Unbounded: A Generative Infinite Game of Character Life Simulation

Jialu Li, Yuanzhen Li, Neal Wadhwa, Yael Pritch,
David E. Jacobs, Michael Rubinstein, Mohit Bansal, Nataniel Ruiz

What’s generative infinite game?

- Generative: All the game environments, character states, character actions are encapsulated with generative models.

What's generative infinite game?

- Generative: All the game environments, character states, character actions are encapsulated with generative models.
- Infinite: "played for the purpose of continuing the play", with no fixed boundaries and evolving rules.

Main Components

- Real-time controllable text-to-image generation model capable of maintaining character consistency and environment consistency throughout the generative game

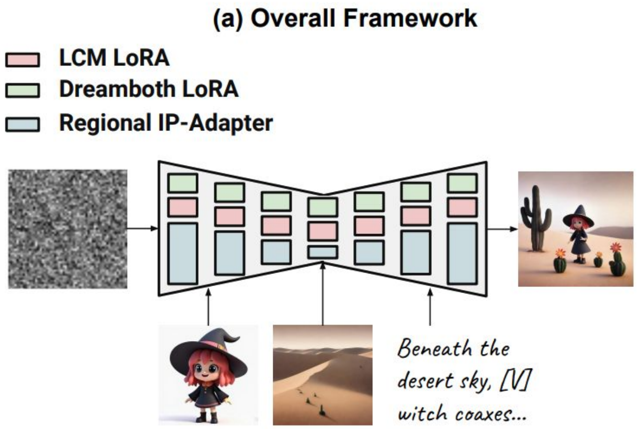
Main Components

- Real-time controllable text-to-image generation model capable of maintaining character consistency and environment consistency throughout the generative game
- A large language model acts as game engine for interactive user experience

Controllable Text-to-Image Generation

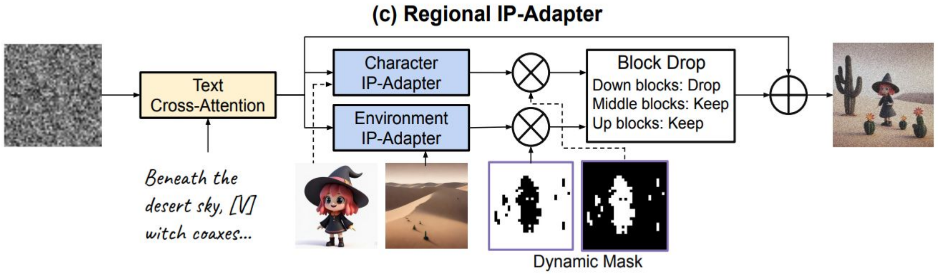
- Latent consistency model for few-step inference

- Dreambooth for character consistency



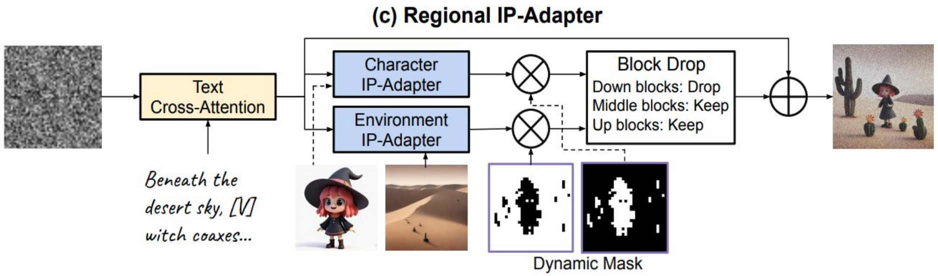
Controllable Text-to-Image Generation

- Dynamic regional IP-Adapter with block drop to mitigate interference.



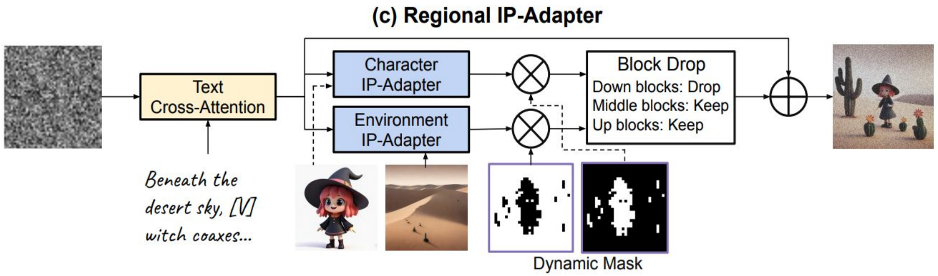
Controllable Text-to-Image Generation

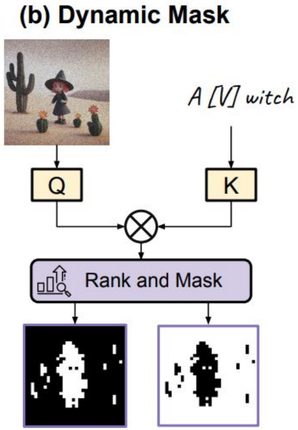
- Dynamic regional IP-Adapter with block drop to mitigate interference.
- Regional: Apply condition to relative region instead of full image.



Controllable Text-to-Image Generation

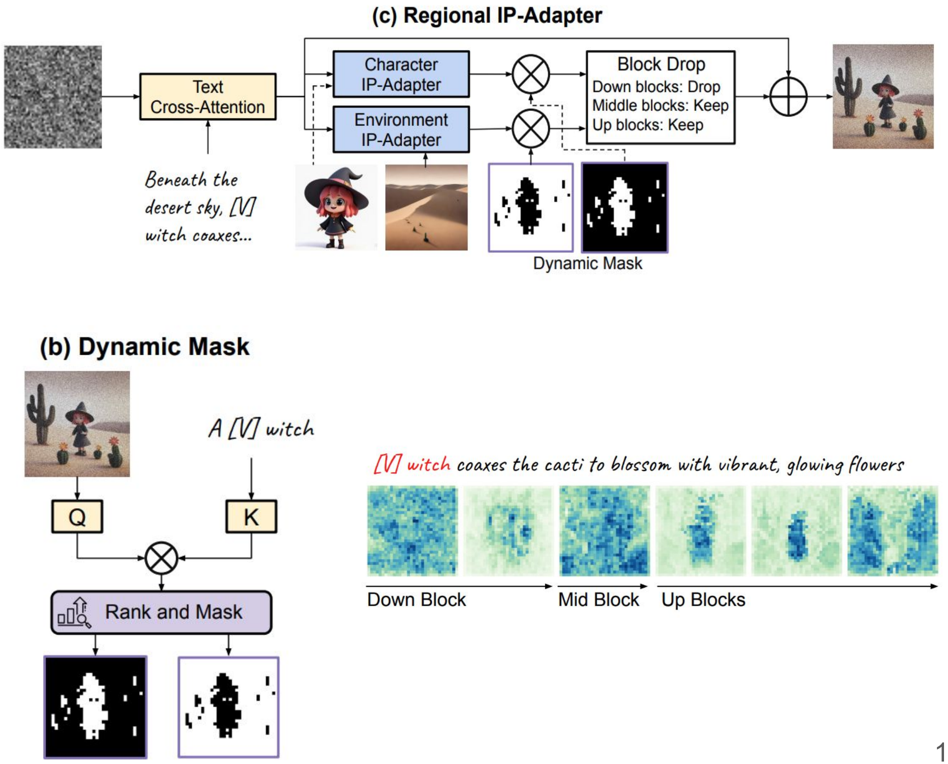
- Dynamic regional IP-Adapter with block drop to mitigate interference.
- Regional: Apply condition to relative region instead of full image.
- Dynamic: The region mask is dynamically extracted from cross-attention layers.





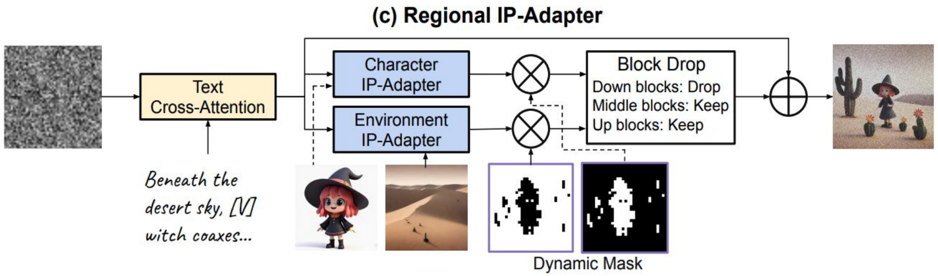
Controllable Text-to-Image Generation

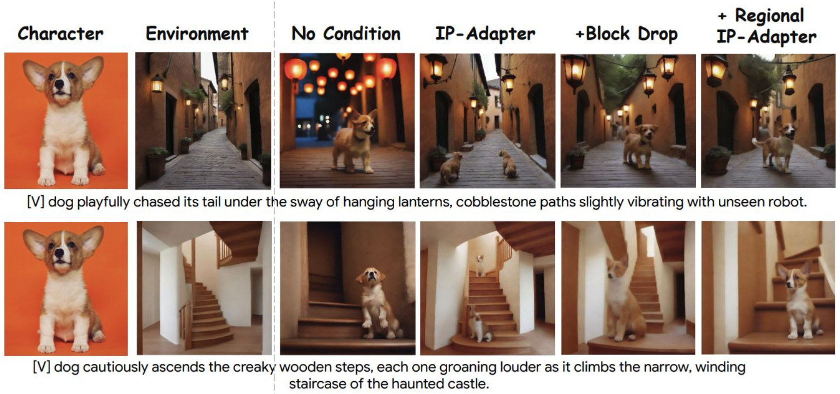
- Dynamic regional IP-Adapter with block drop to mitigate interference.
- Regional: Apply condition to relative region instead of full image.
- Dynamic: The region mask is dynamically extracted from cross-attention layers.
- Block Drop: Drop blocks with low quality masks



Controllable Text-to-Image Generation

- Dynamic regional IP-Adapter with block drop to mitigate interference.
- Regional: Apply condition to relative region instead of full image.
- Dynamic: The region mask is dynamically extracted from cross-attention layers.
- Block Drop: Drop blocks with low quality masks





Evaluation -- Comparison with Previous Approach

- Evaluation dataset: 5,000 (character image, environment description, text prompt) triplets
- Evaluation metrics: CLIP, DINO, DreamSim

Evaluation -- Comparison with Previous Approach

- Evaluation dataset: 5,000 (character image, environment description, text prompt) triplets
- Evaluation metrics: CLIP, DINO, DreamSim

Methods	**Environment Consistency**			**Character Consistency**			**Semantic Alignment**
	CLIP-I^E ↑	DINO^E ↑	DreamSim^E ↓	CLIP-I^C ↑	DINO^C ↑	DreamSim^C ↓	CLIP-T ↑
IP-Adapter (Ye et al., 2023)	0.470	0.381	0.595	0.366	0.139	0.832	0.168
IP-Adapter-Instruct (Rowles et al., 2024)	0.334	0.151	0.832	0.246	0.124	0.872	0.098
StoryDiffusion (Zhou et al., 2024c)	0.528	0.257	0.733	0.629	0.464	0.545	**0.242**
Ours	**0.563**	**0.322**	**0.675**	**0.676**	**0.470**	**0.488**	**0.242**

- Our approach consistently outperforms previous approach in maintaining environment consistency and character consistency, while achieving comparable performance in maintaining semantic alignment.

Evaluation -- Effectiveness of Regional IP-Adapter

No.	Block Drop	Regional IP-Adapter	Scale	Environment Consistency			Character Consistency			Alignment
				CLIP-I^E ↑	DINO^E ↑	DreamSim^E ↓	CLIP-I^C ↑	DINO^C ↑	DreamSim^C ↓	CLIP-T ↑
1.	✗	✗	1.0	0.123	0.111	0.885	0.073	0.024	0.973	0.034
2.	✓	✗	1.0	0.414	**0.331**	**0.647**	0.337	0.147	0.832	0.149
3.	✓	✓	1.0	**0.563**	0.322	0.675	**0.676**	**0.470**	**0.488**	**0.242**
4.	✗	✗	0.5	0.470	0.381	0.595	0.366	0.139	0.832	0.168
5.	✓	✗	0.5	**0.577**	**0.332**	**0.640**	0.627	0.374	0.575	**0.252**
6.	✓	✓	0.5	0.549	0.263	0.726	**0.705**	**0.514**	**0.450**	0.246

- Adding block drop improves both environment and character consistency compared with multi-IP-Adapter.

Evaluation -- Effectiveness of Regional IP-Adapter

No.	Block Drop	Regional IP-Adapter	Scale	Environment Consistency			Character Consistency			Alignment
				CLIP-I^E ↑	DINO^E ↑	DreamSim^E ↓	CLIP-I^C ↑	DINO^C ↑	DreamSim^C ↓	CLIP-T ↑
1.	✗	✗	1.0	0.123	0.111	0.885	0.073	0.024	0.973	0.034
2.	✓	✗	1.0	0.414	**0.331**	**0.647**	0.337	0.147	0.832	0.149
3.	✓	✓	1.0	**0.563**	0.322	0.675	**0.676**	**0.470**	**0.488**	**0.242**
4.	✗	✗	0.5	0.470	0.381	0.595	0.366	0.139	0.832	0.168
5.	✓	✗	0.5	**0.577**	**0.332**	**0.640**	0.627	0.374	0.575	**0.252**
6.	✓	✓	0.5	0.549	0.263	0.726	**0.705**	**0.514**	**0.450**	0.246

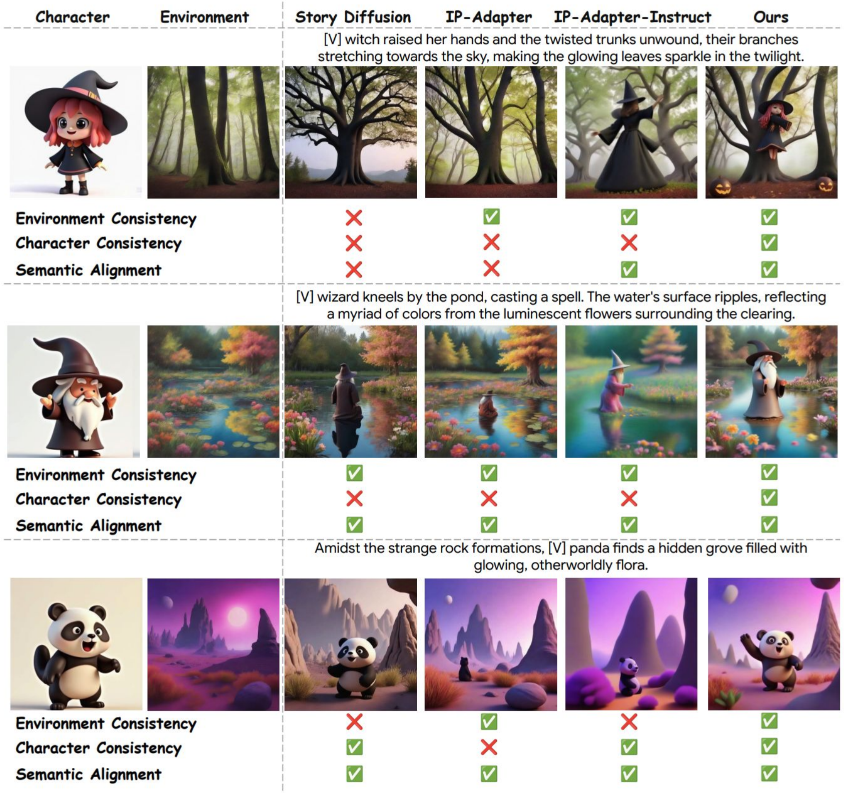
- Adding block drop improves both environment and character consistency compared with multi-IP-Adapter.
- Our regional IP-Adapter enhances character consistency and text alignment while maintaining comparable performance in environment consistency.

Evaluation -- Effectiveness of Regional IP-Adapter

No.	Block Drop	Regional IP-Adapter	Scale	Environment Consistency			Character Consistency			Alignment
				CLIP-I^E ↑	$DINO^E$ ↑	$DreamSim^E$ ↓	CLIP-I^C ↑	$DINO^C$ ↑	$DreamSim^C$ ↓	CLIP-T ↑
1.	✗	✗	1.0	0.123	0.111	0.885	0.073	0.024	0.973	0.034
2.	✓	✗	1.0	0.414	**0.331**	**0.647**	0.337	0.147	0.832	0.149
3.	✓	✓	1.0	**0.563**	0.322	0.675	**0.676**	**0.470**	**0.488**	**0.242**
4.	✗	✗	0.5	0.470	0.381	0.595	0.366	0.139	0.832	0.168
5.	✓	✗	0.5	**0.577**	**0.332**	**0.640**	0.627	0.374	0.575	**0.252**
6.	✓	✓	0.5	0.549	0.263	0.726	**0.705**	**0.514**	**0.450**	0.246

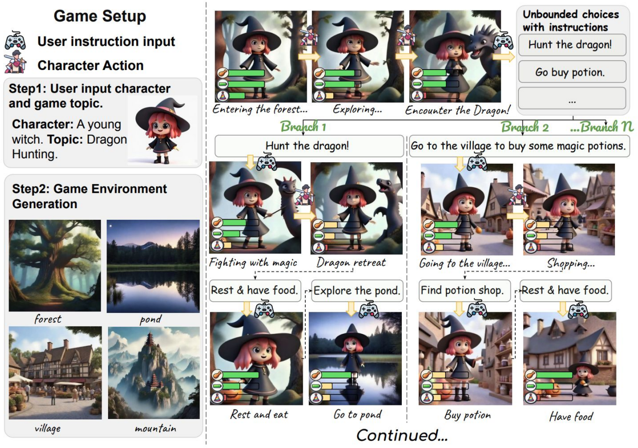
- Injecting IP-Adapter with different scale can effectively balance character consistency and environment consistency.

Evaluation -- Qualitative Examples



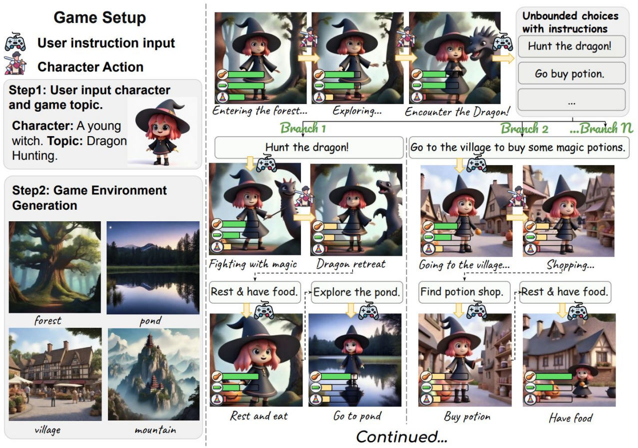
LLM as Game Engine

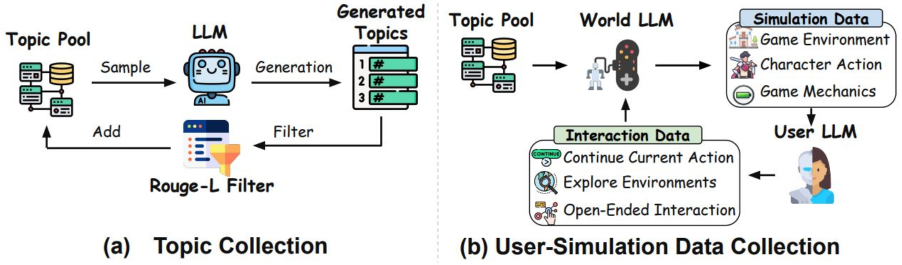
- Game Engine
  - Environment control
  - Coherent story generation
  - Game mechanics
  - Prompt rewriting



LLM as Game Engine

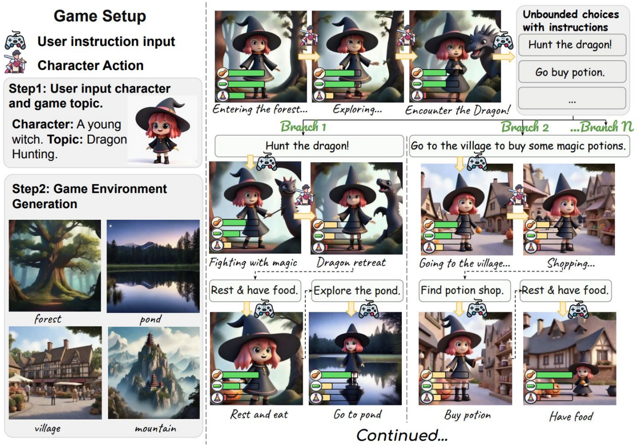
- Game Engine
  - Environment control
  - Coherent story generation
  - Game mechanics
  - Prompt rewriting
- Distill Gemma-2B for real-time interactive response

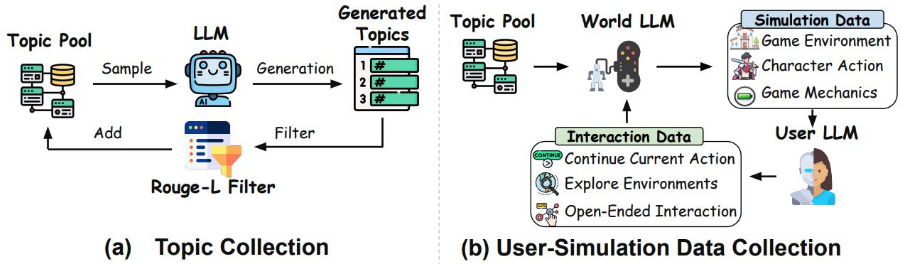




LLM as Game Engine

- Game Engine
  - Environment control
  - Coherent story generation
  - Game mechanics
  - Prompt rewriting
- Distill Gemma-2B for real-time interactive response
- Data collection
  - Diverse topic collection
  - Multi-round interaction data between the world simulation LLM and the user LLM.





Evaluation -- Comparison with Large Models

- Evaluation dataset: 100 user-simulator interaction samples
- Evaluation metrics: GPT-4 as judge

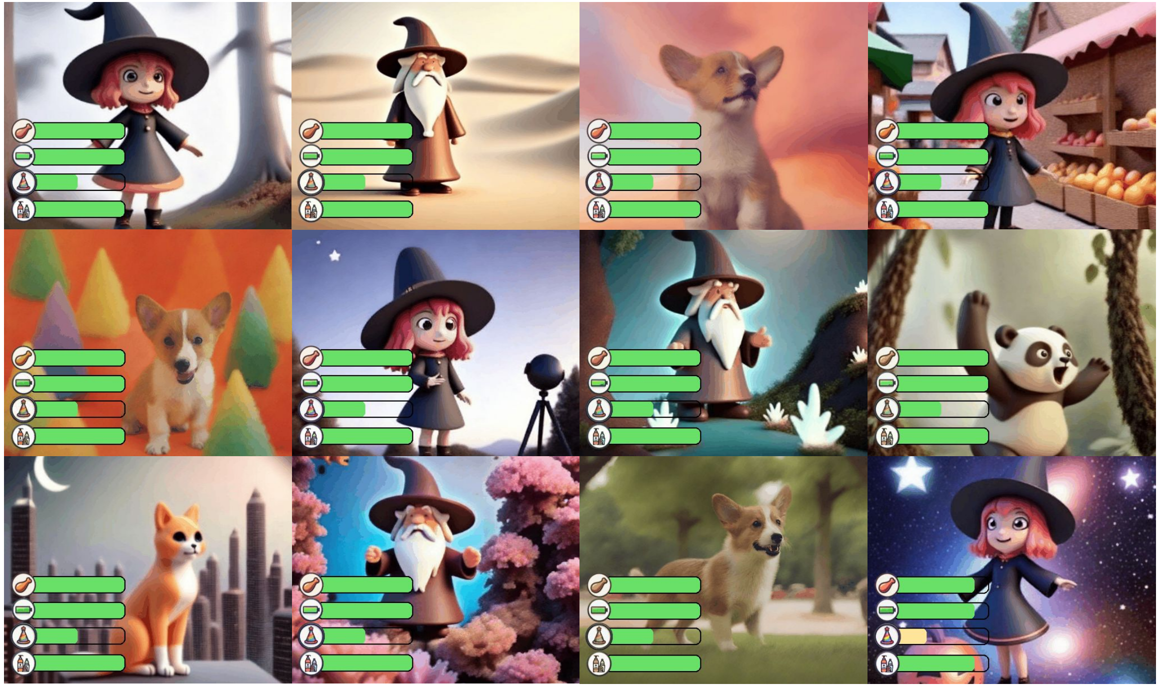
Evaluation -- Comparison with Large Models

- Evaluation dataset: 100 user-simulator interaction samples
- Evaluation metrics: GPT-4 as judge

Model	Overall		State Update		Environment Relevance		Story Coherence		Instruction Following	
	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours
Gemma-2B (Team et al., 2024)	6.22	**7.44**	5.60	**7.47**	6.12	**7.94**	6.34	**7.57**	6.43	**7.67**
Gemma-7B (Team et al., 2024)	6.80	**7.39**	6.29	**7.43**	7.07	**7.91**	6.90	**7.48**	6.89	**7.53**
Llama3.2-3B (Meta, 2024)	7.21	**7.50**	6.86	**7.38**	7.63	**7.93**	7.36	**7.56**	7.31	**7.67**
Ours-1k	7.65	**7.82**	7.50	**7.74**	8.10	**8.19**	7.78	**7.93**	7.82	**7.97**
GPT-4o (OpenAI, 2023)	**7.76**	7.68	**7.69**	7.66	**8.20**	8.10	**7.95**	7.82	**7.85**	7.82

- Distillation with 5k data achieves comparable performance to GPT-4o.

Summary



Thank you!

Project Website: https://generative-infinite-game.github.io/

If you have any questions, please contact jialuli@cs.unc.edu.