

Lecture 3

Data generation for multimodal LLMs

Zinkovich Viktoriia

Special thanks to **Anton Konushin** for the slides and lecture content



Introduction: Course Plan



Day 1
Image modality

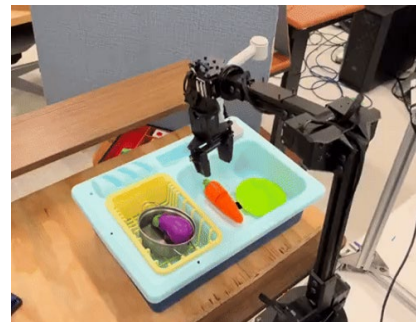
Day 3
Data generation
in MLLMs



Day 5
3D models



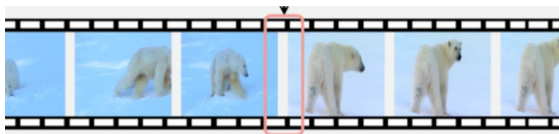
Day 2
Video modality



Day 4
Action modality

Recap: Lecture #2

- 1 Discussed **popular benchmarks** for video modality (PerceptionTest, Video-MMMU, MLVU) → long context testing

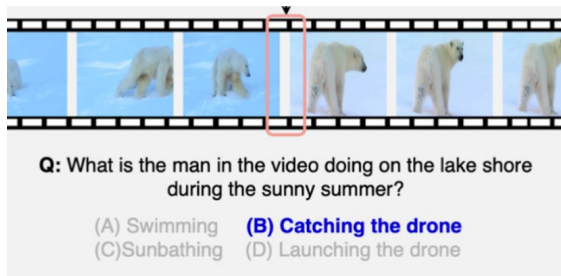


Q: What is the man in the video doing on the lake shore during the sunny summer?

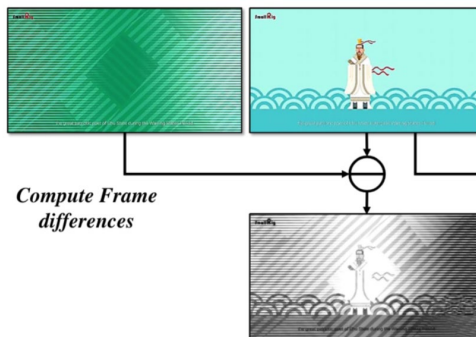
- (A) Swimming **(B) Catching the drone**
(C) Sunbathing (D) Launching the drone

Recap: Lecture #2

1 Discussed **popular benchmarks** for video modality (PerceptionTest, Video-MMMU, MLVU) → long context testing

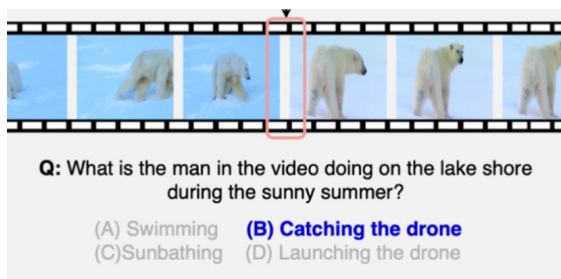


2 Described **several SOTA LLM modes** for video processing, including LLaVA-OneVision, VideoLLaMA3, AdaReTake

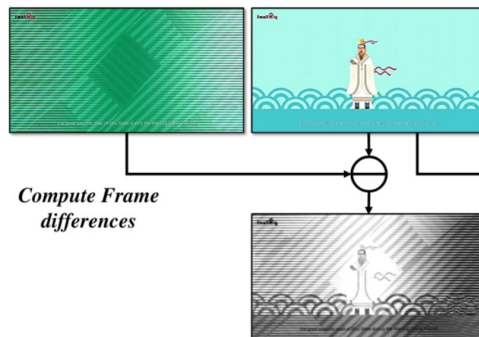


Recap: Lecture #2

1 Discussed **popular benchmarks** for video modality (PerceptionTest, Video-MMMU, MLVU) → long context testing



2 Described **several SOTA LLM modes** for video processing, including LLaVA-OneVision, VideoLLaMA3, AdaReTake



3 Analyzed models for **any - to - any modality**, that utilize next-prediction token paradigm



Lecture Plan

- 1 Evolution steps of **Generative** AI
- 2 How **LLM / MLLMs** can help us to improve GenAI?



1

Introduction

Steps of GenAI Evolution

What lies in the power of
GenAI?



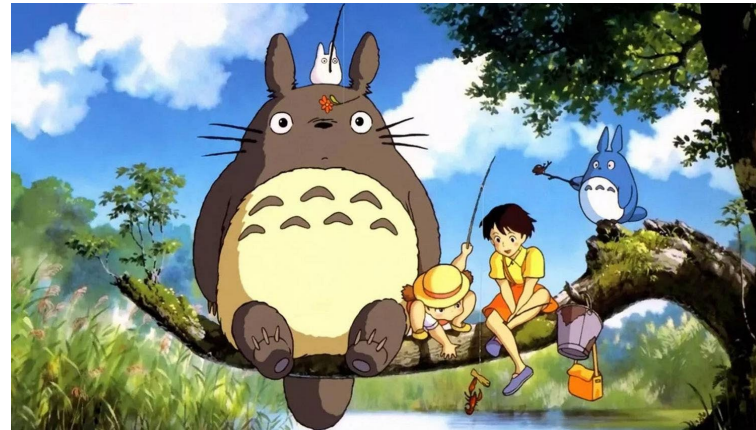
Introduction: GenAI

New generation method from the OpenAI in **Ghibli style** (1 M users in 1 hour)



Introduction: GenAI

Ghibli style refers to the distinctive visual aesthetic developed by Studio Ghibli, a renowned Japanese animation studio — **Hayao Miyazaki**, 1985



Introduction: GenAI

Generation is **not constant** —latent representation problems



Introduction: GenAI

What lies in the **power of Generative AI** ?

Introduction: GenAI

What lies in the **power of Generative AI** ?

significant generative pre-training → 1) synthetic datasets + 2) GPU parallelization



NO, you cannot understand the meaning of a text without explicitly evaluating its linguistic constituents and defining grammar rules!



haha gpus go brrrrrrrr



Introduction: GenAI Evolution

Image generation: from fixed data distribution to arbitrary guided generation



StyleGAN, 2019

Introduction: GenAI Evolution

Image generation: from fixed data distribution to arbitrary guided generation



StyleGAN, 2019



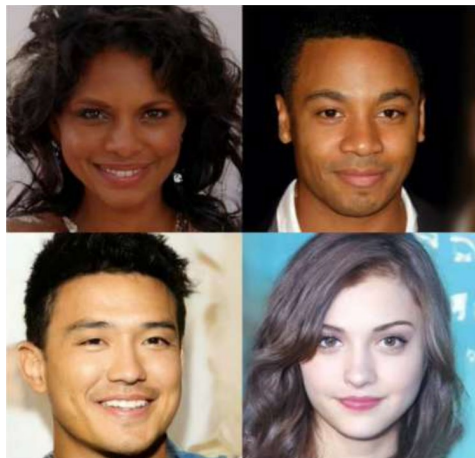
DDPM, 2020

Introduction: GenAI Evolution

Image generation: from fixed data distribution to arbitrary guided generation



StyleGAN, 2019



DDPM, 2020



Guided Diffusion, 2021

Introduction: GenAI Evolution

Single domain generation → open domain → interactive / intervalled generation

VAE (Dec 2013), GAN (Jun 2014), Normalizing flow (May 2015), DDPM (Dec 2020) ...

DALL-E (Feb 2021), LDM (Dec 2021), DALL-E 2 (Apr 2022), Imagen (May 2022) ...

DALL-E 3 (Sept 2023), DreamLLM (Sept 2023), mini DALL-E 3 (Oct 2023), CoDi-2 (Nov 2023) ...

2013 - 2020



GAN on MNIST



GAN on CIFAR-10



DDPM on LSUN



DDPM on CelebA-HQ

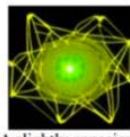
2021 - 2022



A painting of a squirrel eating a burger (LDM)



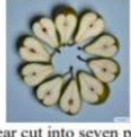
An animal half mouse half octopus (LDM)



A slightly conscious neural network (LDM)



A wine glass on top of a dog (Imagen)



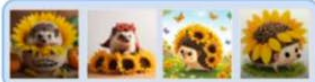
A pear cut into seven pieces arranged in a ring (Imagen)



A wine glass on top of a dog (Imagen)

2023 -

My 5 year-old keeps talking about a "super-dupersunflower hedgehog" -- what does it look like?



Could you design some stickers for it?



Interactive and interleaved generation of DALL-E 3

Introduction: Parameter Growth, T2I

Stable Diffusion

~1B parameters

Based on 3D U-Net with
transformer units



Introduction: Parameter Growth, T2I

Stable Diffusion

~1B parameters

Based on 3D U-Net with
transformer units



Stable Diffusion 3

8B parameters

Transformers



Introduction: Parameter Growth, T2I

Stable Diffusion

~1B parameters

Based on 3D U-Net with
transformer units



Stable Diffusion 3

8B parameters

Transformers



FLUX

12B parameters

Transformers



Introduction: Parameter Growth, T2V

VideoCrafter2 [2024-01]

1.8B parameters

Based on 3D U-Net with
transformer units



Introduction: Parameter Growth, T2V

VideoCrafter2 [2024-01]

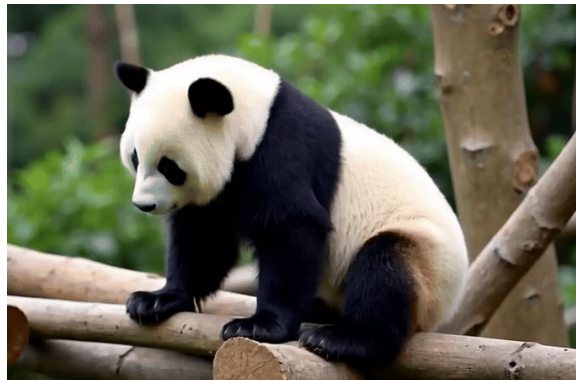
1.8B parameters

Based on 3D U-Net with
transformer units

Open -Sora [2024 -03]

5.8B parameters

Transformers



Introduction: Parameter Growth, T2V

VideoCrafter2 [2024-01]

1.8B parameters

Based on 3D U-Net with
transformer units

Open -Sora [2024 -03]

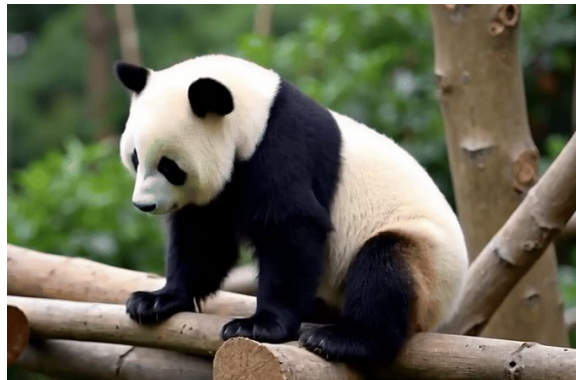
5.8B parameters

Transformers

CogVideo [2024 -09]

5B parameters

Transformers



MLLM/ LLM Integration

Large-scale retraining on poorly labeled data helped models in all modalities, but mostly different types of models **developed independently**

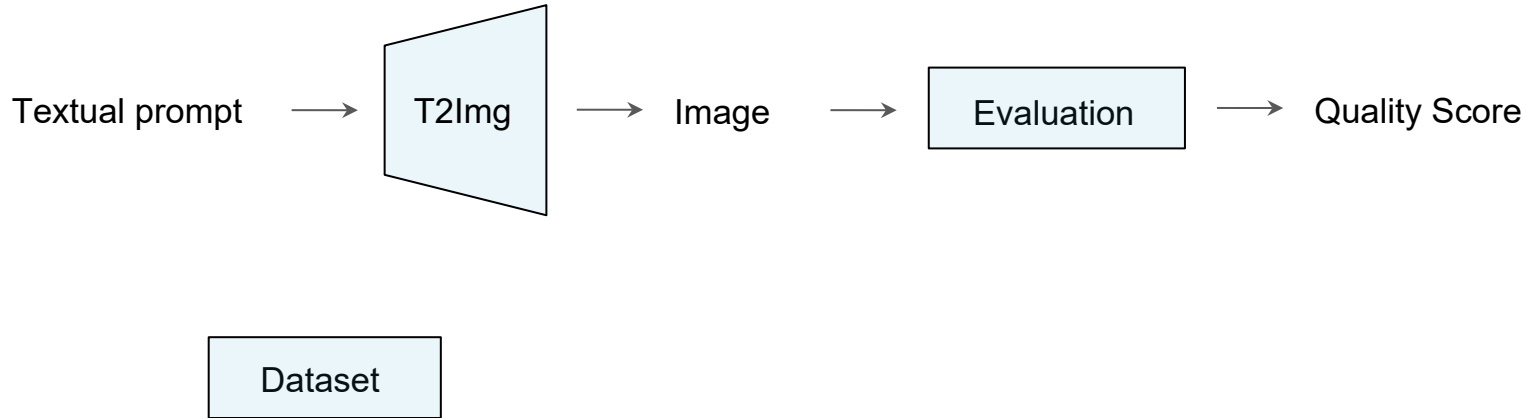


MLLM/ LLM Integration

What is **missing** in the scheme?

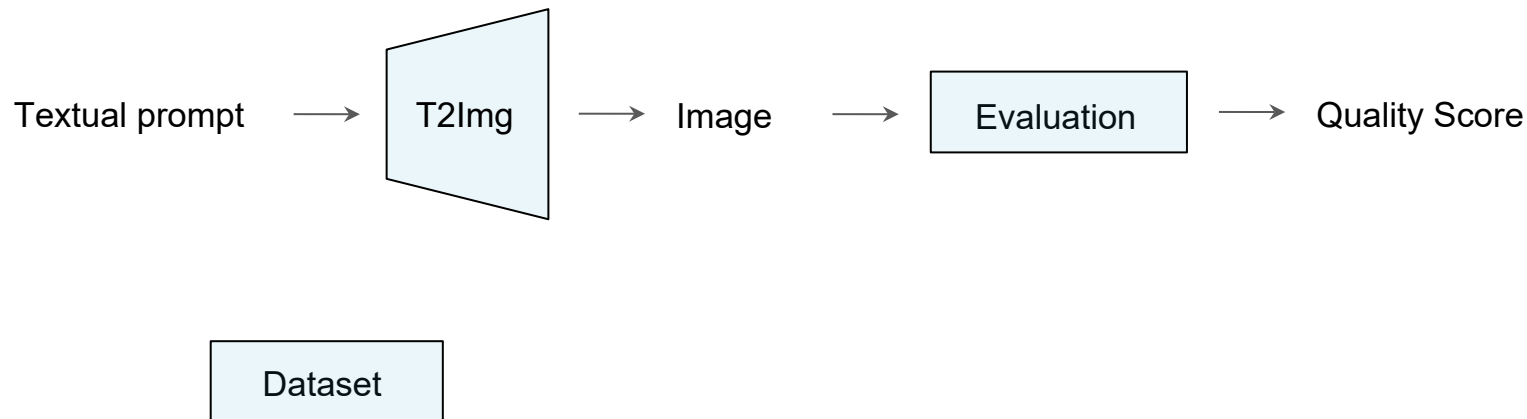


MLLM/ LLM Integration



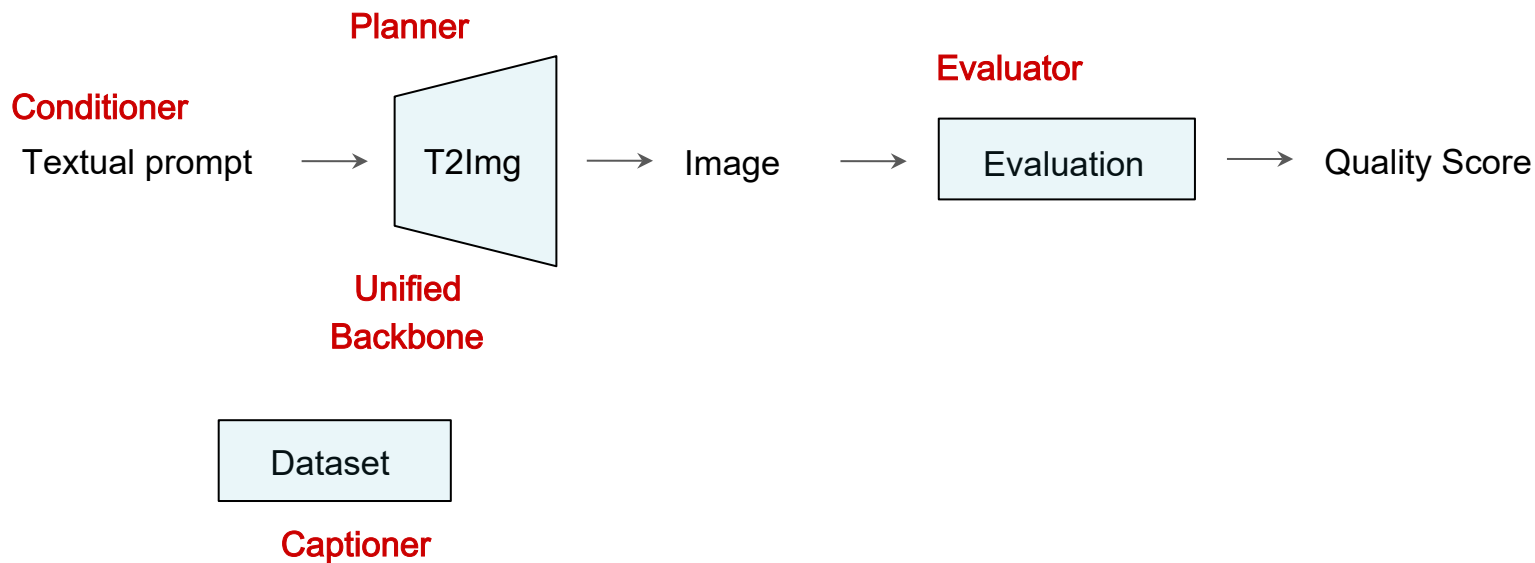
MLLM/ LLM Integration

How **MLLM / LLM can be integrated** in the scheme below to enhance it?



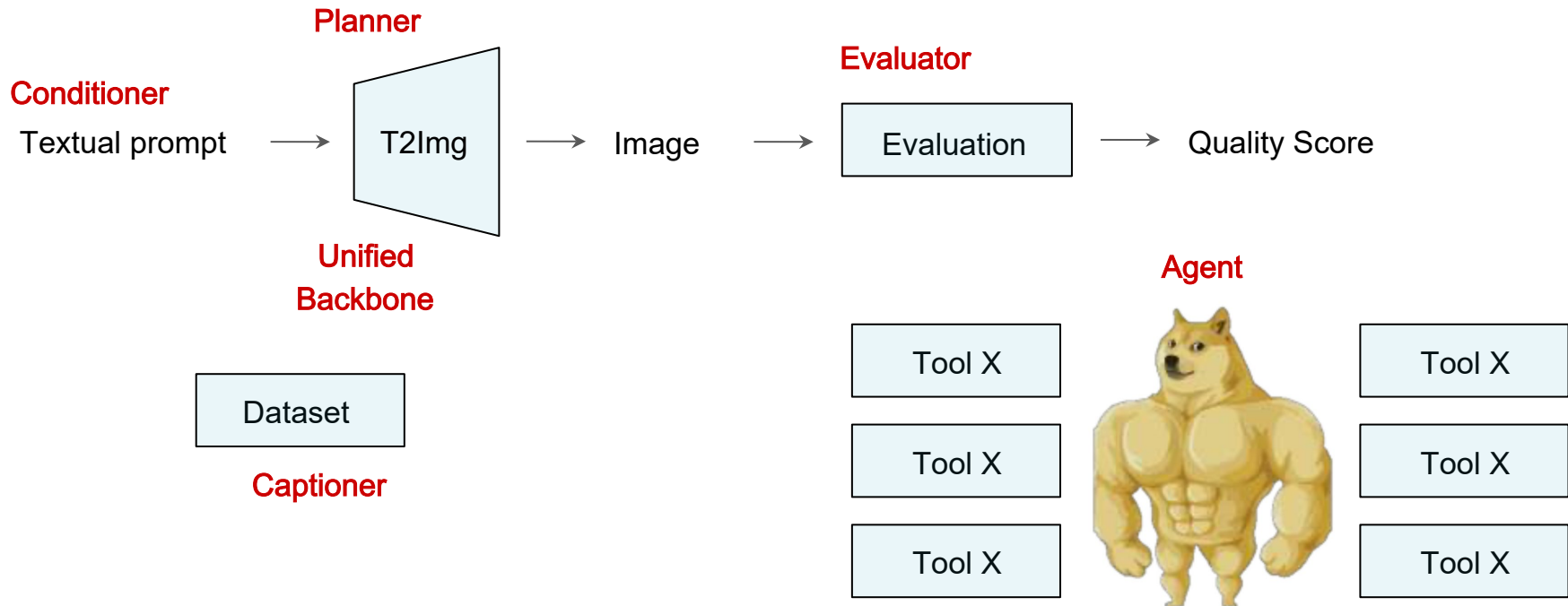
MLLM/ LLM Integration

How **MLLM / LLM** can be integrated in the scheme below to enhance it?



MLLM/ LLM Integration

How MLLM / LLM can be integrated in the scheme below to enhance it?



2

LLM as Unified Backbone

LLM as a unified backbone
for generative AI

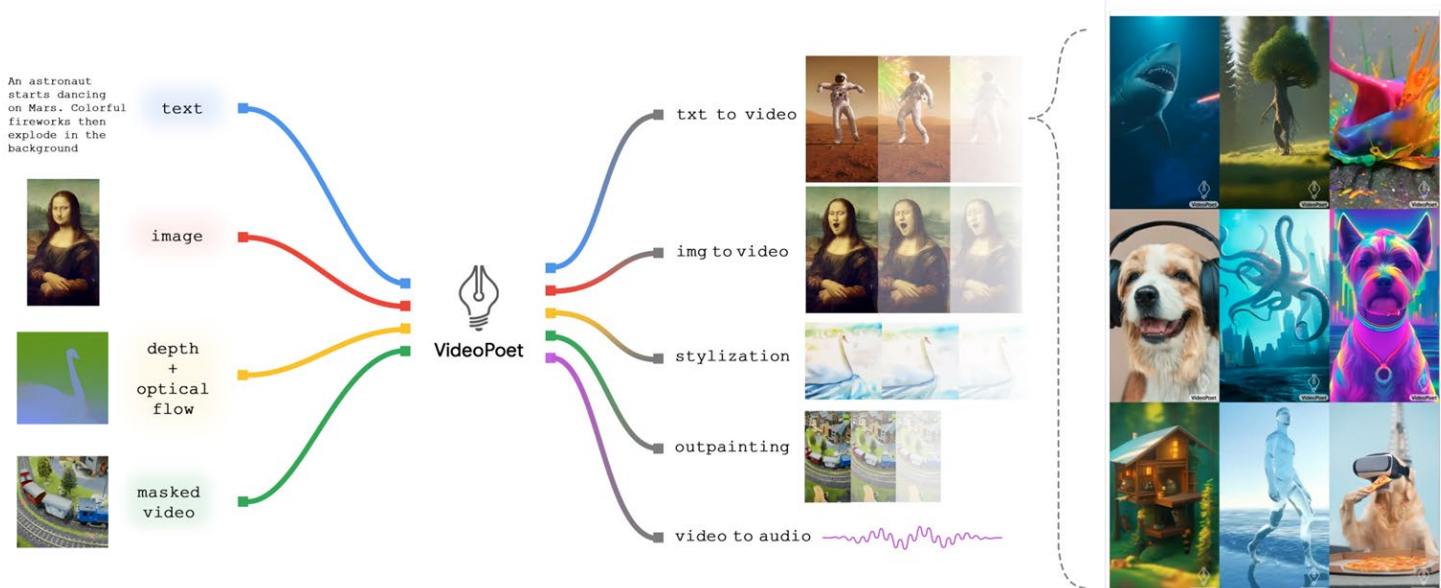


Unified Backbone: VideoPoet

What is the **main idea of VideoPoet** ?

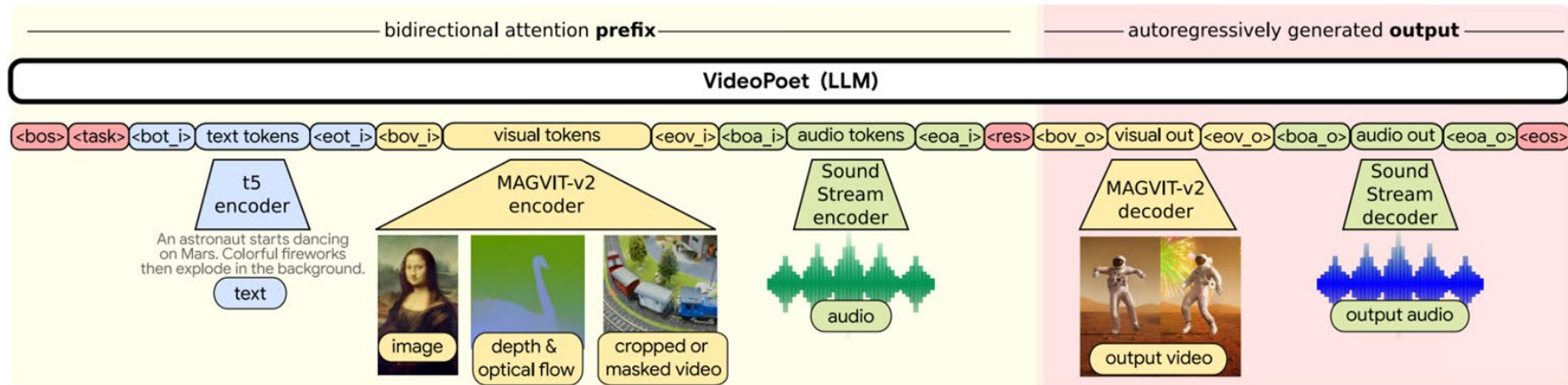
Unified Backbone: VideoPoet

Utilize LLM to preserve **autoregressive token prediction paradigm**, tokenize each modality with its own tokenizer



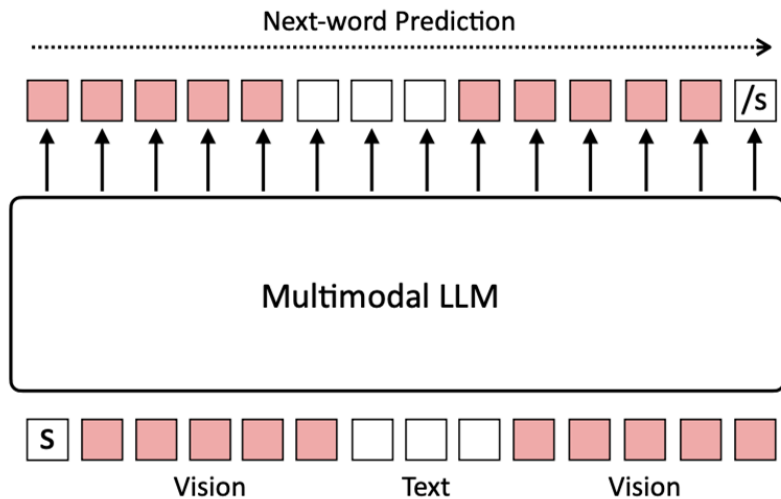
Unified Backbone: VideoPoet

Utilize LLM to preserve **autoregressive token prediction paradigm**, tokenize each modality with its own tokenizer



Unified Backbone: SEED-LLaMA

- Introduce **SEED**, an elaborate **image tokenizer** that empowers LLMs with the ability to SEE and Draw at the same time
- **SEED-LLaMA** has exhibited compositional emergent abilities such as **multi-turn in-context** multimodal generation

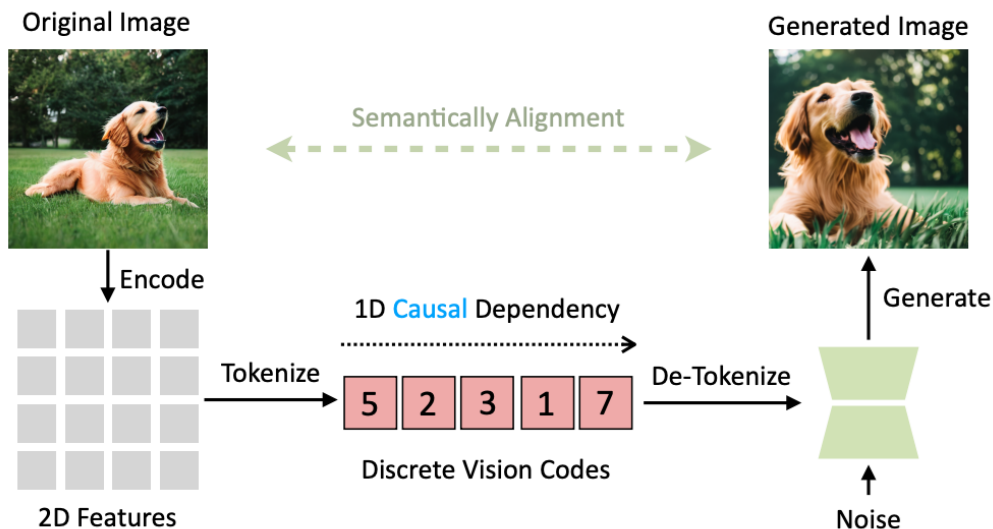


(b) Multimodal Autoregression with SEED visual tokens

SEED-LLaMA: SEED Tokenizer

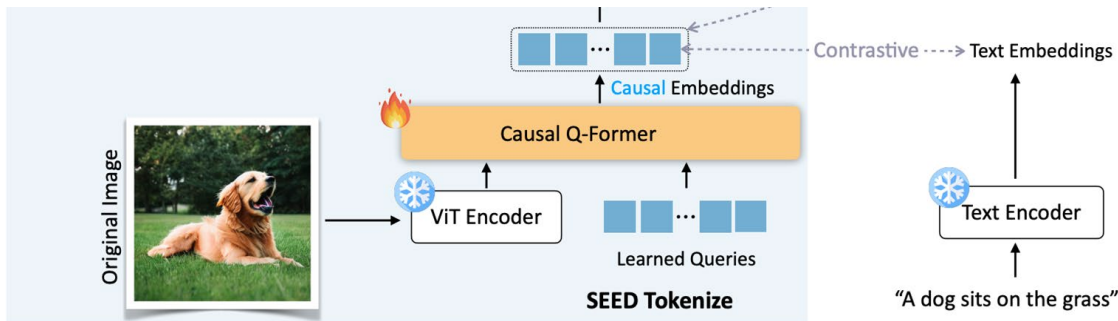
VQ-based image tokenizer that produces **discrete visual codes** with:

1. **1D causal** dependency (*unidirectional attention in dominant LLMs*)
2. **high-level** semantics for both visual comprehension and generation tasks (*share weights and training objectives*)



(a) SEED Visual Tokenizer

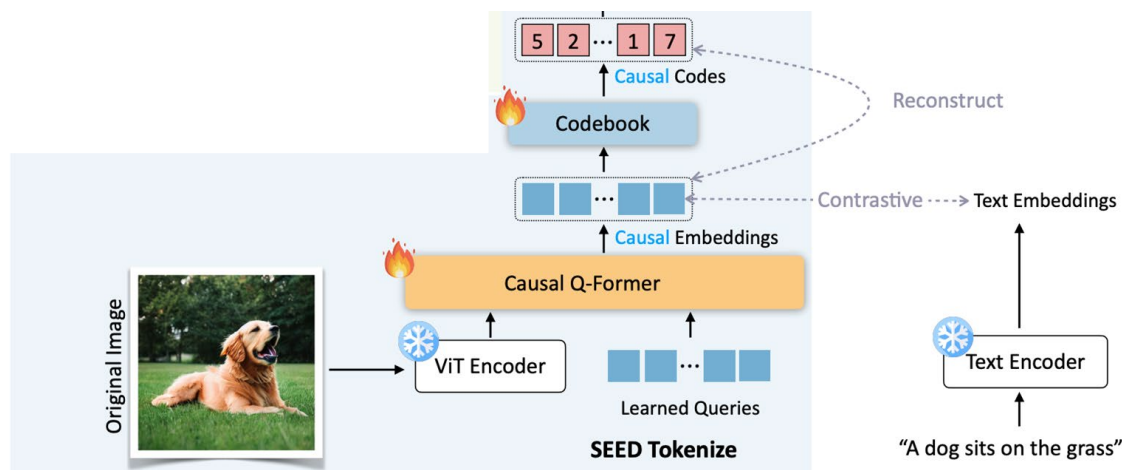
Unified Backbone: SEED-LLaMA



First train a **Causal Q-Former** to convert 2D raster-ordered features (16×16) produced by the ViT encoder into a sequence of **32 causal embeddings**

Self-attention with **previous embeddings** in query

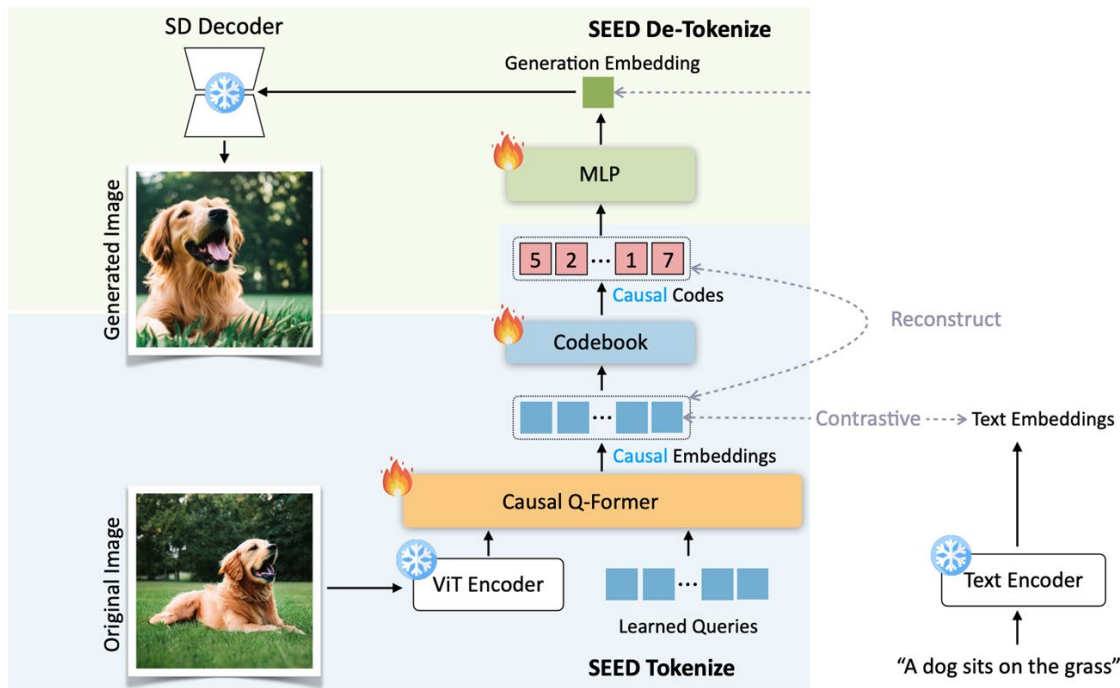
Unified Backbone: SEED-LLaMA



Train a **visual codebook** to discretize the causal embeddings to quantized visual codes (32 tokens) with causal dependency

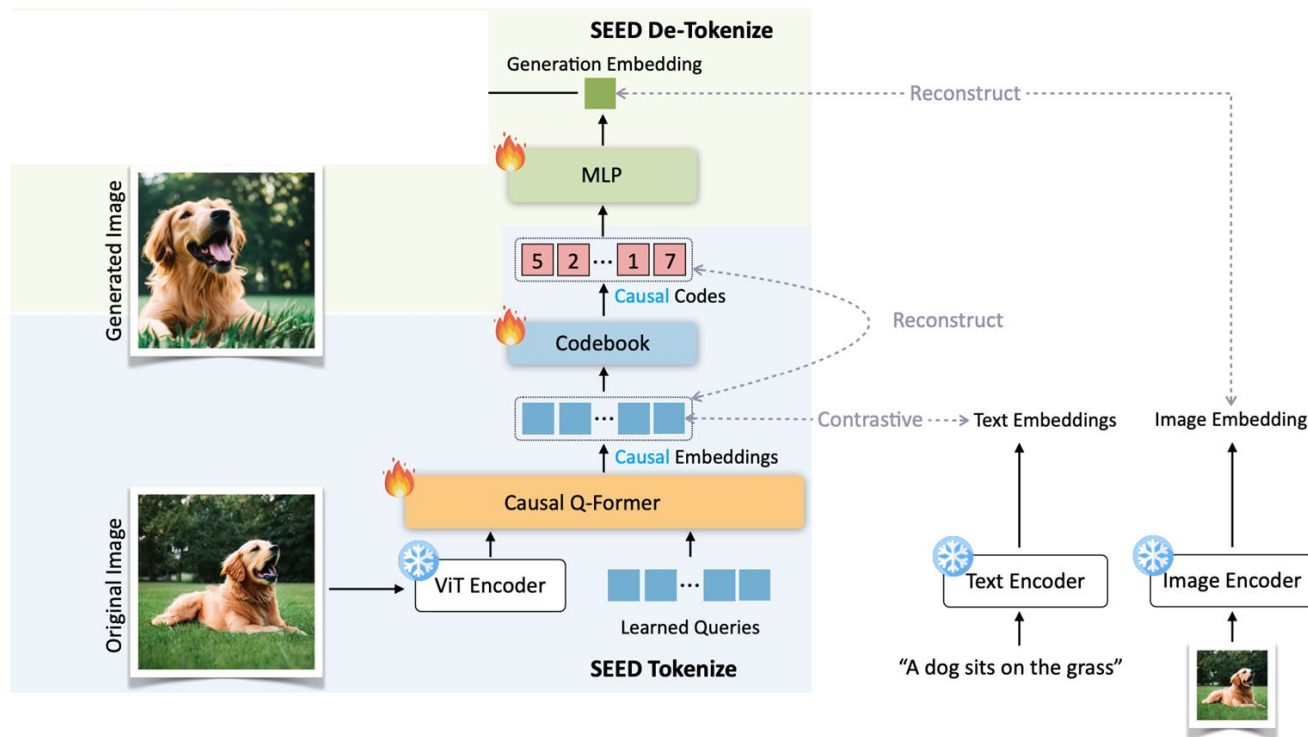
Employ a **multi-layer Transformer** (decoder), to **reconstruct** the continuous causal embeddings

Unified Backbone: SEED-LLaMA



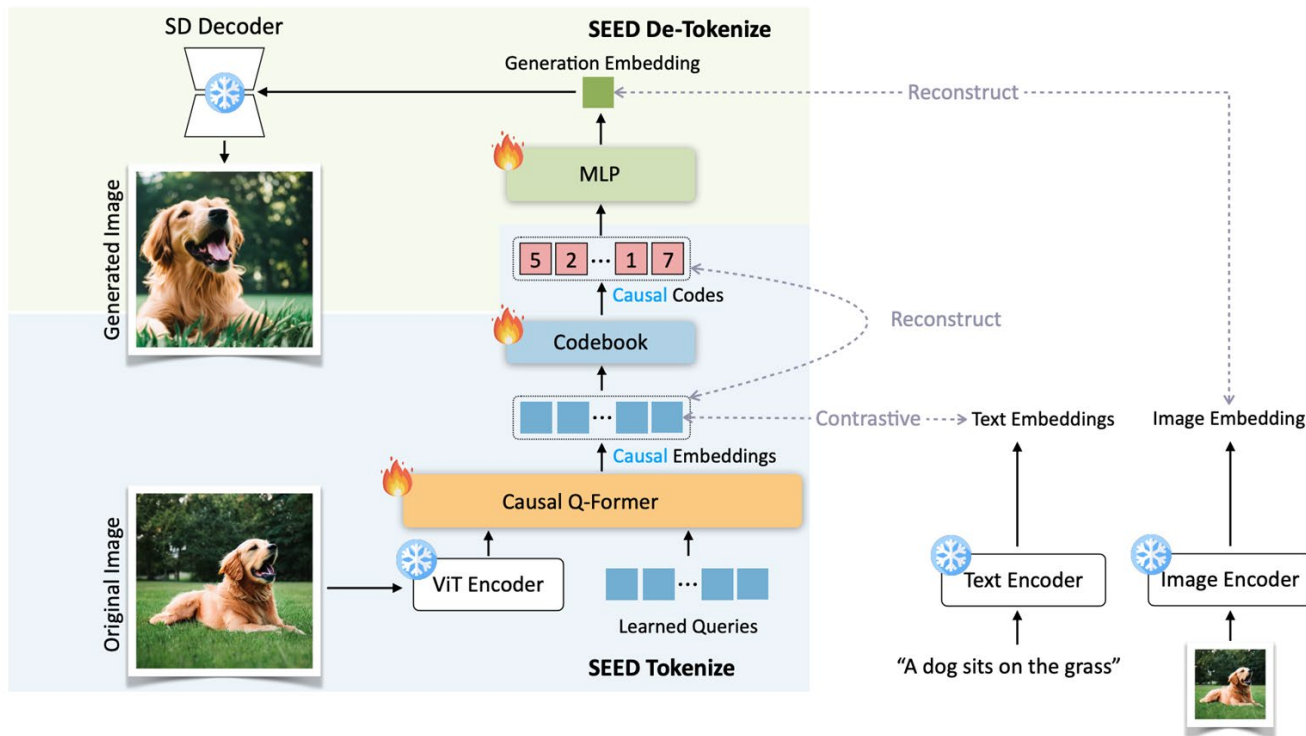
Further employ a **MLP** to reconstruct the image embedding (1 token)

Unified Backbone: SEED-LLaMA



Minimize MSE between
image and generation
embeddings

Unified Backbone: SEED-LLaMA




During inference, the generation embedding are fed into **SD-UNet** to decode realistic images


Unified Backbone: SEED-LLaMA


Reconstruction images of SEED tokenizer (i.e., original image \rightarrow SEED tokenize \rightarrow causal visual codes \rightarrow SEED de-tokenize \rightarrow reconstructed image)




Unified Backbone: SEED-LLaMA


Chatting with SEED-LLaMA 


 Last year, I visited the Golden Gate Bridge and was deeply impressed. Let me show you a picture.





I regret not bringing my dog Lucky along for the visit. This is Lucky.





 Can you generate an image of Lucky sitting in front of the Golden Gate Bridge?




 It looks great! Can you make Lucky wear sunglasses?



 Lucky looks cool in this image. Can you turn it into an oil painting?




 I love this painting. Can you create a story about Lucky's adventure for this painting?

Once upon a time, Lucky was on a mission to find the perfect spot to enjoy the view of the Golden Gate Bridge. He wandered around the park, sniffing and exploring, until he finally found a patch of grass where he could lay down and take in the beauty of the bridge. As he sat there, he put on his sunglasses to protect his eyes from the sun and admired the view.

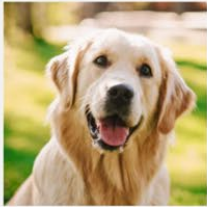
Unified Backbone: SEED-LLaMA

Chatting with SEED-LLaMA


Last year, I visited the Golden Gate Bridge and was deeply impressed. Let me show you a picture.




I regret not bringing my dog Lucky along for the visit. This is Lucky.




Can you generate an image of Lucky sitting in front of the Golden Gate Bridge?



It looks great! Can you make Lucky wear sunglasses?



Lucky looks cool in this image. Can you turn it into an oil painting?



I love this painting. Can you create a story about Lucky's adventure for this painting?

Once upon a time, Lucky was on a mission to find the perfect spot to enjoy the view of the Golden Gate Bridge. He wandered around the park, sniffing and exploring, until he finally found a patch of grass where he could lay down and take in the beauty of the bridge. As he sat there, he put on his sunglasses to protect his eyes from the sun and admired the view.

Why is background **inconsistent** ?

Unified Backbone: Janus-Pro



“Одна голова хорошо, а две – лучше”



Janus-Pro: Architecture

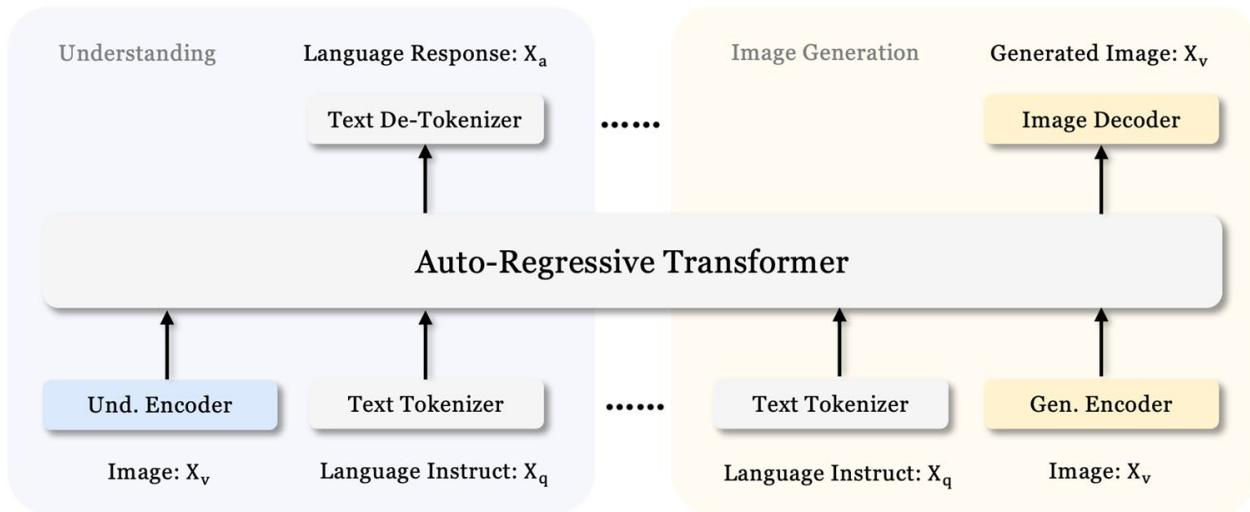


Multimodal understanding

—
SigLIP encoder to extract high-dimensional semantic features from images

Visual generation

—
VQ tokenizer to convert images into discrete IDs



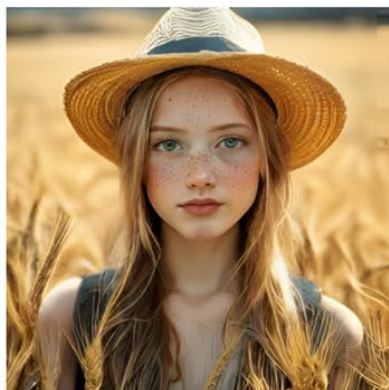
Janus-Pro: Architecture



Text-To-Image Generation



A golden retriever lying peacefully on a wooden porch, with autumn leaves scattered around.



A young woman with freckles wearing a straw hat, standing in a golden wheat field.



A single drop of water clinging to a green leaf, with sunlight creating a faint rainbow prism.



An ancient stone bridge arching over a crystal-clear mountain stream, surrounded by lush greenery.

General Knowledge



Can you introduce the background story of this cake?



This cake is themed around the classic cartoon character **Tom and Jerry**. The cake features the characters Tom (the cat) and Jerry (the mouse) in a playful setting, with a backdrop that includes the words "Tom & Jerry". The cake itself is decorated with fondant shaped like cheese, which is a recurring theme in the cartoon series. The design also includes a small fondant house with a door, adding to the whimsical and playful theme of the cake.

3

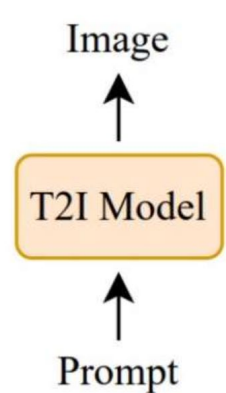
LLM as Planner

LLM as planners for
generative AI

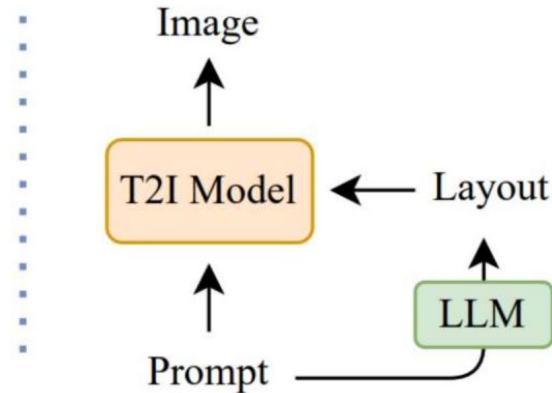


Planner: Layout for Spatial Location

- Layouts/ Bounding boxes for **spatial location planning**
- **Semantic descriptions**
 - Motion descriptions for temporal planning
 - Patch descriptions for Image scale planning
 - Entity descriptions for character/ object design across different video clips



(a) General T2I models



(b) Layouts as intermediate results

LayoutGPT: Visual Planning

[2D Numerical Reasoning] *There are three elephants standing beside a pool of water.*

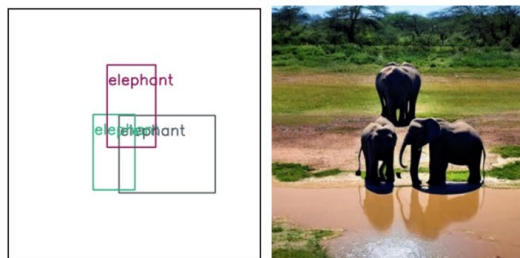
StableDiffusion
(v2.1)



Attend-and-Excite



LayoutGPT
+ GLIGEN

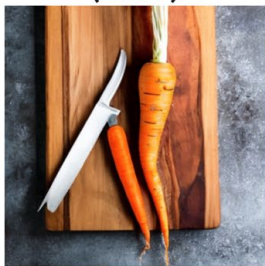


LayoutGPT, a method to compose in-context visual demonstrations to **enhance** the **visual planning** skills of LLMs

LayoutGPT: Visual Planning

[2D Spatial Reasoning] *A carrot and some onion next to a knife on a cutting board.*

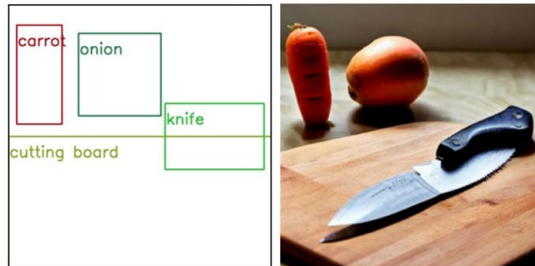
StableDiffusion
(v2.1)



Attend-and-Excite



LayoutGPT
+ GLIGEN



LayoutGPT outperforms text-to-image models/ systems by **20-40%** and achieves comparable performance as human users in designing visual layouts for **numerical** and **spatial correctness**

LayoutGPT: Methodology

Input Condition C_j

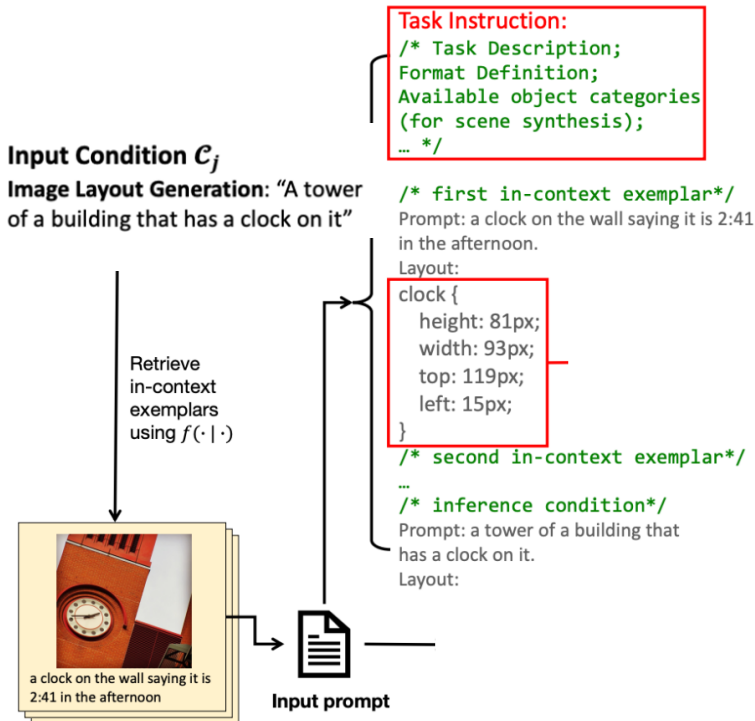
Image Layout Generation: "A tower of a building that has a clock on it"

Retrieve
in-context
exemplars
using $f(\cdot | \cdot)$



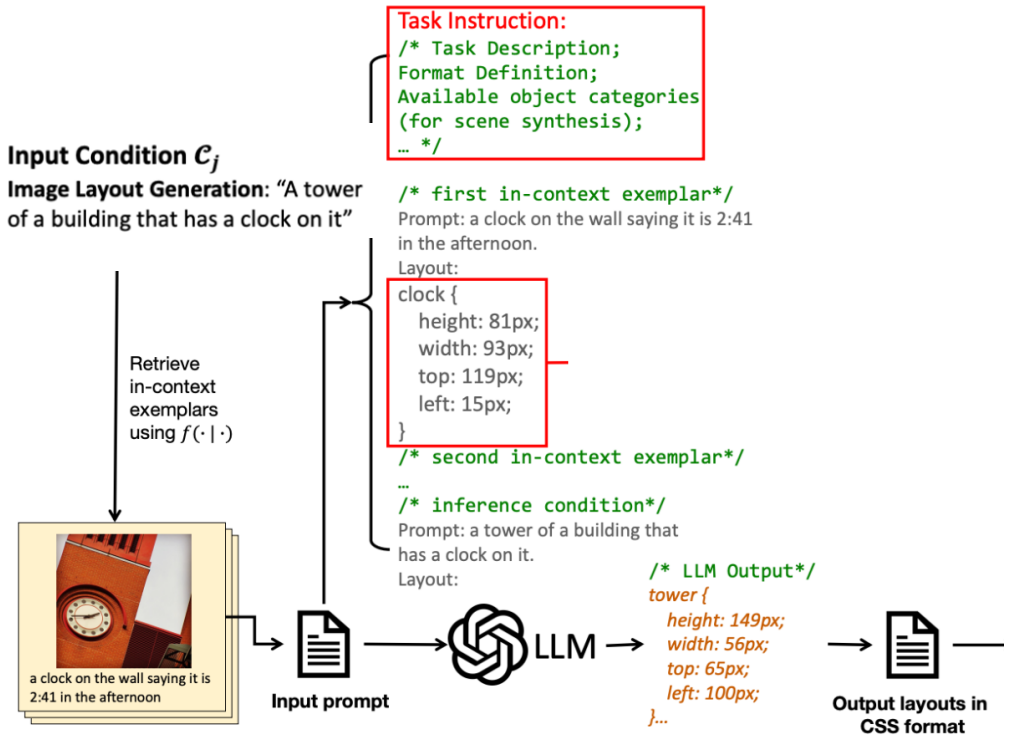
For 2D text-conditioned image layout generation, we adopt the **CLIP model** to extract **text features** and measure the cosine similarity between them

LayoutGPT: Methodology



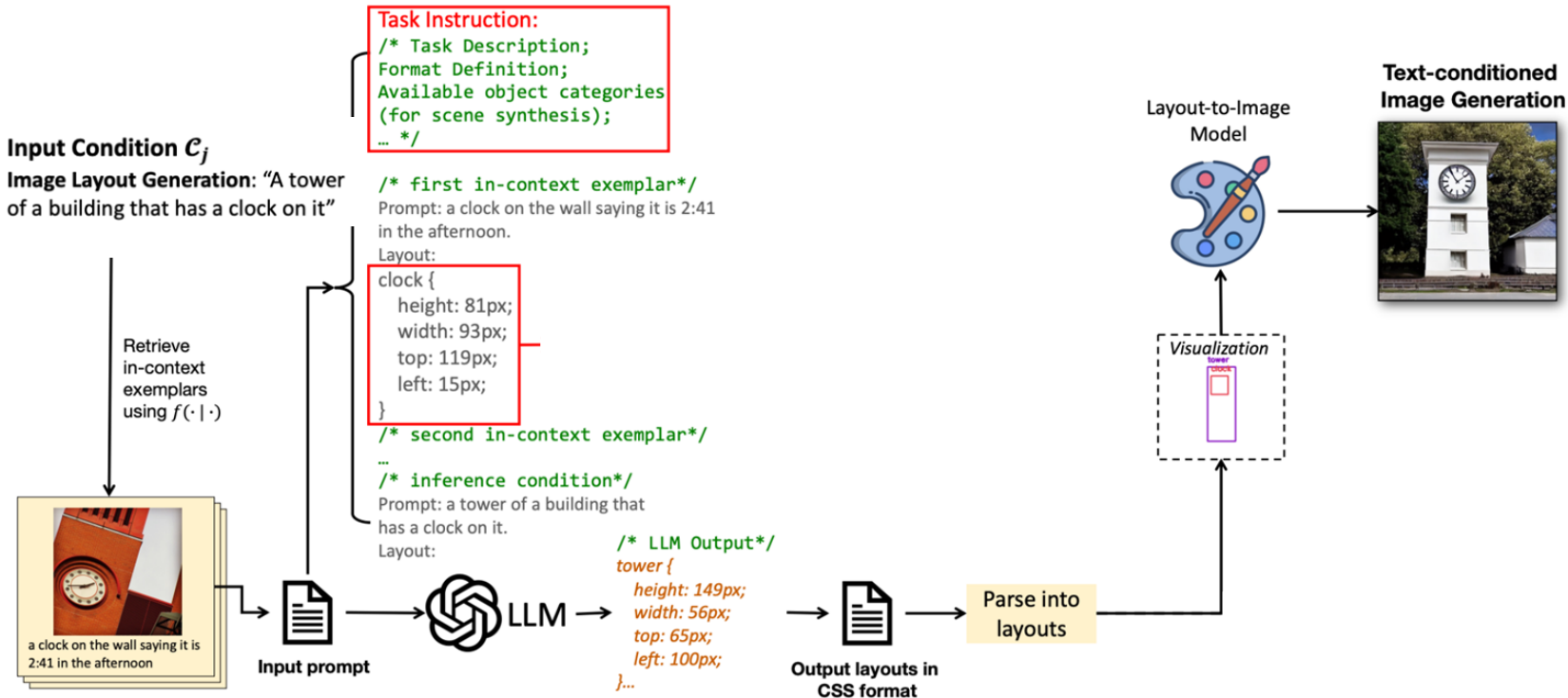
As LLMs like GPT3.5/4 are trained with **code snippets**, formatting image/scene layouts in **CSS structures** potentially enhances the LLMs' interpretation of the spatial meaning

LayoutGPT: Methodology



As LLMs like GPT3.5/4 are trained with **code snippets**, formatting image/scene layouts in **CSS structures** potentially enhances the LLMs' interpretation of the spatial meaning

LayoutGPT: Methodology



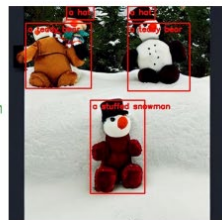
LayoutGPT: Examples

Qualitative comparison between
Stable Diffusion, LayoutGPT, and
human annotations regarding
numerical (top row) and **spatial**
reasoning (bottom row) skills

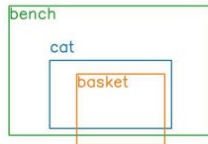
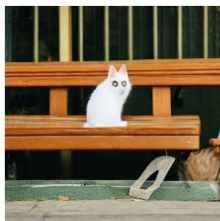
Stable Diffusion



LayoutGPT LayoutGPT+GLIGEN Human+GLIGEN



Two teddy bears and a stuffed snowman wearing hats (Numerical)



A cat is sitting on a basket under a bench (Spatial)

ROI Ctrl: Examples

Models struggle to accurately associate **positional** and **attribute** information with **multiple instances** → regional instance control (ROI Ctrl)

cookies	a frontier face of chihuahua	strawberry
muffin with blueberries on it	blueberries	fried chicken wings
a Shiba Inu with smiling face	golden poodle puppies	a crusty white bread

Grid test for instance control



(a) DALL-E 3

(b) ROI Ctrl (SD2)

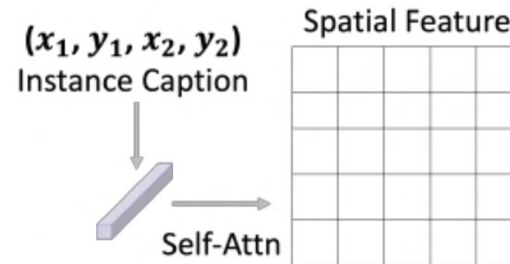
ROI Ctrl: Prior Methods

The goal of Instance Control is to “tell” the diffusion network, via its cross-attention, **where** and **what** to draw

Feature map = 16 x 16 feature map

Desired instance = “Place apple in Box A”

Box A = $(x_1=10, y_1=15)$ to $(x_2=26, y_2=31)$



(a) ROI Injection via Embedding

fusing box coordinate embeddings with
instance caption embeddings

ROI Ctrl: Prior Methods

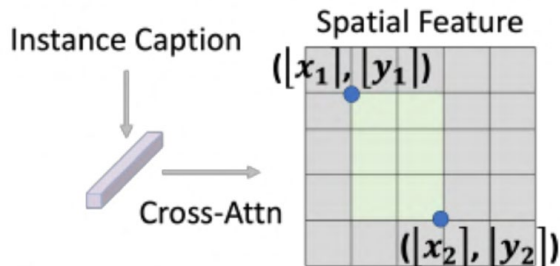
The goal of Instance Control is to “tell” the diffusion network, via its cross-attention, **where** and **what** to draw

Feature map = 16 x 16 feature map

Desired instance = “Place apple in Box A”

Box A = (x₁=10, y₁=15) to (x₂=26, y₂=31)

$$M_A(i, j) = \begin{cases} 1, & \text{if } 10 \leq i \leq 26 \text{ and } 15 \leq j \leq 31, \\ 0, & \text{otherwise.} \end{cases}$$



(b) ROI Injection with Attention Mask

use masked crossattention to isolate each
ROI

ROI Ctrl: Methodology

Global Caption: Three people stand near the lake.

ROI-1: A woman with white dress. ROI-2: A man with a black suit. ROI-3: A man with armors.



w/. ROI Ctrl



w/o. ROI Ctrl

Region Control in Generation
(ROI-Ctrl Alone)

Global Caption: Two dogs and a cat on the grass.

ROI-1: A dog with orange fur. ROI-2: A cat with white fur. ROI-3: A dog with grey fur.



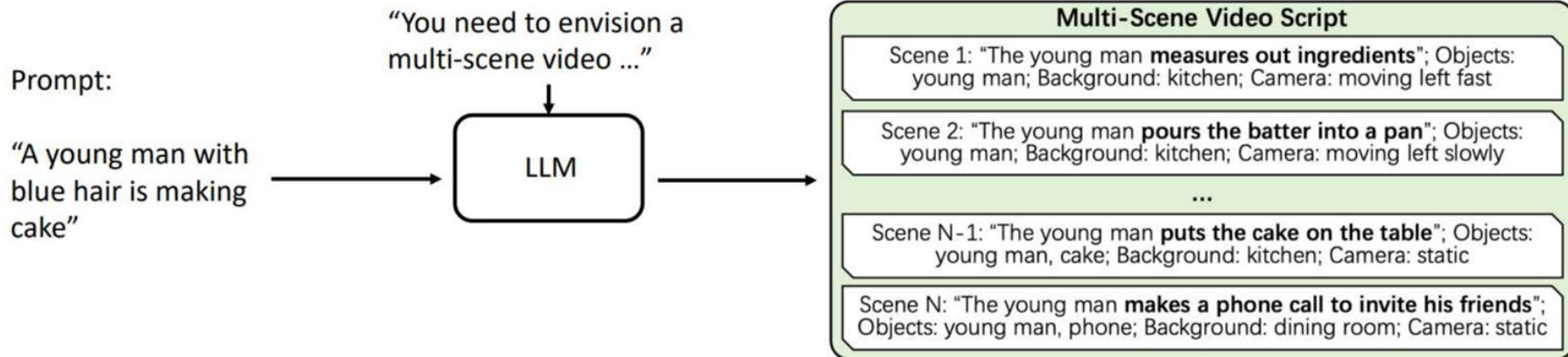
w/. ROI Ctrl



w/o. ROI Ctrl

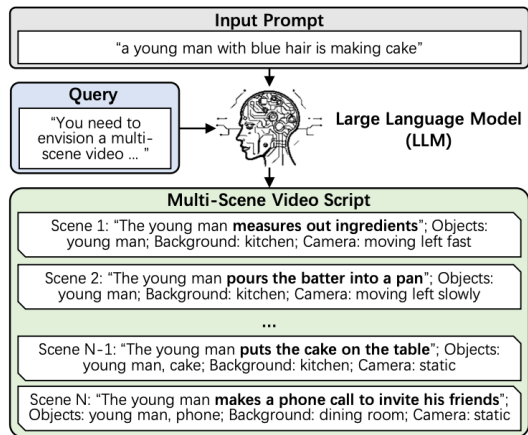
LLM as Planner: Scene Decomposition

Prompt decomposition in multiple prompts



Video Studio: Almost Agent -System!

(1) Multi-scene video script generation

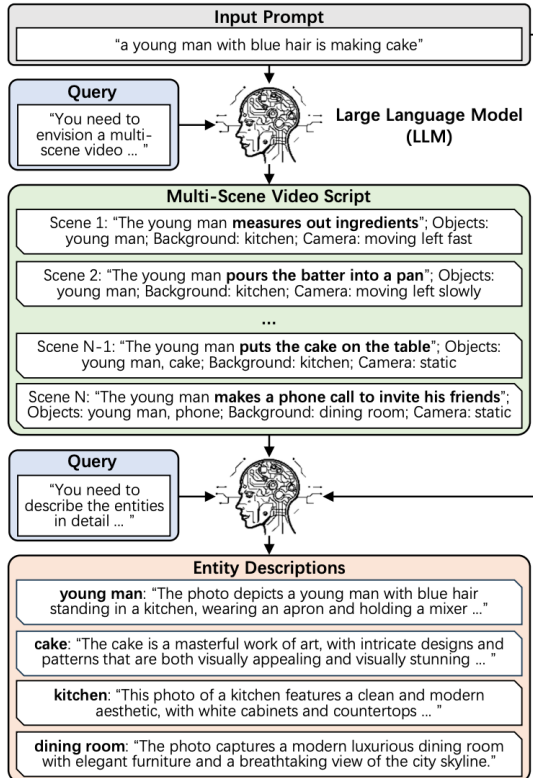


Video Script for each scene:

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

Video Studio: Almost Agent -System!

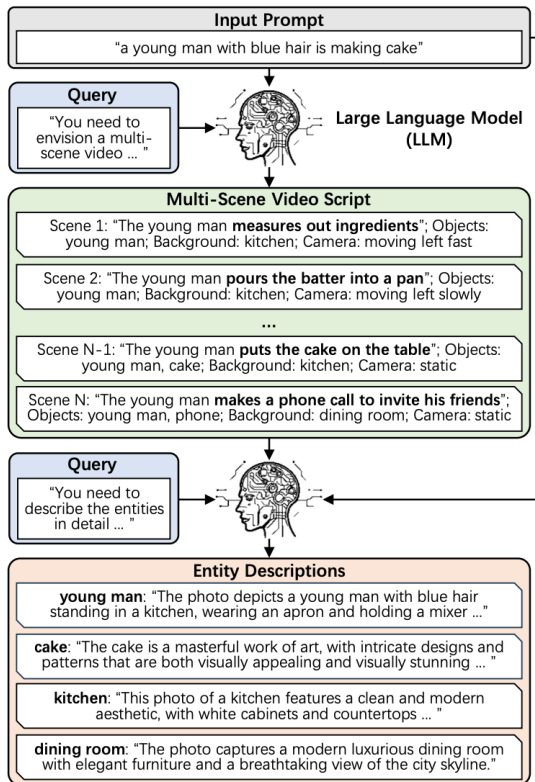
(1) Multi-scene video script generation



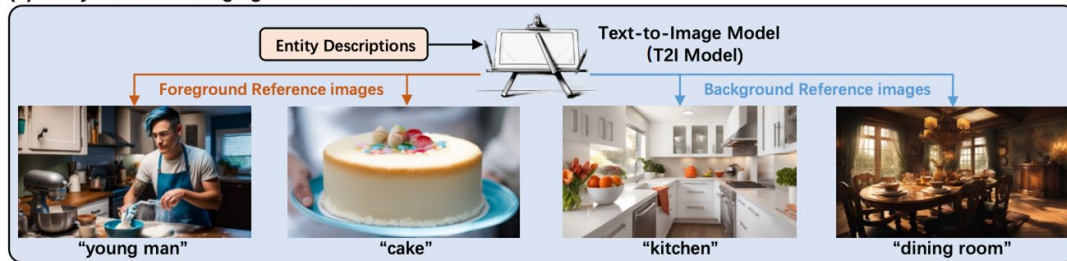
Detailed description for
each common entity

Video Studio: Almost Agent -System!

(1) Multi-scene video script generation

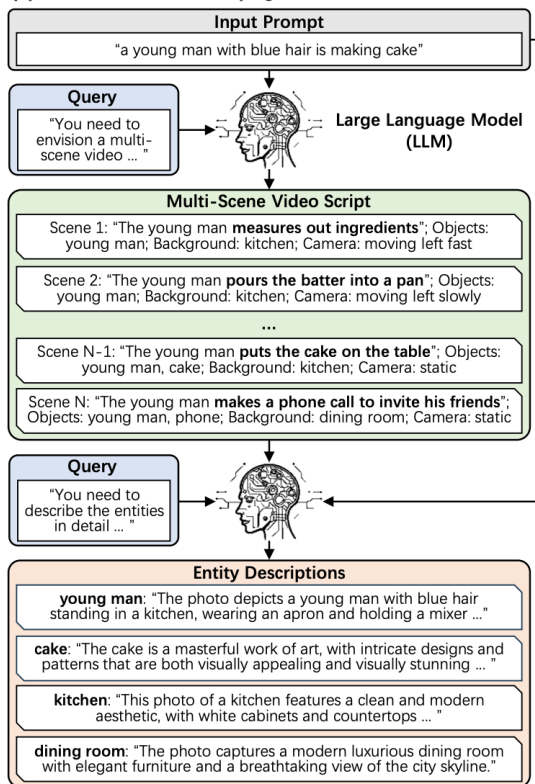


(2) Entity reference image generation

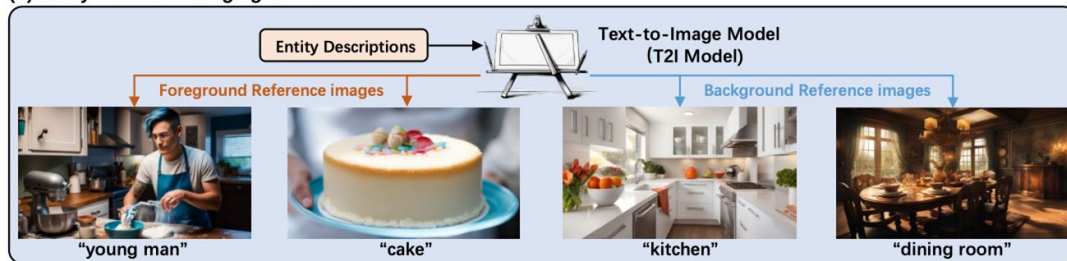


Video Studio: Almost Agent-System!

(1) Multi-scene video script generation



(2) Entity reference image generation



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



"young man"



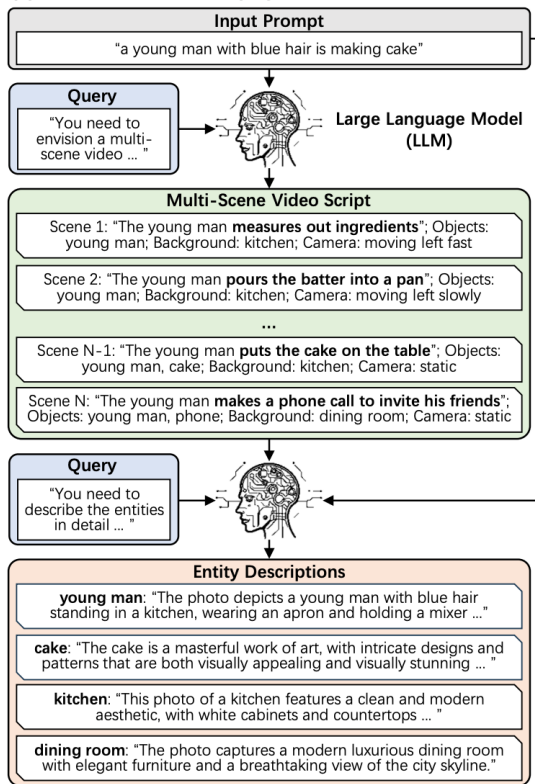
"kitchen"



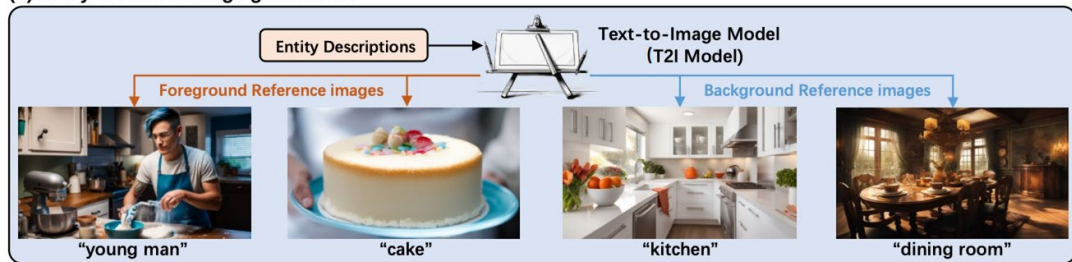
Video Scene 1

Video Studio: Almost Agent-System!

(1) Multi-scene video script generation



(2) Entity reference image generation



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



VideoStudio: Almost Agent -System!

Real Reference Images:



input prompt:

The cat lies in the room

The cat lies in the driving car

The cat plays in the flowers

VideoStudio: Almost Agent-System!

input prompt:

The cat lies in the room

The cat lies in the driving car

The cat plays in the flowers

Real Reference Images:



VideoStudio: Almost Agent -System!

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

Real Reference Images:



VideoStudio: Almost Agent -System!

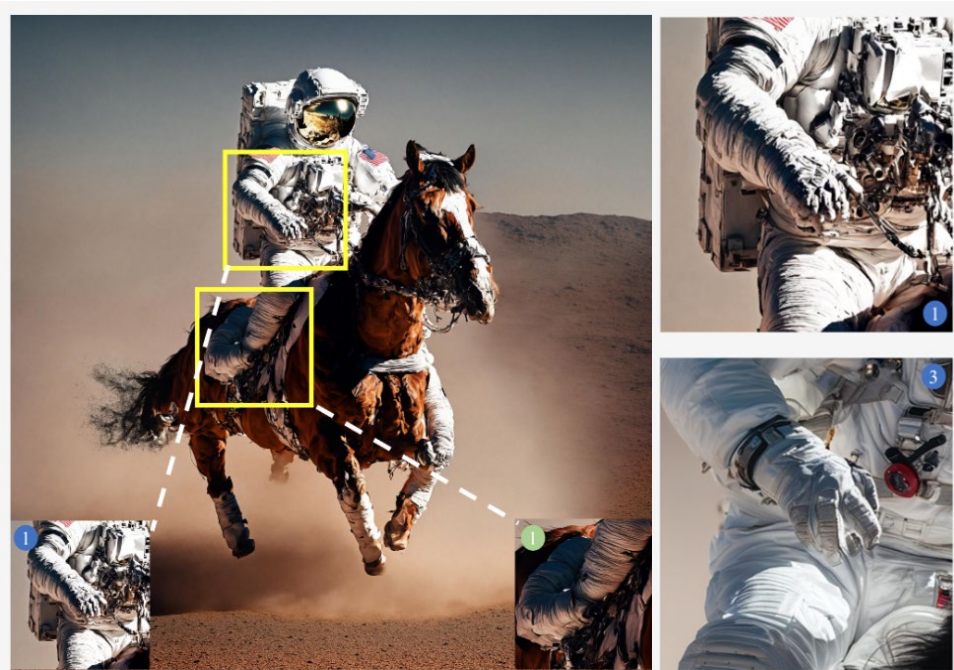
input prompt:

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

Real Reference Images:

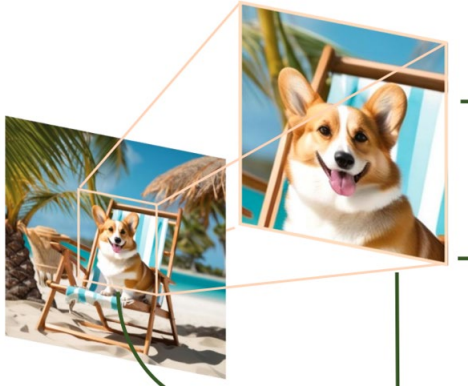


LLM as Planner: Hierarchical Prompts

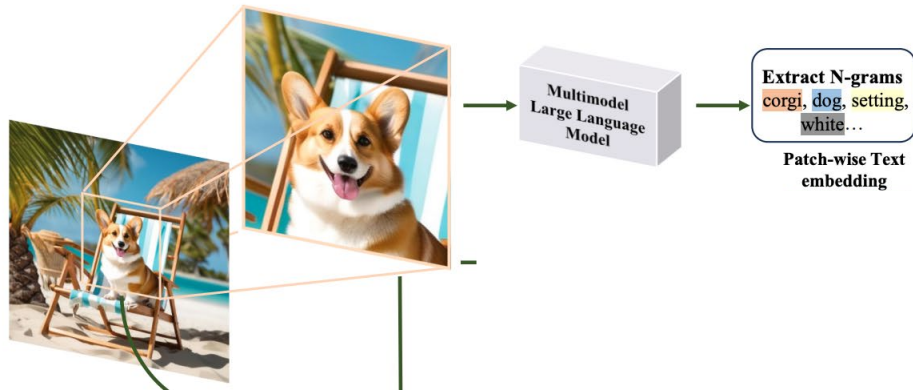


Models often struggle with issues of **object repetition** and **structural artifacts** especially when scaling to 4K resolution and higher

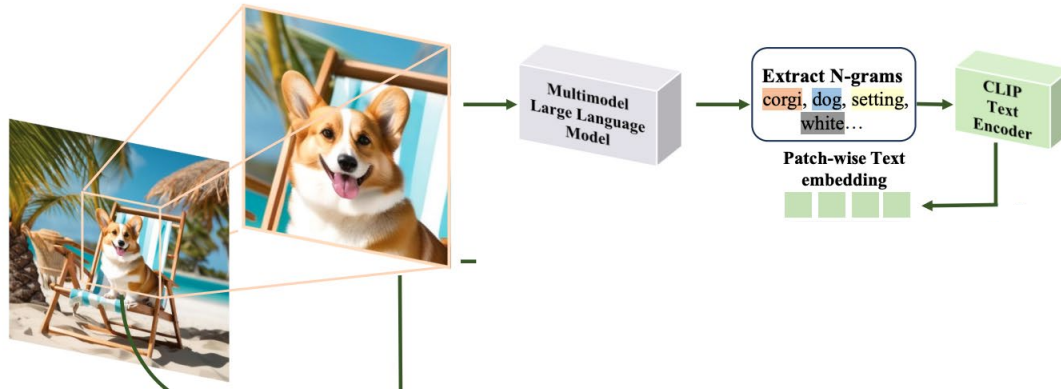
HiPrompt: Local Patches



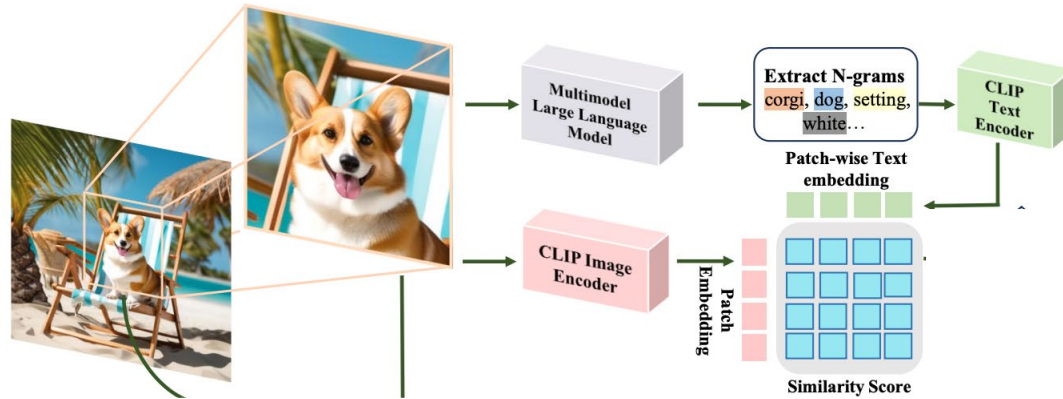
HiPrompt: Local Patches



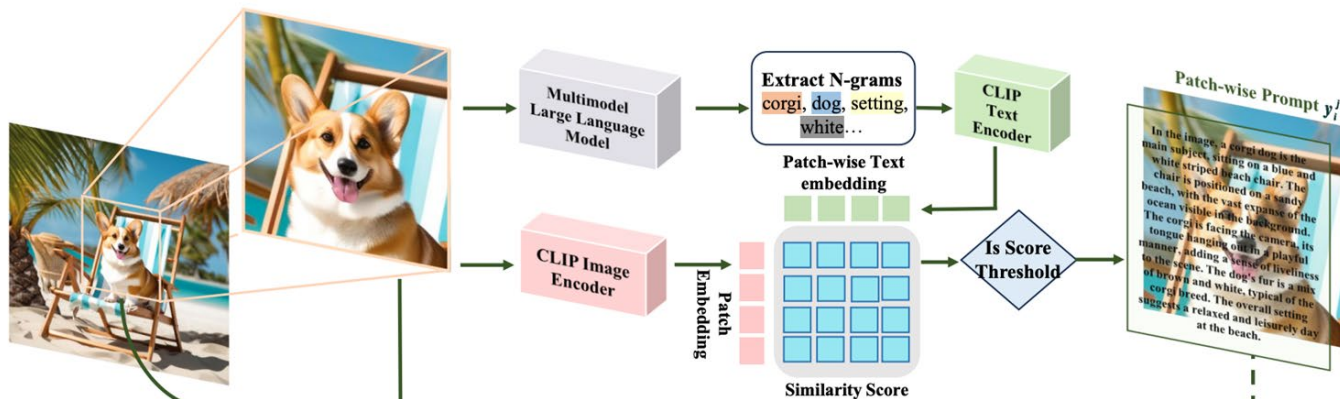
HiPrompt: Local Patches



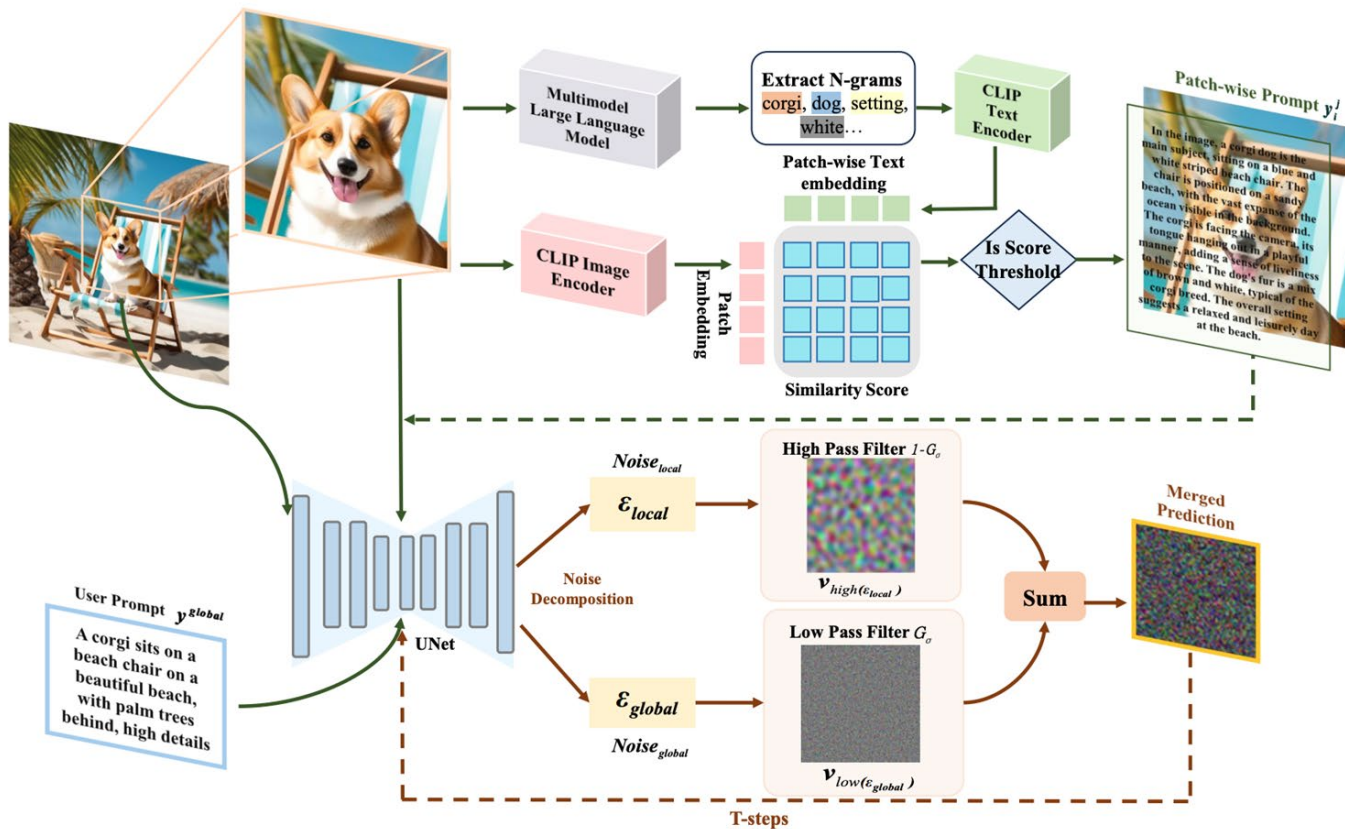
HiPrompt: Local Patches



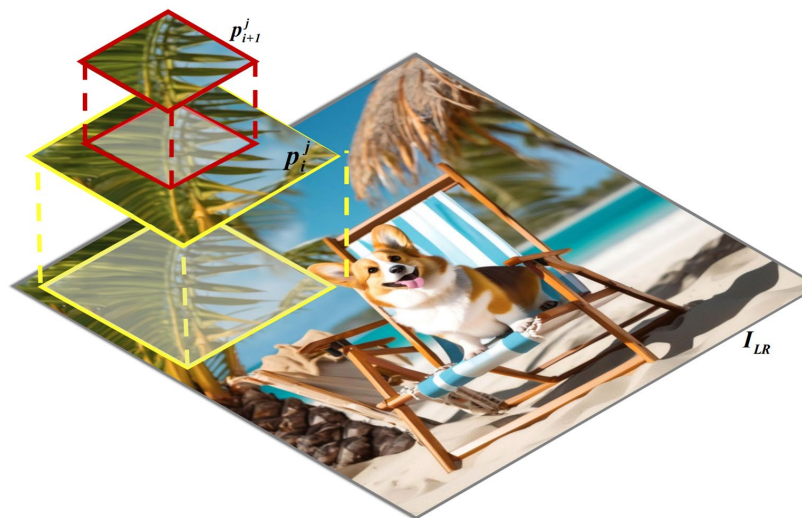
HiPrompt: Local Patches



HiPrompt: Local Patches



HiPrompt: Local Patches



Global description

A corgi sits on a beach chair on a beautiful beach, with palm trees behind, high details

Local description

A detailed image of a lush green palm leaf, with a vibrant yellow hue, set against a clear blue sky, taken from a close-up perspective

HiPrompt: Local Patches



Visual comparison between **DemoFusion** (Du et al. 2024) and **HiPrompt**

HiPrompt: Local Patches



Examples of
HiPrompt at
various higher
resolutions




4

LLM as Captioner

LLM as a captioner for
generative AI



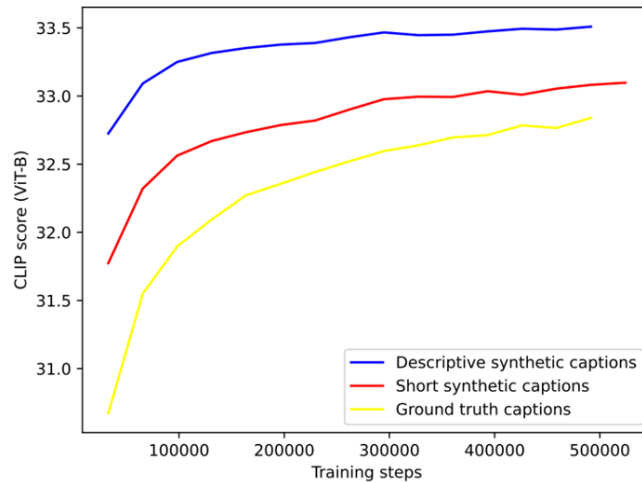
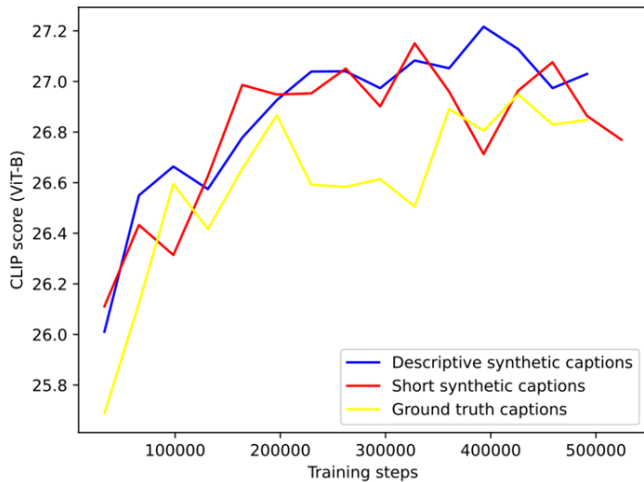
DALL-E 3: Improved Dataset

Image	Alt Text	SSC	DSC
	<p>now at victorian plumbing.co.uk</p>	<p>a white modern bathtub sits on a wooden floor.</p>	<p>this luxurious bathroom features a modern freestanding bathtub in a crisp white finish. the tub sits against a wooden accent wall with glass-like panels, creating a serene and relaxing ambiance. three pendant light fixtures hang above the tub, adding a touch of sophistication. a large window with a wooden panel provides natural light, while a potted plant adds a touch of greenery. the freestanding bathtub stands out as a statement piece in this contemporary bathroom.</p>
	<p>is he finished...just about!</p>	<p>a quilt with an iron on it.</p>	<p>a quilt is laid out on an ironing board with an iron resting on top. the quilt has a patchwork design with pastel-colored strips of fabric and floral patterns. the iron is turned on and the tip is resting on top of one of the strips. the quilt appears to be in the process of being pressed, as the steam from the iron is visible on the surface. the quilt has a vintage feel and the colors are yellow, blue, and white, giving it an antique look.</p>
	<p>23 (19 of 30) 1200</p>	<p>a jar of rhubarb liqueur sitting on a pebble background.</p>	<p>rhubarb pieces in a glass jar, waiting to be pickled. the colors of the rhubarb range from bright red to pale green, creating a beautiful contrast. the jar is sitting on a gravel background, giving a rustic feel to the image.</p>

- Human descriptions are **very simple** and noisy
- People usually don't **describe the position of objects**, important parts (like the kitchen sink)
- VLM can be trained in **very detailed image descriptions**

DALL-E 3: Improved Dataset

CLIP scores for text-to-image models trained on different caption types



DALL-E 3: Examples



ShareGPT4V: Synthetic Dataset



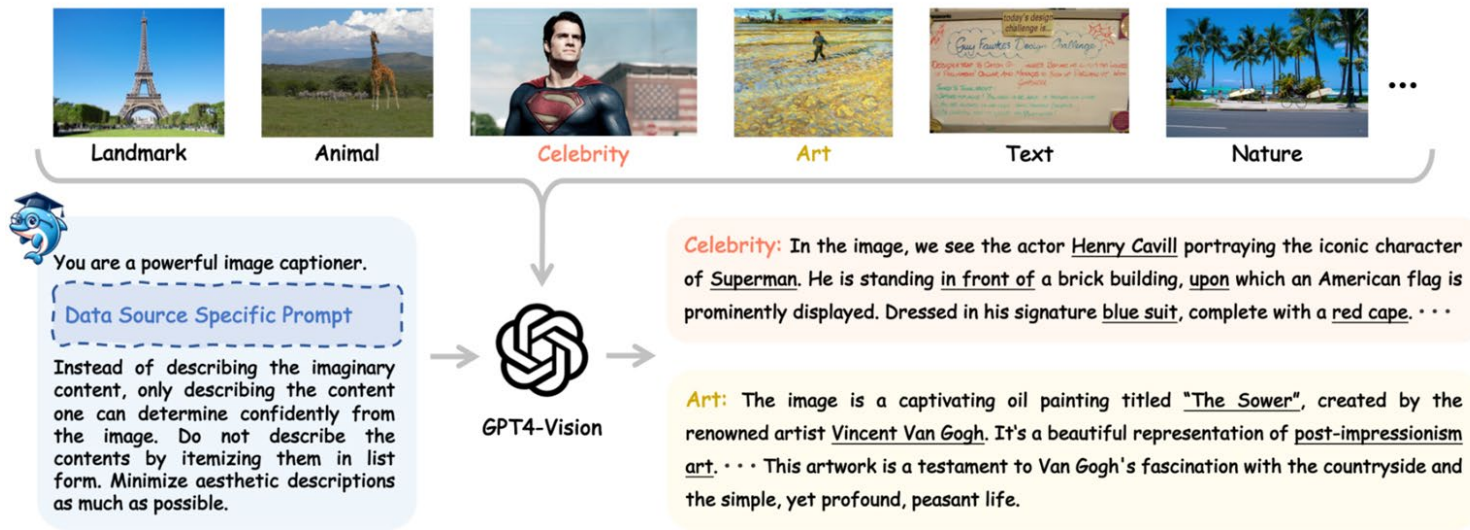
Comparison of widely -used caption datasets and **ShareGPT4V**
'Avg.' column shows the average character number of the caption

Dataset Name	Image Source	Visible	Captioned by	Samples	Avg.
COCO-Caption	COCO	✓	Human	118K	52
BLIP-LCS	LCS	✓	BLIP	558K	54
LLaVA-23K	COCO	x	GPT4	23K	609
ShareGPT4V	LCS, COCO, etc.	✓	GPT4-Vision	100K	942
ShareGPT4V-PT	LCS, COCO, etc.	✓	Share-Captioner	1,246K	826

ShareGPT4V: Syntetic Dataset



Need clear dataset: all objects that occur in the picture are not made up (no conclusions / evaluations of aesthetic)



ShareGPT4V: Syntetic Dataset



COCO:

Young children standing on a platform waiting for a train to arrive.
Adults and children watching a train slowly leave.
A family near a railroad track watching the train pass.
People waiting on a platform as a train pulls up.
A train station with a green train on the tracks and children waiting for it to go by.

LLaVA: At a train station, a group of people, including both young children and adults, are standing on a platform waiting for a train to arrive. The train is already present on the tracks, partially visible on the right side of the image. Some of the people watch the train closely, while others seem to be patiently anticipating its departure.

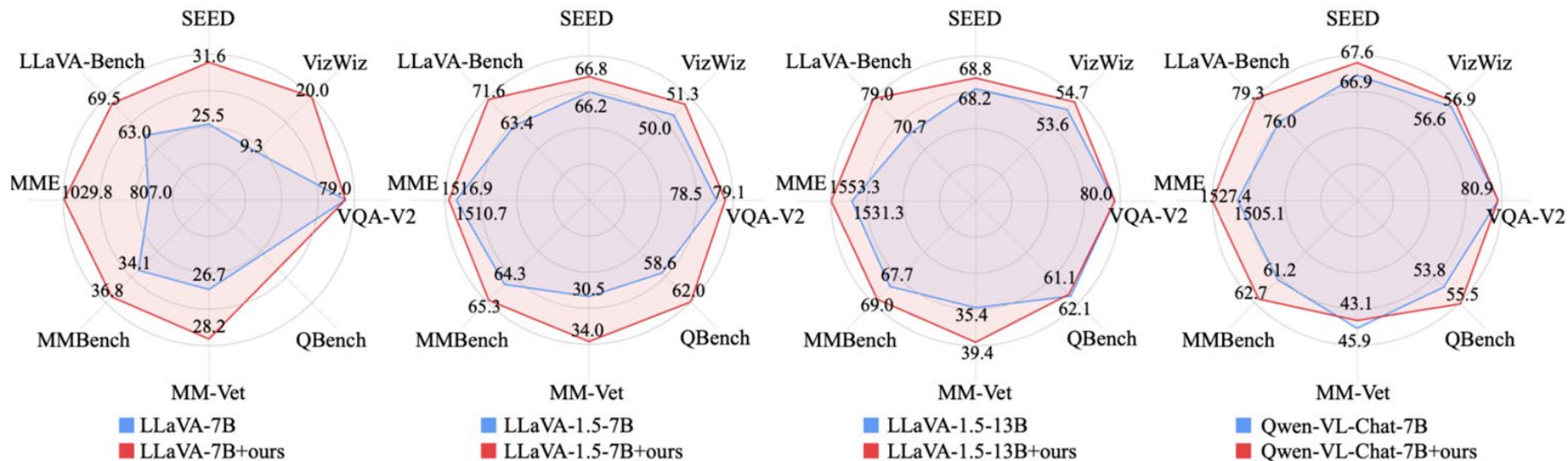
There is a total of eight individuals waiting for the train, with one child in the middle of the platform and the others scattered around. A backpack can be found on the far left side of the platform, **suggesting that someone may have set it down while waiting.**

ShareGPT4V: The image captures a moment at a train station. A **green** train is on the tracks, moving away from the platform labeled "Platform 2". The train's motion is observed by people standing on the platform, their attention drawn to the departing vehicle.

A **red** signal stands on the tracks, its vibrant color contrasting with the green of the train. Signs are prominently displayed around the platform. One warns "Beware of Trains", a cautionary reminder of the powerful machines that frequent these tracks. Another sign advises "Please Use The Footbridge To Cross The Line. If You Require Assistance Please Ask The Station Staff", guiding passengers to safely navigate the station.

The sky above is a **clear blue**, providing a serene backdrop to the bustling activity of the station. **In the distance, trees can be seen**, adding a touch of nature to this man-made setting. The image is a snapshot of everyday life at a train station, capturing both its routine operations and its inherent charm.

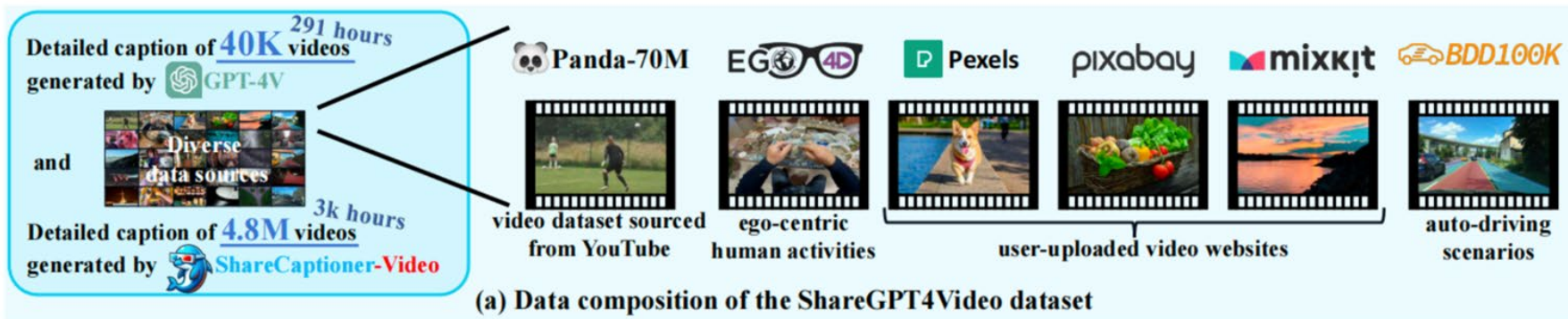
ShareGPT4V: Synthetic Dataset



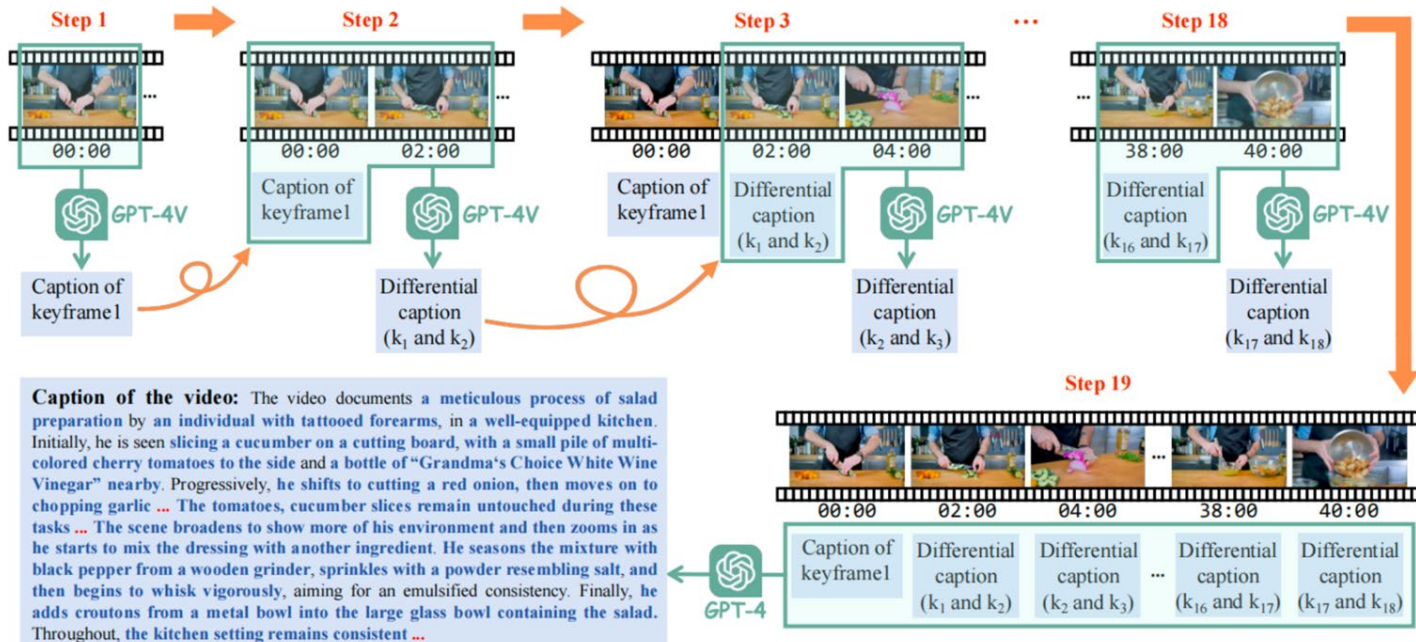
ShareGPT4 Video: Synthetic Dataset



Authors decided not to stop! Published analogous dataset for video



ShareGPT4 Video: Syntetic Dataset



ShareGPT4 Video: Synthetic Dataset



The video captures the spectacle of a continuous fireworks show against the backdrop of a starry night sky. It commences with a burst of vibrant reds, greens, purples, and yellows that paint the heavens and cast shimmering reflections upon the water below. As the display progresses, the fireworks evolve, transitioning from the initial array to a focus on radiant oranges, yellows, and fiery reds. These explosions form captivating clusters at the heart of the sky, ascending in breathtaking formations accompanied by trailing plumes of smoke, adding a dramatic flourish to the visual narrative. Throughout the duration, the fireworks maintain their dynamic allure, their patterns and positions evolving to underscore the ongoing spectacle. Meanwhile, the mirrored reflections on the water's surface faithfully echo the colors and shapes above, further enhancing the mesmerizing and ever-changing nature of the display.



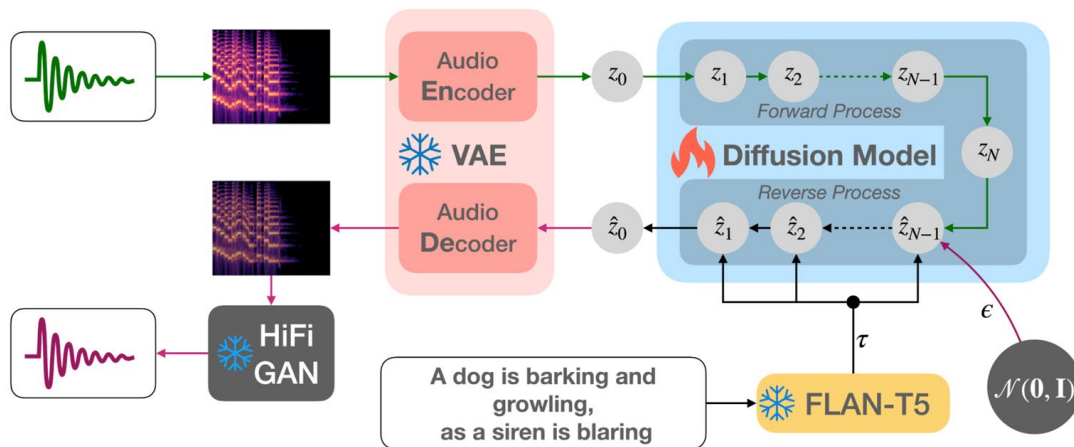
5

LLM as Conditioner

LLM as a conditioner for
generative AI



Let's Tango: Text2Audio Model



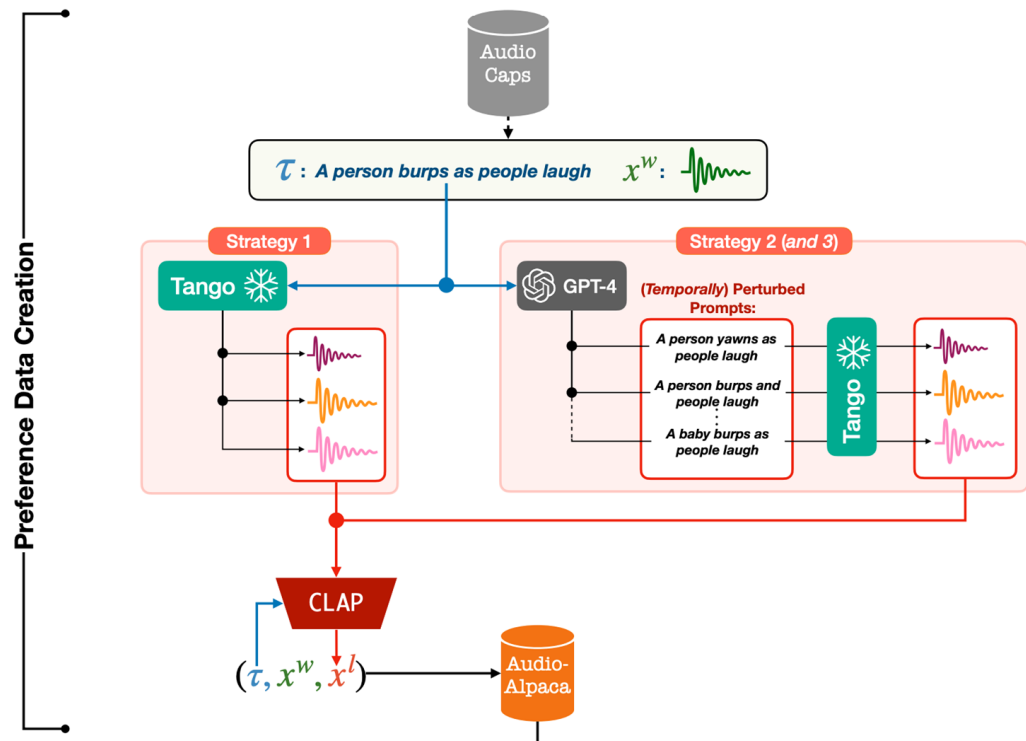
Describe the sound of
the ocean



A man is speaking in a
huge room

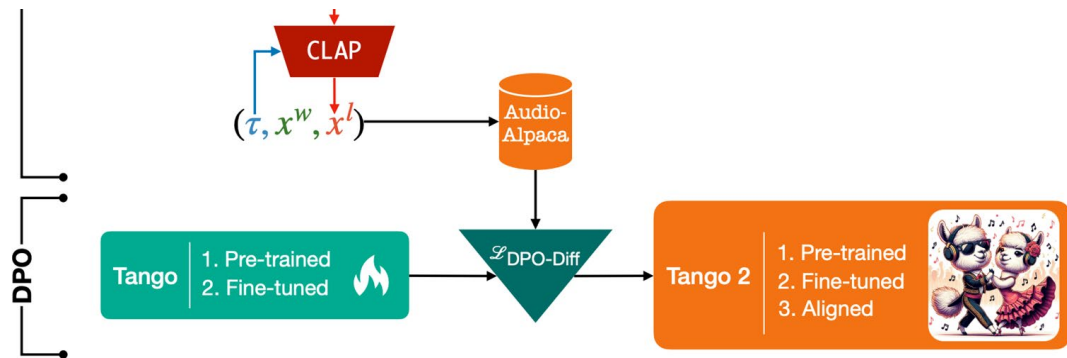


LLM as Conditioner: Tango-2



- Take the original Tango
- Make “augmentation” of data
- Generate audio options
- Sort them by quality using an external model
- Tune Tango using DPO

LLM as Conditioner: Tango-2



- Take the original Tango
- Make “augmentation” of data
- Generate audio options
- Sort them by quality using an external model
- **Fine-tune Tango**

6

LLM as Evaluator

LLM as evaluators for
generative AI



LLM as Evaluator: DreamSync

Prompt: "a little girl wearing a bright yellow dress and a copper crown is riding a badger through a field of flowers."



LLM as Evaluator: DreamSync

Prompt: "a little girl wearing a bright yellow dress and a copper crown is riding a badger through a field of flowers."



 Language Model

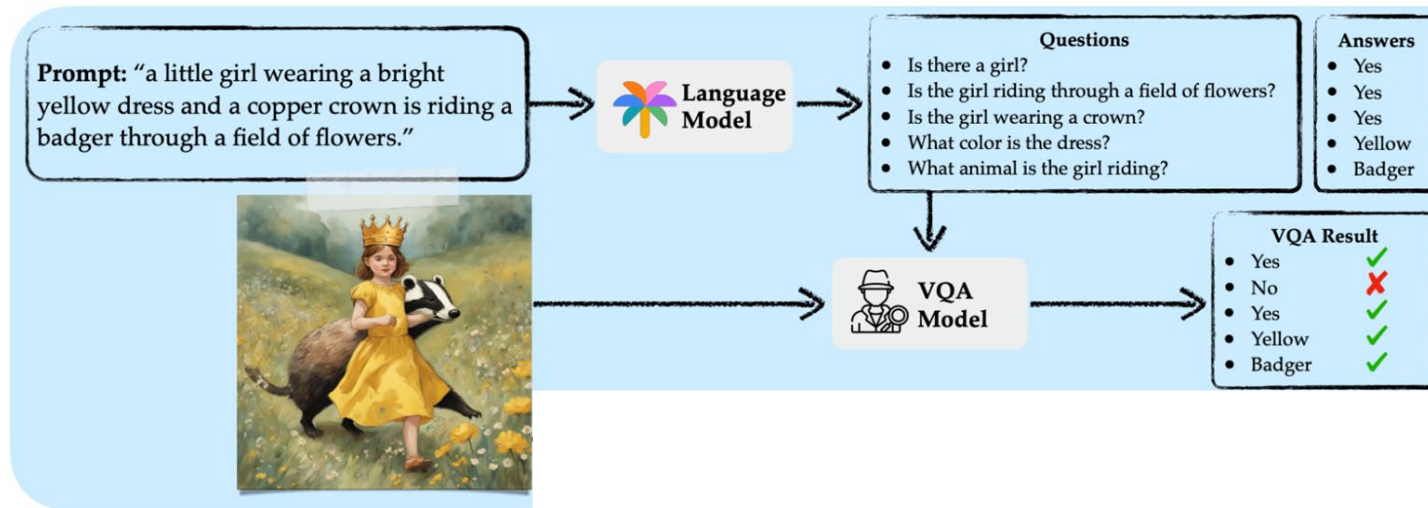
Questions

- Is there a girl?
- Is the girl riding through a field of flowers?
- Is the girl wearing a crown?
- What color is the dress?
- What animal is the girl riding?

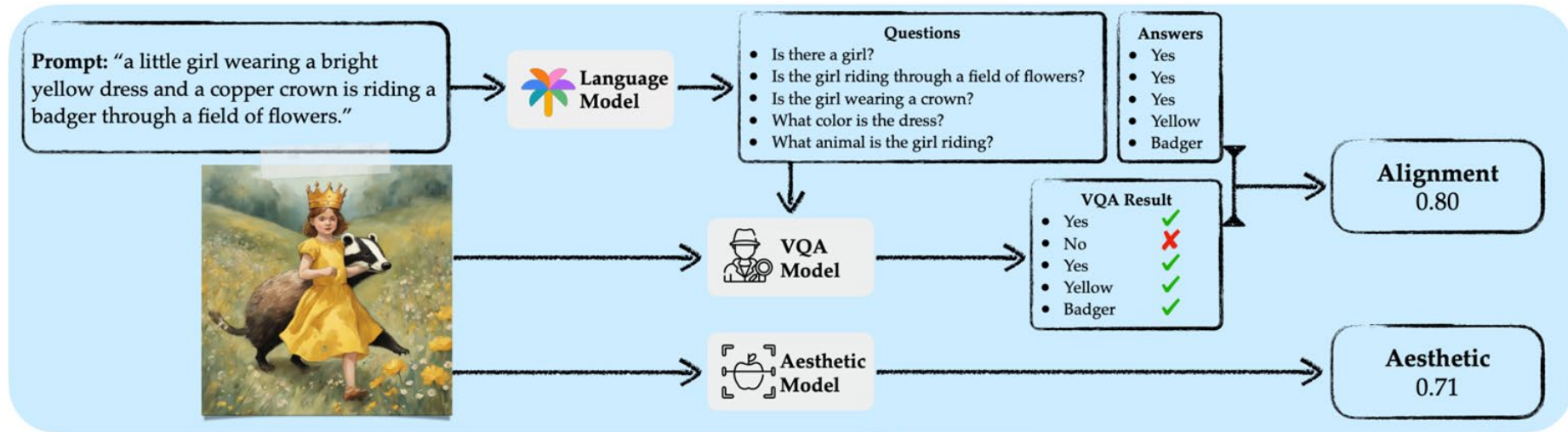
Answers

- Yes
- Yes
- Yes
- Yellow
- Badger

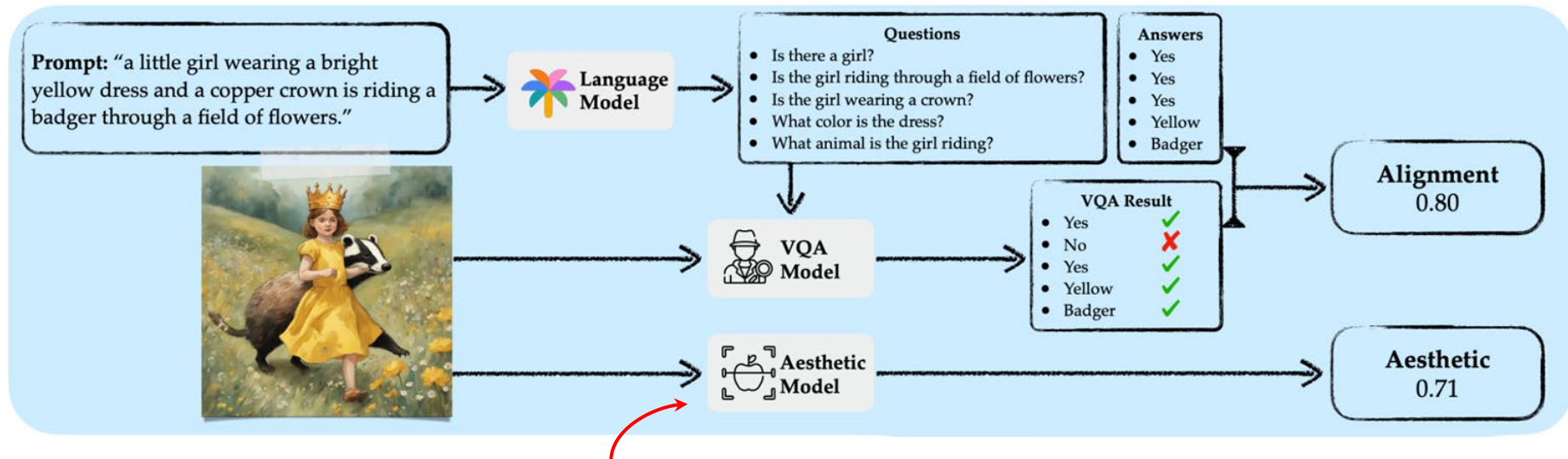
LLM as Evaluator: DreamSync



LLM as Evaluator: DreamSync

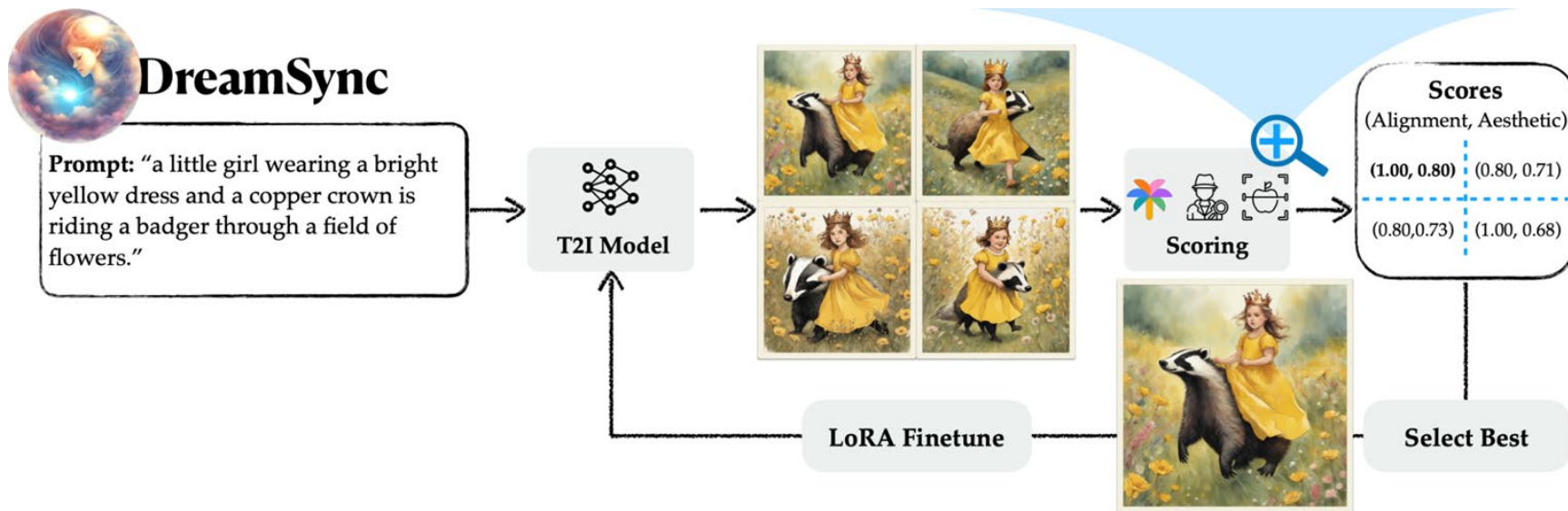


LLM as Evaluator: DreamSync



VILA (Learning Image Aesthetics from User Comments) – VLM, similar to CLIP, trained on a subset of LAION and an AVA dataset of 250K images, marked up with ratings and **comments on aesthetic**

LLM as Evaluator: DreamSync



LLM as Evaluator: DreamSync

SDXL



SDXL + DreamSync



An apple sitting on a chair made of marshmallow,
in a park, at sunrise



One wooden heart and one marble heart

7

LLM as Agents

LLM as agents for generative
AI for **end-to-end** pipeline



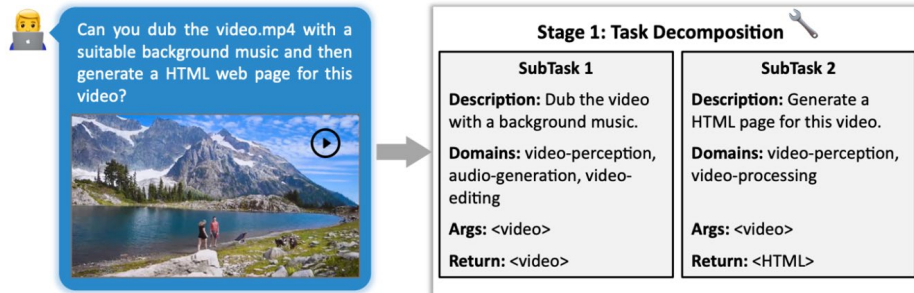
LLM as Agent: ControlLLM



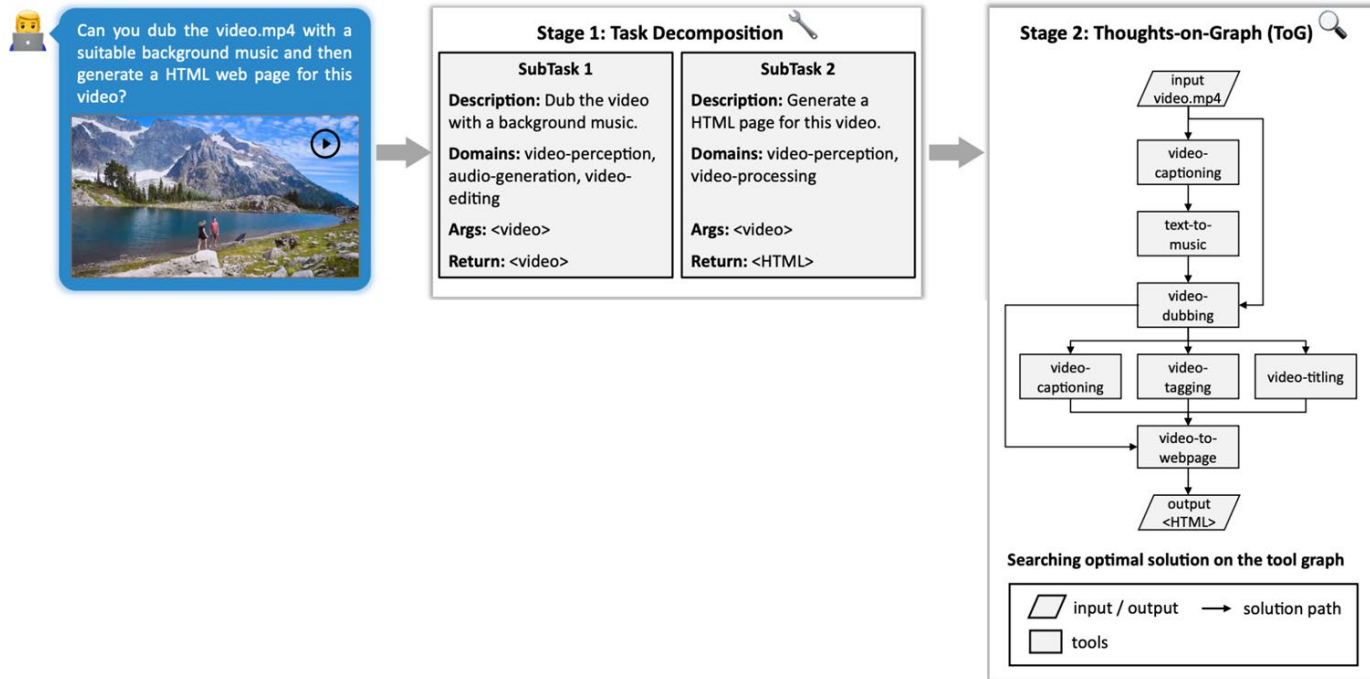
Can you dub the video.mp4 with a suitable background music and then generate a HTML web page for this video?



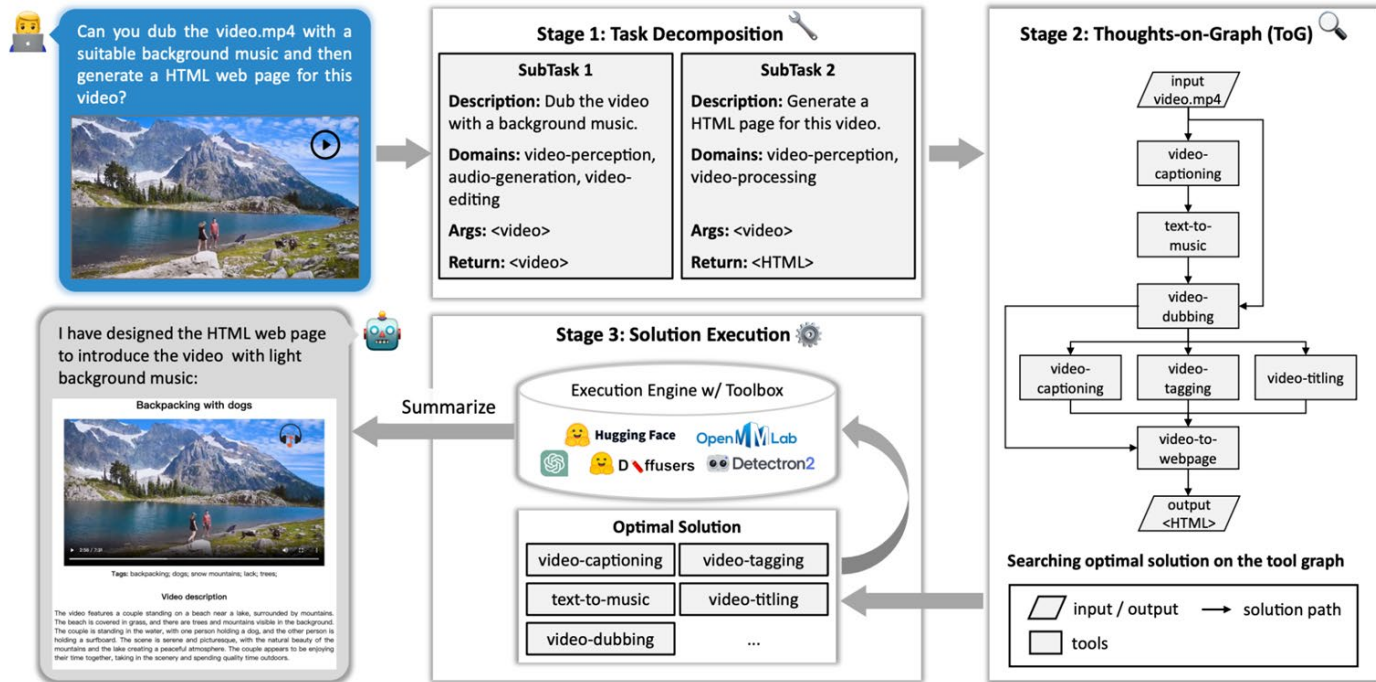
LLM as Agent: ControlLLM



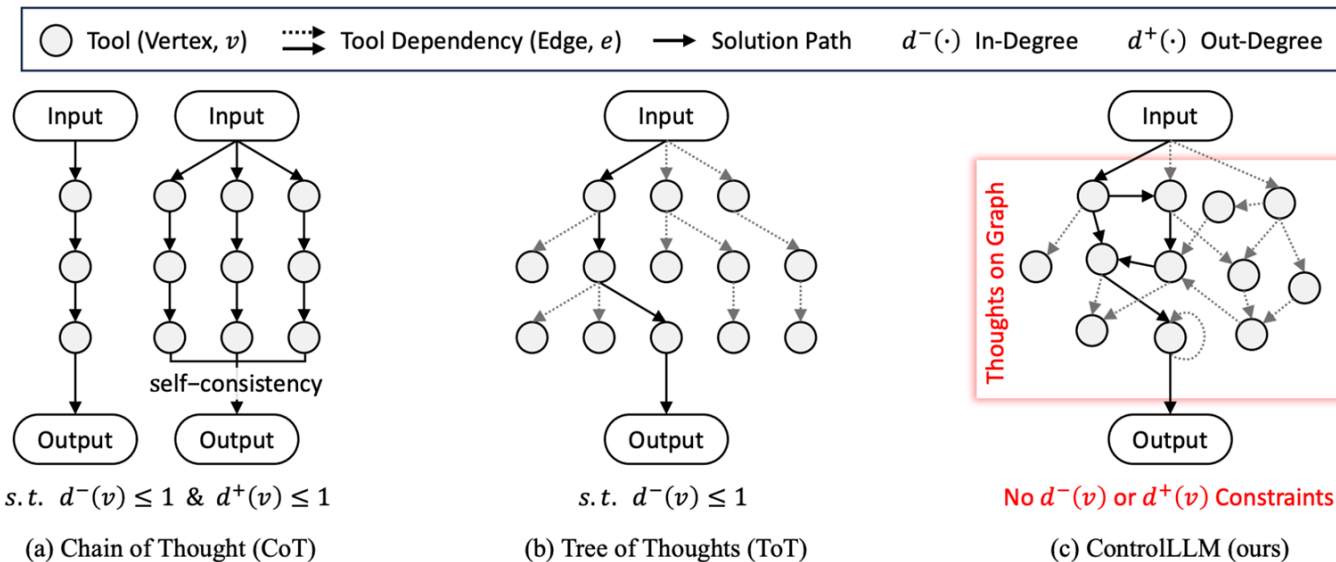
LLM as Agent: ControlLLM



LLM as Agent: ControlLLM



LLM as Agent: ControlLLM



ControlLLM

explores possible **outgoing edges**,
 evaluates which edge **best**
advances toward
 the final goal,
prunes dead-end
 branches

LLM as Agent: Movie Agent



After the catastrophe, Ne Zha and Ao Bing's souls survived, but their bodies were on the verge of shattering. Tai Yi planned to use the seven-colored lotus to restore them. However, Sheng Gongbao led the dragon clan and other sea demons to besiege Chentang Pass, seeking revenge on Ne Zha. In the fierce battle, Ne Zha accidentally shattered Ao Bing's body. To save him, Ne Zha decided to travel to Yuxu Palace and take the Immortal Ascension Exam to obtain a way to revive Ao Bing....

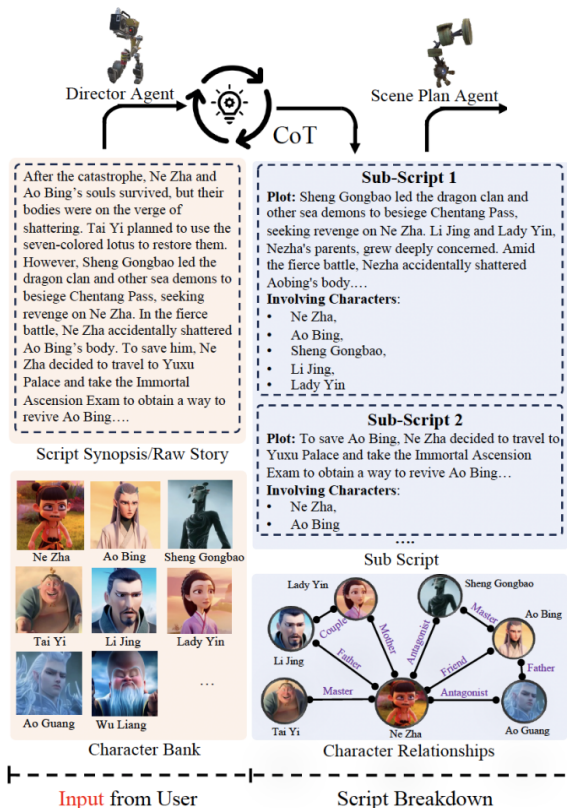
Script Synopsis/Raw Story



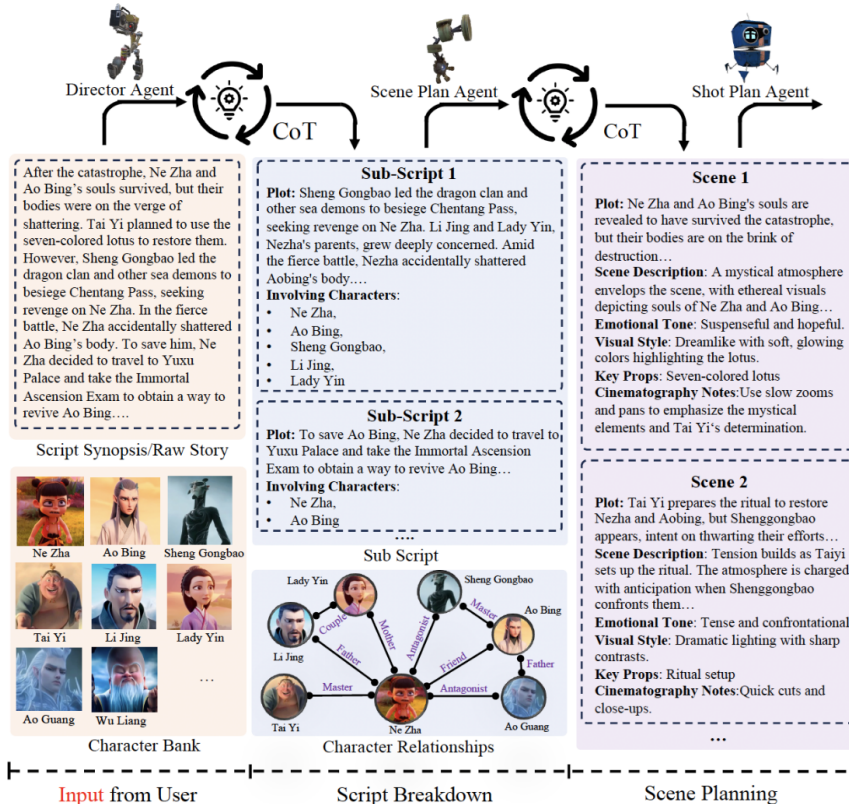
Character Bank

Input from User

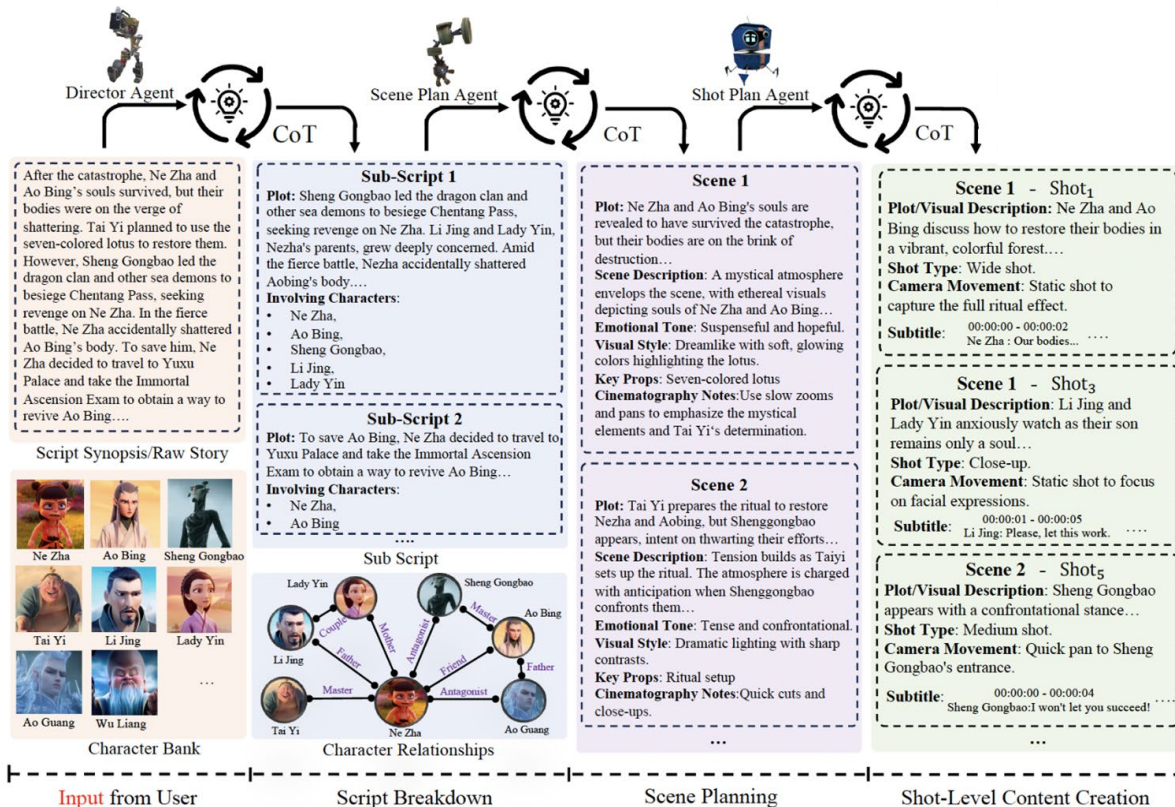
LLM as Agent: Movie Agent



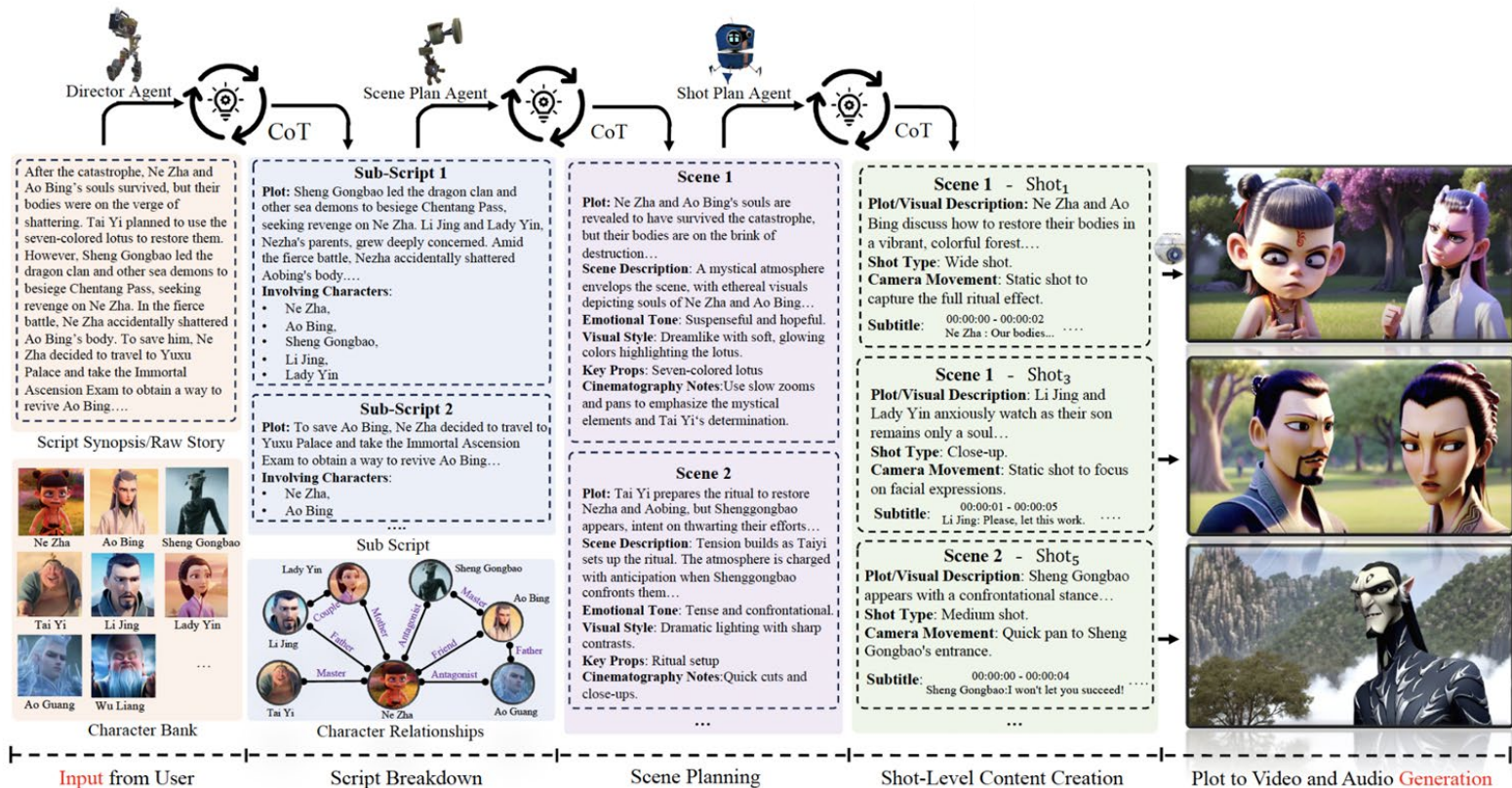
LLM as Agent: Movie Agent



LLM as Agent: Movie Agent



LLM as Agent: Movie Agent




LLM as Agent: Movie Agent

Input from User


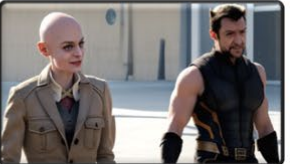


After a mysterious temporal anomaly disrupts reality, Deadpool finds himself unexpectedly entangled with Paradox, a rogue time traveler manipulating history for his own gain. As Deadpool attempts to fix the chaos in his usual reckless manner, Emma, a brilliant strategist with a hidden past, joins the mission, offering both intellect and combat skills. Meanwhile, Wolverine, drawn into the conflict...

Script Synopsis/Raw Story



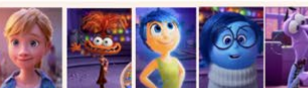
Character Bank

Plot, Video and Audio Generation





<p style="text-align: center;">Scene 1 - Shot₁</p> <p>Plot: Deadpool looked around cautiously...</p> <p>Camera Movement: Tracking around the room to highlight tech.</p> <p style="text-align: center;">00:00:00 - 00:00:02</p> <p>Subtitle: Deadpool: Wow, you got more tech than Stark...</p>	<p style="text-align: center;">Scene 2 - Shot₂</p> <p>Plot: Emma and Wolverine work together against Paradox's device...</p> <p>Camera Movement: Spin around shot...</p> <p style="text-align: center;">00:00:02 - 00:00:04</p> <p>Subtitle: Emma: Distract him while I disable this</p>	<p style="text-align: center;">Scene 2 - Shot₃</p> <p>Plot: Deadpool frantically uses random tools amidst chaos...</p> <p>Camera Movement: Dynamic pans and quick zooms.</p> <p style="text-align: center;">00:00:00 - 00:00:03</p> <p>Subtitle: Gong bao: Onward to Chentang Pass!</p>	<p style="text-align: center;">Scene 3 - Shot₅</p> <p>Plot: Close-up of Wolverine's face, eyes showing turmoil...</p> <p>Camera Movement: Static to focus on Wolverine's expression...</p> <p style="text-align: center;">00:00:00 - 00:00:04</p> <p>Subtitle: Wolverine: What... what is happening to me?</p>
			

As Riley enters her teenage years, her mind's Headquarters undergoes a sudden transformation, welcoming new Emotions. Anxiety takes charge, believing that constant worry will help Riley navigate adolescence, while Envy fuels self-doubt. Joy, Sadness, Anger, Fear, and Disgust struggle to adapt as Anxiety's influence grows, sidelining their roles...

Script Synopsis/Raw Story



Character Bank

<p style="text-align: center;">Scene 1 - Shot₁</p> <p>Plot: Joy looks towards Riley with a reassuring smile, while Riley's face reflects gratitude and calm...</p> <p>Camera Movement: Static Camera.</p> <p style="text-align: center;">00:00:00 - 00:00:03</p> <p>Subtitle: Joy: We'll always be here for you.</p>	<p style="text-align: center;">Scene 1 - Shot₂</p> <p>Plot: Joy confronts Envy at the center of Headquarters, trying to reassert control...</p> <p>Camera Movement: Quick cuts with handheld swaying...</p> <p style="text-align: center;">00:00:00 - 00:00:02</p> <p>Subtitle: Joy: Envy, stop disrupting everything!</p>	<p style="text-align: center;">Scene 2 - Shot₃</p> <p>Plot: Anger and Fear are discussing how each can contribute positively to Riley's emotional balance.</p> <p>Camera Movement: Gentle pan capturing.</p> <p style="text-align: center;">00:00:01 - 00:00:03</p> <p>Subtitle: Anger: Stay calm</p>	<p style="text-align: center;">Scene 5 - Shot₂</p> <p>Plot: Disgust and Anxiety are at the console, coordinating their efforts...</p> <p>Camera Movement: Gentle pan capturing.</p> <p style="text-align: center;">00:00:01 - 00:00:05</p> <p>Subtitle: Disgust: Let's make sure everything stays in good taste.</p>
			

LLM as Agent: Movie Agent, Task 1

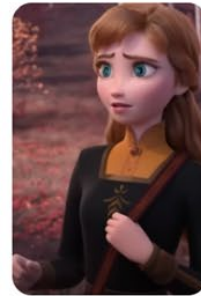
Script Synopsis to Movie Generation

Anna, Elsa, Kristoff, Olaf, and Mattias embark on a journey to uncover the truth behind the mysterious voice calling Elsa. As they travel to the enchanted forest, they discover that the past holds secrets about their kingdom and Elsa's powers. Mattias, a loyal Arendelle soldier trapped in the forest for years, helps them navigate the tensions between Arendelle and the Northuldra people...

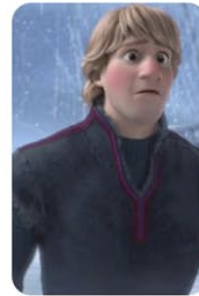
Script Synopsis/Raw Story



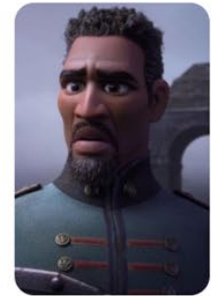
Elsa



Anna



Kristoff



Mattias

LLM as Agent: Movie Agent, Task 2

Sub-Script 1: The Call and the Journey Begins

Plot: Anna, Elsa, Kristoff, and Olaf embark on a journey toward a mysterious voice calling Elsa. Along the way, supernatural weather linked to Elsa's emotions challenges them. In the enchanted forest, they encounter frozen relics of a past conflict between Arendelle and the Northuldra. Elsa's magic calms elemental spirits blocking their path, and they uncover cryptic symbols tied to Elsa's childhood memories.

Involving Characters: Elsa, Anna, Kristoff, and Olaf.

Emotional Tone: Curiosity and determination.

Key Themes: Adventure, discovery, and mystery.

Scene 1 - Shot 1

Plot: Elsa stands alone in her ice dress, listening to a mysterious, ethereal voice. Her expression is a mix of curiosity and concern as she looks towards the distant horizon.

Involving Characters: Elsa.

Shot Type: Close-up.

Camera Movement: Static shot.

Subtitle: {Elsa: I can hear it again... the voice. It's calling me.}



Scene 1 - Shot 3

Plot: A wide shot reveals the enchanted forest bathed in ethereal light. Glowing plants and shimmering air create a magical atmosphere, emphasizing the vastness and wonder of the setting.

Involving Characters:

Shot Type: Wide shot.

Camera Movement: Static shot.

Subtitle: {}



Scene 1 - Shot 4

Plot: Anna and Elsa continue their conversation, now focusing on the maps. They lean over a large map, pointing out key locations and discussing the route.

Involving Characters: Elsa, Anna.

Shot Type: Medium close-up.

Camera Movement: Dolly-in.

Subtitle: {Anna: This path should lead us to the mountains. Elsa: Let's mark it clearly.}



Scene 1 - Shot 6

Plot: Anna, Elsa and Kristoff set out through a dense forest, their travel gear visible. Wide shot emphasizes the vast and challenging landscape they face.

Involving Characters: Elsa, Anna, Kristoff.

Shot Type: Wide shot.

Camera Movement: Static shot.

Subtitle: {Anna: Here we go, everyone. Stick together. Kristoff: Stay close, it's easy to get lost here.}



LLM as Agent: Movie Agent, Task 3

Process with **MovieAgent** (GPT4-o + ROICtrl + VALL-E X + Mango Talking Human)

Anna, Elsa, Kristoff, Olaf, and Mattias embark on a journey to uncover the truth behind the mysterious voice calling Elsa. As they travel to the enchanted forest, they discover that the past holds secrets about their kingdom and Elsa's powers. Mattias, a loyal Arendelle soldier trapped in the forest for years, helps them navigate the tensions between Arendelle and the Northuldra people...

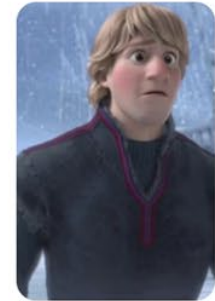
Script Synopsis/Raw Story



Elsa



Anna



Kristoff

LLM as Agent: Movie Agent, Task 4

Scene 1 - Shot 1

Plot: Elsa stands alone in her ice dress, listening to a mysterious, ethereal voice. Her expression is a mix of curiosity and concern as she looks towards the distant horizon.

Involving Characters: Elsa.

Subtitle:

Elsa: I can hear it again... the voice. It's calling me. I have to follow it... I have to know the truth.



Scene 6 - Shot 2

Plot: Group shot of Anna and Kristoff sharing a meal with representatives, smiles and nods exchanged as they reach a moment of understanding.

Involving Characters: Kristoff, Anna.

Anna: We can learn so much from each other. We just need to listen and work together.

Kristoff: Anna, this is dangerous. Are you sure about this?

Anna: I know it won't be easy, but we can't turn back now. Trust me, Kristoff.

Kristoff: You never stop surprising me... Okay, I trust you.

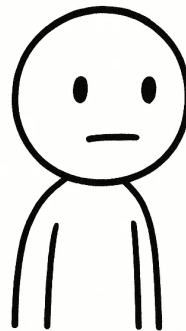


Conclusions

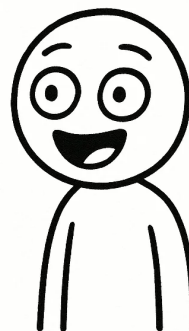
- 1 LLMs have found many applications in **generation tasks**
- 2 Due to more a **priori information** , we can generate images using LLM. However, still a **lack of locality** : small changes in the picture greatly perturb the picture as a whole
- 3 Prompt decomposition, LLM as evaluator, improved conditioning, and for sure it **will all expand further**

this is the end of
the lecture

ME



THEM



tomorrow is the
another lecture

