Make A Game: A Novel Paradigm for Interactive Game Rendering

Anonymous ICME submission

***Abstract*—Driven by the growing demand for immersive and personalized gaming experiences, existing game video generation models often lack robust multi-modal control and suffer from spatial misalignment or autoregressive drift, thereby necessitating a more advanced real-time interactive solution.To this end, we propose Make-A-Game(MAG), the first DiT-based interactive game video generation model that unifies spatial and non-spatial alignment control signals within the UC-3DMMAttn module and employs Action Prompt Blocks(APBs) for precise, real-time manipulation of in-game entities. MAG follows a three-stage training approach, enabling it to produce high-resolution, drift-free, and temporally coherent content by dynamically modulating character actions while preserving detailed environments. Extensive experiments demonstrate MAG's superiority in controllability, video fidelity, and stability, providing new avenues for industrial-scale game development and broader interactive media systems. Thus, MAG represents a transformative step in video generation, leveraging the DiT architecture to integrate multi-modal controls seamlessly. The project will be available at https://icme2025sub.github.io/MAG/**

***Index Terms*—Interactive Game Video Generation, DiT Architecture, Unified Control, Multi-Modal Alignment**

I. Introduction

The video game industry has experienced exponential growth, fueled by rising demand for personalized, immersive content; however, traditional development that relies on manually coded game engines proves ever more prohibitive in cost and time. Concurrently, breakthroughs in video generation models [1] have demonstrated impressive multi-modal generative capabilities, naturally prompting their integration into modern game development to enhance efficiency and creativity(check Appendix A for more).

Recent research explores neural network models for game simulation and interaction, but these efforts often remain bound by Unet-based architectures with limited scalability and awkward integration of multi-modal control signals, such as the incompatibility between spatial alignment-based control methods (e.g., ControlNet [2]) and non-spatial alignment-based control methods (e.g., IPAdapter [3]). Moreover, precise control of in-game entities remains elusive, as moving characters typically interfere with background rendering, exacerbated by autoregressive drift [4]. Thus, a model that unifies diverse control paradigms with high scalability stands as a critical need for realistic and interactive game simulation.

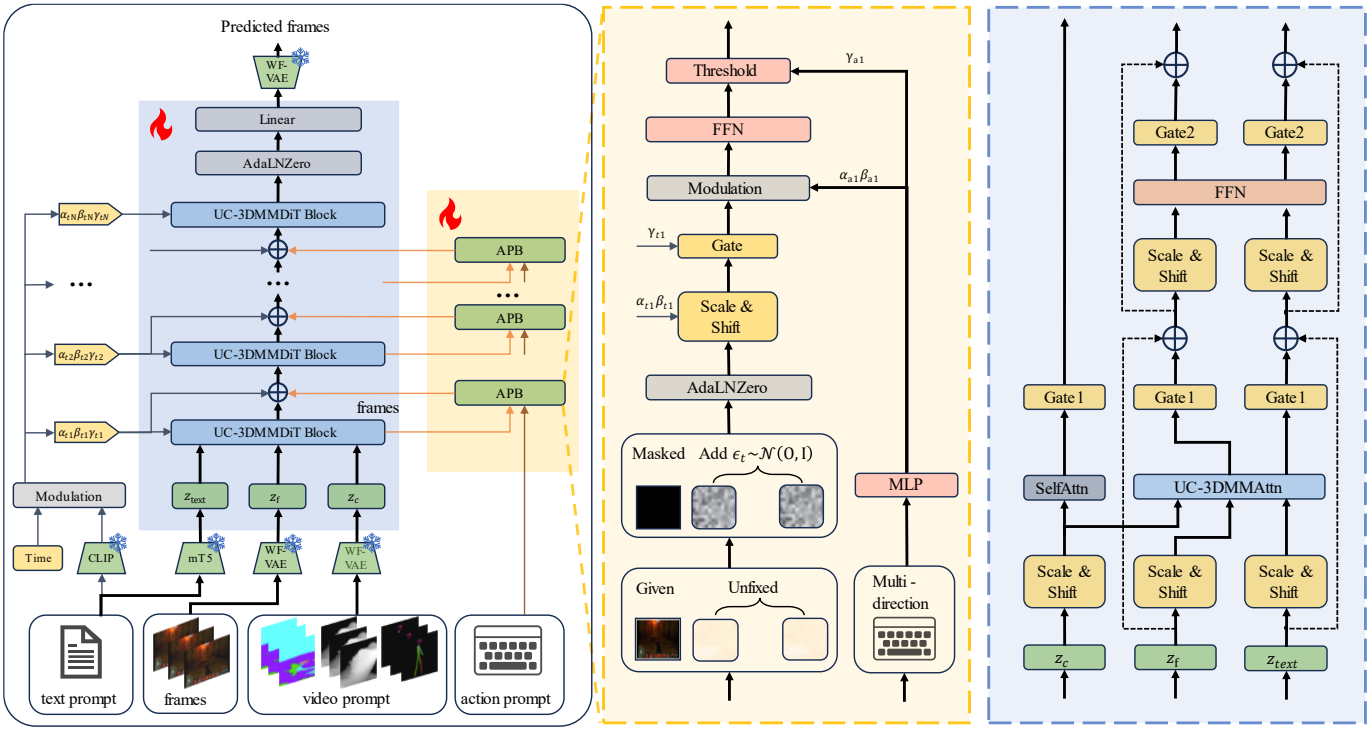
To address these challenges, we introduce **Make-A-Game (MAG)**, the first interactive game simulation model that seamlessly combines the DiT architecture with multi-modal control signals. MAG follows a "backbone + lightweight branch" paradigm: its backbone, comprising stacked Unified Control 3D MMDiT (UC-3DMMDiT) blocks, handles complex and controllable environments, while lightweight Action Prompt Blocks (APBs), inserted at strategic intervals, offer precise manipulation of in-game entities. Key to the backbone is our Unified Control 3DMMAttention (UC-3DMMAttn), which unifies spatial and non-spatial alignment-based control signals within the DiT framework, transcending prior architectures reliant on inserted layers or isolated auxiliary branches. The APBs further decouple entity movements from overall scene generation, mitigating autoregressive drift and refining entity control.

TABLE I
Comparison of MAG and Other Game Generation Models

	Genie [5]	MarioVGG [6]	GameNGen [4]	Oasis [7]	**MAG (ours)**
Architecture	TRM	Unet	Unet	TRM	TRM
Playable (real time)	×	×	✓	✓	✓
Infinite length	×	×	✓	✓	✓
HD (720P)	×	×	×	✓	✓
Unified control	×	×	×	×	✓

Our three-stage training pipeline and corresponding data preparation strategy enable efficient learning of MAG. In the first stage, we fine-tune a foundational model on large-scale gaming images [8,9]. Next, we train the MAG backbone on curated image-to-video datasets like CogVideoX [10]. Finally, we freeze the backbone and optimize the lightweight APB modules using high-quality gameplay motion video data. During inference, MAG autoregressively generates content, with APBs optional for accurate entity control. Experiments confirm that MAG offers state-of-the-art controllability while maintaining visual fidelity and mitigating autoregressive drift.

MAG thus presents a transformative advancement in interactive game video generation, delivering both efficient environmental control and precise entity manipulation as shown in Table I. Although further refinements are necessary for full-scale deployment in industrial pipelines, MAG's modular design and the generality of the DiT architecture signal promising applicability across broader controllable systems beyond gaming.



(a) MAG Overview (b) Action Prompt Block(APB) (c) UC-3DMMDiT Block

Fig. 1. Overview architecture (a) of MAG and its backbone (c) and plug-and-play branches (b).

In summary, our contributions can be summarized as:

- We extend the DiT architecture to interactive game generation, proposing MAG—a novel "backbone + lightweight branch" approach that achieves high-quality and versatile control.
- We introduce UC-3DMMAttn, which unifies spatial and non-spatial alignment-based control signals in the DiT backbone, elevating the fidelity of environment generation.
- We design lightweight APB modules, which deliver plug-and-play entity control while alleviating background drift from motion interference.
- We propose a three-stage training and data preparation pipeline for MAG, attaining SOTA control performance across multiple relevant metrics.

II. Methods

In this section, we will introduce the proposed Make-A-Game (MAG) model, including its overall model architecture with inputs and outputs (II-A), the design details of its three key modules (Encoder Composition, UC-3DMMDiT block, and APB) (II-B), and the training and inference pipeline of MAG (II-C).

A. Overview of MAG Architecture

MAG is the first to extend the DiT model for interactive game scene generation, showcasing advanced multi-modal control. As illustrated in Fig. 1(a), its architecture adopts a "backbone plus detachable branch" design and processes four inputs: video frames, a video prompt, a text prompt for environmental control, and an action prompt for character behavior. The first three inputs are encoded into latent representations via the encoder composition and processed by the backbone, composed of stacked UC-3DMMDiT blocks, before decoding into predicted frames. The action prompt is mapped by the lightweight APB module into small feature space offsets at each timestep, which are added to the backbone to enable real-time interactive character control. The macro working principle of MAG is formally defined as:

$$z_{gen} = f_D(f_{UC}(f_E(z_f, z_c, z_{text})) + APB(a)), \quad (1)$$

where z_f, z_c, z_{text}, and z_a represent the latent embeddings of the video frames, video prompt, text prompt, and action prompt respectively. Below, we will introduce the three main components of MAG in detail.

B. Three Core Components of MAG

a) Encoder Composition: MAG utilizes a combination of three pretrained encoders—CLIP [11], mT5 [12], and WFVAE [13]—to compress tri-modal inputs (excluding the action prompt) into latent spaces and decode the generated outputs back into pixel space. Specifically, WFVAE [13] is employed to transform video frames and the video prompt into multi-level Haar wavelet representations, directly encoding the low-frequency energy components into latent variables z_{video} and z_{vp} via the equation(with further elaboration in Appendix B):

$$S_{ijk}^{(l)} = S^{(l-1)} * (f_i \otimes f_j \otimes f_k), \quad (2)$$

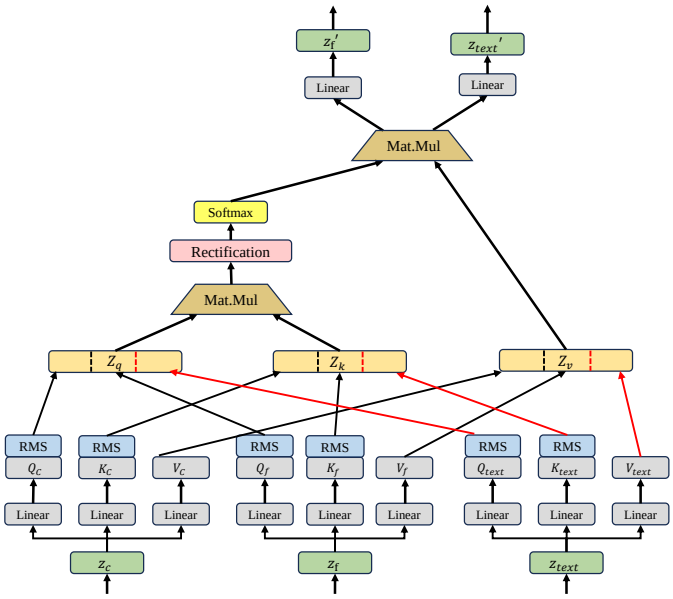


Fig. 2. The detailed architecture of UC-3DMMAttn adopts a unified approach to handle the flow of control information.

thereby enhancing both compression efficiency and reconstruction fidelity. Concurrently, the text prompt is encoded into z_{text} using the mT5 [12] encoder, which ensures an optimal balance between MAG's multilingual comprehension of gaming environments and parameter scalability, while the CLIP encoder processes the text prompt in parallel. The latent representation derived from CLIP is incorporated into the timestep embedding, and post-rectification, generates global pooling factors α_{ti}, β_{ti}, and γ_{ti} for $i = 1, 2, \ldots, N$. These factors, inspired by FLUX.1dev [14], rectify UC-3DMMDiT blocks and APB within subsequent adaptive LayerNormZero layers, ensuring precise semantic alignment between the backbone and auxiliary branches in the generative space.

b) Unified Control-3DMMDiT Block: Building upon MAG's potent yet parameter-efficient control mechanism, the UC-3DMMDiT block (Fig. 1(c)) constitutes the pioneering video generation architecture that integrates both spatially aligned and non-spatially aligned control tasks within a unified framework; specifically, we retain the conventional treatment of z_f and z_{text} in MMDiT while seamlessly amalgamating the control information z_c as an additional component in the concatenated stream $Z = [z_f; z_{\text{text}}; z_c]$, which is subsequently subjected to adaptive 3DRoPE and again concatenated for attention computation:

$$\text{UC-3DMMAttn}(Z) = \text{softmax}\left(\frac{Q_Z K_Z^\top}{\sqrt{d_Z}} + \Delta_\gamma\right) V_Z, \quad (3)$$

where the UC-3DMMAttn structure is shown in Fig. 2, and we introduce a rectification module that modulates the bias control and attention scaling to flexibly govern the contribution of z_c while ensuring zero-initialization control.

To accomplish the differential modulation of attention across varying modalities, we define the control bias Δ_γ piecewise, thereby adjusting attention weights among distinct token types:

$$\Delta_\gamma[i,j] = \begin{cases} 0, & \text{if } (i,j) \in \mathcal{T} \cup \mathcal{T}', \\ \log(1+\gamma), & \text{if } (i,j) \in \mathcal{C}, \\ 0, & \text{otherwise}, \end{cases} \quad (4)$$

where $\mathcal{T}$ and $\mathcal{T}'$ represent the attention spans of z_{text} and z_f tokens, respectively, $\mathcal{C}$ indicates the interaction region of z_c and z_f, and γ modulates the intensity of conditional control. We parameterize this bias matrix through a zero-convolution mechanism to enable zero-initialization:

$$\Delta_\gamma[i,j] = Z(f_\gamma(i,j); \Theta_z), \quad (5)$$

where

$$f_\gamma(i,j) = \begin{cases} \log(1+\gamma), & \text{if } (i,j) \in \mathcal{C}, \\ 0, & \text{otherwise}, \end{cases} \quad (6)$$

and the zero-convolution operation $Z(\cdot; \Theta_z)$ is defined by

$$Z(I; \{W, B\})_{p,i} = B_i + \sum_j I_{p,j} W_{i,j}, \quad (7)$$

where $W = 0$ and $B = 0$ are initially set to ensure $\Delta_\gamma = 0$, thereby nullifying the influence of the conditional token at the onset of training; however, as training proceeds, the non-zero components of Δ_γ gradually acquire the ability to capture and convey the conditioning effects from z_c onto z_f. Consequently, this unified control paradigm enables the direct involvement of z_c in the multi-modal attention mechanism [15] without a separate processing branch, thereby offering a streamlined architectural design underpinned by an elegant mathematical formulation.

c) Action Prompt Block (APB): The APB component serves as the core mechanism for interactive control, adaptively modulating motion and viewpoint predictions of MAG over successive frames in response to user keyboard inputs (action prompts), while simultaneously preserving the controllable generative capability of the backbone UC-3DMMDiT. This functionality is achieved by modeling the conditional distribution $p(V_{1:n} \mid T, C_{1:n})$, where $C_{1:n}$ represents the user-provided keyboard input sequence, and is implemented through three integrated phases: input frame preprocessing, fusion modulation, and influence intensity control.

Input Frame Preprocessing To effectively distinguish context frames from target prediction frames and enable tailored diffusion strategies, the APB employs a masking function $M(i)$ on the feature representation z_t at each time step, while introducing mild Gaussian noise into the initial frames to mitigate temporal accumulation drift. Denoting ϵ_t as the noise vector and α as the noise control intensity, this process is mathematically formalized as:

$$\tilde{z}_t = (1 - \alpha M(i)) \odot z_t + \alpha M(i) \odot \epsilon_t, \quad (8)$$

where $M(i) = 0$ for time steps $i \leq x$ (corresponding to the given initial frames) and $M(i) = 1$ otherwise, followed by normalization of the resulting feature $\tilde{z}_t$ and its propagation to the subsequent stage.

TABLE II
QUANTITATIVE COMPARISON OF MAG AND OTHER VIDEO GENERATION MODELS

Model	Video Quality					Control Performance						
	Temporal Quality			Frame Quality		UC-3DMMAttn			APB		Text Prompts	
	TF↑	MS↑	DD↑	FID↓	AQ↑	SR↑	Scene↑	Color↑	SC↑	BC↑	OC↑	AS↑
Pika-1.0$_{\text{arXiv24}}$ [16]	96.89	97.25	97.47	246.2	47.58	22.34	25.14	61.02	65.98	72.14	88.71	25.88
Show-1$_{\text{IJCV23}}$ [17]	95.58	<u>98.07</u>	98.30	157.4	44.50	53.40	26.20	57.11	70.24	<u>82.43</u>	<u>90.02</u>	27.15
Pyramid Flow$_{\text{arXiv24}}$ [18]	91.24	92.11	90.58	<u>134.5</u>	48.14	58.60	23.68	57.44	<u>81.12</u>	75.44	**91.20**	25.16
Gen-3$_{\text{Online24}}$ [19]	<u>97.06</u>	96.55	<u>99.19</u>	**91.8**	<u>60.10</u>	24.40	26.10	63.29	78.54	76.11	88.01	26.77
CogVideoX-5B$_{\text{arXiv24}}$ [10]	96.17	96.47	96.87	191.9	**70.90**	<u>62.10</u>	25.84	61.82	75.12	70.14	85.20	26.60
OpenSora-Plan1.2$_{\text{arXiv24}}$ [20]	94.22	95.12	97.44	217.3	50.45	60.12	28.44	65.22	72.20	73.13	84.50	27.44
OpenSora1.2$_{\text{Online24}}$ [21]	93.12	94.14	98.24	225.6	49.80	59.44	<u>30.12</u>	<u>67.02</u>	71.40	72.42	83.02	26.92
T2V-Turbo$_{\text{NeurIPS24}}$ [22]	95.44	94.88	96.14	378.4	52.20	61.18	24.22	60.80	73.22	72.98	85.10	<u>28.02</u>
Lavie$_{\text{IJCV24}}$ [23]	96.80	97.50	97.12	278.8	48.90	58.44	26.84	59.10	77.44	79.20	86.44	27.90
MAG (ours)	**97.20**	**98.20**	**99.58**	215.8	51.24	**63.50**	**31.22**	**68.14**	**86.20**	**84.50**	<u>90.70</u>	**28.50**

Fusion Modulation Within this module, scaling and shift parameters $\alpha(a_t)$ and $\beta(a_t)$, derived from the keyboard input vector a_t, modulate the latent features $z_{\mathcal{F}}$ of the video frames via standard normalization and modulation:

$$\hat{z} = \alpha(a_t) \odot \frac{(z_{\mathcal{F}} - \mu)}{\sigma} + \beta(a_t), \tag{9}$$

where μ and σ denote the mean and standard deviation of $z_{\mathcal{F}}$, respectively, and $\odot$ signifies element-wise operations, thereby permitting flexible integration of action prompts in the latent space for real-time adjustments to character poses and camera perspectives.

Influence Intensity Control A Multi-Layer Perceptron (MLP) predicts a gating coefficient $g_t \in (0, 1)$ to reconcile visual smoothness with responsive dynamics, thus adjusting the modulation intensity:

$$\begin{aligned} z_t^{(\text{next})} &= \Gamma^{(l)}(z_t, a_t; g_t) \\ &= g_t \left[\alpha(a_t) \odot \frac{(z_t - \mu)}{\sigma} + \beta(a_t) \right] + z_t(1 - g_t), \end{aligned} \tag{10}$$

where $\Gamma^{(l)}$ designates the operation in the l-th APB layer; a larger g_t induces more pronounced modulation, whereas a smaller g_t promotes smoother transitions. By injecting the output features into the backbone UC-3DMMDiT Block, the system adaptively modulates character and scene generation in real time, achieving a balance between visual coherence and interactive responsiveness.

C. Training and Inference

To achieve robust visual generation and precise controllability, MAG employs a three-stage training pipeline. First, FLUX.1 dev's base weights are fine-tuned on a large static 3D game image dataset, adopting methods from Opensora [21] to inherit generative capabilities while incorporating game scene-specific inductive biases. Next, inspired by CogVideoX [10], data preprocessing (filtering and re-captioning) and batching of multi-resolution image and video data are performed for training. Finally, the APB is integrated into the frozen UC-3DMMDiT backbone and fine-tuned on a curated dataset emphasizing character motion, enhancing motion control for interactive scene generation. During inference, MAG autoregressively models $p(f_{x_{t+1}} \mid f_{x_{1:t}}, c, a)$, where the UC-3DMMDiT Block generates the next frame $f_{x_{t+1}}$ (world state, styles, and camera movements) from prior frames $f_{x_{1:t}}$ and multi-modal controls c, while the APB uses player inputs a to modulate character actions, enabling real-time, controllable, and coherent scene generation. Details are in Appendix D.

III. EXPERIMENT

A. Quantitative Comparison

a) Evaluation Metrics: To evaluate MAG systematically, we use two metric categories: video quality and control performance. Video quality focuses on perceptual fidelity independent of control, while control performance assesses alignment with user instructions. Based on VBench++ [24], we evaluate temporal flickering, motion smoothness, and dynamic degree for cross-frame consistency, alongside FID and aesthetic quality for frame-level fidelity. For control performance, we analyze spatial relationships, scene, and color alignment to assess UC-3DMMAttn, examine subject and background consistency to evaluate APB's effectiveness, and measure overall consistency and appearance style for text-video correspondence. Details are in Appendix C and E.

b) Results: We conducted a quantitative comparison of MAG's performance against several well-known video generation models. Following VBench++ [24]'s methodology, we carefully designed prompts tailored for evaluating game scenarios. For each model, 50 samples were tested, and the average scores were reported. To ensure fairness, all videos were generated at 720p resolution, and each test consisted of

Fig. 3. Qualitative comparison of MAG and other commercial products or Game-Specific generation models. Check more clearer results in Appendix G.

60 consecutive frames. The quantitative results are shown in Table II.

Table II highlights MAG's superior performance in generating high-quality videos, excelling in temporal coherence and dynamic fluidity critical for game videos. MAG achieves state-of-the-art temporal quality metrics, demonstrating exceptional cross-frame consistency and smooth motion. Although its frame-wise quality (e.g., FID) is competitive but not the best, and its aesthetic quality lags slightly behind models like CogVideoX-5B [10], MAG outperforms in control performance, achieving top scores in spatial/non-spatial alignment, subject/background consistency, and stylistic alignment, showcasing the APB module's effectiveness in maintaining coherence with user prompts and stylistic requirements.

Overall, while there is room for improvement in frame-wise fidelity and aesthetics, MAG's advantages in temporal quality and control performance make it a highly competitive solution for interactive game video generation. Detailed metric explanations are in Appendix E.

B. Qualitative Comparison

While quantitative metrics evaluate MAG's performance, certain qualitative aspects, such as the autonomous generation of essential game elements (e.g., mini-maps, health bars, ammo counters) and the avoidance of distortions or frame drift during movement [4], resist numerical assessment. A qualitative comparison (Appendix F, Fig. 3) shows that while QingYing [25], JiMeng [26], Hailuo [27], and Kling [28] fail to integrate key FPS components or perspectives, CogVideoX [10] includes a first-person view and a mini-map but omits health bars and ammo counters. DIAMOND [29], though incorporating a reticle, suffers from frame drift, whereas MAG achieves superior stability and fidelity by seamlessly integrating all core FPS elements, producing smooth and genre-accurate content. More results are in Appendix G and I.

TABLE III
ABLATION RESULTS FOR ENVIRONMENTAL CONTROL

Method	Video Quality			Ctrl. Performance	
	TF↑	FVD↓	FID↓	SC↑	BC↑
w/o APB	97%	1518	378.6	61%	27%
MAG	97%	917	215.8	63%	84%

C. Ablation Study

We conducted two groups of ablation experiments to verify the environmental control performance of UC-3DMMAttn and the action control performance of APB in MAG.

a) Environmental Control Performance: To assess the environmental control capabilities of UC-3DMMAttn, we conducted two experiments: one with the video prompt stream entirely removed, eliminating control information, and another using a 3DMMAttn variant inspired by IPAdapter [3], where control information is added directly for non-spatially aligned tasks (details in Appendix H). Testing 50 samples per setting with identical prompts, results in Table III show that UC-3DMMAttn outperforms the simple addition method in both

TABLE IV
ABALATION RESULTS FOR ACTION CONTROL

Method	Video Quality			Ctrl. Performance	
	TF↑	MS↑	DD↑	SR↑	Color↑
w/o control	96%	95%	99%	11%	41%
w/o spatially aligned	97%	98%	98%	59%	42%
MAG	97%	98%	99%	63%	68%

generation quality and spatial alignment, demonstrating its versatility in balancing spatial and non-spatial tasks during game video generation.

b) Action Control Performance: The detachable, plug-and-play design of APB allowed action control validation without adding network layers, simply detaching APB to use text prompts for controlling character actions. As shown in Table IV, removing APB drastically reduced action generation accuracy, with metrics like subject consistency (SC) and background consistency (BC) experiencing significant declines, alongside visual issues such as character morphing and background deformation. In contrast, models with APB retained precise action control and background integrity, highlighting APB's role in enhancing execution, stability, and flexible, interactive control over game characters.

IV. CONCLUSION

We propose MAG, the first model to extend the DiT architecture for interactive game video generation, integrating UC-3DMMAttn and APB to produce high-quality, drift-free visuals with efficient control over environments and character actions. Future work includes scaling MAG for improved performance and exploring its potential for broader applications in interactive system simulation and content generation.

REFERENCES

[1] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun, "Sora: A review on background, technology, limitations, and opportunities of large vision models," 2024.

[2] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," .

[3] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," 2023.

[4] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter, "Diffusion models are real-time game engines," 2024.

[5] 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.

[6] Virtuals Protocol, "Video game generation: A practical study using mario," 2024, Preprint.

[7] Decart and Spruce Campbell Julian Quevedo, Quinn McIntyre, "Oasis: A universe in a transformer," 2024.

[8] Hrishav Bakul Barua, Kalin Stefanov, KokSheik Wong, Abhinav Dhall, and Ganesh Krishnasamy, "Gta-hdr: A large-scale synthetic dataset for hdr image reconstruction," *arXiv preprint arXiv:2403.17837*, 2024.

[9] Lebin Zhou, Kun Han, Nam Ling, Wei Wang, and Wei Jiang, "Gameir: A large-scale synthesized ground-truth dataset for image restoration over gaming content," 2024.

[10] 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.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[12] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, June 2021, pp. 483–498, Association for Computational Linguistics.

[13] Zongjian Li, Bin Lin, Yang Ye, Liuhan Chen, Xinhua Cheng, Shenghai Yuan, and Li Yuan, "Wf-vae: Enhancing video vae by wavelet-driven energy flow for latent video diffusion model," *arXiv preprint arXiv:2411.17459*, 2024.

[14] Black Forest Labs, "Flux.1dev: Text to image model," https://blackforestlabs.ai/flux-1-tools/, 2024.

[15] Chang Liu, Henghui Ding, Yulun Zhang, and Xudong Jiang, "Multi-modal mutual attention and iterative interaction for referring image segmentation," *IEEE Transactions on Image Processing*, vol. 32, pp. 3054–3065, 2023.

[16] Leijie Wang, Nicolas Vincent, Julija Rukanskaitė, and Amy X. Zhang, "Pika: Empowering non-programmers to author executable governance policies in online communities," 2024.

[17] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou, "Show-1: Marrying pixel and latent diffusion models for text-to-video generation," *arXiv preprint arXiv:2309.15818*, 2023.

[18] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin, "Pyramidal flow matching for efficient video generative modeling," 2024.

[19] Anastasis Germanidis, "Introducing Gen-3 Alpha: A New Frontier for Video Generation," https://runwayml.com/research/introducing-gen-3-alpha, June 2024, Accessed June 17, 2024.

[20] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, Tanghui Jia, Junwu Zhang, Zhenyu Tang, Yatian Pang, Bin She, Cen Yan, Zhiheng Hu, Xiaoyi Dong, Lin Chen, Zhang Pan, Xing Zhou, Shaoling Dong, Yonghong Tian, and Li Yuan, "Open-sora plan: Open-source large video generation model," *arXiv preprint arXiv:2412.00131*, 2024.

[21] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You, "Open-sora: Democratizing efficient video production for all," March 2024.

[22] Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang, "T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback," in *Advances in neural information processing systems*, 2024.

[23] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al., "Lavie: High-quality video generation with cascaded latent diffusion models," *IJCV*, 2024.

[24] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu, "Vbench++: Comprehensive and versatile benchmark suite for video generative models," *arXiv preprint arXiv:2411.13503*, 2024.

[25] Zhipu QingYan, "Zhipu QingYan: A Dialogue Model with Trillion Parameters," Accessed on 2024-12-24.

[26] Shenzhen Lianmeng Technology Co., Ltd., "Jimeng AI - A One-Stop AI Creation Platform," 2024, Accessed on 2024-12-24.

[27] MiniMax, "Conch AI — A Productivity Product under MiniMax, Your AI Partner, 10x Work and Study Efficiency," Accessed on 2024-12-24.

[28] Kling AI, "Kling," https://klingai.kuaishou.com/, June 2024, Accessed June 6, 2024.

[29] 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.