AAA games, requires substantial human resources and extensive development time. Traditional game engines like *Unreal* and *Unity* rely heavily on pre-made assets and fixed logic scripts, which not only limits game creation to predefined scenes and plots, but also means players will eventually exhaust all content. Even in open-world games like *The Legend of Zelda: Breath of the Wild*, while offering extensive freedom, players will ultimately experience all predetermined content. How to rapidly create high-quality and innovative games at scale while minimizing human costs remains a critical challenge for the entire gaming industry.

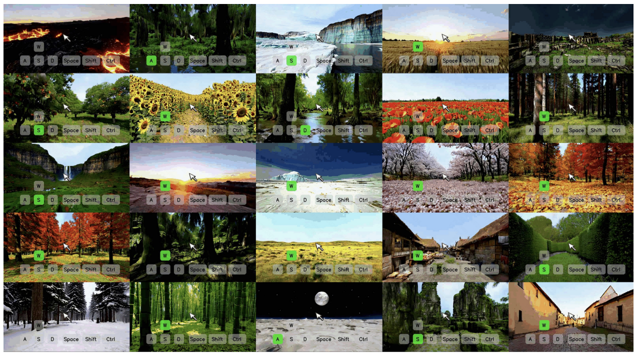


Figure 1. GameFactory [92]'s ability to generalize action control from Minecraft to open-domain scenarios.

We propose Generative Game Engine (GGE) as a next-generation solution that dynamically generates both assets and logic. This paradigm shift offers several key advantages: (1) lower development costs for game studios through automated content generation; (2) reduced entry barriers for individual developers by eliminating the need for extensive asset creation; and (3) truly open-world experiences with unlimited, dynamically generated content that provides endless unique gameplay experiences.

Building upon recent advances in video generation, we propose Interactive Generative Video (IGV) as a promising foundation for GGE implementation. As illustrated in Fig. 4 (a), IGV is not the entirety of GGE. From a definitional perspective, IGV is viewed from a technical angle, while GGE is viewed from an application angle. Specifically, IGV represents video generation technology that supports interactive user input control, while GGE represents a game engine that utilizes generative AI to create games. IGV, as a potential realization of GGE, offers four key advantages: (1) powerful generalizable generative capabilities, (2) physics-aware world modeling, (3) user-controlled generation for interactive experiences, and (4) leveraging vast video data for training. In the following subsections, we elaborate on these advantages in detail.

3.1. Generalizable Generation for Unlimited Games

Video generation models excel at creating not just high-quality content, but novel and diverse game content. Pretrained on vast real-world video collections, these mod-

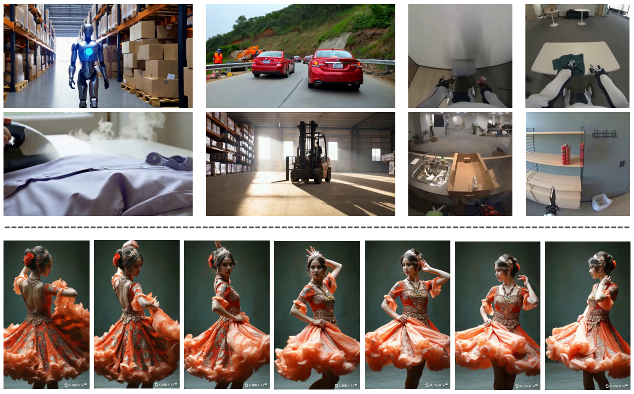


Figure 2. Physics-aware generation capabilities of video models. Top: Examples from Cosmos [55] demonstrating physical understanding in diverse scenarios including robotics, autonomous driving, manufacturing, and home environments. Bottom: Human motion examples generated by Kling [38].

els develop comprehensive understanding of visual elements and relationships. Their novelty manifests in two aspects: (1) generalization ability to transfer skills to unprecedented scenarios, as shown by GameFactory [92] generating action-controllable videos in open-domain settings (Fig. 1), and (2) compositional creativity to combine learned elements innovatively, demonstrated by Sora [56]'s "origami undersea" scenes[1]. This compositional capability has become a key research focus, with dedicated evaluation benchmarks [31, 68] measuring such abilities.

3.2. Physics-aware World Modeling

Video generation models demonstrate potential in understanding the inherent rules of the real world, particularly physical knowledge [56, 85, 88]. During training, to ensure accurate video prediction, these models naturally learn implicit physical priors embedded in training videos. These priors encompass various common physical phenomena, including gravity, elasticity, explosions, collisions, as well as complex motion patterns of humans and animals. While traditional game engines typically rely on predefined physical formulas, manual annotations, or motion capture, IGV leverages its learned physical priors to directly generate physically plausible content. This capability significantly simplifies game engine design and reduces the technical expertise required from developers, thereby enhancing game production efficiency. As shown in Figure 2, the generated video examples demonstrate IGV's physics-aware capabilities, highlighting its potential value for game development.

3.3. Interactive Generation with User Control

Precise control of visual generation models has made progress [13, 53, 57, 94]. Current video generation models support various control signals essential for gaming interactions. These control capabilities excel in intuitive operations such as camera viewpoint adjustment [4, 22, 28, 79] and character movement control [30]. Such precise and responsive control enables players to interact with generated content, creating engaging gaming experiences. With rapid development, more control signal types are being supported, further expanding interactive possibilities [34, 71]. Fig. 3 demonstrates IGV's strong interactive control capabilities and validates its potential for game development.

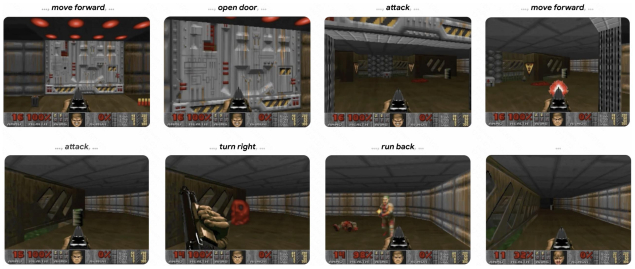


Figure 3. GameNGen [73] shows interactive gameplay in generated videos.

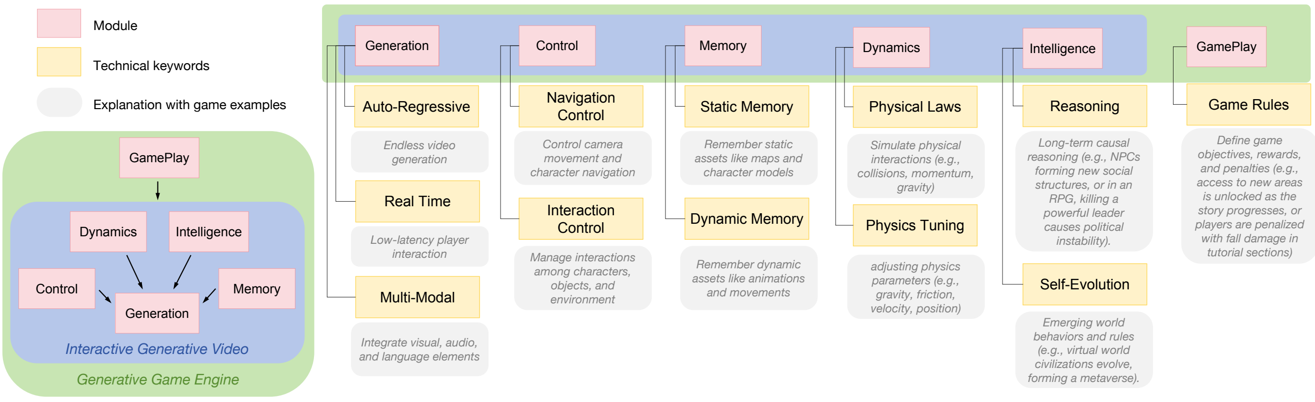
3.4. Video Data Accessibility Enables Scaling

Video data offers unique advantages for training generative game engines through its accessibility and unified representation format [88]. Unlike traditional game engines that require various heterogeneous assets (3D models, textures, animations, etc.) with manual effort, videos are widely available across internet platforms and continuously growing through social media and streaming services. Moreover, videos naturally capture diverse real-world phenomena and human experiences, enabling models to learn comprehensive world knowledge through large-scale training. This abundant video data facilitates training powerful video generation models at scale, while using video as a unified representation simplifies the development process by avoiding the complexity of managing different asset formats, making it an ideal foundation for generative game development.

4. Framework of Generative Game Engine

We decompose our IGV-centered Generative Game Engine into six functional modules. Fig. 4 (a) demonstrates the relationships between these modules, while Fig. 4 (b) presents their key components and explanation. IGV consists of five core modules: First, the **Generation** module represents the basic generative capability of the video generation model. Four extension modules are built upon it: the **Control** module supports different modal control signals and is key to achieving interactivity; the **Memory** module maintains historical generation content from both dynamic and static aspects, crucial for ensuring temporal consistency; the **Dynamics** module models the internal rule logic of the game's virtual world, especially physical rules; while the **Intelligence** module enables advanced capabilities includ-

[1] https://www.youtube.com/watch?v=KGcLSTFEgSk


Module
Technical keywords
Explanation with game examples
GamePlay
Dynamics
Intelligence
Control
Generation
Memory
Interactive Generative Video
Generative Game Engine
Generation
Control
Memory
Dynamics
Intelligence
GamePlay
Auto-Regressive
Endless video generation
Real Time
Low-latency player interaction
Multi-Modal
Integrate visual, audio, and language elements
Navigation Control
Control camera movement and character navigation
Interaction Control
Manage interactions among characters, objects, and environment
Static Memory
Remember static assets like maps and character models
Dynamic Memory
Remember dynamic assets like animations and movements
Physical Laws
Simulate physical interactions (e.g., collisions, momentum, gravity)
Physics Tuning
adjusting physics parameters (e.g., gravity, friction, velocity, position)
Reasoning
Long-term causal reasoning (e.g., NPCs forming new social structures, or in an RPG, killing a powerful leader causes political instability).
Self-Evolution
Emerging world behaviors and rules (e.g., virtual world civilizations evolve, forming a metaverse).
Game Rules
Define game objectives, rewards, and penalties (e.g., access to new areas is unlocked as the story progresses, or players are penalized with fall damage in tutorial sections)

(a) Framework of Generative Game Engine

(b) Technical Keywords of Each Module in Generative Game Engine

Figure 4. Proposed framework of Generative Game Engine (GGE). (a) Architecture and interactions between modules of GGE. (b) Technical keywords and their explanation of each module.

ing causal reasoning and self-evolution. These five modules, through their video interface, create an independent virtual world with its own emergent properties and behaviors. However, a virtual world alone does not constitute a complete game experience, as games require external rules that embody game designers' intentions, providing players with clear objectives and feedback that create gaming enjoyment. Therefore, we propose an additional **GamePlay** module based on IGV, which serves as the key differentiator between GGE and IGV and is responsible for implementing these external rule logic within the virtual game world.

4.1. Generation

❑ **Concept.** The Generation Module handles video generation, the fundamental functionality of IGV. While ensuring basic video generation requirements like visual quality and motion coherence, this module encompasses three crucial functionalities to achieve optimal interactive experience: (1) **Streaming Generation** enables continuous video synthesis with frame-level control frequency. This supports endless procedural worlds in *No Man's Sky* where players can seamlessly explore for hundreds of hours, real-time weather and day-night cycles in *Red Dead Redemption 2* that evolve continuously, and instant response to rapid player inputs in rhythm games like *Beat Saber* where every frame matters. (2) **Real-time Processing** facilitates low-latency interaction with users. This is essential in competitive games like *Counter-Strike*, *Forza Motorsport*, and *League of Legends* where instant visual feedback is crucial. (3) **Multi-modal Generation** complements the video output with other modalities like text and audio. This includes dynamic music that responds to gameplay in *Journey*, positional audio cues for enemy locations in *PUBG*, ambient sound effects in *Minecraft*, and real-time dialogue subtitles in *Mass Effect*.

❑ **Technical Approaches and Future Directions.**

(1) **Streaming Generation**:

Diffusion-based methods [56] excel at generating high-quality visual content. A straightforward way to achieve streaming generation is to use different noise levels across frames. The variable noise levels mechanism means that later frames (with higher noise) can depend on previous frames (with lower noise), implementing autoregressive generation. Representative methods like Diffusion Forcing [10, 18, 65, 69, 90] have been widely used in game video generation [15, 21, 73, 92].

Next token prediction offers another approach to autoregressive video generation [39, 77], though its visual quality currently lags behind diffusion methods. However, its potential for integration with LLM, which could enable strong causal reasoning abilities [82, 96], makes it a promising direction.

Recent attempts to combine diffusion models with next token prediction aim to maintain quality while modeling frame causality [19, 43]. While these hybrid approaches show promise, they are still in early stages and their potential to surpass established diffusion-based methods remains to be seen.

(2) **Real-time Generation**:

Recent works have demonstrated promising advances in efficient video generation through various algorithmic techniques. These include lightweight model distillation [36], ODE-based diffusion step reduction [76], high-compression VAEs [11], and causal architectures like CausVid [90] with distribution matching distillation and Cosmos [55] with Medusa speculative decoding, key-value caching, and tensor parallelism. These advances, coupled with hardware optimizations like GPU parallelization and quantization, suggest that real-time video generation will soon be accessible to game developers on common hardware. While large companies can provide cloud computing support, we expect future personal developers to run these models on accessible machines.

(3) **Multi-modal Generation**:

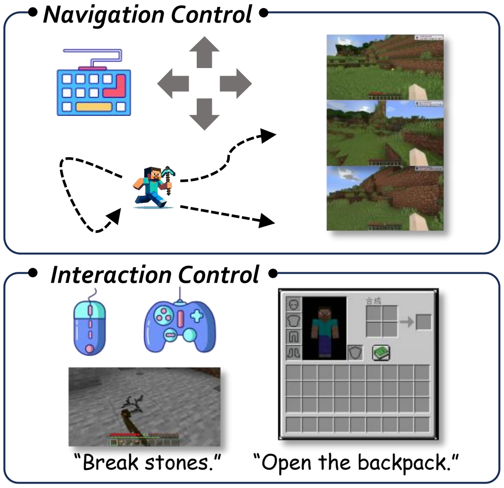


Figure 5. The Control module manages player control through two aspect: navigation and interaction control.

One approach is to develop unified large multimodal models that support understanding and generation across multiple modalities including text, vision, audio, human motion, depth maps, and so on. Recent works have started exploring this direction [39, 59, 82, 96], though significant challenges remain. An alternative strategy is to first develop specialized large models for individual modalities [5, 20, 32, 84] before integrating them into a unified system. Specifically, this requires designing pipeline relationships between different expert models within the unified system. For example, language models generate video generation instructions, the generated videos then serve as input for audio models to produce corresponding audio signals, ultimately leading to a unified output.

4.2. Control

❑ **Concept.** As shown in Fig. 5, the control module manages user control of the virtual world through two aspects: (1) **navigation control** enables players to navigate and explore the virtual world through camera and character movement. For example, in racing games, players use arrow keys or "WASD" for acceleration, braking, and steering, while in open-world games, players typically use "WASD" for character movement, mouse for camera rotation, and space bar for jumping or climbing. (2) **interaction control** allows players to manipulate objects within the virtual environment. For instance, in construction games, players use left mouse clicks to select and place buildings, right clicks to rotate structures, and keyboard shortcuts like 'E' to access inventory or 'Q' to demolish objects.

❑ **Technical Approaches and Future Directions.**

The technical implementation of control mechanisms has been well-studied. Common approaches include: (1) Cross Attention [21, 73, 92], where control signals are transformed into conditional features that serve as keys and values, while the video features serve as queries. (2) Another approach uses external Adaptors [9, 79], which directly fuses control features with video features.

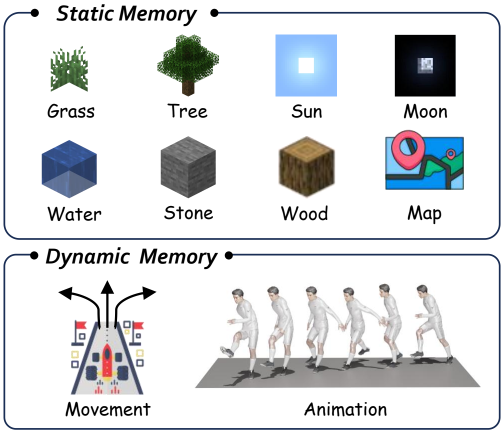


Figure 6. The Memory module consists of static and dynamic memory.

While control is easily mastered in fixed scenes, it should generalize to open-domain scenarios. Some works [16, 21, 92] have leveraged video generation priors for this purpose, but generalizing complex actions with limited control annotations remains challenging and requires further exploration.

Learning control, especially for interaction control, goes beyond mechanical execution and requires understanding the underlying rules of how interactions change the environment (as part of physical laws). Following a data-driven approach, future work aims to collect large-scale datasets [9, 92] and improve the learning of these interaction rules. Physical laws will be further discussed in Sec. 4.4.

For game control signal design, the key design principle is to align with users' gaming intuitions. A promising research direction would be developing more natural control signals that better match human habits, such as using gesture recognition or brain-computer interfaces.

4.3. Memory

❑ **Concept.** Conventional video generation models rely solely on attention mechanisms, struggling to maintain scene layouts, object appearances, and other visual elements in long-duration or large-motion scenarios. As demonstrated in Fig. 6, the Memory module addresses these challenges through two aspects: (1) **static memory** encompasses scene-level and object-level memory, including game maps, buildings, character models, and object appearances. In construction games like *Minecraft* or *SimCity*, the module needs to consistently maintain the structure of player-built constructions; inconsistency in building layouts or designs between frames would severely impact player experience. (2) **dynamic memory** handles short-term motion and behavior patterns, such as character animations, vehicle trajectories, particle effects, and environmental changes like weather transitions. This is crucial in games requiring precise motion consistency, such as fighting games where character movements and attack animations must remain fluid and coherent, or rhythm games where dance movements need to maintain smooth transitions between frames.

❑ **Technical Approaches and Future Directions.**

Current methods mainly rely on attention-based memory, utilizing attention's inherent ability to remember historical frames through cross-attention between historical and predicted frames [15, 73, 81]. However, this approach is unreliable and faces limitations in both precision of memory preservation and limited window size. Another promising solution is using dedicated memory structures, which can be implemented either as implicit high-dimensional features [37] or explicit 3D representations [49–51, 60, 91, 93]. These structures serve as conditional controls for the generation module, ensuring consistent preservation of static elements. The adaptation of these methods as memory mechanisms for game video generation requires further investigation.

4.4. Dynamics

❑ **Concept.** As demonstrated in Fig. 7, the Dynamics Module focuses on two key aspects: (1) **Physical Laws** specifically focuses on comprehending and generating videos that comply with fundamental physics, especially rigid body mechanics including gravity, collision, and acceleration. In racing simulators like *Forza Motorsport*, physics-based puzzle games like *Portal*, and platformers like *Super Mario Odyssey*, where precise physical interactions drive core gameplay mechanics. (2) **Physics Tuning** extends beyond **Physical Laws** by enabling control over physical parameters rather than simply replicating real-world physics. This includes adjusting gravity, friction coefficients, or directly modifying time, velocity, and mass values. In games like *Braid* where time manipulation is core to gameplay, *Superhot* where time moves only when the player moves, and *Control* where physics manipulation powers create unique gameplay experiences.

❑ **Technical Approaches and Future Directions.**

A data-driven approach learns physical laws from large-scale video data [55, 56], though this requires extensive high-quality videos demonstrating diverse physical phenomena [95]. Physics-based memory control offers an alternative by using video generation models as renderers on top of physics simulators [25, 47], ensuring perfect physics compliance but limited to mathematically formulated phenomena. Establishing appropriate benchmarks for evaluating physical accuracy remains crucial [41, 58] to identify limitations and guide improvements. Physics tuning capability, often overlooked in current research, is crucial for models to truly understand and manipulate physical knowledge, and we encourage future research to explore synthetic data with annotated physical parameters as a potential solution.

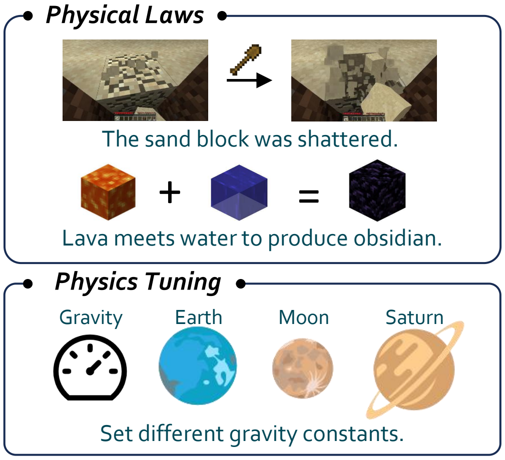

Physical Laws
The sand block was shattered.
+
=
Lava meets water to produce obsidian.
Physics Tuning
Gravity
Earth
Moon
Saturn
Set different gravity constants.

Figure 7. The Dynamics module focuses on physical laws and physics tuning.

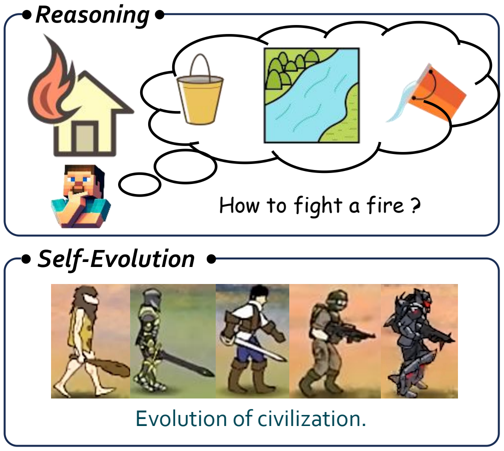


Figure 8. The Intelligence module implements reasoning and self-evolution.

4.5. Intelligence

❑ **Concept.** As Demonstrated in Fig. 8, the Intelligence module implements two key aspects: (1) **Reasoning**: This capability enables long-term causal inference based on initial conditions, creating immersive virtual worlds. For example, the system can predict how a kingdom's economy and social structure might evolve over centuries based on its initial resources and policies, or simulate wildlife migration patterns when environmental conditions change, such as animals seeking new water sources after a river dries up. Similar mechanics can be found in strategy games like *Crusader Kings* and ecosystem simulations like *Planet Zoo*. (2) **Self-Evolution**: This capability goes beyond generating continuous video streams with changing virtual worlds; it enables virtual worlds to continuously develop, evolve, and generate new knowledge, rules, and behaviors through emergent properties. In simulation games, civilizations could naturally emerge and form their own cultures, ecosystems could develop new species, and cities could grow and adapt organically. Such technology could eventually realize a metaverse similar to *The Matrix*, where countless agents and players live in self-evolving virtual worlds.

❑ **Technical Approaches and Future Directions.**

Implementing reasoning capabilities requires video generation models to have a causal structure through autoregessive generation (as discussed in Generation Module in Sec.4.1) and large-scale long-context training [24], similar to large language models. Alternatively, leveraging (multimodal) large language models for causal reasoning alongside video generation models shows promise for unified understanding and generation[82, 96]. Furthermore, if all previously mentioned capabilities including physics understanding, physical simulation, and causal reasoning are successfully implemented and demonstrate powerful performance in the future, we might witness the emergence of remarkable self-evolution capabilities. This convergence of advanced capabilities could potentially lead to truly autonomous virtual worlds, such as metaverses inhabited by countless intelligent agents, or brain-in-a-vat worlds similar to those depicted in *The Matrix*.

4.6. Gameplay

❑ **Concept.** The GamePlay Module builds upon IGV by implementing external **Game Rules**, which are designer-imposed rules such as game objectives, rewards, penalties, and constraints that shape the virtual world's gameplay experience. These include scoring systems in *Tetris*, health and damage systems in *Dark Souls*, mission objectives and reward structures in *Grand Theft Auto*, achievement systems in *Minecraft*, time limits in *Mario*, competitive ranking systems in *League of Legends*, and quest completion rewards in *World of Warcraft*.

❑ **Technical Approaches and Future Directions.**

The implementation of the GamePlay module primarily relies on agent systems empowered by large language models [1, 72] or multimodal large models [46], enabling various gameplay aspects including level design, difficulty scaling, and NPC development. While existing single agents [42, 80] show promise, key research challenges remain in developing unified multi-agent frameworks for game environments. Another practical research direction is exploring how agents and agent systems can enable dynamic, adaptive game rules, including reward and penalty mechanisms. As players progress through games, their skill levels, capabilities, and experience continuously evolve, making it essential to adaptively adjust difficulty levels and reward-penalty systems accordingly.

5. Levels of Generative Game Engine

We propose a five-level maturity model (L0-L4) to evaluate GGE and guide their future development. This framework helps assess current technologies and identify key research directions in GGE. Below we detail each level, with the overview table presented in Supplementary Material.

Level 0: No AI-Assisted Assets Generation. At this foundational level, game engines rely entirely on manually crafted content without any AI-generated elements. All game assets and rules must be pre-designed during the development phase. Classic examples include *Super Mario*, where each level layout is carefully hand-crafted, and *Tetris*, where the game rules and piece designs are fixed. This approach enables precise control but requires heavy resources and restricts players to fixed content.

Level 1: AI-Assisted Assets Generation. Game development combines manual processes with AI-assisted creation of assets and logic during development and gameplay. AI tools generate diverse assets to reduce content creation workload. For instance, in *Cyberpunk 2077*, developers can utilize image generation models like Stable Diffusion [61] to create varied textures for neon billboards, trash piles, and urban details throughout Night City. During gameplay, the engine generates segments, such as unique explosion animations when a player destroys a bridge in an open-world game, or dynamic NPC dialogues in games like *AI Dungeon*. While this approach speeds up development and adds variety, the framework remains pre-designed and needs significant human intervention and curation.

Level 2: Physics-Compliant Interactive World Generation. This level shifts from manual-centric development to interactive video generation, representing AI-Driven Generative Game Engines. The engine continuously generates physics-compliant content based on player interactions in real-time. For example, when a player sets fire to a wooden bridge, the engine dynamically generates not only the realistic blazing effects but also adapts the game world accordingly, such as rerouting enemy paths around the destroyed structure. While many works operate at this level [15, 73, 92], significant improvements are needed in physics understanding, simulation realism, and generalizable interaction.

Level 3: Causal-Reasoning World Simulation. Building on Level 2's physics-compliant generation, which focuses on immediate responses, this level adds causal reasoning across time to address short-term limitations. The engine maintains a world model that understands player actions and logic rules, generating content that reflects long-term cause-and-effect relationships. For example, when a player assassinates a faction leader in Act 1, the engine simulates the resulting political instability, leading to city-wide riots and power struggles that emerge in Act 3. Through this understanding, the game creates storylines where players' early choices shape the world's future development.

Level 4: Self-Evolving World Ecosystem. Building on Level 2's physics generation and Level 3's causal reasoning, as these capabilities continuously advance, the model emerges with self-evolution abilities. The game world becomes a self-evolving ecosystem where complex systems emerge from initial rules and interactions. For example, as the NPC population grows, they autonomously organize into governance structures and establish trade networks, exhibiting emergent social behaviors beyond their initial programming. At this stage, the engine will create virtual worlds similar to those in *Ready Player One* or *The Matrix*, where players can not only play but potentially live within these worlds. This advancement will revolutionize gaming and profoundly impact human society.

6. Alternative Views

> **Alternative View #1**: While GGE represents an automated approach to game content generation, it is worth examining whether it shares the same limitations as Procedural Content Generation (PCG) methods, specifically the tendency to produce repetitive content and the presence of difficult-to-fix bugs that potentially limit their practical applications.

Potential Solution #1: GGE differs fundamentally from Procedural Content Generation (PCG). PCG creates infinite content by randomly combining limited assets and predefined logic rules. In contrast, GGEs learn from massive datasets, acquiring knowledge of unlimited assets and world logic rules. Unlike PCG's meaninglessly repetitive content generation, GGE can create truly diverse content, similar to how AI image generation has enabled diverse, high-quality artworks on Civitai [14]. Additionally, GGEs implicitly model logic rules and leverage control modules for precise control, avoiding PCG limitations such as difficult control, procedural bugs and debugging needs.

> **Alternative View #2**: Given that traditional rendering pipelines in standard game engines offer efficient asset rendering and allow more resources to be allocated to gameplay enhancement, why should we adopt IGV instead of maintaining the traditional approach that prioritizes gameplay dynamics over graphical realism?

Potential Solution #2: Traditional rendering pipelines efficiently handle graphics, allowing developers to focus more resources on gameplay rather than graphics. This raises concerns that IGV might shift too many resources toward visual realism at the expense of gameplay quality, a trade-off that many developers would find unreasonable. However, IGV represents a paradigm shift that enhances both graphical realism and gameplay quality together, rather than trading off one for the other. We analyze this from four aspects: (1) **Complete system**: IGV is not just a generation model, but a system integrating Control, Memory, Dynamics and Intelligence. Beyond improving graphics, it enhances gameplay through, for example, personalized game content, infinite explorable experiences, and intelligent NPC behaviors. (2) **Enhanced gaming experience**: IGV enables dynamic, customized, and infinitely explorable experiences, which traditional game development with triangle rendering and added gameplay logic cannot easily achieve. (3) **Enhanced creative freedom**: IGV's vitual world generation capabilities free developers from focusing on graphics, allowing them to concentrate on gameplay design. Its controllability enables developers to freely exercise creativity in designing more innovative gameplay experiences. (4) **Positive industry impact**: IGV's efficiency and capabilities accelerate game development and lower entry barriers. This attracts more developers, resulting in more creative games and enriching the gaming experience industry-wide.

We also address an additional alternative view regarding the concerns about GGE costs in Supplementary Material.

7. Ethical Issues

Several key ethical issues need to be carefully considered in the development and application of GGE: copyright concerns (determining ownership and protection of AI-generated content), security issues (preventing the generation of harmful content), creativity concerns (whether AI enhances or limits human creative expression), democratization implications (the impact of lowering barriers to game creation), and labor concerns (potential effects on gaming industry workers). These critical issues require thorough discussion and resolution, which we address in detail in Supplementary Material.

8. Conclusion

In this position paper, we have presented Interactive Generative Video (IGV) as a promising foundation for next-generation game engine. We proposed a comprehensive framework with six essential modules and established a five-level maturity model (L0-L4) to guide future research and development. Through our analysis, we demonstrated that IGV's unique capabilities in content generation, physics simulation, and interactive control make it an ideal candidate for revolutionizing the gaming industry. We believe this work provides a clear roadmap for advancing game engine technology while identifying key challenges and opportunities for future exploration.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. 7

[2] Luma AI. Luma ai. https://lumalabs.ai/, 2024. 1, 2

[3] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In Thirty-eighth Conference on Neural Information Processing Systems, 2024. 1, 2

[4] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. arXiv preprint arXiv:2503.11647, 2025. 2, 3

[5] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22861–22872, 2024. 5

[6] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. arXiv preprint arXiv:2405.04233, 2024. 1, 2

[7] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In Forty-first International Conference on Machine Learning, 2024. 1

[8] Shaofei Cai, Zihao Wang, Xiaojian Ma, Anji Liu, and Yitao Liang. Open-world multi-task control through goal-aware representation learning and adaptive horizon prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13734–13744, 2023. 2

[9] Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. arXiv preprint arXiv:2411.00769, 2024. 1, 5

[10] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. arXiv preprint arXiv:2407.01392, 2024. 2, 4

[11] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. arXiv preprint arXiv:2410.10733, 2024. 4

[12] Jingye Chen, Yuzhong Zhao, Yupan Huang, Lei Cui, Li Dong, Tengchao Lv, Qifeng Chen, and Furu Wei. Model as a game: On numerical and spatial consistency for generative games. arXiv preprint arXiv:2503.21172, 2025. 1

[13] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. arXiv e-prints, pages arXiv–2305, 2023. 3

[14] Civitai. Civitai. https://civitai.com/, 2022. 8

[15] Etched Decart. Oasis: A universe in a transformer. https://oasis-model.github.io/, 2024. 1, 2, 4, 6, 7

[16] Google DeepMind. Genie 2: A large-scale foundation world model. https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/, 2024. 1, 2, 5

[17] Google DeepMind. Veo 2: Our state-of-the-art video generation model. https://deepmind.google/technologies/veo/veo-2/, 2024. 1, 2

[18] Chaorui Deng, Deyao Zhu, Kunchang Li, Shi Guang, and Haoqi Fan. Causal diffusion transformers for generative modeling. arXiv preprint arXiv:2412.12095, 2024. 4

[19] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. arXiv preprint arXiv:2412.14169, 2024. 2, 4

[20] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. In Advances in Neural Information Processing Systems, 2023. 5

[21] 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. 1, 2, 4, 5

[22] Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. In ICLR, 2025. 2, 3

[23] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. arXiv preprint arXiv:2503.18938, 2025. 1

[24] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. arXiv preprint arXiv:2503.19325, 2025. 7

[25] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. arXiv preprint arXiv:2501.03847, 2025. 6

[26] 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. arXiv preprint arXiv:2504.08388, 2025. 1

[27] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In European Conference on Computer Vision, pages 330–348. Springer, 2025. 2

[28] 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. 2, 3

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 2020. 2

[30] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8153–8163, 2024. 3

[31] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems, 36:78723–78747, 2023. 3

[32] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems, 36:20067–20079, 2023. 5

[33] Haobin Jiang, Junpeng Yue, Hao Luo, Ziluo Ding, and Zongqing Lu. Reinforcement learning friendly vision-language model for minecraft. In European Conference on Computer Vision, pages 1–17. Springer, 2025. 2

[34] Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. Fulldit: Multi-task video generative foundation model with full attention. arXiv preprint arXiv:2503.19907, 2025. 3

[35] Anssi Kanervisto, Dave Bignell, Linda Yilin Wen, Martin Grayson, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Tabish Rashid, Tim Pearce, Yuhan Cao, et al. World and human action models towards gameplay ideation. Nature, 638(8051):656–663, 2025. 1

[36] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. Bk-sdm: A lightweight, fast, and cheap version of stable diffusion. In European Conference on Computer Vision, pages 381–399. Springer, 2025. 4

[37] Seung Wook Kim, Yuhao Zhou, Jonah Philion, Antonio Torralba, and Sanja Fidler. Learning to simulate dynamic environments with gamegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1231–1240, 2020. 6

[38] Kling. Kling ai: Next-generation ai creative studio. https://app.klingai.com/, 2024. 1, 2, 3

[39] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125, 2023. 2, 4, 5

[40] 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

[41] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. arXiv preprint arXiv:2502.20694, 2025. 6

[42] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. Advances in Neural Information Processing Systems, 36: 51991–52008, 2023. 7

[43] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. arXiv preprint arXiv:2406.11838, 2024. 4

[44] Antonios Liapis, Georgios N Yannakakis, and Julian Togelius. Designer modeling for sentient sketchbook. In 2014 IEEE Conference on Computational Intelligence and Games, pages 1–8. IEEE, 2014. 2

[45] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022. 2

[46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024. 7

[47] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In European Conference on Computer Vision, pages 360–378. Springer, 2024. 6

[48] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003, 2022. 2

[49] Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. You see it, you got it: Learning 3d creation on pose-free videos at scale. arXiv preprint arXiv:2412.06699, 2024. 6

[50] Willi Menapace, Stéphane Lathuilière, Aliaksandr Siarohin, Christian Theobalt, Sergey Tulyakov, Vladislav Golyanik, and Elisa Ricci. Playable environments: Video manipulation in space and time. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3584–3593, 2022.

[51] Willi Menapace, Aliaksandr Siarohin, Stéphane Lathuilière, Panos Achlioptas, Vladislav Golyanik, Sergey Tulyakov, and Elisa Ricci. Promptable game models: Text-guided game simulation via masked diffusion models. ACM Transactions on Graphics, 43(2):1–16, 2024. 6

[52] Panagiotis Migkotzidis and Antonios Liapis. Susketch: Surrogate models of gameplay as a design assistant. IEEE Transactions on Games, 14(2):273–283, 2021. 2

[53] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In Proceedings of the AAAI conference on artificial intelligence, 2024. 3

[54] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 18444–18455, 2023. 2

[55] NVIDIA. Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575, 2025. 3, 4, 6

[56] OpenAI. Creating video from text. https://openai.com/index/sora/, 2024. 1, 2, 3, 4, 6

[57] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. arXiv preprint arXiv:2408.06070, 2024. 3

[58] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. arXiv preprint arXiv:2410.18072, 2024. 6

[59] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. arXiv preprint arXiv:2205.06175, 2022. 5

[60] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. arXiv preprint arXiv:2503.03751, 2025. 6

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022. 7

[62] Runway. Runway : Tools for human imagination. https://runwayml.com/, 2024. 1, 2

[63] Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. arXiv preprint arXiv:2504.08685, 2025. 1, 2

[64] Gillian Smith, Jim Whitehead, and Michael Mateas. Tanagra: Reactive planning and constraint solving for mixed-initiative level design. IEEE Transactions on computational intelligence and AI in games, 3(3):201–215, 2011. 2

[65] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. arXiv preprint arXiv:2502.06764, 2025. 2, 4

[66] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems, 2019. 2

[67] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. International Conference on Learning Representations, 2021. 2

[68] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. arXiv preprint arXiv:2407.14505, 2024. 3

[69] Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, SiYu Zhou, Qian He, and Jing Liu. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. arXiv preprint arXiv:2503.07418, 2025. 4

[70] The Movie Gen team. Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720, 2024. 1, 2

[71] Yao Teng, Enze Xie, Yue Wu, Haoyu Han, Zhenguo Li, and Xihui Liu. Drag-a-video: Non-rigid video editing with point-based interaction. arXiv preprint arXiv:2312.02936, 2023. 3

[72] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023. 7

[73] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. arXiv preprint arXiv:2408.14837, 2024. 1, 2, 3, 4, 5, 6, 7

[74] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025. 1, 2

[75] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. 2023. Comment: Project website and open-source codebase: https://voyager. minedojo. org/Cited on, page 33, 2023. 2

[76] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. arXiv preprint arXiv:2312.09109, 2023. 4

[77] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869, 2024. 2, 4

[78] Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. arXiv preprint arXiv:2302.01560, 2023. 2

[79] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In ACM SIGGRAPH 2024 Conference Papers, 2024. 2, 3, 5

[80] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. arXiv preprint arXiv:2308.08155, 2023. 7

[81] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. arXiv preprint arXiv:2504.12369, 2025. 1, 6

[82] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528, 2024. 4, 5, 7

[83] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter:

Animating open-domain images with video diffusion priors, 2023. 2

[84] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10371–10381, 2024. 5

[85] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. arXiv preprint arXiv:2310.06114, 2023. 1, 3

[86] Mingyu Yang, Junyou Li, Zhongbin Fang, Sheng Chen, Yangbin Yu, Qiang Fu, Wei Yang, and Deheng Ye. Playable game generation. arXiv preprint arXiv:2412.00887, 2024. 1

[87] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In ACM SIGGRAPH 2024 Conference Papers, pages 1–12, 2024. 2

[88] Sherry Yang, Jacob C Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Position: Video as the new language for real-world decision making. In Proceedings of the 41st International Conference on Machine Learning, 2024. 1, 3

[89] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072, 2024. 1, 2

[90] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast causal video generators. arXiv preprint arXiv:2412.07772, 2024. 4

[91] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6658–6667, 2024. 6

[92] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos, 2025. 1, 2, 3, 4, 5, 7

[93] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048, 2024. 6

[94] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF international conference on computer vision, pages 3836–3847, 2023. 3

[95] Qi Zhao, Xingyu Ni, Ziyu Wang, Feng Cheng, Ziyan Yang, Lu Jiang, and Bohan Wang. Synthetic video enhances physical fidelity in video synthesis. arXiv preprint arXiv:2503.20822, 2025. 6

[96] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039, 2024. 4, 5, 7