



Mask²DiT: Dual Mask-based Diffusion Transformer for Multi-Scene Long Video Generation

Tianhao Qi^{1,2} Jianlong Yuan^{2†} Wanquan Feng² Shancheng Fang^{3*} Jiawei Liu² SiYu Zhou²
Qian He² Hongtao Xie¹ Yongdong Zhang¹

¹University of Science and Technology of China ²Bytedance Intelligent Creation
³Yuanshi Inc.

qth@mail.ustc.edu.cn {fangsc, htxie, zyd73}@ustc.edu.cn

{yuanjianlong, fengwanquan, liujiawei.cc22, zhousiyu.vladimir, heqian}@bytedance.com



Project Page



Arxiv



Code



Laboratory



Personal Page

Outline



2

1

Motivation

2

Related Work

3

Methodology











4

Experiments

5

Conclusion

➤ Current Status of Video Generation Models

- Wan 2.1 : open-sourced, 81 frames, 16 fps, 5s, single-scene
- Hunyuan Video : open-sourced, 129 frames, 24 fps, 5s, single-scene
- CogVideoX : open-sourced, 49/161 frames, 8/16 fps, 5/10s, single-scene
- Veo2 : close-sourced, 192 frames, 24 fps, 8s, single-scene
- Kling : close-sourced, 121/241 frames, 24 fps, 5/10s, single-scene
- Sora : close-sourced, 150 frames, 30 fps, 5s, single-scene
- Pika 2.2 : close-sourced, 121/241 frames, 24 fps, 5/10s, single-scene
- Gen 3 Alpha : close-sourced, 256 frames, 24 fps, 10s, single-scene
- Hailuo : close-sourced, 141 frames, 25 fps, 5s, single-scene
- Dreamina : close-sourced, 121/241 frames, 24 fps, 5/10s, single/multi-scene(s)

In summary, most mainstream video generation models are limited to producing **single-scene** videos with durations of **up to 10 seconds**.

- Limitations of Current Video Generation Models
 - **Duration Limitation:** Difficult to generate long-duration video content
 - **Single-Shot Limitation:** Lacks multi-scene transitions and narrative capability
 - **Poor Coherence:** Challenging to maintain content coherence and consistency in long videos

The above issues hinder the applicability of current models in commercial or real-world scenarios beyond **entertainment**.

- Motivation
 - How to generate **longer videos**
 - How to enable **multi-scene storytelling with narrative control**
 - How to ensure **temporal consistency across scenes**, as well as **visual-textual alignment within each scene**

Outline



5

1

Motivation

2

Related Work

3

Methodology

4

Experiments

5

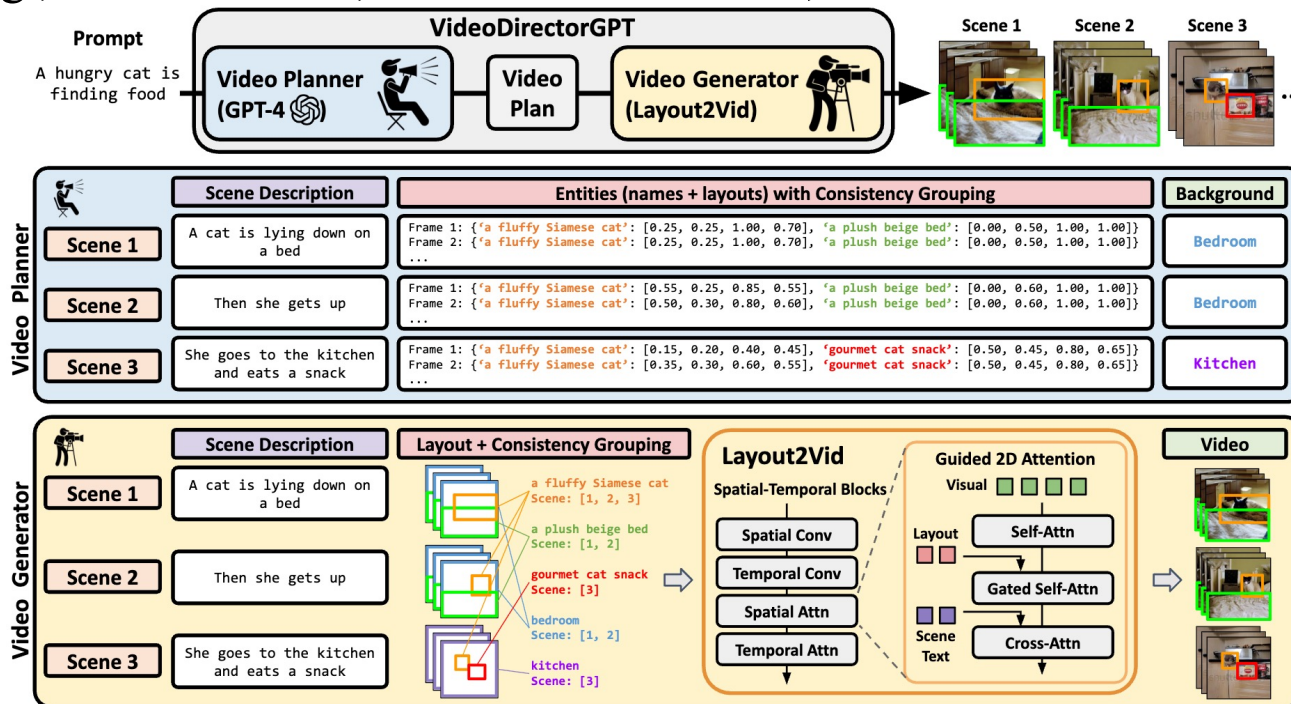
Conclusion

Related Work



6

- LLM/Agent-based Methods
 - Leverage LLM/Agent to create multi-scene scripts
 - Generate each scene independently
 - E.g., VideoStudio^[1], VideoDirectorGPT^[2], Mora^[3]



[1] Lin H, Zala A, Cho J, et al. VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning[C]//First Conference on Language Modeling.

[2] Long F, Qiu Z, Yao T, et al. VideoStudio: Generating Consistent-Content and Multi-Scene Videos[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 468-485.

[3] Yuan Z, Liu Y, Cao Y, et al. Mora: Enabling generalist video generation via a multi-agent framework[J]. arXiv preprint arXiv:2403.13248, 2024.

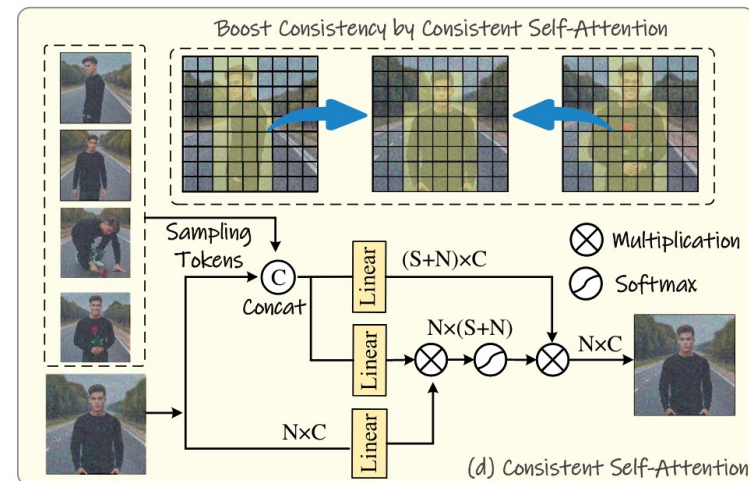
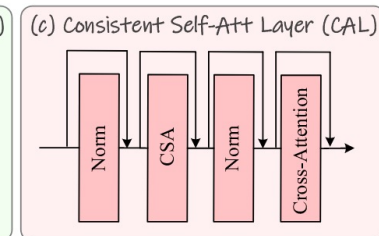
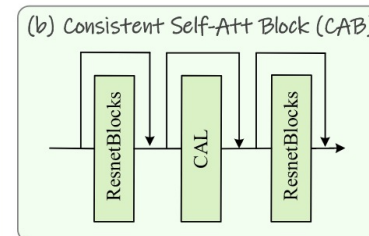
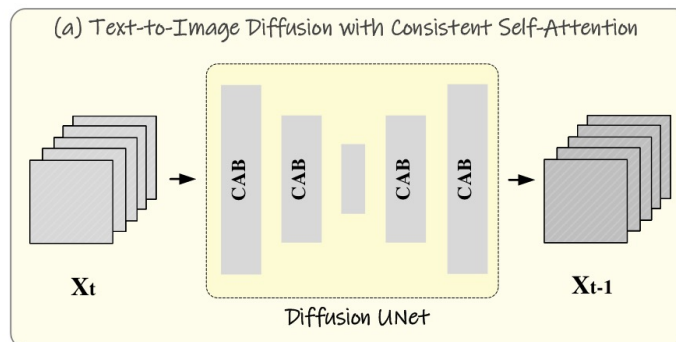
Related Work



7

➤ Keyframe-based Methods

- Leverage text-to-image models to generate one keyframe for each scene
- Connect every keyframe by adopting a pre-trained image-to-video model
- E.g., StoryDiffusion^[4]



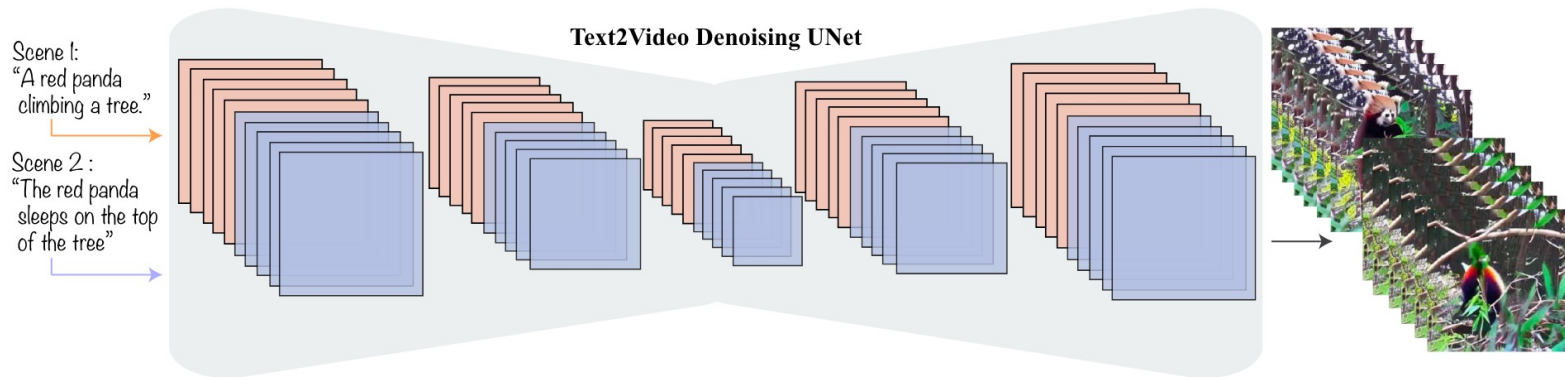
Related Work



8

➤ Finetuning-based Methods

- Inherit the conditional behavior from the pretrained model while enabling multiple conditioning inputs.
- E.g., TALC^[5]



➤ Limitations

- U-Net architecture with limited scalability
- Limited duration
- Inconsistent visual content

Outline



9

1

Motivation

2

Related Work

3

Methodology

4

Experiments

5

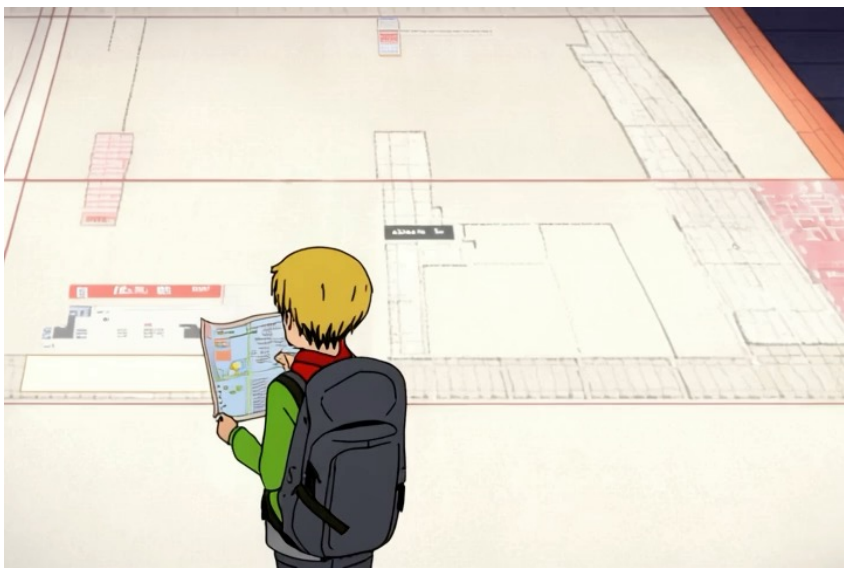
Conclusion

Model Capabilities



10

Fixed-Count Multi-Scene Generation



This video is generated in a single pass using three different text prompts, each guiding a 6-second scene, resulting in an 18-second multi-scene video.

Auto-Regressive Scene Extension

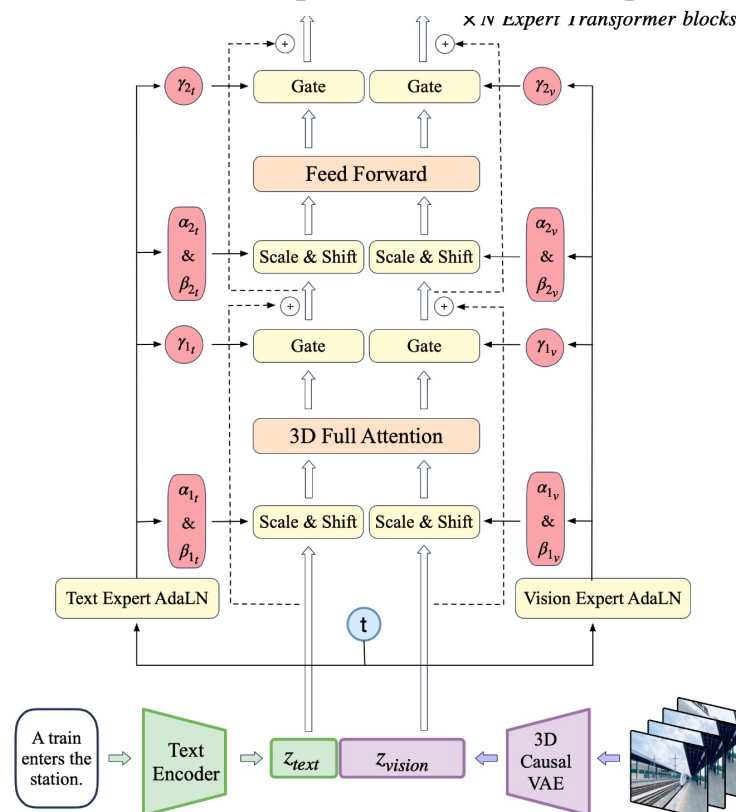


This video demonstrates auto-regressive scene extension, where the model generates the third 6-second scene conditioned on the first two 6-second scenes (12s in total) as context.

➤ Preliminary

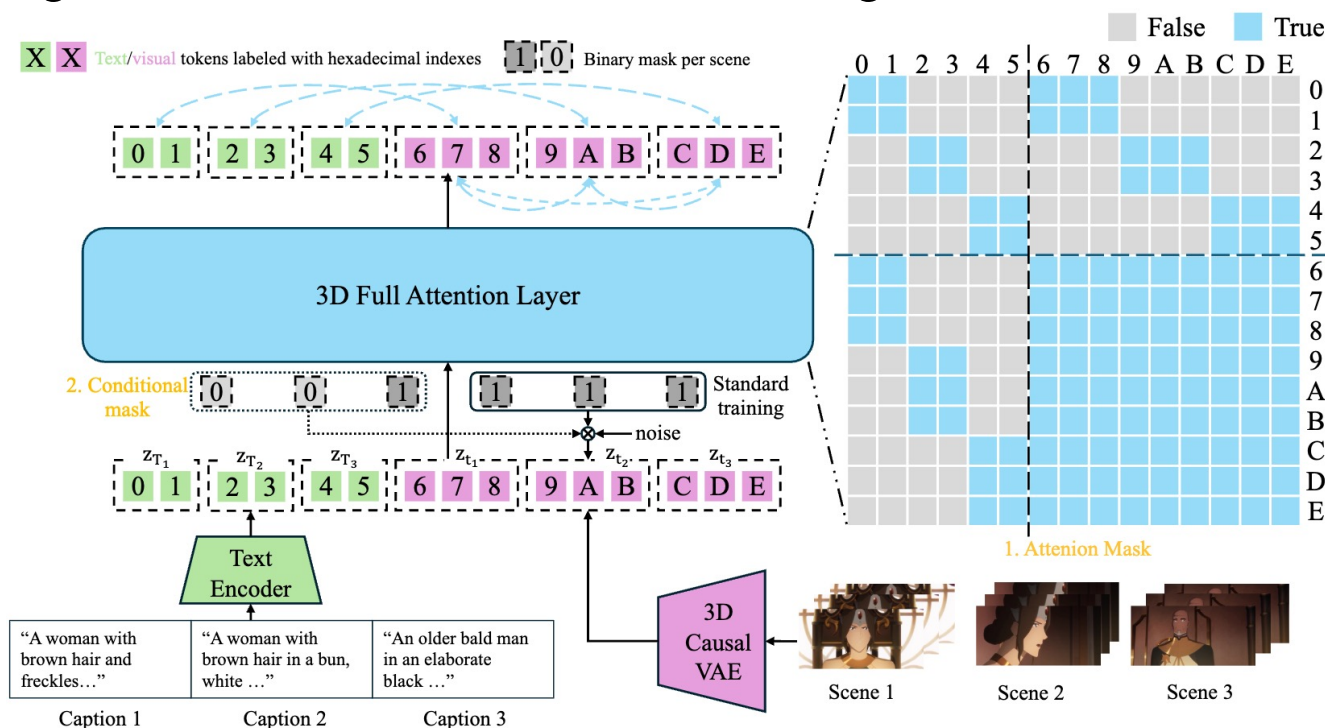
➤ Architecture: CogVideoX^[6]

➤ Objective: $L = \mathbb{E}_{z_V, z_T, \epsilon \sim \mathcal{N}(0,1), t} \left[\|v - v_\theta(z_t, t, z_T)\|_2^2 \right]$



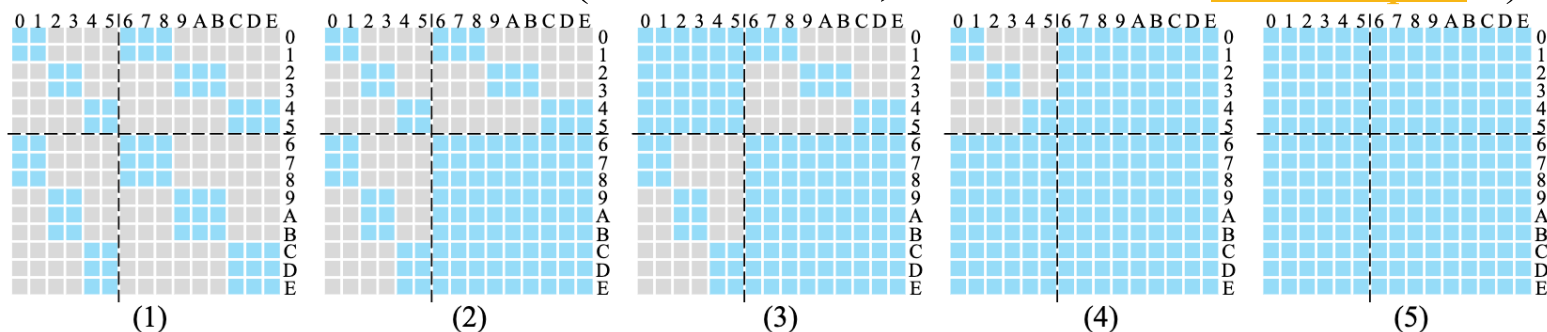
➤ Pipeline

- Core idea: **multi-prompt conditioning**
- Symmetric binary attention mask -> visual-textual alignment within each scene & intra-scene visual coherence
- Segment-level conditional mask -> auto-regressive scene extension



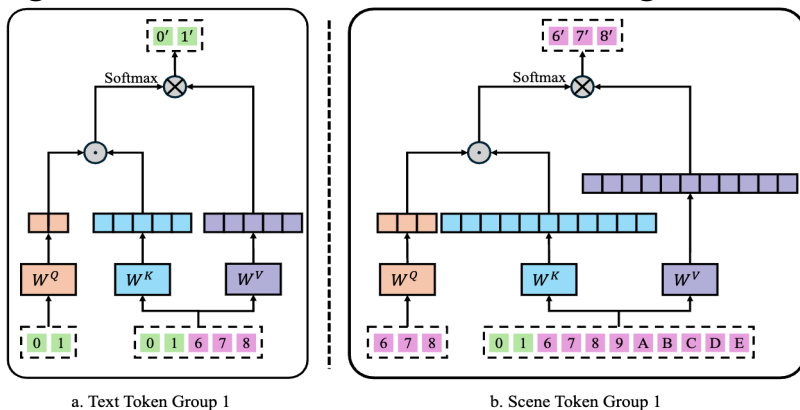
➤ Symmetric Binary Attention Mask

➤ Attention mask variants (**v2**: our choice, **v3**: concurrent [ShotAdapter](#)^[7])



➤ Grouped attention mechanism (**1 self attn with mask** \rightarrow **2n cross attn**)

➤ The total length of the concatenated token sequence is considerable, therefore, the memory coverage of the attention mask becomes significant



➤ Multi-stage Training

➤ Pre-training stage (10000 iterations)

- Pre-training on concatenated $n = 3$ single-scene clips without contextual relationships to adapt to longer sequences
- Select 1 million video samples from Panda70M^[8], sample $n = 3$ single-scene video clips from the dataset and concatenate them

➤ Supervised fine-tuning stage (10000 iterations)

- Fine-tuning on 5000 cartoons videos containing $n = 3$ consecutive scenes with contextual relationships, which are filtered by ViClip^[9] for sufficient inter-clip similarity
- Mixed training settings

Probability $p \rightarrow$ auto-regressive scene extension training $L = \mathbb{E}_{z_V, z_T, \epsilon \sim \mathcal{N}(0,1), t} \sum_{i=1}^n \left[m_{c_i} \cdot \|v - v_\theta(z_{t_i}, t, z_{T_i})\|_2^2 \right]$

Probability $1 - p \rightarrow$ normal training $L = \mathbb{E}_{z_T, z_V, \epsilon \sim \mathcal{N}(0,1), t} \sum_{i=1}^n \left[\|v - v_\theta(z_{t_i}, t, z_{T_i})\|_2^2 \right]$

Batch size: 8

Resolution/Frame: 480x720 / 49x3=147

Learning rate/ Probability p : 1e-5 / 0.5

[8] Chen T S, Siarohin A, Menapace W, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 13320-13331.

[9] Wang Y, He Y, Li Y, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation[J]. arXiv preprint arXiv:2307.06942, 2023.

➤ Multi-stage Training

➤ Pre-training stage (10000 iterations)

- Pre-training on concatenated $n = 3$ single-scene clips without contextual relationships to adapt to longer sequences
- Select 1 million video samples from Panda70M^[8], sample $n = 3$ single-scene video clips from the dataset and concatenate them

➤ Supervised fine-tuning stage (10000 iterations)



****Answering Style**:**

Answers should be comprehensive, conversational, and use complete sentences. The answer should be in English no matter what the user's input is. Provide context when necessary and maintain a certain tone. Begin directly without introductory phrases like "The image/video showcases" "The photo/video captures" "In the first/second video" and so on. For example, say "A woman is on a beach", instead of "A woman is depicted in the image". Vague expressions should be avoided in descriptions. Try to use more specific descriptions instead of terms like "video1" or "video2".

****Note**:**

When describing, first describe the character and scene, then describe the events that occurred in the video, as well as the actions of the characters.

****Output Format**:**

[the first video description]
[the second video description]
[the third video description]

****User Input**:**

Please detailedly describe each video in order and express the same elements in different videos in the same way. When describing the characters, it is necessary to give actor1, actor2, etc. and describe who actor1 and actor2 are.

[8] Chen T S, Siarohin A, Menapace W, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 13320-13331.

[9] Wang Y, He Y, Li Y, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation[J]. arXiv preprint arXiv:2307.06942, 2023.

Outline



16

1

Motivation

2

Related Work

3

Methodology

4

Experiments

5

Conclusion

➤ Evaluation

➤ Evaluation dataset

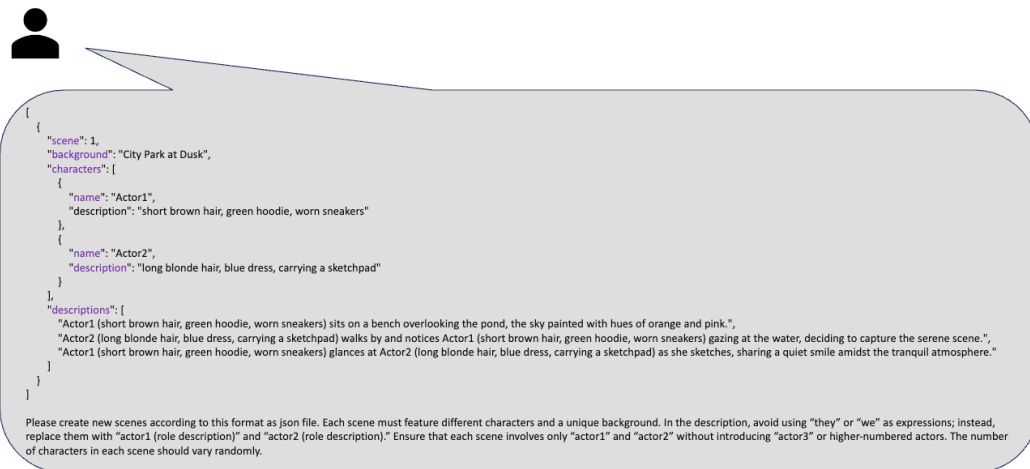


Figure 5. Structure of the evaluation dataset. It contains 50 scenes, each accompanied by three prompts featuring a variable number of characters. ChatGPT [26] was used to generate diverse prompts based on this structure, enabling comprehensive evaluation of video generation models in multi-scene scenarios.

➤ Evaluation metrics

- Visual/Semantic Consistency: ViClip^[9] model
- Sequence Consistency: Average of Visual and Semantic Consistency
- Video Quality: FVD^[10]

[9] Wang Y, He Y, Li Y, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation[J]. arXiv preprint arXiv:2307.06942, 2023.

[10] Ge S, Mahapatra A, Parmar G, et al. On the content bias in fr chet video distance[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 7277-7288.

Experiments



18

➤ Quantitative Results

➤ Performance gain

Mask²DiT outperforms CogVideoX in all aspects, **achieving higher visual, semantic, and sequence consistency**, while producing **more realistic** videos with **lower FVD**.

Model	Visual Con. (%)	Semantic Con. (%)	Sequence Con. (%)	FVD (↓)
CogVideoX-2B	55.01	22.64	38.82	835.35
Ours	70.95	23.94	47.45	720.01
CogVideoX-5B	43.82	20.70	32.26	613.47
Ours	89.21	20.81	55.01	607.64

➤ Comparison with SOTAs

Mask²DiT surpasses existing methods in multi-scene video generation with **the best visual quality** (lowest FVD), **strong visual and sequence consistency**, and **supports auto-regressive scene extension** while preserving coherence across scenes.

Capabilities	Method	Visual Con. (%)	Semantic Con. (%)	Sequence Con. (%)	FVD (↓)
Fixed-Scene Generation	CogVideoX	55.01	22.64	38.82	835.35
	StoryDiffusion + CogVideoX-I2V	69.06	25.59	47.32	905.69
	TALC	67.47	20.25	43.86	1516.59
	VideoStudio	61.28	22.64	41.96	1213.88
	Ours	70.95	23.94	47.45	720.01
AR Scene Extension	Ours	75.33	24.29	49.81	-

Experiments



19

➤ Quantitative Results

➤ User study

Aspect	Visual Consistency↑	Semantic Consistency↑	Video Quality↑	Overall↑
CogVideoX	8.96	12.12	9.23	9.09
StoryDiffusion + CogVideoX-I2V	29.85	28.79	29.23	28.79
TALC	13.43	13.64	12.31	12.12
Ours	47.76	45.45	49.23	50.00

Mask²DiT is **avored by real users**, winning top marks in visual and semantic consistency, video quality, and overall preference across the board.

➤ Ablations on “attention mask variants”

Method	Visual Con. (%)	Semantic Con. (%)	Sequence Con. (%)
v1	58.09	24.69	41.39
v2	68.45	23.82	46.14
v3	77.56	23.47	50.52
v4	93.43	20.73	57.08
v5	90.72	20.60	55.66

Combining qualitative and quantitative results, our attention mask (**v2**) achieves an **optimal balance** between **inter-scene visual consistency** and **intra-scene semantic consistency**.

Experiments



20

Quantitative Results

User study

Aspect	Visual Cons
CogVideoX	8.96
StoryDiffusion + CogVideoX-I2V	29.8
TALC	13.4
Ours	47.7

Mask²DiT is **avored by real users**, consistency, video quality, and over

Ablations on “attention mask”

Method	Visual Con. (%)
v1	58.09
v2	68.45
v3	77.56
v4	93.43
v5	90.72

Combining qualitative and quantitative results, we achieve an **optimal balance** between **inter-scene consistency**.

V1



V2



V3



V4



V5



Experiments



21

➤ Quantitative Results

➤ Ablations on “inter-clip similarity filtering threshold with ViClip^[9]”

Method	Visual Con. (%)	Semantic Con. (%)	Sequence Con. (%)
CogVideoX [38]	55.01	22.64	38.82
Ours + Dataset-0.0	63.94	24.31	44.13
Ours + Dataset-0.6	66.46	23.96	45.21
Ours + Dataset-0.7	67.71	23.98	45.85
Ours + Dataset-0.8	73.22	23.54	48.38

Improved dataset quality **enhances visual and sequential coherence**, but **reduces textual fidelity** by eliminating semantically rich but diverse samples.

➤ Ablations on “Training settings”

Method	Visual Con. (%)	Semantic Con. (%)	Sequence Con. (%)	Aesthetic Quality	Imaging Quality
Pre-training	49.04	26.36	37.70	-	-
SFT	73.22	23.54	48.38	57.44	72.37
Pre-training + SFT	73.73	23.73	48.73	57.68	73.20

Pre-training phase contributes to **enhanced visual quality** and **imaging quality** in the generated videos

Experiments

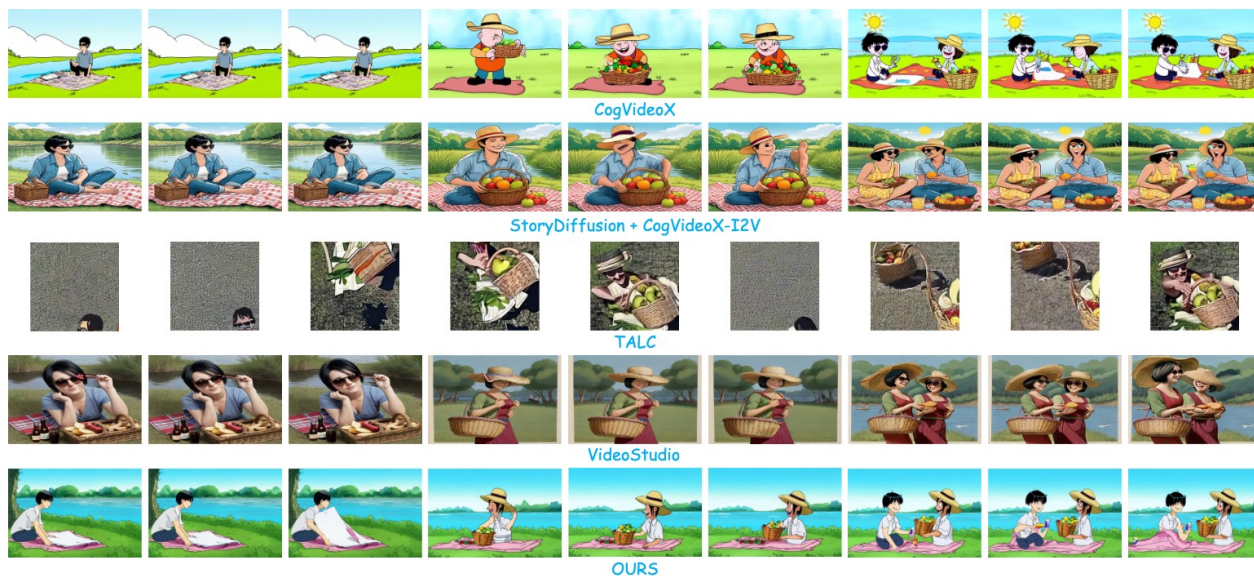


22

➤ Qualitative Results

Mask²DiT delivers significantly **better visual coherence** than SOTA baselines, demonstrating superior consistency in character appearance, background integrity, and overall style across multi-scene videos.

➤ Case 1



(a) "Actor1 (short black hair, wearing sunglasses, spreading out a picnic blanket) sets up by the riverside, watching the gentle flow of the water.", "Actor2 (wearing a straw hat, carrying a basket of fruit) arrives with a cheerful smile, placing the basket on the blanket.", "Actor1 (short black hair, wearing sunglasses, spreading out a picnic blanket) and Actor2 (wearing a straw hat, carrying a basket of fruit) share snacks, laughing as the sun shines brightly on the river."

Experiments

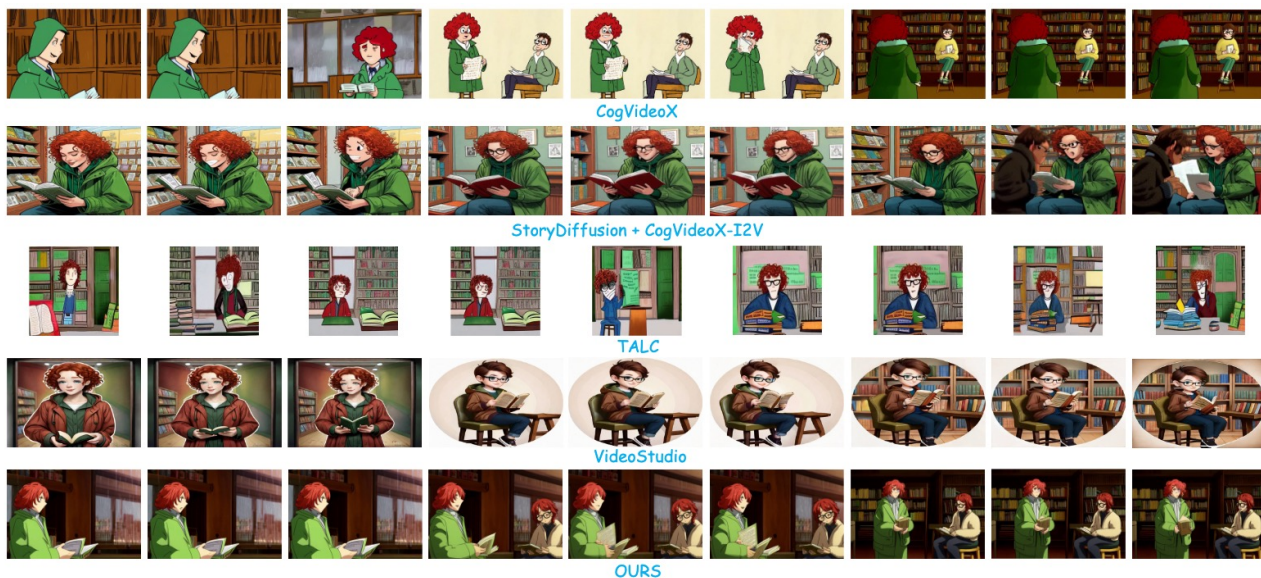


23

➤ Qualitative Results

Mask²DiT delivers significantly **better visual coherence** than SOTA baselines, demonstrating superior consistency in character appearance, background integrity, and overall style across multi-scene videos.

➤ Case 2



(b) "Actor1 (curly red hair, green raincoat, looking at a poetry book) flips through pages with a small smile, the sound of rain tapping against the store's windows.", "Actor2 (wearing glasses, oversized sweater, sitting on a stool reading) notices Actor1 (curly red hair, green raincoat, looking at a poetry book) and glances up with a gentle smile.", "Actor1 (curly red hair, green raincoat, looking at a poetry book) approaches Actor2 (wearing glasses, oversized sweater, sitting on a stool reading), asking a question about the book in her hands, leading to a quiet conversation amid the cozy shelves."

Experiments



24

➤ Qualitative Results

Mask²DiT delivers significantly **better visual coherence** than SOTA baselines, demonstrating superior consistency in character appearance, background integrity, and overall style across multi-scene videos.

➤ Case 3



(c) "Actor1 (brown hair tied back, wearing a green hiking vest, holding a camera) stands at the edge of the clearing, framing a shot of the misty forest.", "Actor2 (wearing a beige hat, blue flannel shirt, sitting on a log) watches Actor1 (brown hair tied back, wearing a green hiking vest, holding a camera) with a smile, appreciating the tranquility of the early morning.", "Actor1 (brown hair tied back, wearing a green hiking vest, holding a camera) lowers the camera and shares a quiet laugh with Actor2 (wearing a beige hat, blue flannel shirt, sitting on a log) as they both take in the serene beauty of the forest."

Experiments



25

➤ Qualitative Results for Auto-Regressive Scene Extension



Outline



26

1

Motivation

2

Related Work

3

Methodology

4

Experiments

5

Conclusion

Conclusion



27

➤ Summary

- Enabling the generation of a fixed number of visually consistent scenes
- Equipping the T2V model with auto-regressive scene extension capabilities to synthesize additional scenes beyond the initial fixed number
- Designing a mixed training task

➤ Future Directions

- Generalizing to real-world videos
- Adapting to variable-length scenes
- Scaling up total video durations
- Supporting diverse video resolutions
- Enhancing motion dynamics