SIGGRAPH ASIA 2024 TOKYO 東京

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

Conference | 3–6 December 2024
Exhibition | 4–6 December 2024
Venue | Tokyo International Forum, Japan

Anim–Director: A Large Multimodal Model Powered Agent for Controllable Animation Video Generation

Yunxin Li, Haoyuan Shi, Baotian Hu, Longyue Wang, Min Zhang
Harbin Institute of Technology, Shenzhen (HITSZ)

Sponsored by acm 
Organized by koelnmesse

Automatic Animation Generator

Task Form

- Given a short story or instruction, the animation-making system generates an animated video for users

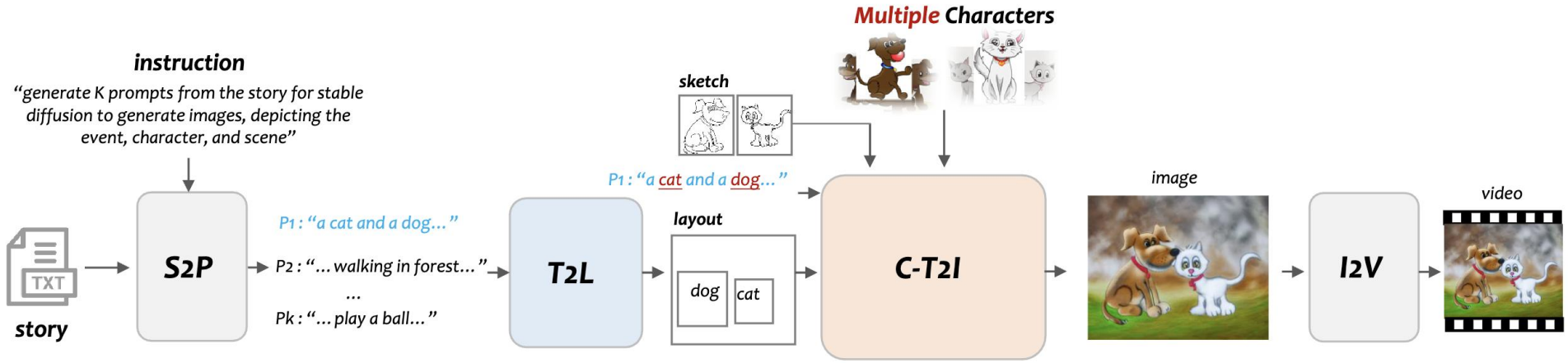
Challenge

- **Complexity in Algorithms**: computer graphics, motion physics, and artificial intelligence.
- **Balancing Control and Automation**: achieving the right balance between control (**user-driven customization**) and automation (**algorithm-driven generation**). Providing too much control can complicate the user interface.
- **Realism and Quality**: generating animations that are visually appealing and realistic is difficult. It involves challenges in **rendering, lighting, camera movement**, and ensuring smooth motion without artifacts or unnatural movements.

Automatic Animation Generator

Previous Method

- Complex pipelines require a lot of **manual intervention** and **high-quality training data**
- Every stage often contains a neural model trained with special human labelling data
- How to **simplify this system, reduce manual intervention and improve production efficiency**?

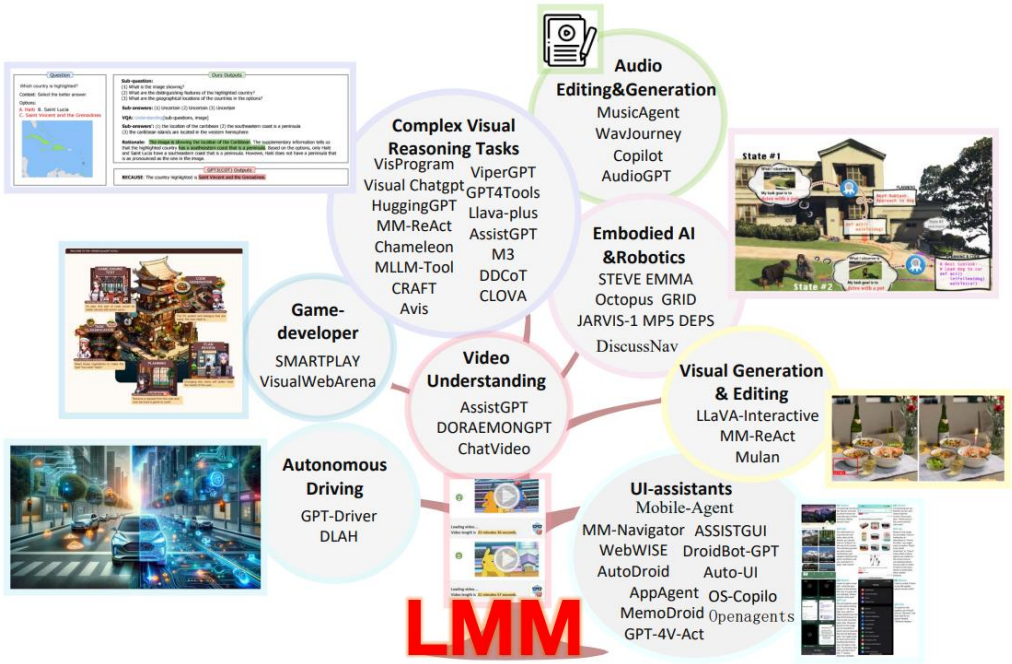


SIGGRAPH ASIA 2024 TOKYO 東京

Anim-Director: LMM-Based Animation Generator

Advantages of Large Multimodal Models (LMMs)

- Powerful **understanding and reasoning capabilities** on Images and videos
- LMMs can undersand **the details** of human instructions and images, with powerful generalization
- LMMs can **replace human** to interact with the external environment



SIGGRAPH ASIA 2024 TOKYO 東京

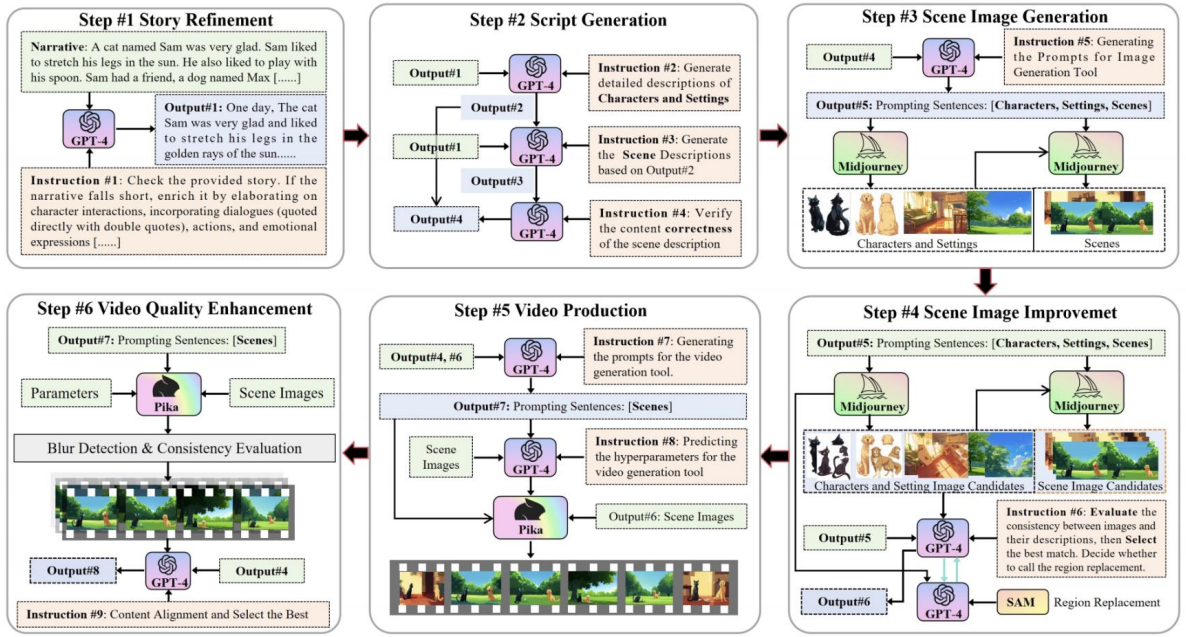
Anim–Director: LMM Powered Animation Agent

Core

- LMM **serves as the central processor** and interact deeply with generative AI tools, effectively taking over roles traditionally performed by humans.

Training-Free

- LMMs interact seamlessly with generative image/video tools to **generate prompts, evaluate visual quality, and select the best one** to optimize the final output



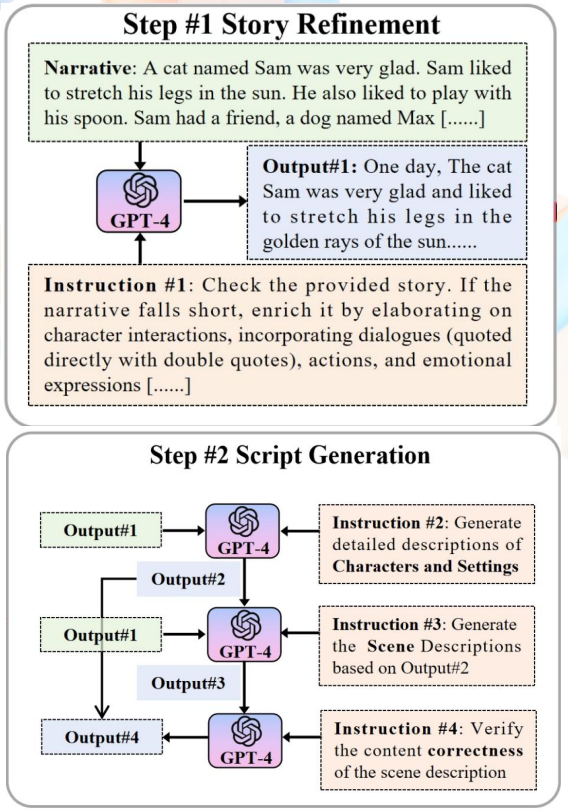
Anim–Director: LMM Powered Animation Agent

Story Refinement

- **Expand a brief story into a detailed and plot-rich narrative** by adding character dialogue and story details.

Script Generation

- **Character and Scene Setting Extraction:** Extract and unify character and scene formats.
- **Structured Scene Description Generation:** Generate detailed scene descriptions based on the extracted characters and settings.
- **Verification and Adjustment:** Ensure scene descriptions align with the story content and make necessary adjustments.



SIGGRAPH ASIA 2024 TOKYO 東京

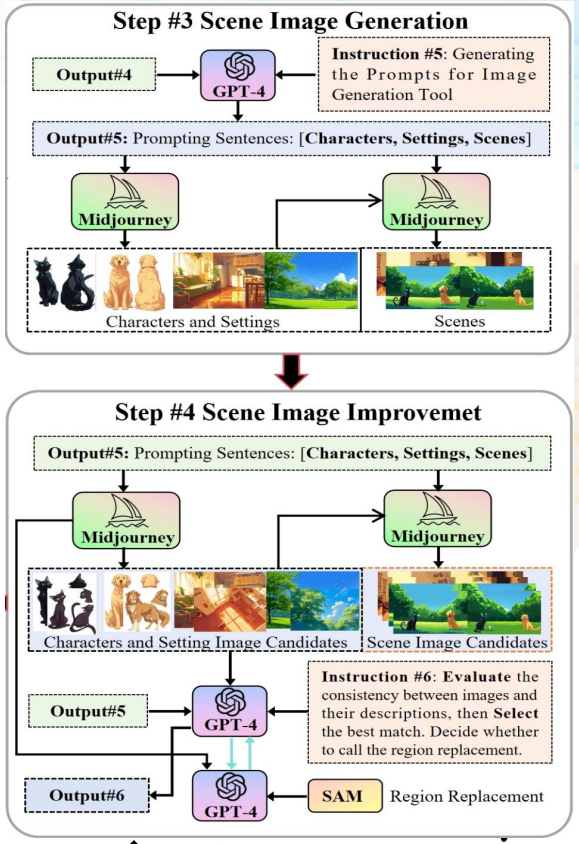
Anim–Director: LMM Powered Animation Agent

Scene Image Generation

- **Generate Scene Images**: Use Midjourney's Text-to-Image method to visualize characters and settings based on GPT-4-generated prompts.

Image Quality Improvement

- **Multiple Candidates Generation**: Generate multiple image candidates. GPT-4 analyzes the generated images to select the best one, ensuring consistency with scene descriptions and reasonable image layout.
- **Consistency Check**: Verify character consistency between generated images and prior character images. This involve using SAM for image segmentation and Midjourney for region replacement. These images are refined and iterated until they pass the consistency check.



Anim–Director: LMM Powered Animation Agent

Video Production

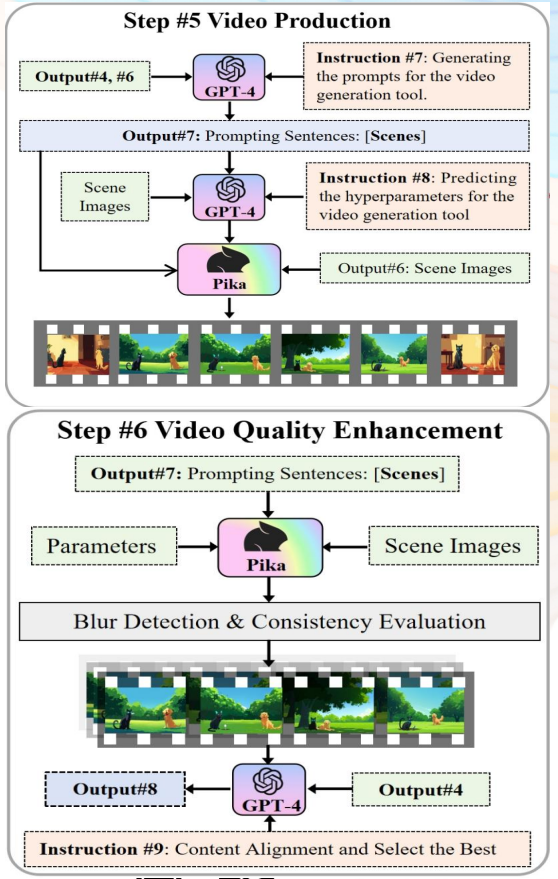
Text + Image -> Video: Combine scene images and text prompts for controllable video generation.

Prompt Generation & Parameter Prediction: GPT-4 creates prompts and predicts optimal hyperparameters for video generation.

Video Quality Enhancement

Generate & Evaluate Candidates: Create ten video candidates for each scene. Use video content metrics (e.g., blur detection, subject and background consistency) to evaluate and rank the top candidates.

Select & Refine: GPT-4 selects the best candidate for each scene based on metrics above. Then all the scene videos are spliced to form the final complete animation video.



Experimental Result

- Anim-Director achieves **the highest performance** in **both image and video** quality generation.
- It excels in **contextual coherence** and **Text-Image similarity.**
- The **deep interaction** between tools and LMM is effective.

Method	Coherence	I-I Sim	T-I Sim
Custom-Diffusion [Kumari et al. 2023]	0.74	0.66	0.28
DPT-T2I [Qu et al. 2024]	0.75	0.65	**0.29**
MidJourney-V6 [2024]	0.76	0.69	0.28
Anim-Director (Ours)	**0.87** ↑ 11%	**0.85** ↑ 16%	**0.29**
Ours w/o Image Improvement	0.82	0.82	0.28

Table 3. Quantitative text-image evaluations on Contextual Coherence (Coherence), Image-Image Similarity (I-I Sim), and Text-Image Similarity (T-I Sim).

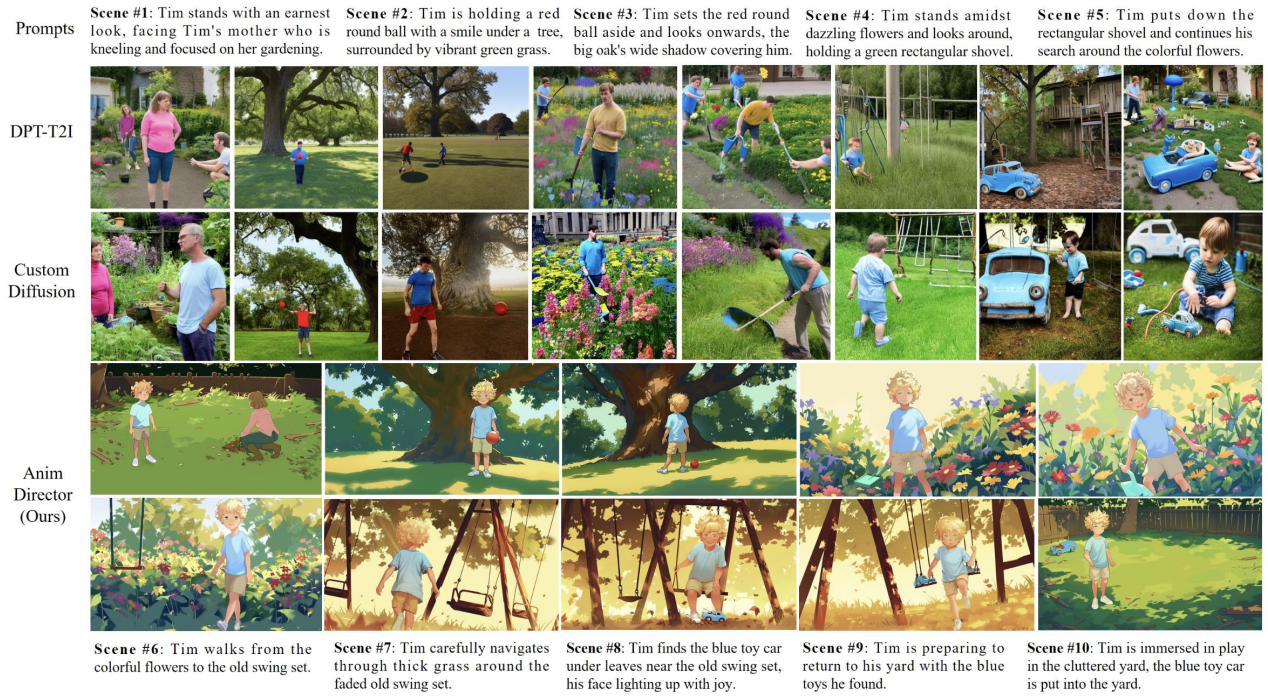
Method	V-Q	Subject	BackG	T-V	Avg.#Len
VideoCrafter1 [Chen et al. 2023b]	0.55	0.71	0.86	0.15	17.3
VideoCrafter2 [Chen et al. 2024]	0.70	0.81	0.88	0.17	17.4
TaleCrafter [Gong et al. 2023]	0.65	0.71	0.79	0.18	17.4
ModelScope [Wang et al. 2023c]	0.70	0.48	0.75	0.15	26.9
AnimateDiff [Guo et al. 2024]	0.71	0.78	0.84	0.16	17.3
SVD [Blattmann et al. 2023a]	0.50	0.81	0.90	0.16	15.1
DynamiCrafter [Xing et al. 2023]	0.72	0.82	0.88	0.18	21.7
Vlogger [Zhuang et al. 2024]	0.72	0.80	0.87	0.17	21.6
Gen-2 [2023]	0.70	0.82	0.88	0.18	42.3
Pika-v2 [2024]	0.66	0.82	0.90	0.18	32.5
Anim-Director (Ours)	**0.74**	**0.86**	**0.93**	0.19	35.0
Ours w/o Video Enhancement	0.67	0.84	0.91	0.18	35.0

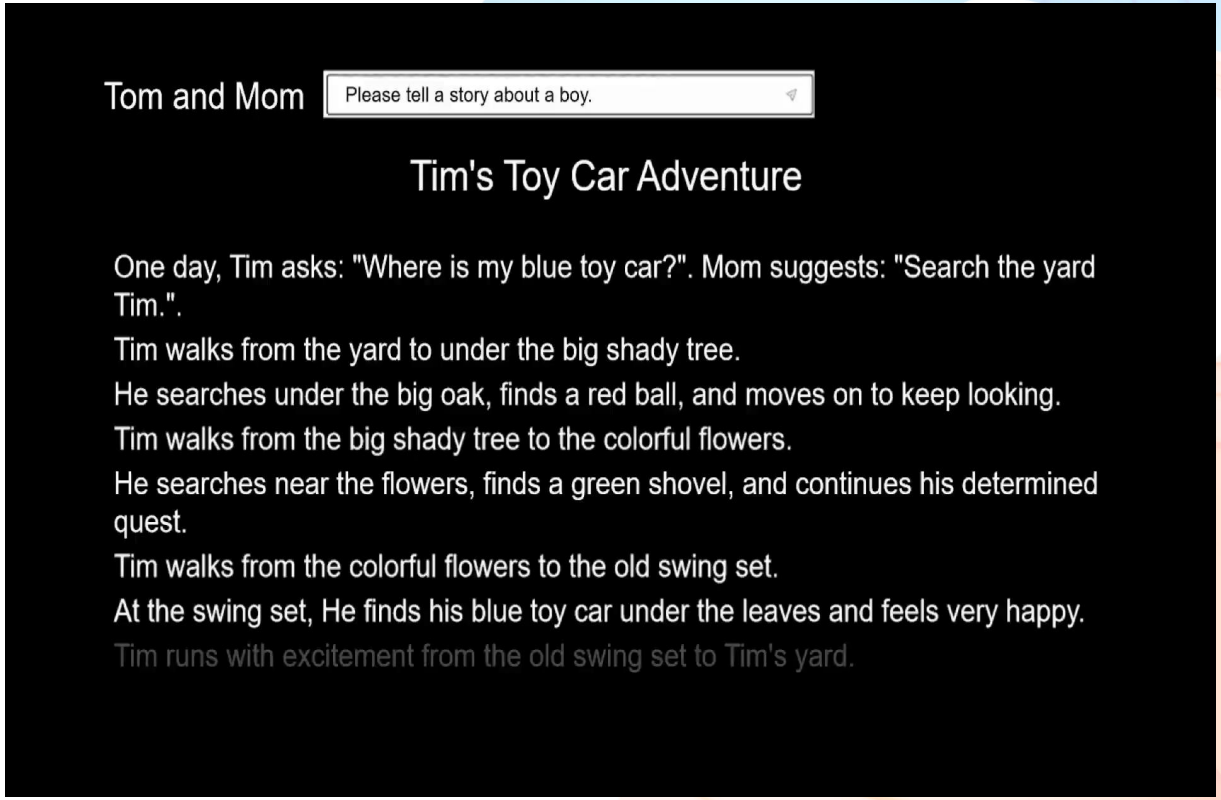
Table 4. Quantitative video quality comparisons on Distortion detection (V-Q), Subject consistency (Subject), Background consistency (BackG), and Text-Video alignment score (T-V). A higher score indicates better performance. Avg.#Len refers to the average duration (second) of videos.

SIGGRAPH ASIA 2024 TOKYO 東京

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

Case: Tim's Toy Car Adventure





Case: Tom and His Mule's Tale of Friendship

Narrative

Tom had a mule that was not much to look at but dearly loved. They worked side by side on the farm, where the mule proved invaluable in carrying loads. One day, the mule injured its leg. Concerned, Tom quickly said, "Oh no! I must repair you, my friend," and bandaged the mule's leg, which soon felt better. Once healed, the duo returned to their farm duties, happier than ever. Tom often thanked his sturdy companion, saying, "Thank you, my friend. You are the best mule." Together, they continued to live happily on the farm.

DPT-T2I

Custom Diffusion

Anim Director (Ours)

SIGGRAPH ASIA 2024 東京

Takeaway Message

- Large Multimodal Models (LMMs) based multimodal agents will help human handle some complex tasks by **interacting with the external environment**, such as GUI-Agents.

- More powerful image and video generation models can further improve the quality of generated long animation videos. Current generative models often produce inferior outcomes, so **we need a more stable and controllable generative model**.

- **Unifying multimodal understanding and generation models**, like Uni-MoE, can streamline the processes of script creation, and image or video generation and editing by using a single, integrated system.

If you have any question, please feel free to contact me

Personal Web: *yunxinli.github.io* *E-mail:* *liyunxin@stu.hit.edu.cn* *Twitter:* *Yunxin Li (@LyxTg)*

SIGGRAPH ASIA 2024 TOKYO 東京

Thanks for your Listening