Position: Interactive Generative Video as Next-Generation Game Engine (Supplementary Materials)

A. Preliminaries

A.1. Video Generation Models

Video Generation Models aim to synthesize temporally consistent video sequences $\mathbf{x} = \{x_0, x_1, \ldots, x_t\}$, where x_i indicates the i-th frame. The video is generated by an autoregressive model [36, 55, 69, 74], a diffusion model [30, 53], or a masked transformer model [9, 82]. With the rise of diffusion models [29, 41, 42, 61, 62], which have now become the mainstream approach in video generation due to its high-quality generation performance, significant progress has been achieved in this field [2, 7, 17, 35, 37, 53, 56, 58, 63, 67, 78].

Conditional Video Generation is formulated as $p(\mathbf{x}|c)$ and it varies depending on the type of control signal c which denotes conditions such as text prompts or other control signal. Approaches such as [25, 48, 73] incorporate images as control signals for the video generator, improving both video quality and temporal relationship modeling. Methods such as Direct-a-Video [77], MotionCtrl [72], and CameraCtrl [27] use camera embedders to adjust camera poses, enabling control over camera movements in generated videos. 3DTrajMaster [21] extends this capability by transforming 2D camera signals into 3D for more advanced control. ReCamMaster [6] re-shoots the source video with novel camera trajectories.

Autoregressive Video Generation Models. Since game videos require variable-length or even infinite-length generation to enable interactive game experience, autoregressive mechanism is necessary. Autoregressive video generation refers to a process where new frames are generated based on previously generated frames, which can be expressed as $p(x_1, x_2, ..., x_n) = \prod_{i-1}^{n} p(x_i|x_1, x_2, ..., x_{i-1})$, where x_i indicates the i-th frame. An intuitive approach is to adopt GPT-like next-token prediction methods [18, 36, 69]; however, this approach often falls short in terms of generation quality. Relying on the exceptional performance of diffusion models, Diffusion Forcing [12, 60] implements autoregression by applying different levels of noise to different frames, allowing the denoising of new frames with higher noise levels while conditioning on previous frames with lower noise levels. Methods [15, 20, 66, 81] leveraging Diffusion Forcing has achieved remarkable results.

A.2. AI-driven Game Applications

Game Video Generation. Previous works have utilized GANs [33, 34, 43] to generate game videos or used NeRF to reconstruct 3D scenes to simulate the game process [44, 45], but often fall short in terms of generation quality. Using the powerful generative capabilities of diffusion models, some works [3, 15, 20, 66, 76] have produced high-quality game videos. However, the content is typically confined to specific preexisting games. Open-domain Methods [16, 20, 81], by utilizing multi-stage training or large-scale training datasets, create new content for video games.

Game Design Assistant. AI-powered design assistants offer numerous advantages throughout the creative process, depending on the tool, the type of AI and the creative workflow. These systems can streamline development, reduce costs, reduce manual effort, improve team collaboration, and even inspire creativity [38]. In the gaming domain, most existing AI-driven design tools primarily assist by auto-completing an ongoing design [59] or generating multiple design suggestions for creators to evaluate [10, 39, 47, 64].

Intelligent Game Agent. Reinforcement learning has long been the predominant approach in this domain. Early efforts explored the use of hierarchical RL [4, 8, 19, 32, 40, 50, 83] in the context of MineRL competitions [26]. However, due to the absence of guidance from prior knowledge, such approaches often struggle to perform effectively on long-horizon tasks. With the advancement of LLM [1, 65], leveraging their prior knowledge to plan long-horizon tasks has shown promising results. Recent advancements in LLM-related research [31, 54, 68, 70, 75, 84] have significantly propelled the progress of agents in long-horizon tasks.

B. Overview Table for Levels of GGE

In Table A, we demonstrate the overview of different maturity levels for GGE.

Level	Name	Technical Features	Application Examples	Category
L0	No AI-Assisted Assets Generation	Manual creation and integration of all game assets and logic.	*Super Mario*: fixed levels; *Tetris*: fixed rules.	Traditional Manual Game Development
L1	AI-Assisted Assets Generation	AI-assisted creation and integration of game assets and logic.	*Cyberpunk 2077*: AI-generated assets; *AI Dungeon*: real-time NPC dialogues.	
L2	Physics-Compliant Interactive World Generation	Real-time **physics-compliant** video generation with user **interactions**, supported by the Dynamics module.	E.g., Player sets fire to wooden bridges, AI dynamically renders blazing spans and rerouted enemy paths	Next-Gen AI-Driven Generative Game Engine
L3	Causal-Reasoning World Simulation	World simulation with **causal reasoning** across time based on L2, incorporating the Intelligence module.	E.g., Killing a faction leader in Act 1 triggers city-wide riots in Act 3.	
L4	Self-Evolving World Ecosystem	Autonomous world **evolution** with **emergent** behaviors based on L2 and L3, requiring advanced Intelligence module.	E.g. NPCs self-organize governments and trade as population increases.	

Table A. Proposed Maturity Levels (L0-L4) of Generative Game Engine. L0-L1 represent traditional manual game development with limited AI assistance, while L2-L4 showcase next-generation game engines featuring video-based world generation.

C. Additional Alternative Views

Alternative View #3: The economic costs of GGE appear to be significant. For instance, the computational overhead is substantial since GGE relies on IGV and LLMs, which are computationally intensive large models. Do these costs prevent IGV-centered GGE from becoming the next generation of game engines? Are these costs we incur for implementing GGE justified by the benefits it brings?

Potential Solution #3: Regarding the computational costs, we believe these can be effectively reduced through technological advancements. Recent works have demonstrated promising advances in efficient autoregressive video generation. On the algorithmic front, CausVid [80] achieves real-time frame generation through distribution matching distillation (DMD) [79], while Cosmos [49] enables real-time generation by combining Medusa speculative decoding, key-value caching, tensor parallelism, and low-resolution adaptation. Additionally, hardware optimizations like GPU parallelization, quantization, and knowledge distillation have significantly accelerated inference speeds for autoregressive models. With ongoing research in efficient models, we believe autoregressive video generation will eventually achieve real-time performance on commonly available hardware accessible to game developers.

Beyond computational costs, other economic considerations include:

- **Data Collection Costs**: These will be mitigated as more open-source datasets like GameGen-X [11] become available. While initial training incurs costs, trained models reduce future asset production costs, leading to overall savings.
- **Licensing Costs**: Generative models will lower the barrier for developers to create their own new IPs. Building a mutually beneficial ecosystem between developers and gaming companies is also advantageous.
- **Safety Control Costs**: While this affects all generative AI, not just IGV, the benefits of incorporating generative AI outweigh these costs, as demonstrated by successful products like Runway [56], Midjourney [46], and ChatGPT [51].

We believe that these costs will not impede GGE's development or future potential. The technology is continuously evolving, with costs decreasing while model capabilities become increasingly powerful, making the benefits more and more significant. This mirrors the trajectory of large language models. Compared to ChatGPT [51] released in 2022, today's LLMs demonstrate stronger performance (like DeepSeek-R1 [22]'s reasoning capabilities and GPT-4o's multimodal generation abilities [52]) while becoming cheaper and more accessible (open-source models like DeepSeek [22] and Qwen [5] now offer performance comparable to commercial models).

While GGE currently faces some cost-related concerns in the short term, these challenges are outweighed by its transformative value. As discussed in Alternative View #2 in Sec. 6, GGE offers significant advantages over traditional game engines, such as personalized gaming experiences, infinitely generated game content, and lowering the barrier to game development so that everyone can become a game designer. These compelling benefits, which are unattainable with traditional game engines, make GGE's economic costs worthwhile to address and overcome.

D. Ethical Issues

Copyright Issues: How should we determine copyright ownership and protect legitimate copyright interests of all parties involved in GGE-generated games?

Copyright protection presents a new challenge in generative AI development, including IGV, which requires significant attention from both technical and legal perspectives. While this is a complex issue, the industry is actively working towards solutions that can foster the mutual development of AI technology and copyright protection. To address these challenges, we propose several approaches:

Training data for IGV should prioritize the use of copyright-free or properly licensed data sources to minimize legal risks. Game developers can build mutually beneficial partnerships with copyright holders to legally obtain data and share the copyright of the created content. For instance, while noting the Studio Ghibli's recent copyright concerns with OpenAI, we observe that Ghibli has successful experiences in collaborating with game companies (e.g., development of game "Ni no Kuni"). Such examples demonstrate the feasibility and value of proper copyright collaboration.

Technically, research works [23, 24] on dataset copyright protection and detection of unauthorized training data usage are progressing, which has positive implications for IGV in game development.

Security Issues: What measures can be implemented to prevent the generation of harmful content by generative models such as IGV?

IGV systems are built upon existing video generation models, thus inheriting their established safety measures. Current commercial video generation services like Runway and Sora have implemented comprehensive safety systems that filter out inappropriate content including violence, pornography, hate speech, and other harmful materials. From a technical perspective, safety measures can be implemented through various approaches: (1) Value alignment [14, 57] through techniques like RLHF during the model post-training phase would establish fundamental safety boundaries. This alignment with human preferences and values can effectively constrain the model's output content; (2) Real-time harmful content detection [28, 71] using VLMs can quickly analyze generated content, identify potential harmful elements, and block inappropriate content in real-time, which is particularly crucial in interactive gaming environments.

Creativity Issues: Can IGV serve as a creative tool, allowing for deep human creativity?

We believe IGV can enhance human creativity for the following reasons: (1) Interactive generative video technology eliminates mechanical and repetitive tasks in game development, such as debugging, writing basic code, and building standard scenes. By automating these uncreative aspects, developers can channel their energy into creative endeavors, focusing on

innovative gameplay design and unique artistic expressions that truly matter to the gaming experience. (2) IGV breaks down technical barriers, making game development accessible to creators from professional studios to independent developers. This democratization enables more diverse voices to enter the gaming industry, each bringing their unique perspectives and creative visions. Like other AIGC applications, it enables creators to realize their ideas without technical constraints, as demonstrated by artist Sofia Crespo's work[1] that blends technology with organic art, showing how AI amplifies creativity.

Democratization Issues: Does democratizing game creation diminish its value?

We believe that democratizing game creation will not diminish its value, but rather enhance the overall value and creativity of the entire field. Here is our analysis and examples:

The democratization of gaming won't diminish creative value. The widespread availability of technology enables more people to enter this field, generating more diverse creative thinking and innovative designs. Creation value lies in innovation and personalization, not just technical difficulty. Through this technology, even ordinary users can create games with unique characteristics and personal style, which holds its own distinctive value. A good example is the opening up of image generation models, which hasn't diminished the value of artistic creation. People with varying levels of professional expertise have shared numerous new artistic works on Civitai [13], which has actually enhanced the creativity in this field and the value of its works.

Labor Issues: How should we view the potential negative impact of highly automated productivity tools like GGE on labor in the gaming industry?

We acknowledge the labor impact concern with generative AI. IGV aims to enhance productivity and creativity rather than replace human workers. We advocate for measures like education and support programs to help industry professionals leverage AI tools, ensuring positive industry transformation.

E. Workflow Integration with GGE

It is important to emphasize that the introduction of generative game engines (GGE) will not lead to a single, rigid game development workflow. We provide below a framework example of how GGE can be incorporated into game development workflows

Phase #1: Pre-production Phase

• **Concept Design**: Define core game elements (gameplay mechanics, story, target audience, art style) through LLM consultation and convert to IGV condition prompts.
• **Prototype Development**: Select suitable base models based on computing power and performance requirements, and develop prototypes using initial prompts for feasibility testing.

Phase #2: Production Phase

• **Asset and Logic Requirements**: Create detailed prompts for specific assets and logic requirements (e.g., character model descriptions, area map sketches, level-up rule systems).
• **Training Data Collection and Model Fine-tuning**: Fine-tune models with targeted game data (e.g., collecting copyright-free space movie/game assets for a space exploration game).

Phase #3: Testing Phase

• **Functionality Testing**: Test prompt-based content generation and screen for harmful content.
• **Compatibility and Performance Testing**: Optimize performance across different devices with necessary algorithm/hardware acceleration.

[1] https://en.wikipedia.org/wiki/Sofia_Crespo

Phase #4: Post-launch Maintenance

• **Content Updates**: Update model parameters and prompts for new content (DLCs, characters, events).
• **Data Analysis and Optimization**: Use player behavior data (with consent) for model fine-tuning and reinforcement learning.

Feasibility Requirements

The successful implementation of this workflow relies on these key factors:
• (1) **Model Capability**: Robust base IGV models that support efficient control, fine-tuning, and fast inference.
• (2) **Data Accessibility**: Well-established data sharing and copyright mechanisms that enable legal and cost-effective access to high-quality training data.
• (3) **Computing Resources**: Accessible AI computing infrastructure, either through local hardware resources or affordable cloud computing services.

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. 1

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

[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

[4] PKU BAAI. Plan4mc: Skill reinforcement learning and planning for open-world minecraft tasks. arXiv preprint arXiv:2303.16563, 2023. 1

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023. 3

[6] 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. 1

[7] 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

[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. 1

[9] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11315–11325, 2022. 1

[10] Megan Charity, Isha Dave, Ahmed Khalifa, and Julian Togelius. Baba is y'all 2.0: Design and investigation of a collaborative mixed-initiative system. IEEE Transactions on Games, 16(1):75–89, 2022. 1

[11] 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. 3

[12] 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. 1

[13] Civitai. Civitai. https://civitai.com/, 2022. 4

[14] Juntao Dai, Tianle Chen, Xuyao Wang, Ziran Yang, Taiye Chen, Jiaming Ji, and Yaodong Yang. Safesora: Towards safety alignment of text2video generation via a human preference dataset. Advances in Neural Information Processing Systems, 37:17161–17214, 2024. 3

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

[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

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

[18] 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. 1

[19] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022. 1

[20] 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

[21] 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. 1

[22] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025. 3

[23] Junfeng Guo, Yiming Li, Lixu Wang, Shu-Tao Xia, Heng Huang, Cong Liu, and Bo Li. Domain watermark: Effective and harmless dataset copyright protection is closed at hand. Advances in Neural Information Processing Systems, 36:54421–54450, 2023. 3

[24] Junfeng Guo, Yiming Li, Ruibo Chen, Yihan Wu, Heng Huang, et al. Zeromark: Towards dataset ownership verification without disclosing watermark. Advances in Neural Information Processing Systems, 37:120468–120500, 2024. 3

[25] 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. 1

[26] William H Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. Minerl: A large-scale dataset of minecraft demonstrations. arXiv preprint arXiv:1907.13440, 2019. 1

[27] 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. 1

[28] Lukas Helff, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. Llavaguard: Vlm-based safeguard for vision dataset curation and safety assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8322–8326, 2024. 3

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 2020. 1

[30] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303, 2022. 1

[31] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In International conference on machine learning, pages 9118–9147. PMLR, 2022. 1

[32] 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. 1

[33] 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. 1

[34] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5820–5829, 2021. 1

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

[36] 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. 1

[37] 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

[38] Antonios Liapis. Searching for sentient design tools for game development. Phd thesis, IT University of Copenhagen, 2015. 1

[39] 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. 1

[40] Zichuan Lin, Junyou Li, Jianing Shi, Deheng Ye, Qiang Fu, and Wei Yang. Juewu-mc: Playing minecraft with sample-efficient hierarchical reinforcement learning. arXiv preprint arXiv:2112.04907, 2021. 1

[41] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022. 1

[42] 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. 1

[43] Willi Menapace, Stephane Lathuiliere, Sergey Tulyakov, Aliaksandr Siarohin, and Elisa Ricci. Playable video generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10061–10070, 2021. 1

[44] 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. 1

[45] 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. 1

[46] Midjourney. Midjourney. https://www.midjourney.com/home, 2022. 3

[47] Panagiotis Migkotzidis and Antonios Liapis. Susketch: Surrogate models of gameplay as a design assistant. IEEE Transactions on Games, 14(2):273–283, 2021. 1

[48] 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. 1

[49] NVIDIA. Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575, 2025. 2

[50] Junhyuk Oh, Satinder Singh, Honglak Lee, and Pushmeet Kohli. Zero-shot task generalization with multi-task deep reinforcement learning. In International Conference on Machine Learning, pages 2661–2670. PMLR, 2017. 1

[51] OpenAI. Introducing chatgpt. https://openai.com/index/chatgpt/, 2022. 3

[52] OpenAI. Introducing 4o image generation. https://openai.com/index/introducing-4o-image-generation/, 2022. 3

[53] OpenAI. Creating video from text. https://openai.com/index/sora/, 2024. 1

[54] Yiran Qin, Enshen Zhou, Qichang Liu, Zhenfei Yin, Lu Sheng, Ruimao Zhang, Yu Qiao, and Jing Shao. Mp5: A multi-modal open-ended embodied system in minecraft via active perception. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16307–16316, 2024. 1

[55] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems, 32, 2019. 1

[56] Runway. Runway : Tools for human imagination. https://runwayml.com/, 2024. 1, 3

[57] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22522–22531, 2023. 3

[58] 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

[59] 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. 1

[60] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. arXiv preprint arXiv:2502.06764, 2025. 1

[61] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems, 2019. 1

[62] 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. 1

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

[64] Maya Grace Torii, Takahito Murakami, and Yoichi Ochiai. Lottery and sprint: Generate a board game with design sprint method on autogpt. In Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play, pages 259–265, 2023. 1

[65] 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. 1

[66] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. arXiv preprint arXiv:2408.14837, 2024. 1

[67] 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

[68] 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. 1

[69] 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. 1

[70] 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. 1

[71] Zhenting Wang, Shuming Hu, Shiyu Zhao, Xiaowen Lin, Felix Juefei-Xu, Zhuowei Li, Ligong Han, Harihar Subramanyam, Li Chen, Jianfa Chen, et al. Mllm-as-a-judge for image safety without human labeling. arXiv preprint arXiv:2501.00192, 2024. 3

[72] 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. 1

[73] 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. 1

[74] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157, 2021. 1

[75] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Haoran Tan, Chencheng Jiang, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. Octopus: Embodied vision-language programmer from environmental feedback. In European Conference on Computer Vision, pages 20–38. Springer, 2025. 1

[76] 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

[77] 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. 1

[78] 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

[79] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6613–6623, 2024. 2

[80] 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. 2

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

[82] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10459–10469, 2023. 1

[83] Enshen Zhou, Yiran Qin, Zhenfei Yin, Yuzhou Huang, Ruimao Zhang, Lu Sheng, Yu Qiao, and Jing Shao. Minedreamer: Learning to follow instructions via chain-of-imagination for simulated-world control. arXiv preprint arXiv:2403.12037, 2024. 1

[84] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. arXiv preprint arXiv:2305.17144, 2023. 1