GameGen-𝕏: Interactive Open-world Game Video Generation

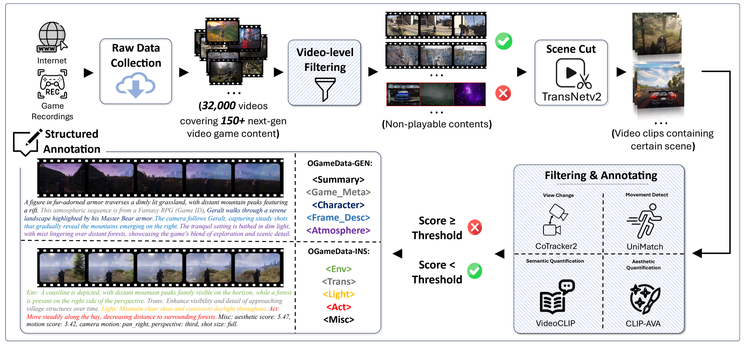
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Haoxuan Che*, Xuanhua He*, Quande Liu#, Cheng Jin, Hao Chen#

* *Indicates co-first authorship; # indicates co-corresponding authors*

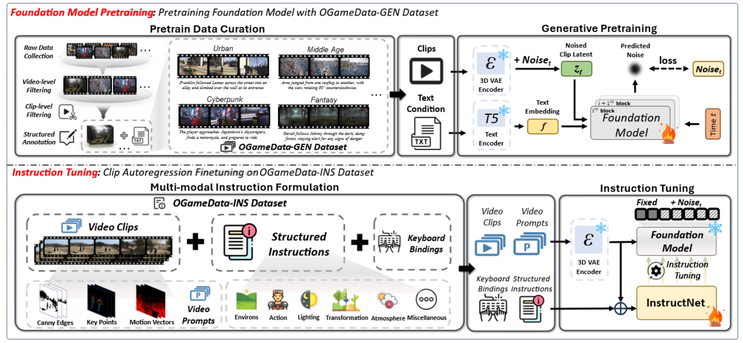
ICLR

OGameData Construction Pipeline



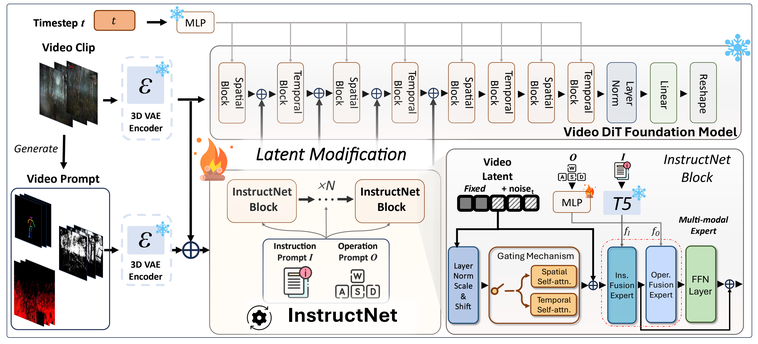
- **OGameData**: A large-scale dataset with 1 million high-res game video clips and 600+ words/min caption density, designed for video generation and interactive control.
- **OGameData-GEN** for generative training and **OGameData-INS** for instruction-based interactive control, providing detailed scene and character annotations.
- Excels in video-text alignment, offering high-quality metadata for 150+ games through precise filtering, segmentation, and annotation.

Two-stage GameGen-X Training Framework



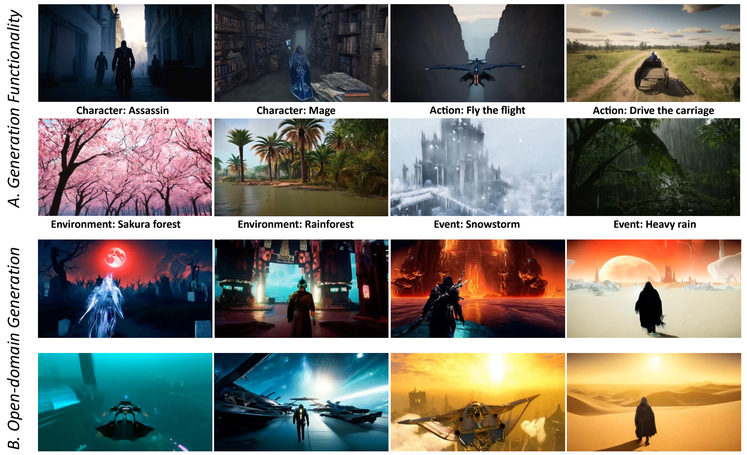
- **GameGen-X** employs a two-stage training framework: first, pretraining the foundation model for game content generation, followed by fine-tuning for interactive control.
- **Foundation Model Pretraining** uses video clips with text conditions to train the model, while **Instruction Tuning** refines it for clip autoregression and multi-modal instruction control.
- This framework enables generation of open-domain game content and dynamic interactive video control, combining text-to-video generation with player input modification.

GameGen-X Model Architecture



- **GameGen-X** integrates a Video DiT Foundation Model with **InstructNet** for generating and controlling game content through multi-modal prompts.
- The architecture uses 3D-VAE encoders to process video clips and prompts, followed by a series of spatial and temporal blocks for video generation.
- **InstructNet** modifies video latents based on structured text instructions and operation prompts, enabling interactive control over scene dynamics and character movements.

GameGen-X for Game Content Generation and Interactive Control



Functionality and Pipeline

- **GameGen-X** supports open-domain game video generation with rich diversity in styles, characters, environments, and actions. (*Fig. A*)
- It enables fine-grained functionality control, allowing users to specify character and environment elements such as actions and events. (*Fig. B*)
- **Interactive control** is supported via structured text prompts and keyboard inputs, enabling scene and character manipulation to respond to users' control signals. (*Fig. C*)
- In practice, users first generate an initial video clip from any prompt (e.g., "foggy desert at sunset"), then interactively modify it using control inputs or visual cues, which simulates the gameplay procedure.

Quantitative Experiment

- **GameGen-X** simultaneously supports high-quality generation and interactive control, meeting the demands of content creation.
- The results demonstrate that it effectively handles diverse scenarios, from complex scene to responsive character manipulation.
- **Ablation studies** confirm the importance of OGameData and InstructNet, with notable drops in performance when key components are removed.

Table 2: Generation Performance Evaluation (* denotes key metric for generation ability)

Method	Resolution	Frames	FID* ↓	FVD* ↓	TVA* ↑	UP* ↑	MS ↑	DD ↑	SC ↑	IQ ↑
Mira (Zhang et al. (2023))	480p	60	360.9	2254.2	0.27	0.25	0.98	0.62	0.94	0.63
OpenSora-Plan1.2 (Lab & etc. (2024))	720p	102	407.0	1940.9	0.38	0.43	0.99	0.42	0.92	0.39
CogVideoX-5B (Yang et al. (2024))	480p	49	316.9	1310.2	0.49	0.37	0.99	**0.94**	0.92	**0.53**
OpenSora1.2 (Zheng et al. (2024b))	720p	102	318.1	1016.3	0.50	0.37	0.98	0.90	0.87	0.52
GameGen-X (Ours)	720p	102	**252.1**	**759.8**	**0.87**	**0.82**	0.99	0.80	0.94	0.50

Table 3: Control Performance Evaluation (* denotes key metric for control ability)

Method	Resolution	Frames	SR-C* ↑	SR-E* ↑	UP ↑	MS ↑	DD ↑	SC ↑	IQ ↑
OpenSora-Plan1.2 (Lab & etc. (2024))	720p	102	26.6%	31.7%	0.46	0.99	0.72	**0.90**	0.51
CogVideoX-5B (Yang et al. (2024))	480p	49	23.0%	30.3%	0.45	0.98	0.63	0.85	**0.55**
OpenSora1.2 (Zheng et al. (2024b))	720p	102	21.6%	14.2%	0.17	0.99	**0.97**	0.84	0.45
GameGen-X (Ours)	720p	102	**63.0%**	**56.8%**	**0.71**	0.99	0.88	0.88	0.44

Table 4: Ablation Study for Generation Ability

Method	FID ↓	FVD ↓	TVA ↑	UP ↑	MS ↑	SC ↑
w/ MiraData	303.7	1423.6	0.70	0.48	0.99	0.94
w/ Short Caption	303.8	**1167.7**	0.53	0.49	0.99	0.94
w/ Progression	294.2	1169.8	0.68	0.53	0.99	0.93
Baseline	**289.5**	1181.3	**0.83**	**0.67**	0.99	**0.95**

Table 5: Ablation Study for Control Ability.

Method	SR-C ↑	SR-E ↑	UP ↑	MS ↑	SC ↑
w/o Instruct Caption	31.6%	20.0%	0.34	0.99	0.87
w/o Decomposition	32.7%	23.3%	0.41	0.99	0.88
w/o InstructNet	12.3%	17.5%	0.16	0.98	0.86
Baseline	**45.6%**	**45.0%**	**0.50**	0.99	**0.90**

Conclusion

- **GameGen-X** is the first diffusion transformer model tailored for open-world game video generation with multi-modal interactive control capabilities.
- By leveraging the large-scale **OGameData** and a two-stage training framework, the model effectively unifies content generation and controllable video continuation.
- This work demonstrates the feasibility of automated game content creation, shows the possibility or future research on data-driven and user-guided game design.