

HiDream.ai

VideoStudio: Generating Consistent-Content and Multi-Scene Videos

Fuchen Long, Zhaofan Qiu, Ting Yao and Tao Mei

HiDream.ai Inc.



Video Diffusion Model

Input Prompt: A young man with blue hair is making cake



ModelScopeT2V



AnimateDiff



LAVIE



VideoCrafter2



CogVideoX

Modern approaches normally focus on video generation in a single-scene scenario

- Luo *et al.* VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. In CVPR, 2023.
- Guo *et al.* Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In ICLR, 2024.
- Wang *et al.* LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models. arXiv:2309.15103, 2023.
- Chen *et al.* VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. arXiv:2401.09047, 2024.
- Yang *et al.* CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. arXiv:2408.06072, 2024.

Multi-Scene Video Generation

Input Prompt: A young man with blue hair is making cake



Measures out ingredients



Pours the batter into a pan



Stirs the batter in the pan



Puts the cake on the table



Makes a phone call to invite friends



In outside of house to wait his friends

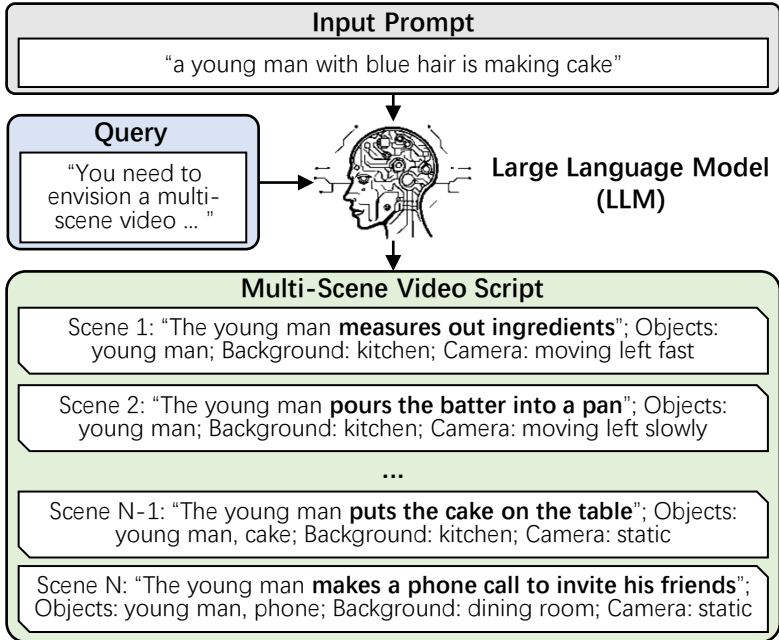
Challenge & Solution

- Challenge of Multi-Scene Video Generation
 - Establish the logic across different events
 - Guarantee the consistency of content (e.g., object or person)

- VideoStudio
 - Large Language Models (LLMs) for logic arrangement
 - Exploring reference image as the link for visual alignment

VideoStudio

(1) Multi-scene video script generation

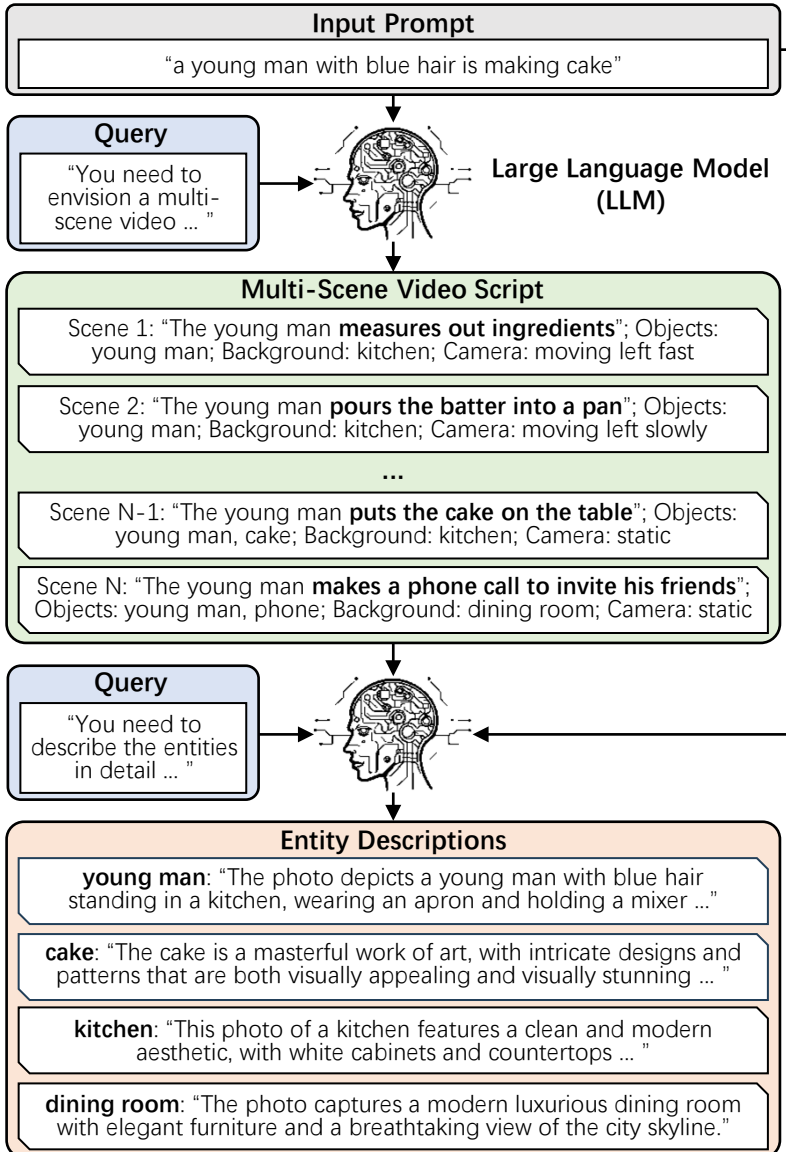


Video Script for each scene:

1. Scene prompt
2. Foreground entity
3. Background entity
4. Camera movement

VideoStudio

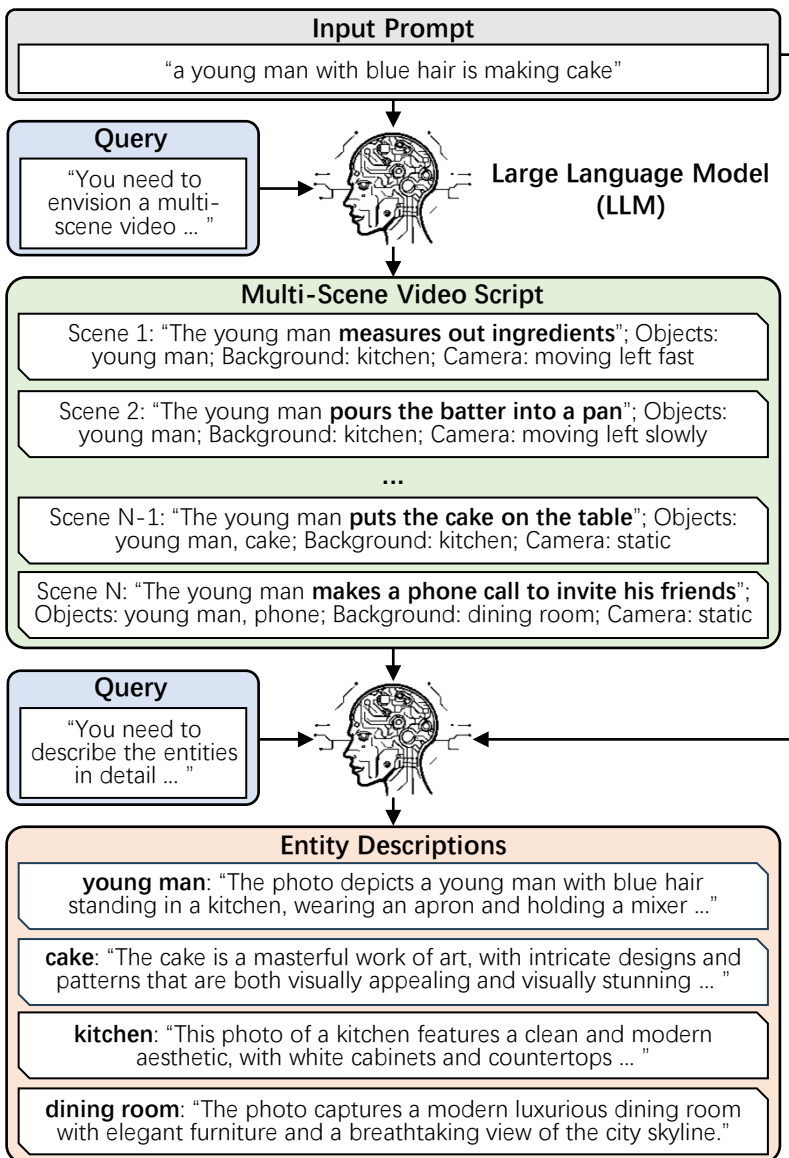
(1) Multi-scene video script generation



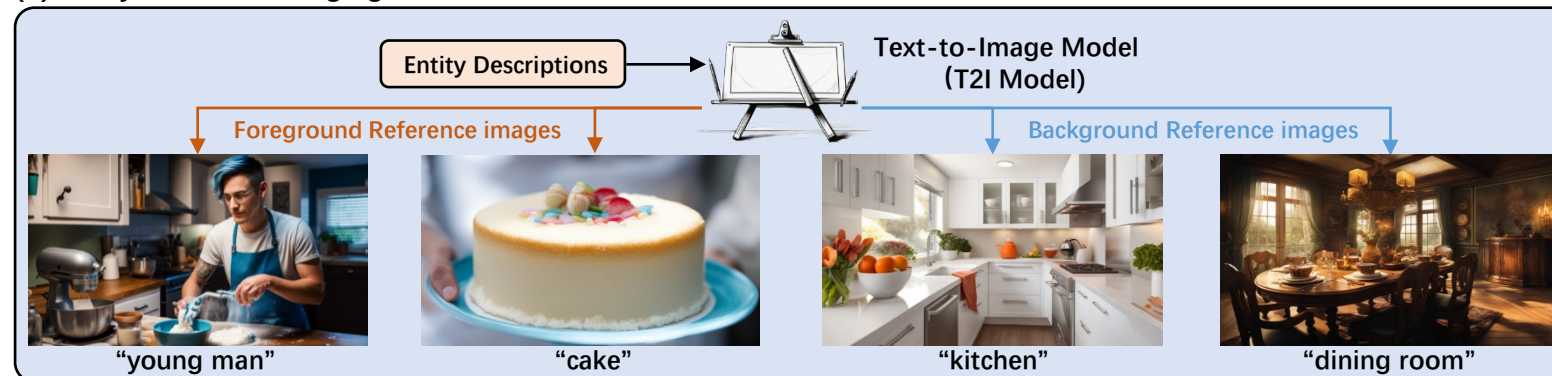
Detailed description for each common entity

VideoStudio

(1) Multi-scene video script generation

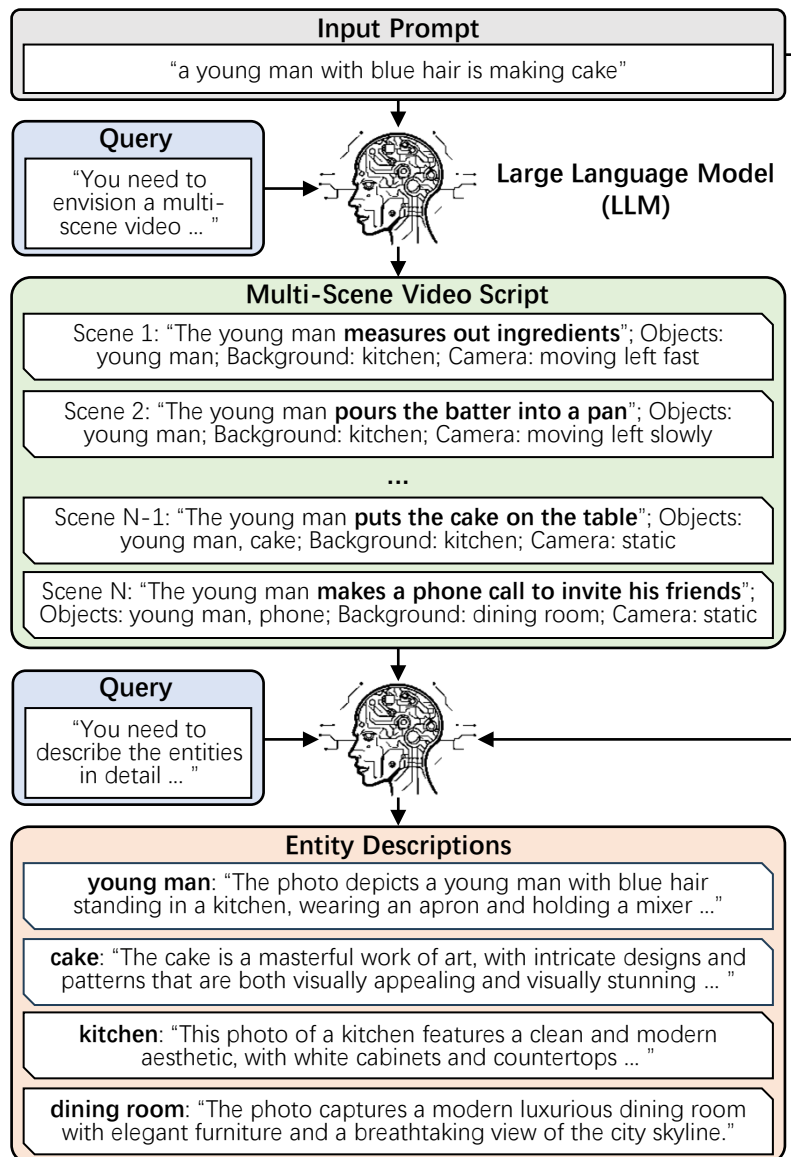


(2) Entity reference image generation

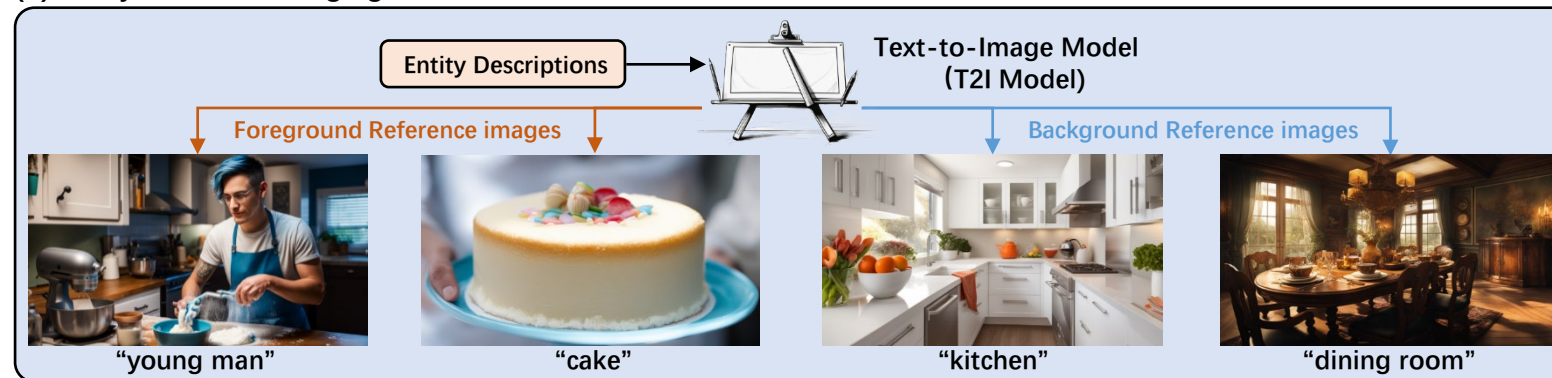


VideoStudio

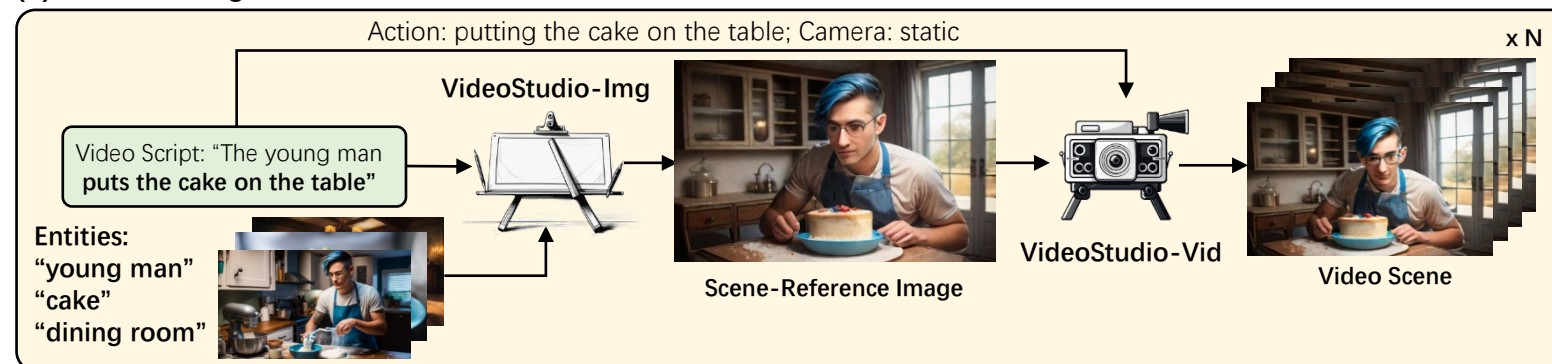
(1) Multi-scene video script generation



(2) Entity reference image generation

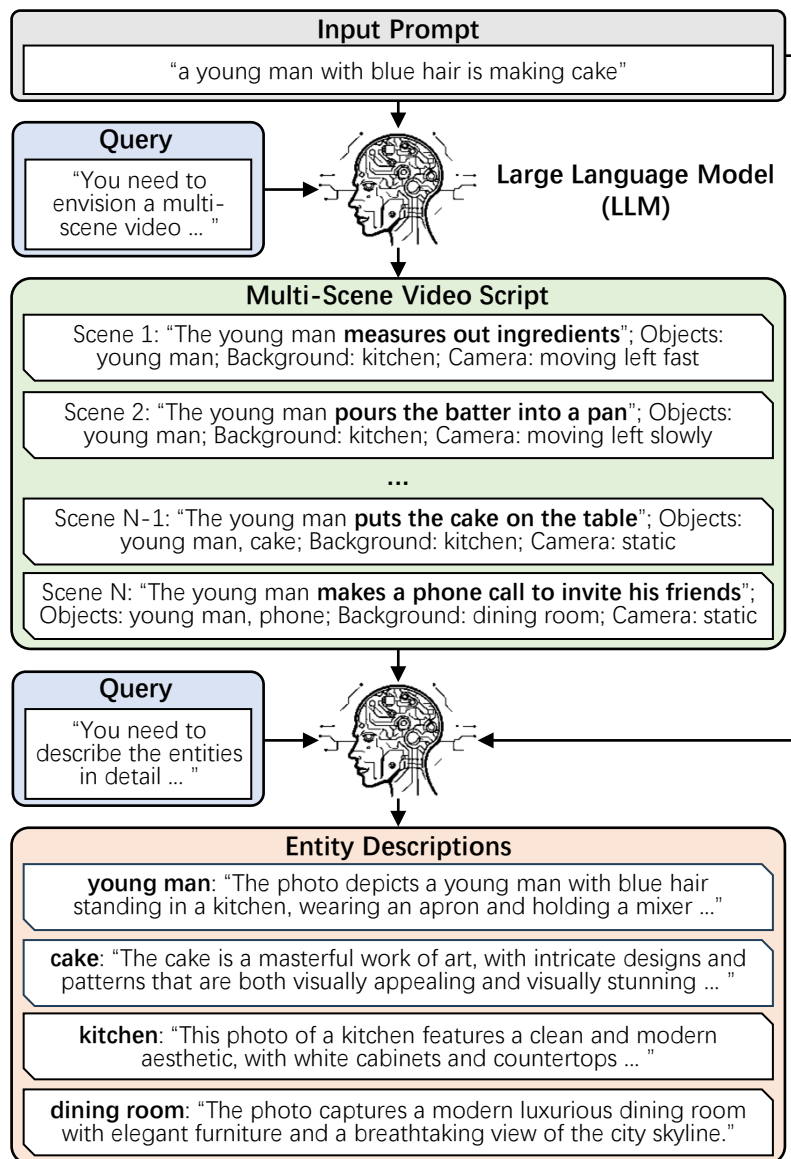


(3) Video scene generation

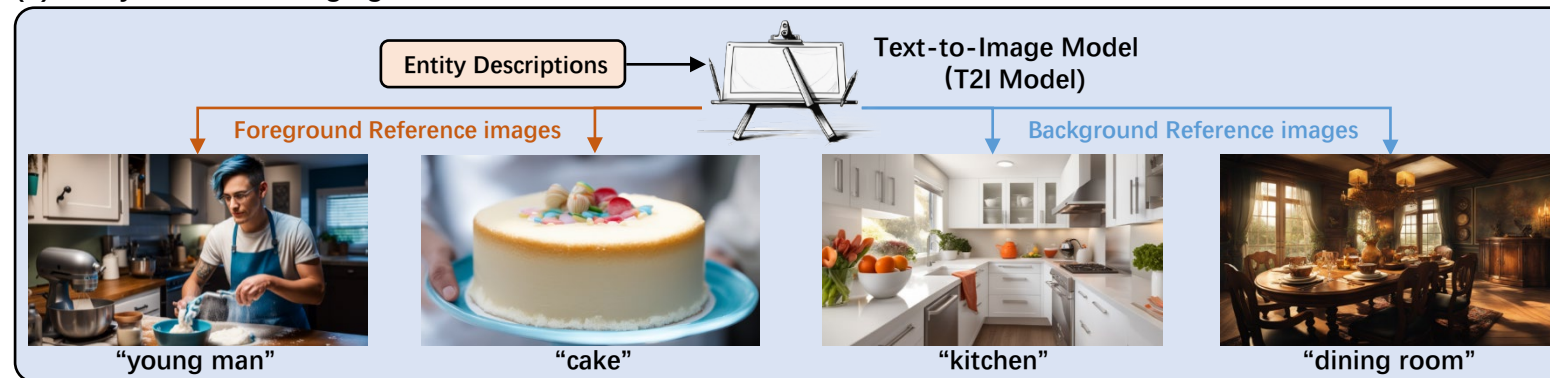


VideoStudio

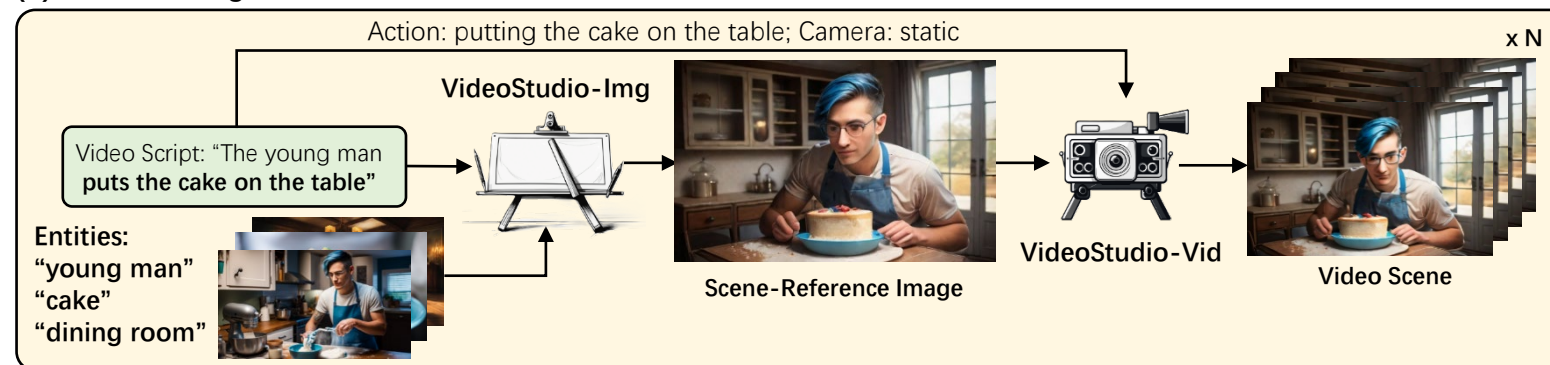
(1) Multi-scene video script generation



(2) Entity reference image generation



(3) Video scene generation

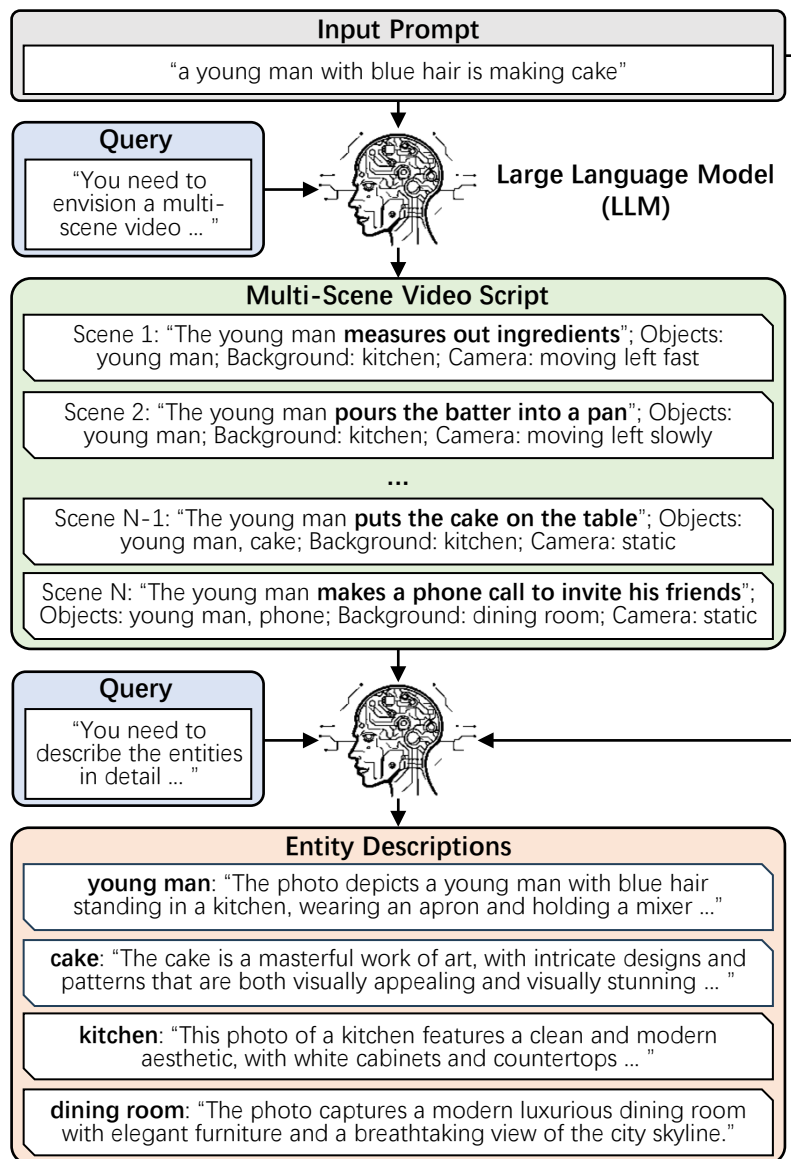


Scene 1: "The young man **measures out ingredients**"

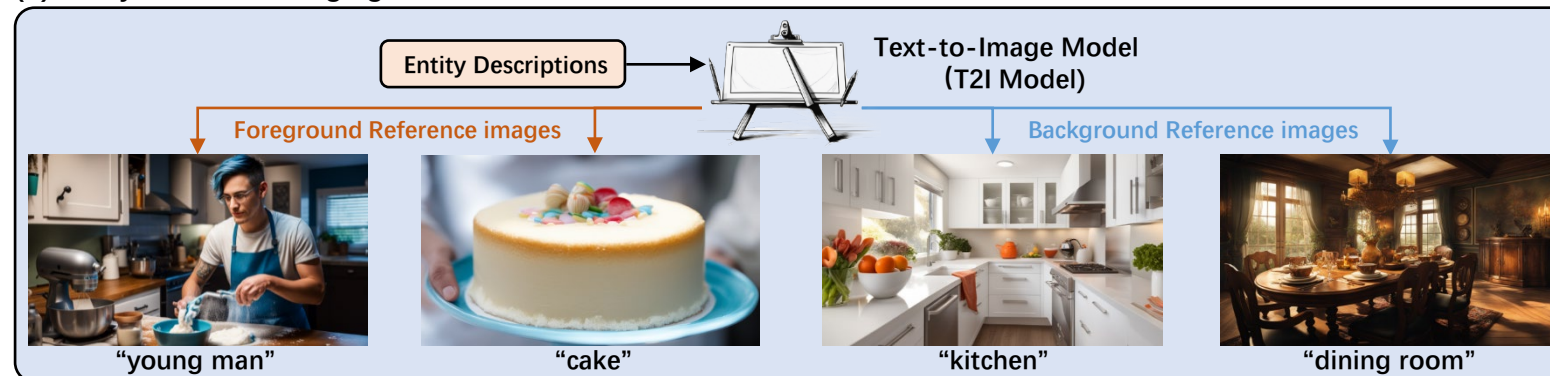


VideoStudio

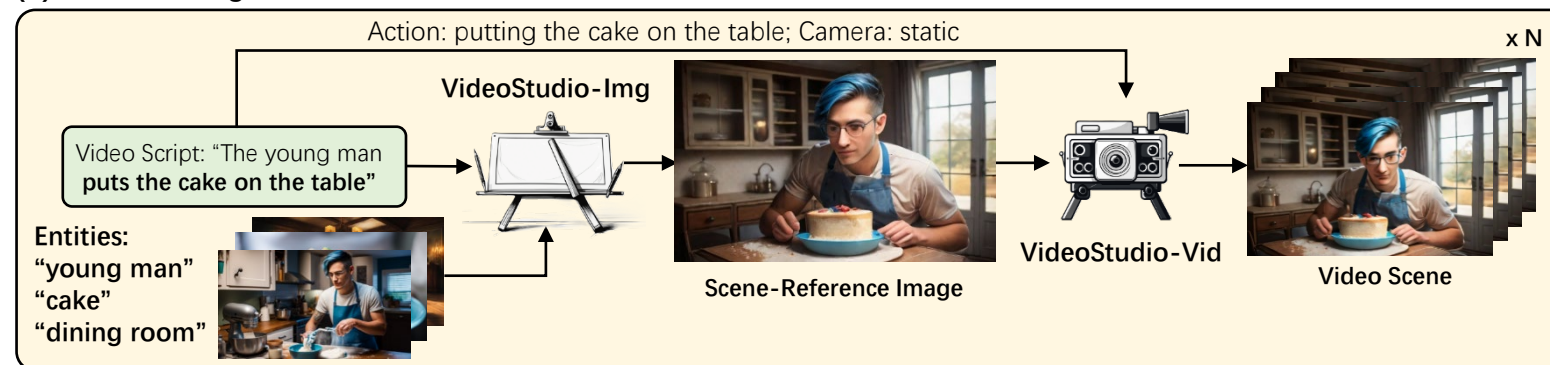
(1) Multi-scene video script generation



(2) Entity reference image generation



(3) Video scene generation

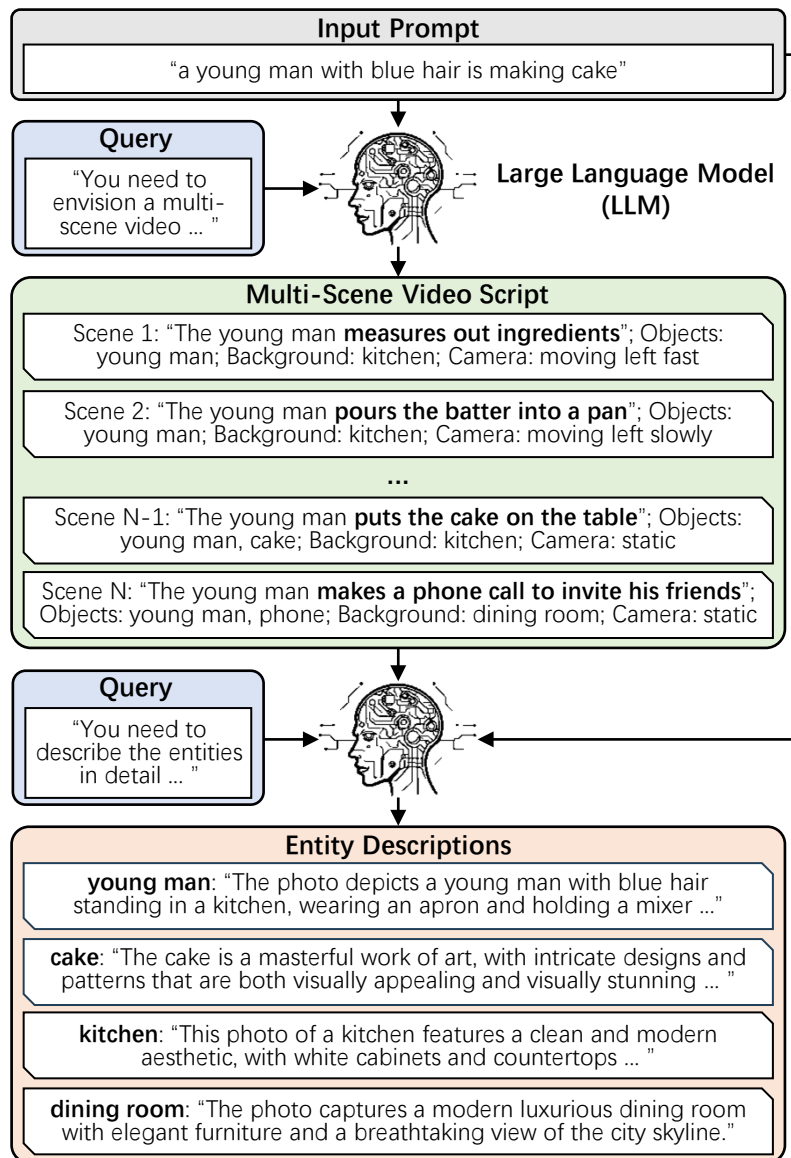


Scene 2: "The young man pours the batter into a pan"

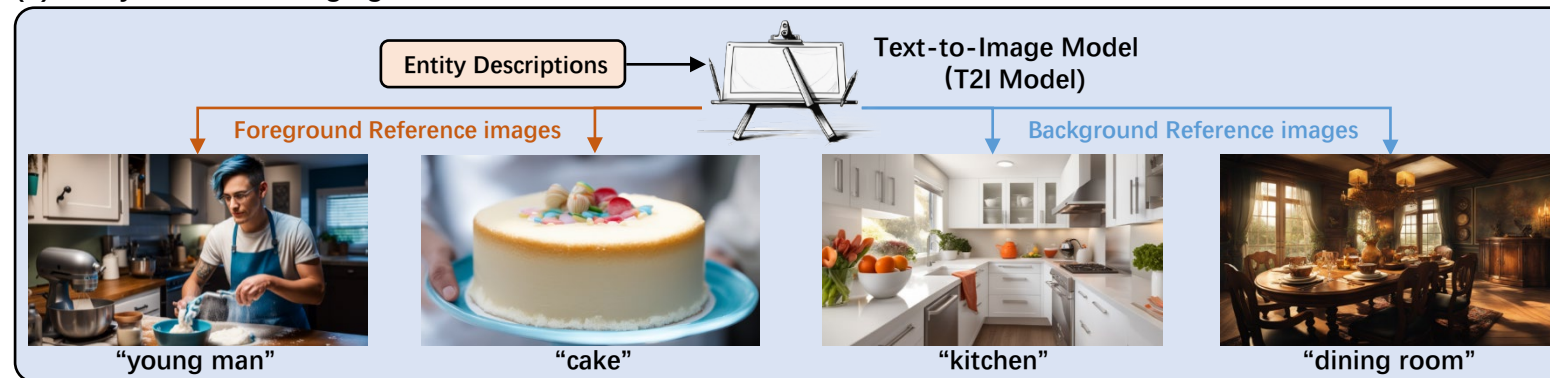


VideoStudio

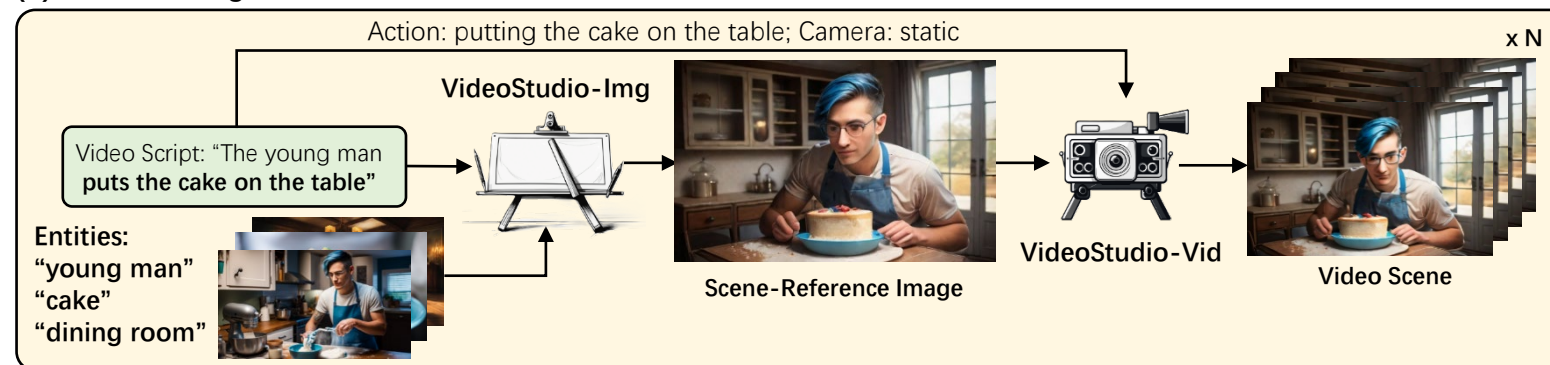
(1) Multi-scene video script generation



(2) Entity reference image generation



(3) Video scene generation

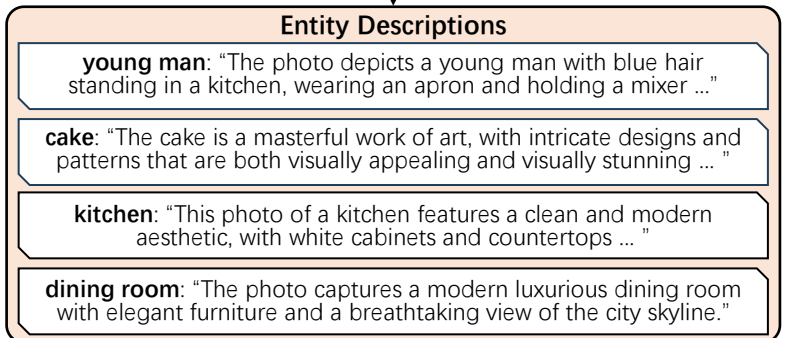
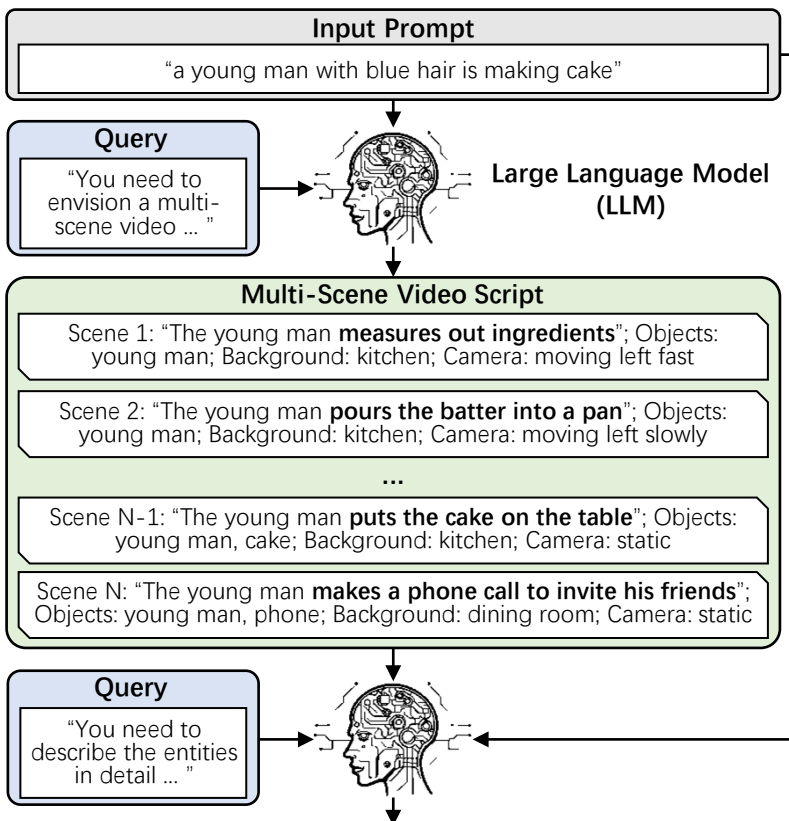


Scene 3: "The young man stirs the batter in the pan"

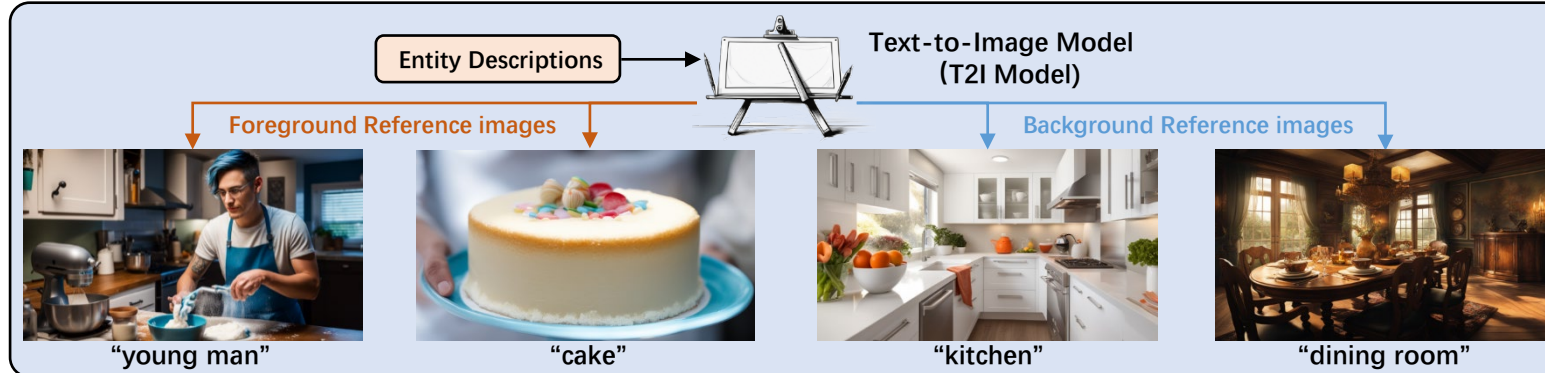


VideoStudio

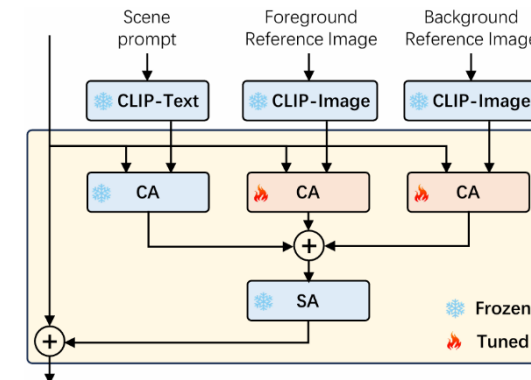
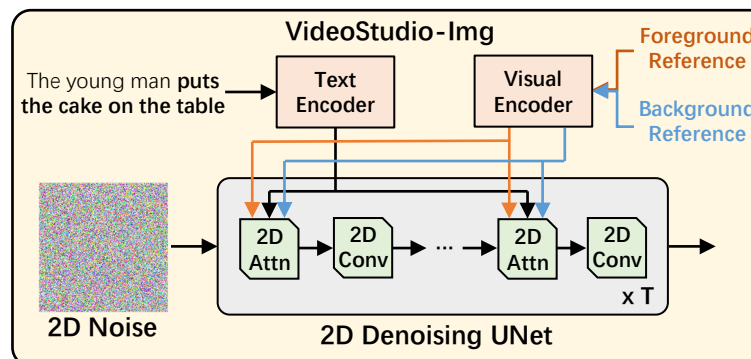
(1) Multi-scene video script generation



(2) Entity reference image generation

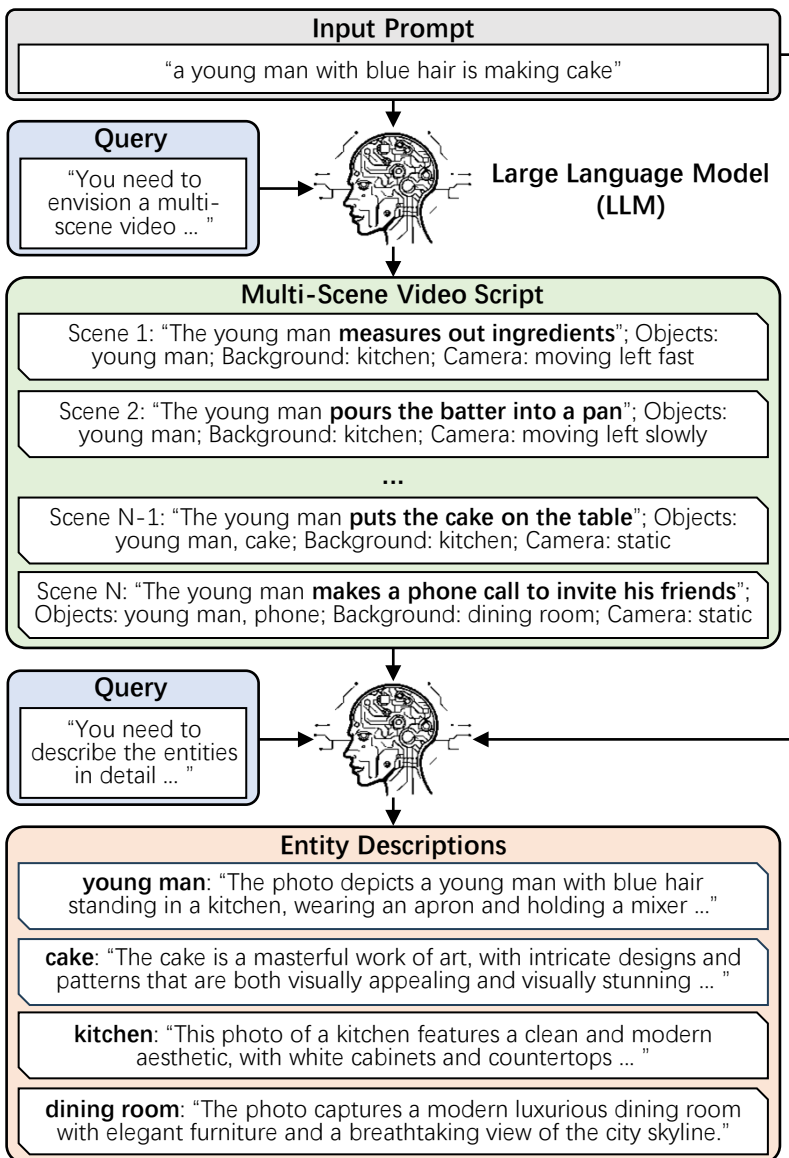


(3) Video scene generation

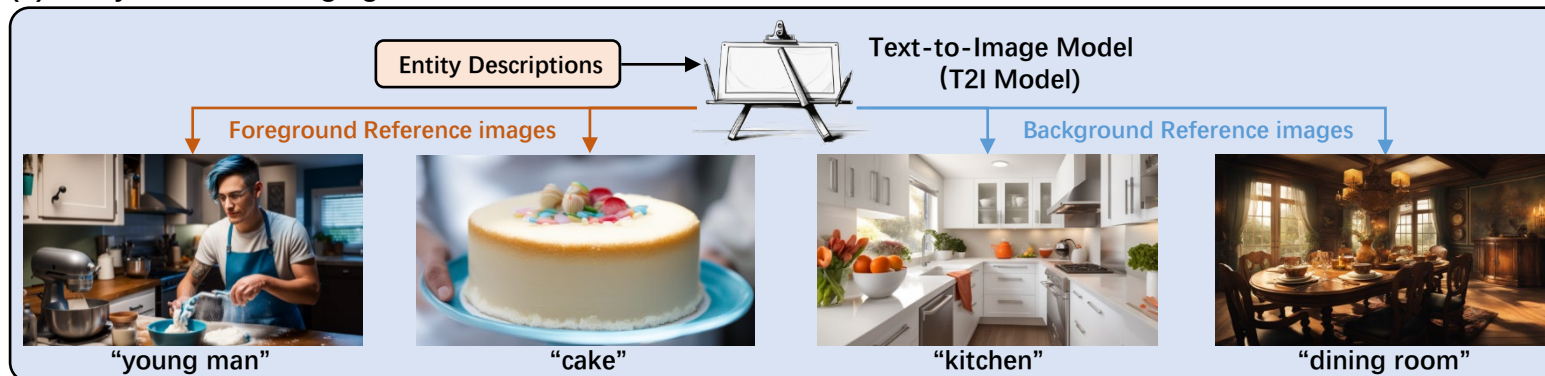


VideoStudio

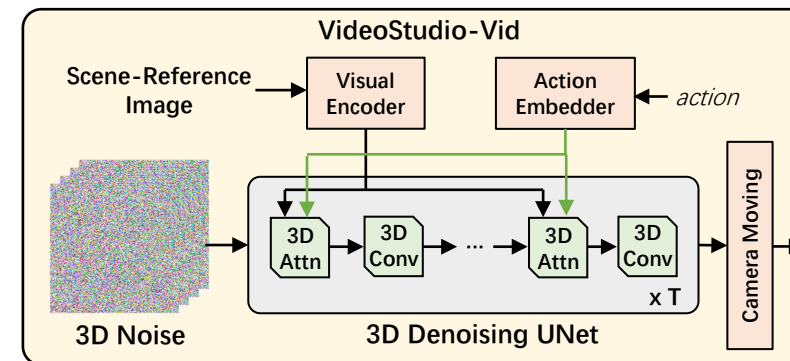
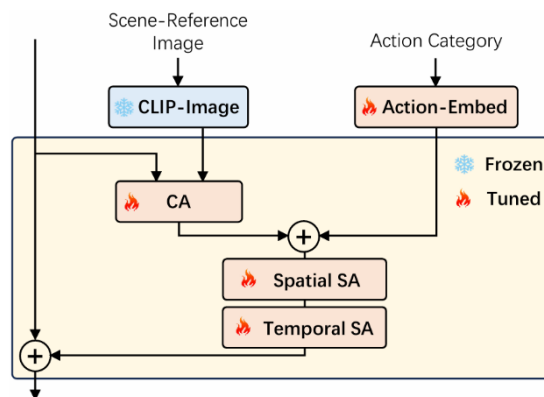
(1) Multi-scene video script generation



(2) Entity reference image generation



(3) Video scene generation



Experiments

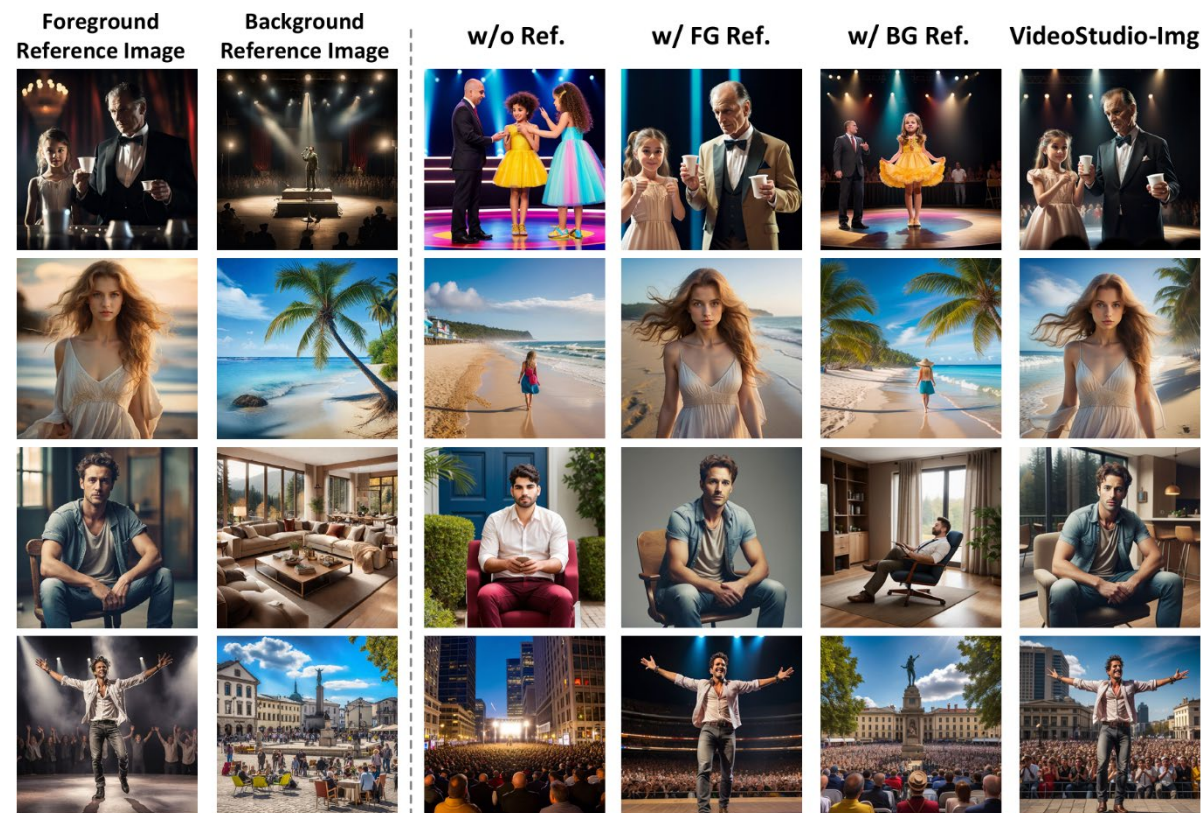
- Datasets
 - Training Data of VideoStudio-Img
 - LAION-2B: 2B image-text pairs
 - Training Data of VideoStudio-Vid
 - WebVid-10M: 10M video-text pairs
 - HD-VG-130M: Subset of 20M video-text pairs
 - Evaluation Data
 - WebVid-10M: 1,024 video-text pairs in validation set
 - MSR-VTT: 2,048 video clips with sentence annotation in validation set
 - ActivityNet Captions: 165 event captions as multi-scene prompts
 - Coref-SV: 100 prompts (10 episode captions \times 10 real-word entities)

Performance Comparison

- Evaluation on VideoStudio-Img

Performances on MSR-VTT

Input References		FG-SIM	BG-SIM	CLIPSIM
FG Ref.	BG Ref.			
w/o Ref.		0.5162	0.4131	0.3001
IP-Adapter		63		
✓		0.7116	0.4035	0.2910
	✓	0.5128	0.5059	0.2954
VideoStudio-Img				
✓		<u>0.7919</u>	0.4393	0.2982
	✓	0.5362	<u>0.5742</u>	<u>0.3002</u>
✓	✓	0.8102	0.5861	0.3023



VideoStudio-Img well aligns visual contents in the foreground and background reference images

Performance Comparison

- Evaluation on VideoStudio-Vid

Performances on WebVid-10M

Approach	FVD (\downarrow)	Frame Consis. (\uparrow)
RF+VideoCrafter ⁵	293.3	97.9
RF+I2VGen-XL ⁶⁸	254.9	97.6
RF+VideoComposer ⁵⁶	231.0	95.9
RF+DynamiCrafter ⁵⁹	176.8	97.5
RF+SVD ²	153.0	98.7
RF+VideoStudio-Vid ⁻	157.3	98.5
RF+VideoStudio-Vid	116.5	98.8

Performances on MSR-VTT

Approach	RF	FID (\downarrow)	FVD (\downarrow)
CogVideo ¹⁷		23.6	-
MagicVideo ⁷⁰		-	998
Make-A-Video ⁴⁶		13.2	-
VideoComposer ⁵⁶		-	580
VideoDirectorGPT ⁺ ²⁶		12.2	550
ModelScopeT2V ⁵⁴		11.1	550
SD+VideoStudio-Vid		11.9	381
RF+VideoCrafter ⁵	✓	45.0	339
RF+I2VGen-XL ⁶⁸	✓	37.4	264
RF+VideoComposer ⁵⁶	✓	31.3	208
RF+DynamiCrafter ⁵⁹	✓	26.1	196
RF+SVD ²	✓	15.3	172
RF+VideoStudio-Vid	✓	10.8	133

Performance Comparison

- Evaluation on Multi-Scene Video Generation

Performances on ActivityNet Captions

Approach	FID (↓)	FVD (↓)	Scene Consis. (↑)
ModelScopeT2V ⁵⁴	18.1	980	46.0
VideoDirectorGPT ²⁶	16.5	805	64.8
VideoStudio w/o Ref.	17.3	624	50.8
VideoStudio	13.2	395	75.1

Performances on Coref-SV

Approach	CLIPSIM (↑)	Scene Consis. (↑)
ModelScopeT2V ⁵⁴	0.3021	37.9
VideoDirectorGPT ²⁶	-	42.8
VideoStudio w/o Ref.	0.3103	40.9
VideoStudio	0.3304	77.3

VideoStudio achieves the best visual consistency across scenes

Multi-Scene Video Examples

• Coref-SV

Input prompt:

- There is a house and many trees
- Cat puts cherry on a pie. Cat is done with the pie.
- Cat puts pie on the table. Cat looks very happy. There are bread, a book, apples, and a pie on the table.
- Cat tastes pie and Cat thinks it is delicious. Cat turns over the page.
- Cat marvels at the picture on the book. Cat eats a piece of pie.



Input prompt:

- Mouse is looking for something in Mouse's library.
- Mouse is standing on the ladder and Mouse is finding something on the bookshelf.
- Mouse found the book. Mouse climbs down a ladder.
- Mouse looks at the book and questions himself.
- Mouse came up with an idea and Mouse decides to make something.



Input prompt:

- Teddy-Bear is reading a book, turning the page by his right paw.
- In the story, Teddy-Bear wears an armor, holding a sword and riding on a white horse.
- There are trees outside the window. Teddy-Bear cheers by raising his both paws
- The clock is running fast. Teddy-Bear is reading a book.



Input prompt:

- There is a mountain covered with snow and trees on it. Mouse is reading a book.
- Mouse is holding a book and makes a happy face.
- Mouse looks happy and talks
- Mouse is holding a flower by her right paw.
- Mouse is smiling and talking while holding a flower on her right paw.
- Mouse is ripping a petal from the flower.
- Mouse is pulling petals off the flower.



Multi-Scene Video Examples

- Real images as entity reference image

Input prompt:

- The cat lies in the room
- The cat lies in the driving car
- The cat plays in the flowers

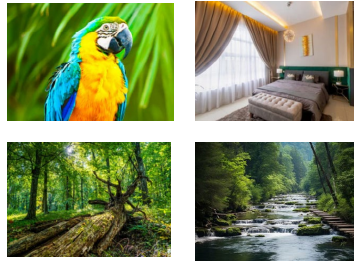
+



Input prompt:

- The parrot stands in the bedroom
- The parrot stands in the forest
- The parrot stands in front of the river

+



Input prompt:

- The motorcyclist stays in the town
- The motorcyclist is riding on the road under the sunset
- The motorcyclist is ridding on the moon

+



Thanks!

longfc.ustc@gmail.com

Model & Code: <https://github.com/FuchenUSTC/VideoStudio>

Project Page: <https://vidstudio.github.io/>

