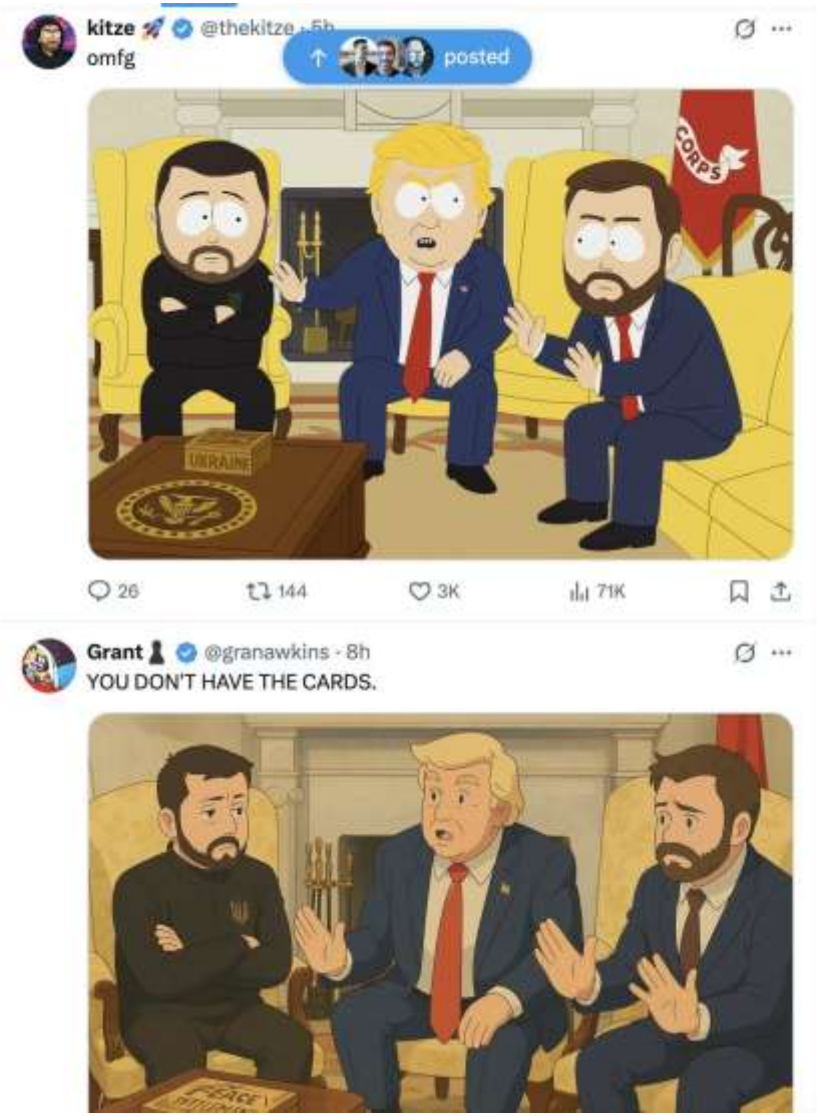
Курс «Мультимодальные БЯМ»

Тема №3

«БЯМ и Генерация»

Антон Конушин

In Ghibli style



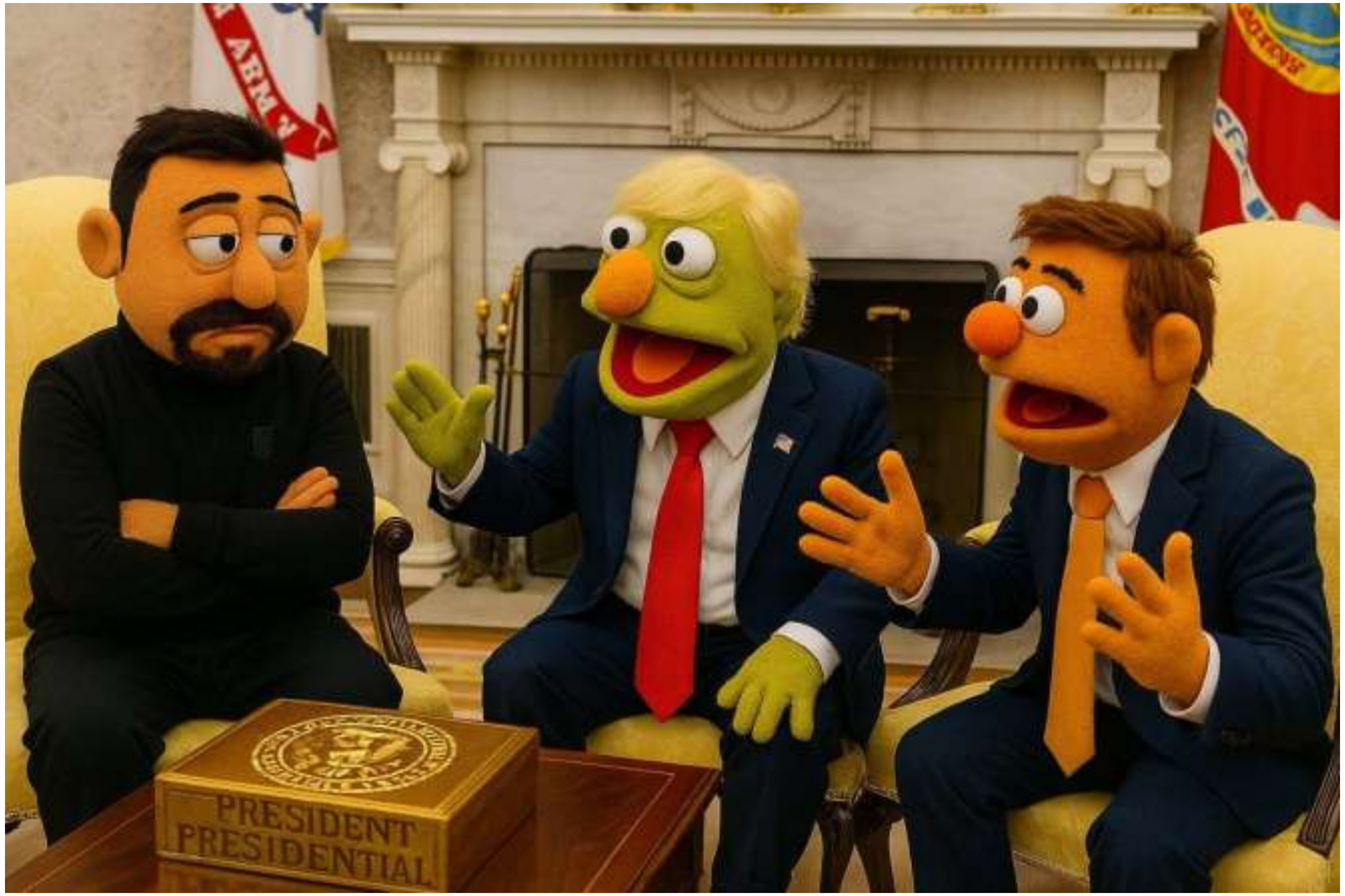
kitze 🚀 @thekitze · 5h

omfg

26 144 3K 71K

Grant @granawkins · 8h

YOU DON'T HAVE THE CARDS.



https://www.lesswrong.com/posts/vmcWCsoyPNsGz82nt/fun-with-gpt-4o-image-generation

In Ghibli style

НАУЧНЫЙ ПРЕМИЯ ГОДА
НАУЧНЫЙ ПРЕМЕР ГОДА
НАУЧНЫЙ ПРЕМЬЕР ГОДА
НАУЧНЫЙ ПРЕМИЯ ГОДА

Что лежит в основе мощи GenAI?

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

haha gpus go brrrrrrrr

- Достижения БЯМ (GPT), генерации картинок, видео и 3D моделей (Stable Diffusion, Flux, Sora, …) можно объяснить масштабному генеративному предобучению

Технологические волны предобучения

Image Generation: From GANs to Diffusion Models

StyleGAN 2019

DDPM 2020

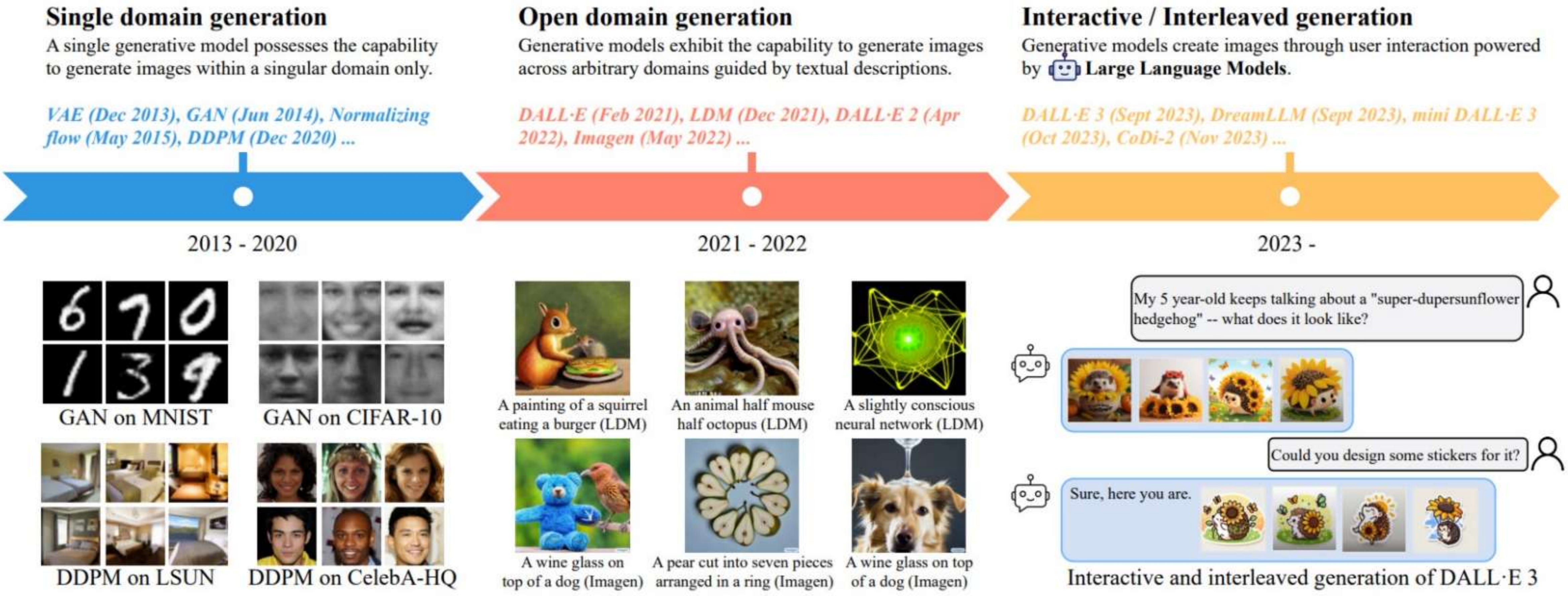
(Denoising Diffusion Probabilistic Models)

Guided Diffusion 2021

(Diffusion Beat GANs)

Генерация изображений

Single domain -> Open Domain -> Free-form / Interactive



Single domain generation

A single generative model possesses the capability to generate images within a singular domain only.

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

2013 - 2020

GAN on MNIST

GAN on CIFAR-10

DDPM on LSUN

DDPM on CelebA-HQ

Open domain generation

Generative models exhibit the capability to generate images across arbitrary domains guided by textual descriptions.

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

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)

Interactive / Interleaved generation

Generative models create images through user interaction powered by **Large Language Models**.

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

2023 -

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

Could you design some stickers for it?

Sure, here you are.

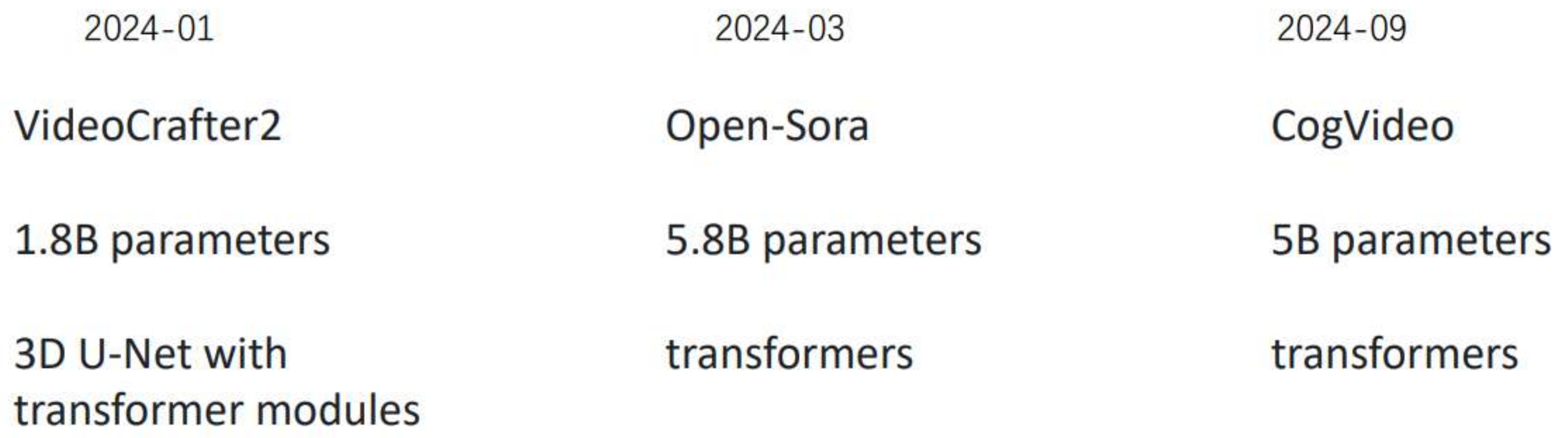
Interactive and interleaved generation of DALL·E 3

Рост параметров моделей

Increasing model parameters of text-to-image models

Stable Diffusion	Stable Diffusion 3	Flux
~1B parameters	8B parameters	12B parameters
U-Net with transformer modules	transformers	transformers

Рост параметров моделей



2024-01	2024-03	2024-09
VideoCrafter2	Open-Sora	CogVideo
1.8B parameters	5.8B parameters	5B parameters
3D U-Net with transformer modules	transformers	transformers

Интеграция моделей

- Масштабное предобучение помогло генеративным моделям во всех модальностях
- Как их можно интегрировать? Как БЯМ / ВБЯМ могут помочь генерации в других модальностях?

БЯМ / ВБЯМ

- GPT4
- LLaMA
- LLaVA

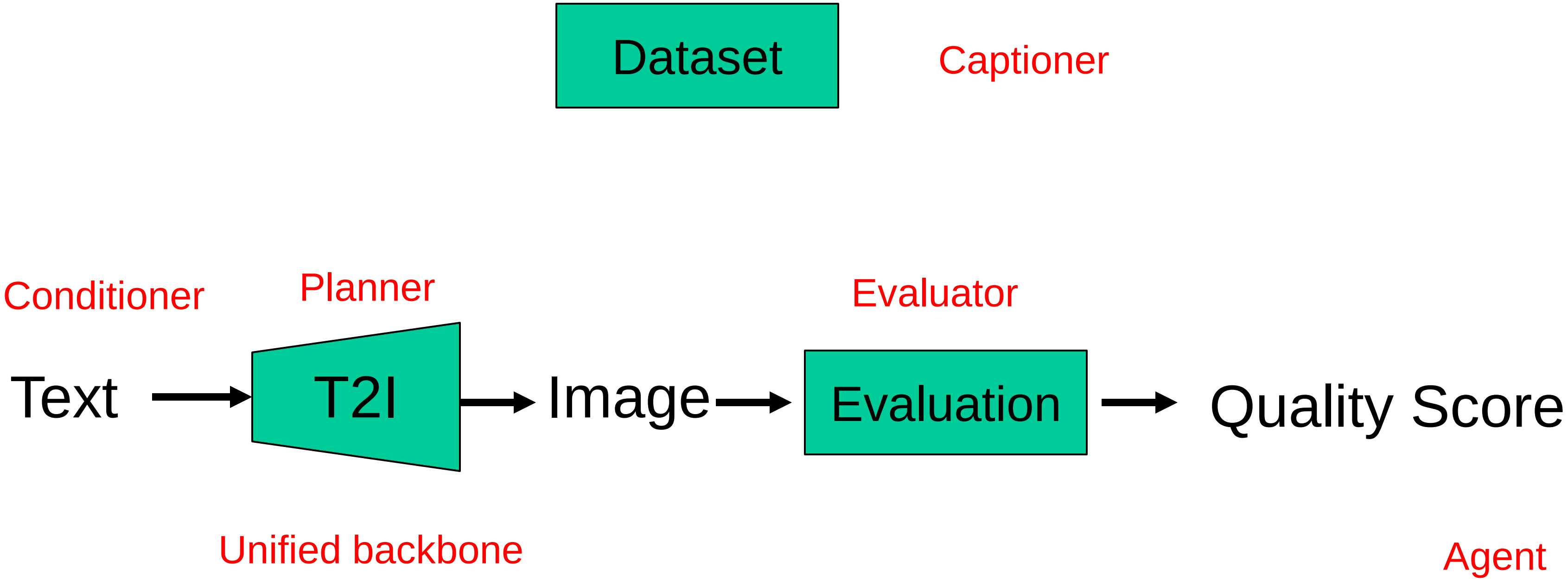
Text2Image

- StableDiffusion
- FLUX

Text2Video

- VideoCrafter
- ModelScope
- AnimateDiff
- Open-Sora
- Cogvideo

Общая схема генерации и место БЯМ



Dataset

Captioner

Conditioner

Planner

Evaluator

Text → T2I → Image → Evaluation → Quality Score

Unified backbone

Agent

Tool X

Tool X

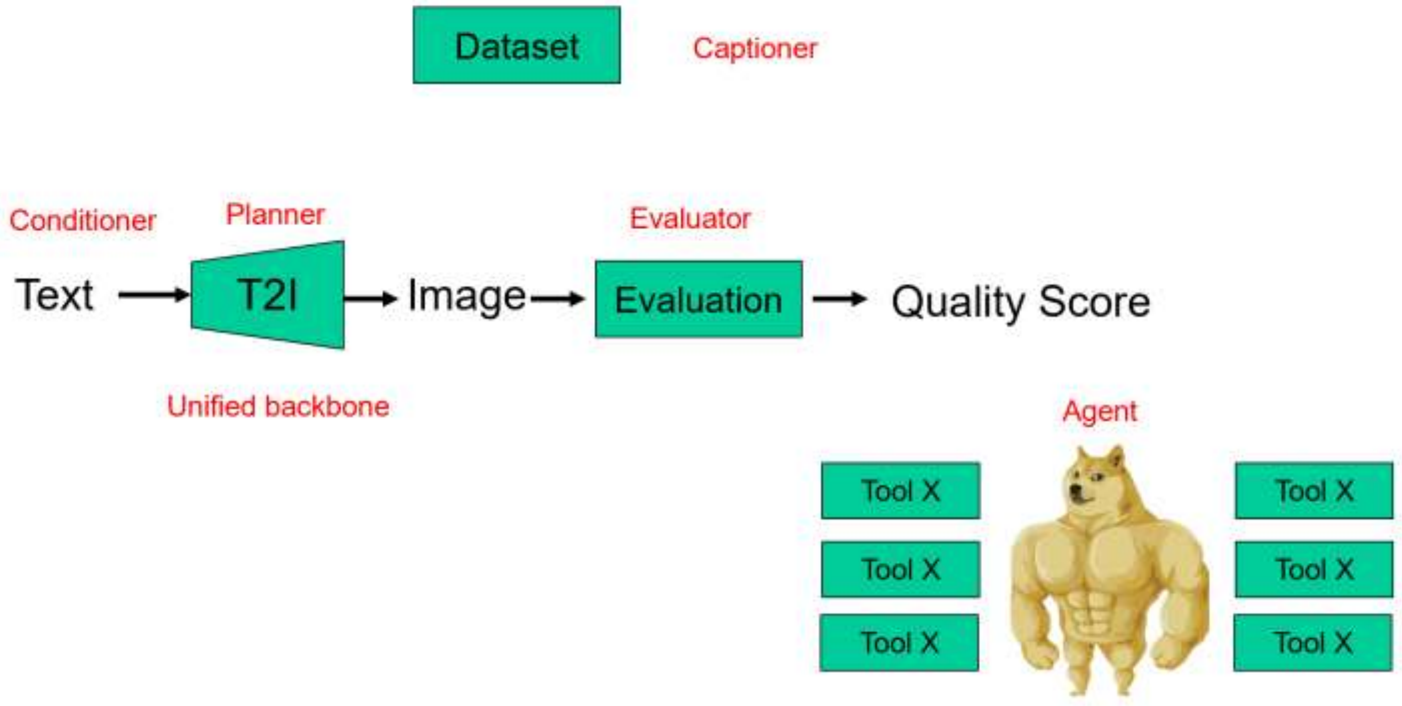
Tool X

Tool X

Tool X

Tool X

Место БЯМ в генерации

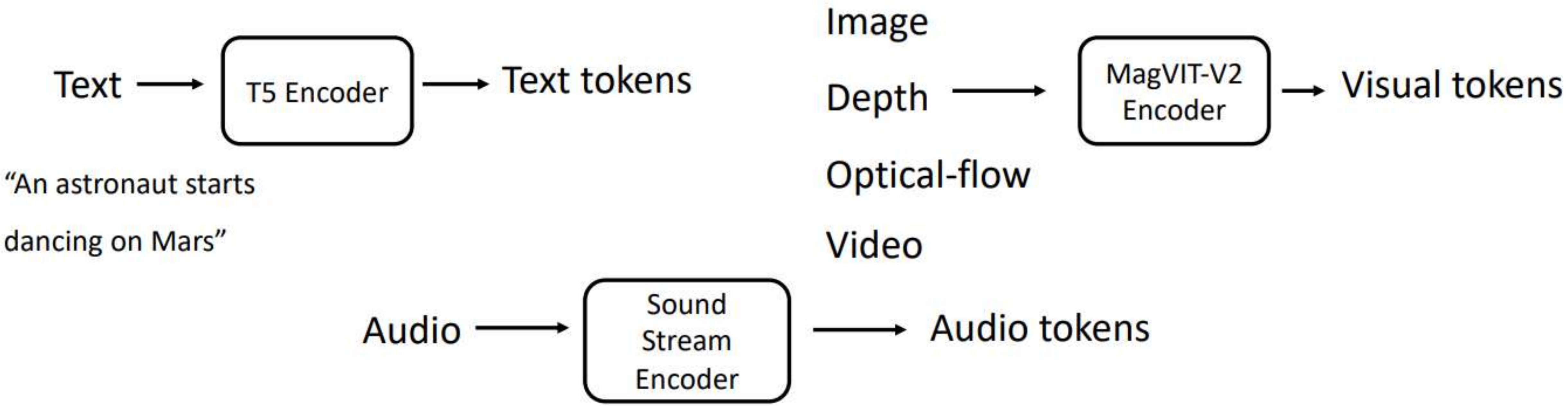


- Unified backbone
- Planner
- Captioner
- Conditioner
- Evaluator
- Agent

БЯМ как универсальная основа

БЯМ как универсальная основа

- Все модальности удобно отобразить в пространство дискретных токенов
- Воспользуемся архитектурой БЯМ для предсказания следующих токенов



Text → T5 Encoder → Text tokens

"An astronaut starts dancing on Mars"

Image

Depth → MagVIT-V2 Encoder → Visual tokens

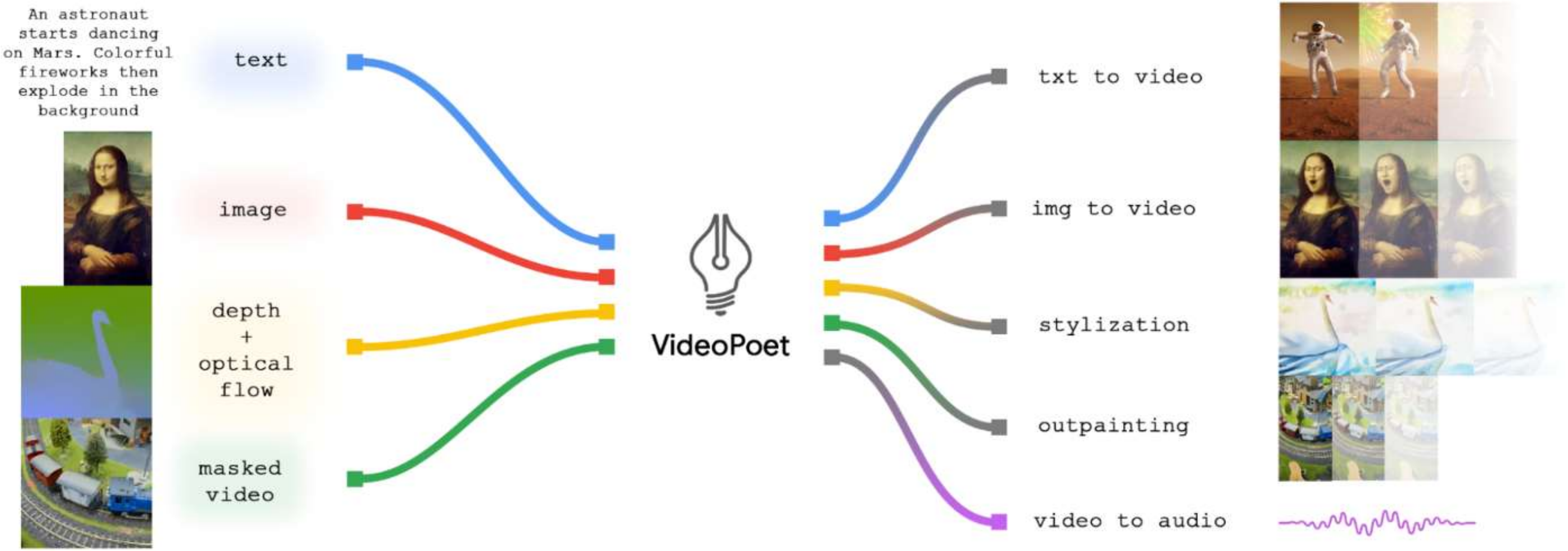
Optical-flow

Video

Audio → Sound Stream Encoder → Audio tokens

(2023-12) VideoPoet: A Large Language Model for Zero-Shot Video Generation [ICML 2024]

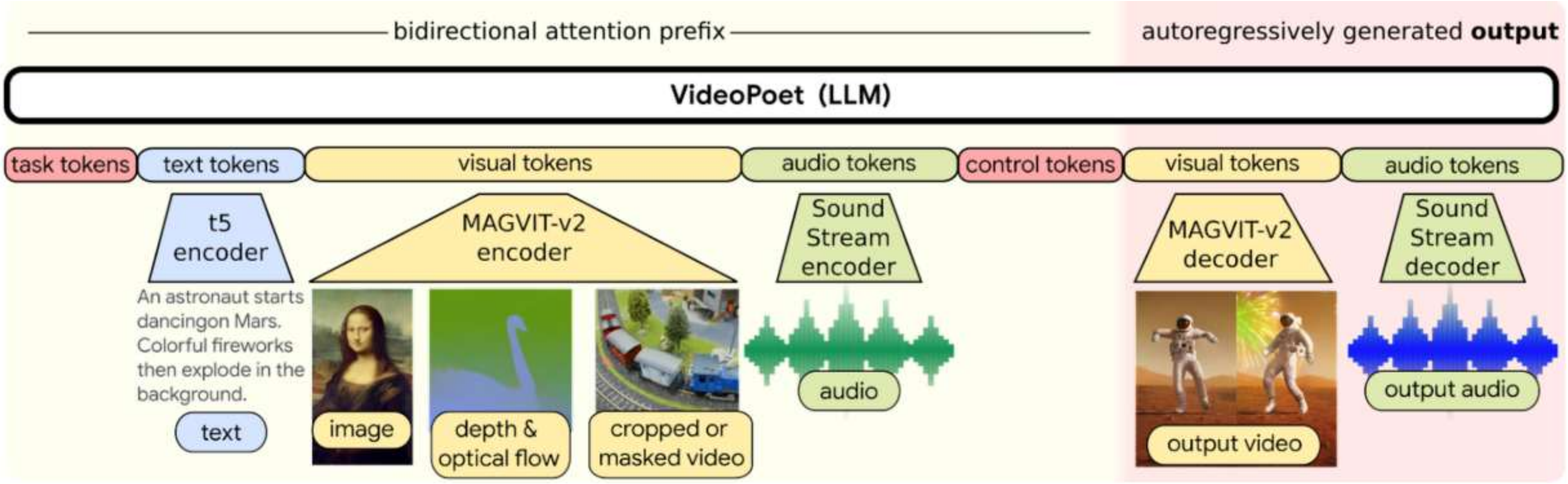
Схема VideoPoet



(2023-12) VideoPoet: A Large Language Model for Zero-Shot Video Generation [ICML 2024]

Схема VideoPoet

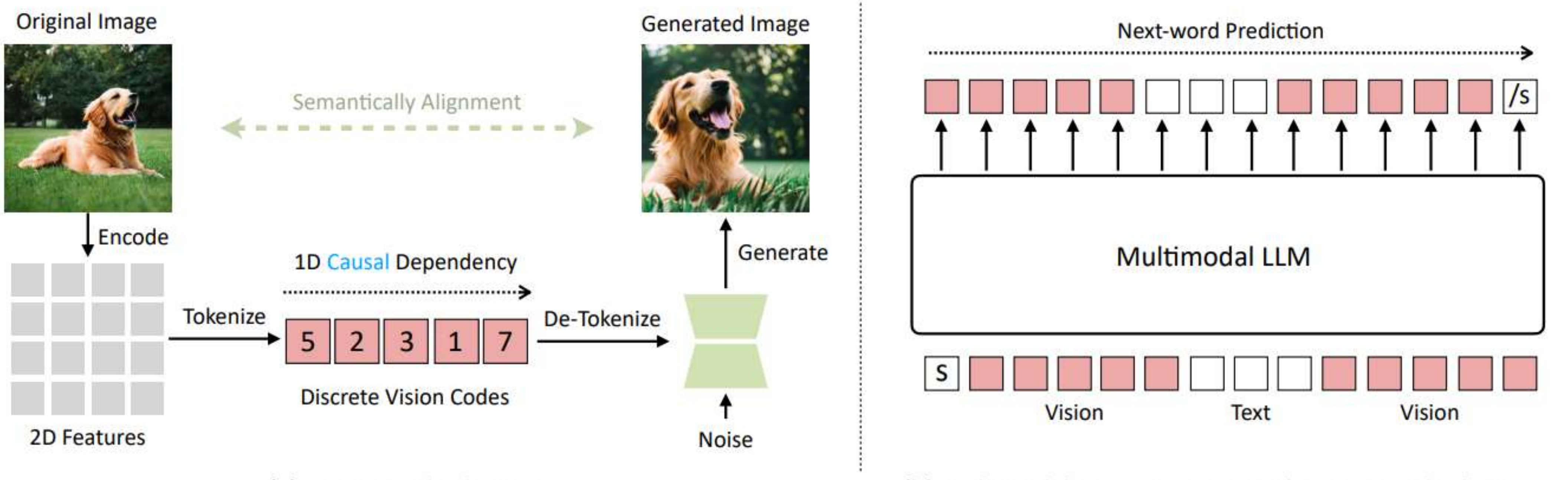
- Авторегрессионно предсказываем токены
- Декодируем их с помощью доменных декодеров



(2023-12) VideoPoet: A Large Language Model for Zero-Shot Video Generation [ICML 2024]

Токенизаторы и де-токенизаторы

Чтобы БЯМ было удобнее работать с картиночными токенами, они должны) иметь (1) 1D casual зависимость (2) высокоуровневую семантику



(a) SEED Visual Tokenizer

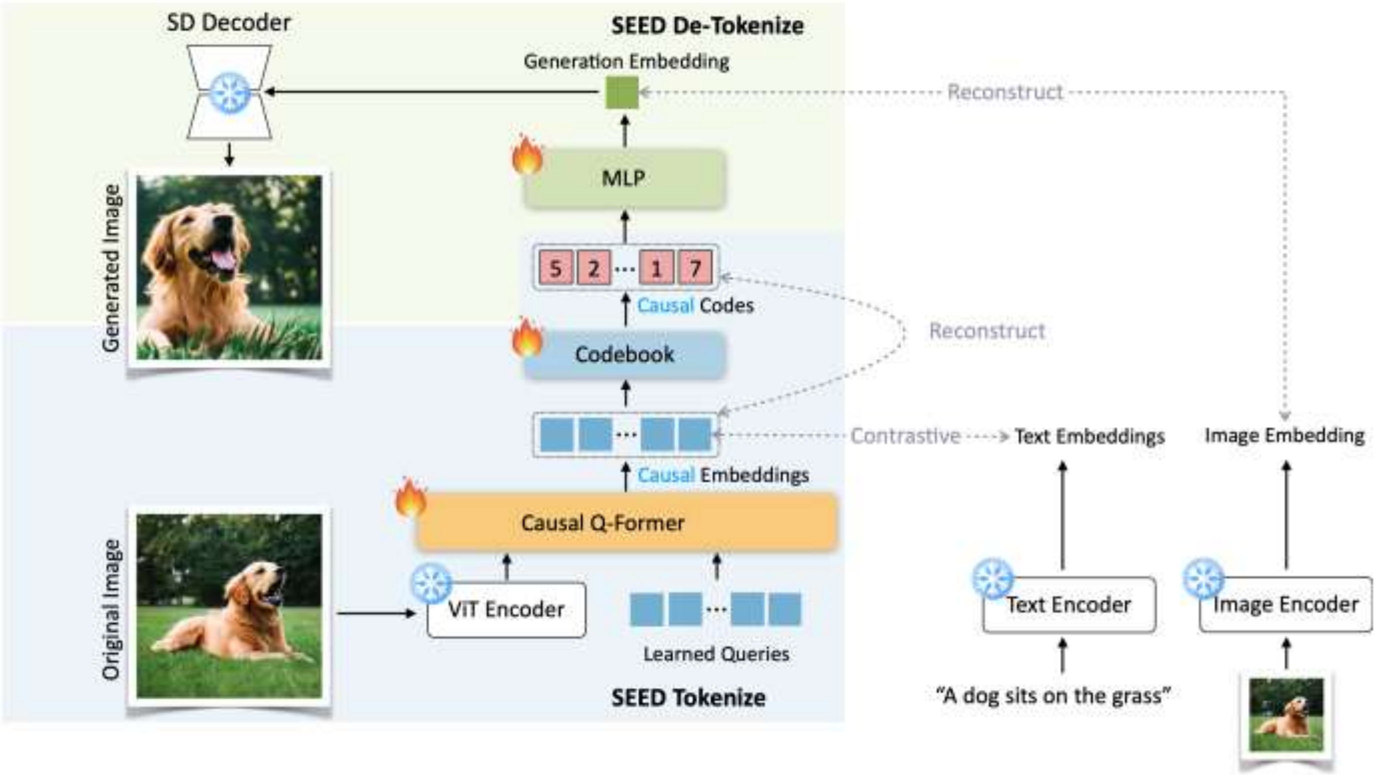
(b) Multimodal Autoregression with SEED visual tokens

Figure 2: (a) SEED is a discrete image tokenizer, producing quantized visual codes with 1D causal dependency and high-level semantics. (b) With SEED tokenizer, LLM is able to perform scalable multimodal autoregression on interleaved visual and textual data with next-word-prediction objective.

(2023-10) [Making LLaMA See and Draw with SEED Tokenize][ICLR 2024]

Токенизаторы и де-токенизаторы

- Схема SEED



Training stages

1. Causal Q-Former,convert visual features to causal embeddings, contrastive loss
2. Train a visual codebook to discretize the causal embeddings to discrete codes, maximize causal embedding cosine similarity

(2023-10) Making LLaMA See and Draw with SEED Tokenize [ICLR 2024]

Токенизаторы и де-токенизаторы

- Примеры работы

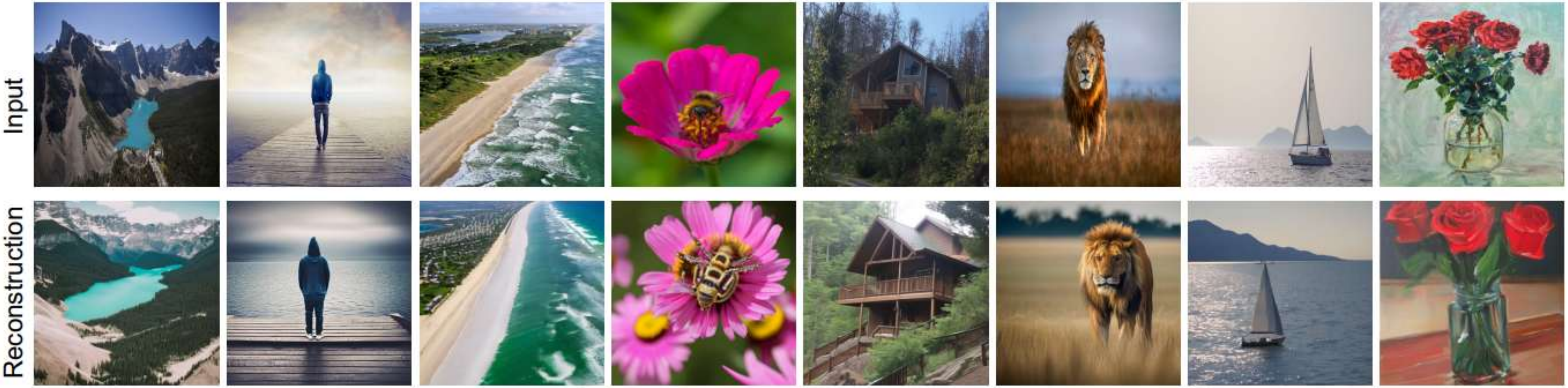
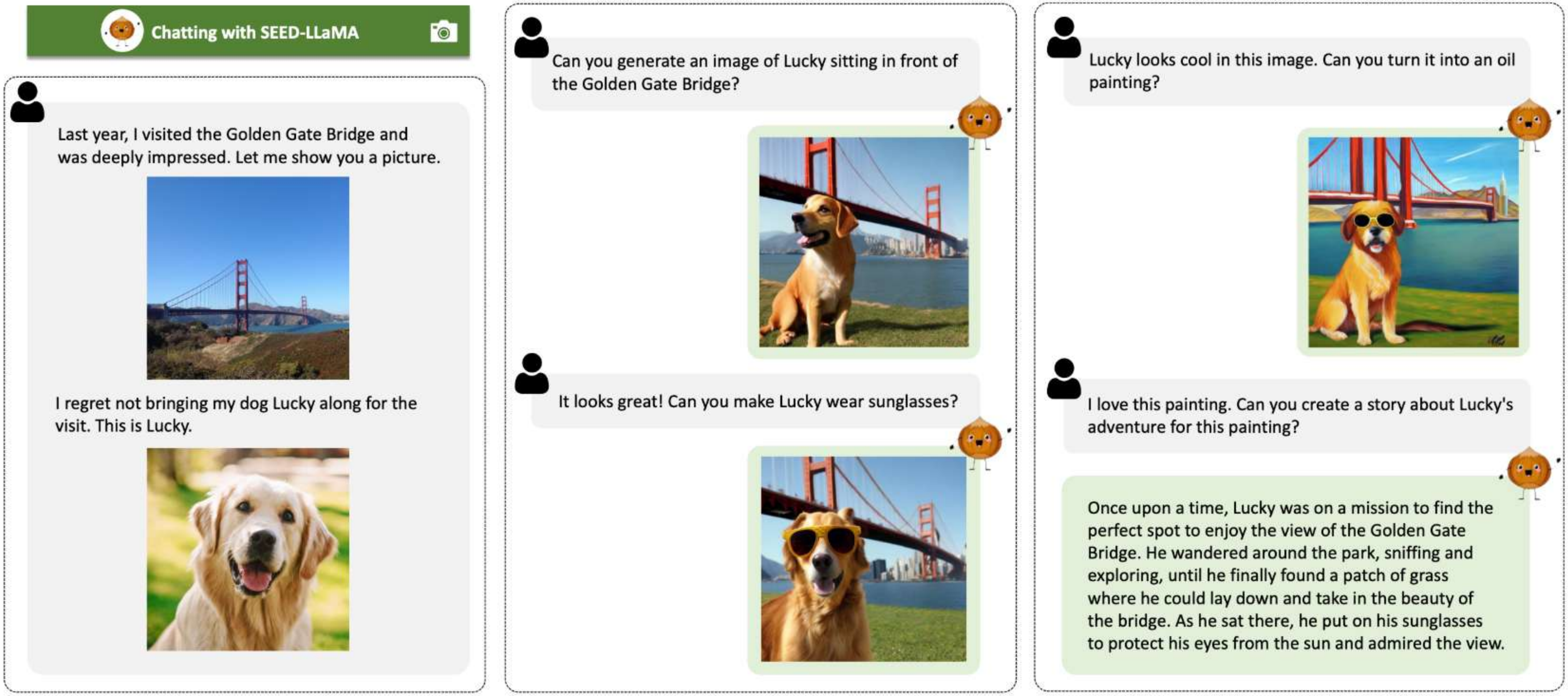


Figure 5: Reconstruction images of SEED tokenizer (*i.e.*, original image → SEED tokenize → causal visual codes → SEED de-tokenize → reconstructed image).

(2023-10) [Making LLaMA See and Draw with SEED Tokenize](#)[ICLR 2024]



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.

(2023-10) Making LLaMA See and Draw with SEED Tokenize [ICLR 2024]

Двухголовый Янус от DeepSeek

Одна голова хорошо, а две – лучше!

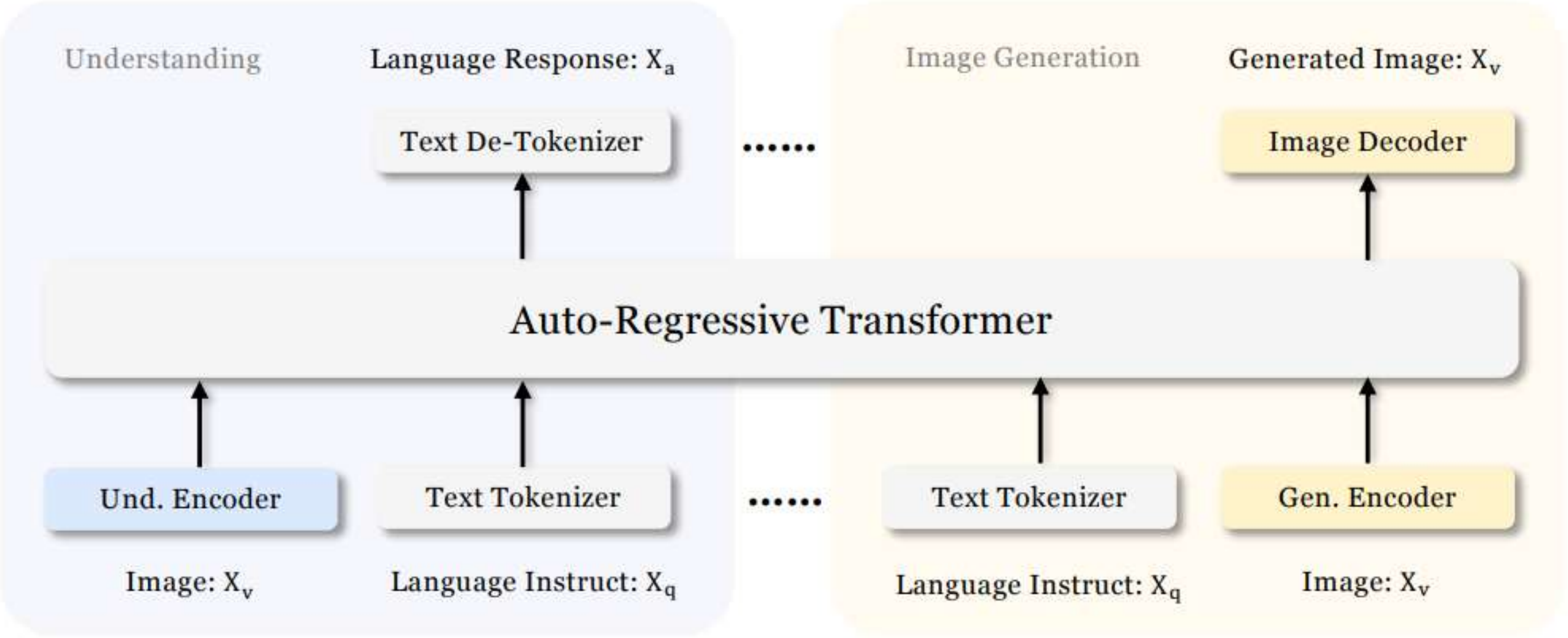


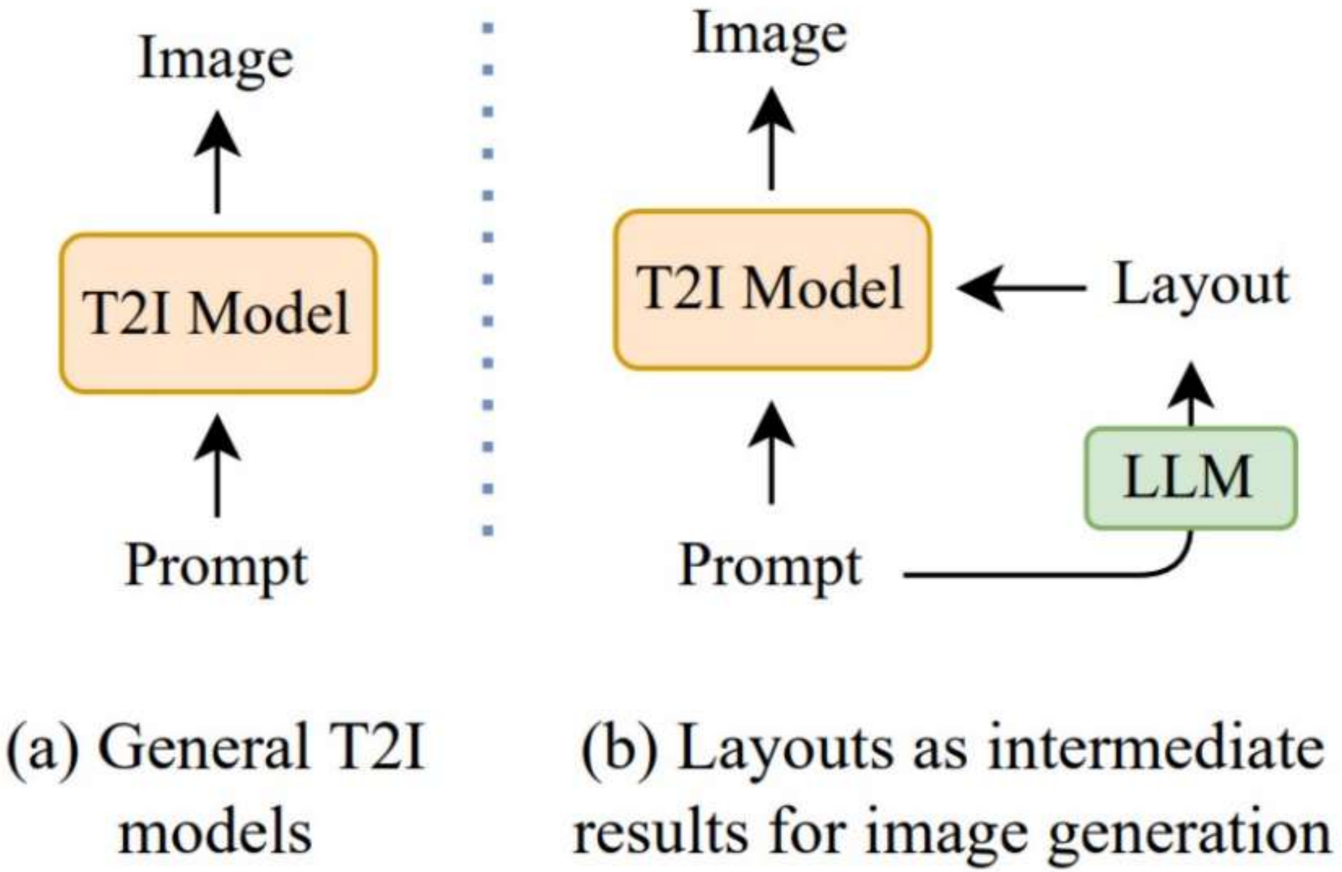
Figure 3 | **Architecture of our Janus-Pro.** We decouple visual encoding for multimodal understanding and visual generation. “Und. Encoder” and “Gen. Encoder” are abbreviations for “Understanding Encoder” and “Generation Encoder”, respectively. Best viewed on screen.

(2025-01) [Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling]

БЯМ-планировщик

БЯМ-планировщик

- 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



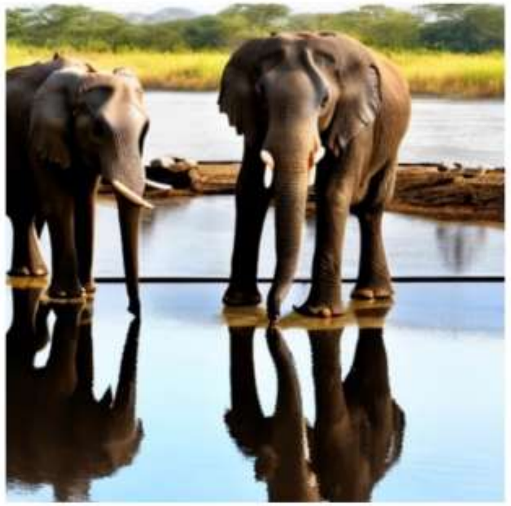
(a) General T2I models

(b) Layouts as intermediate results for image generation

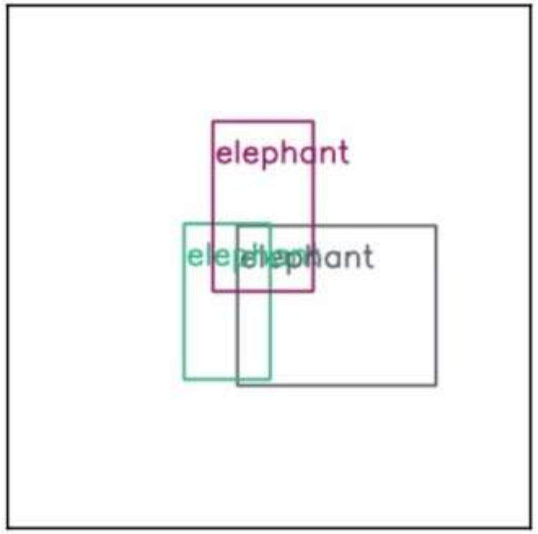
Планирование пространственного расположения

- Проверка правильного количества объектов

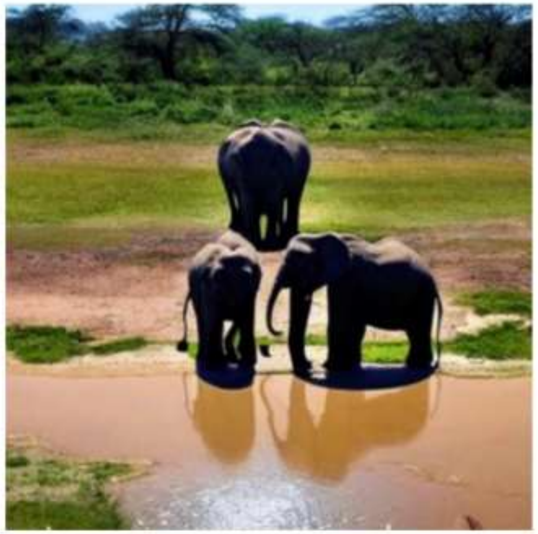
“There are three elephants standing beside a pool of water. ”



Text-to-Image



Text-to-Layout

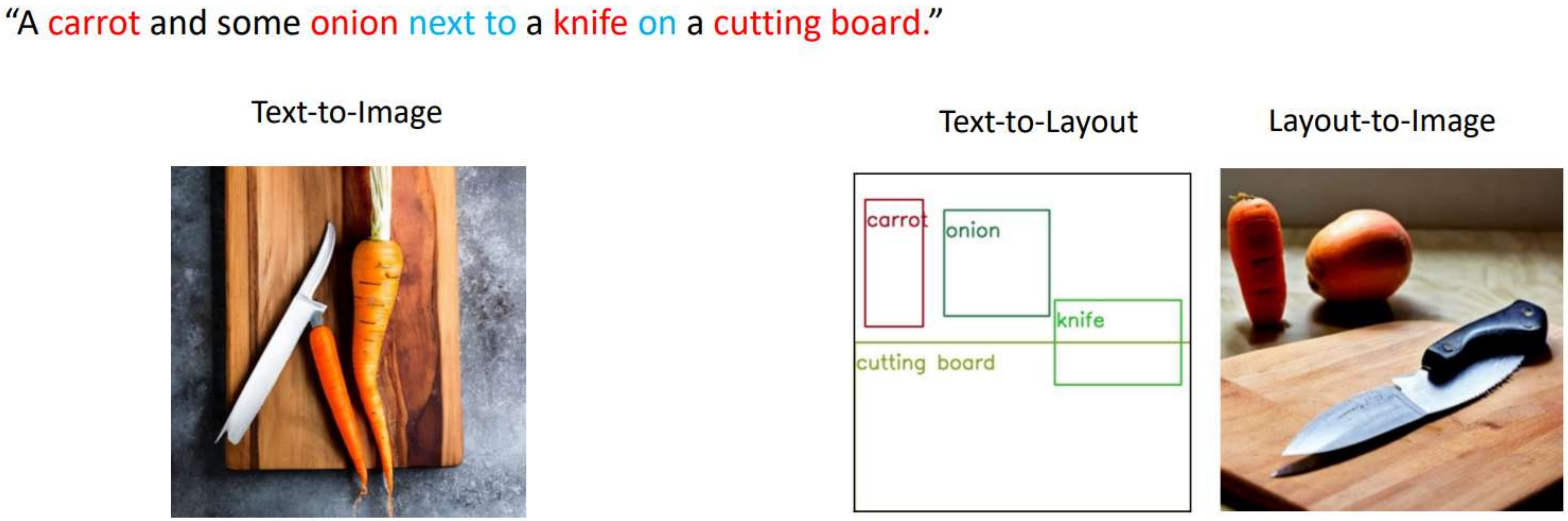


Layout-to-Image

[NeurIPS 2023] LayoutGPT: Compositional Visual Planning and Generation with Large Language Models

Планирование пространственного расположения

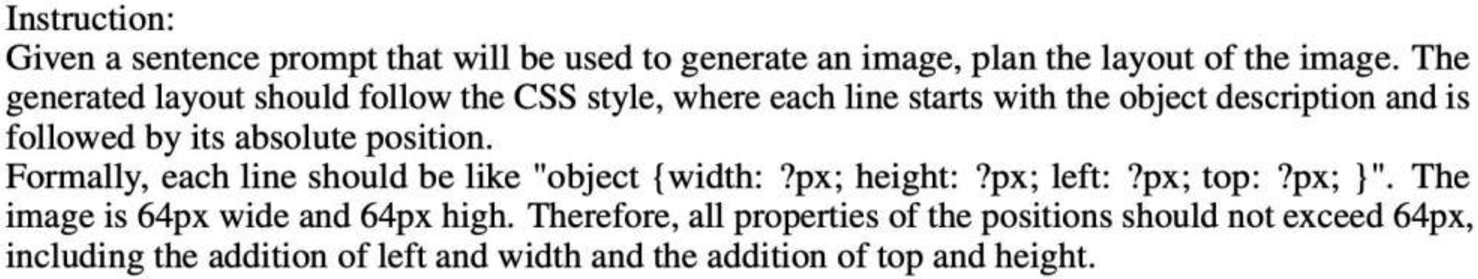
- Обеспечение правильного взаимного расположения объектов



[NeurIPS 2023] LayoutGPT: Compositional Visual Planning and Generation with Large Language Models

Планирование пространственного расположения

Task instruction



Instruction:
Given a sentence prompt that will be used to generate an image, plan the layout of the image. The generated layout should follow the CSS style, where each line starts with the object description and is followed by its absolute position.
Formally, each line should be like "object {width: ?px; height: ?px; left: ?px; top: ?px; }". The image is 64px wide and 64px high. Therefore, all properties of the positions should not exceed 64px, including the addition of left and width and the addition of top and height.

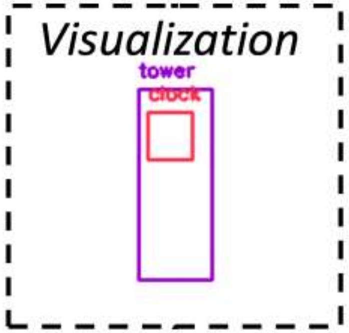
In-context learning Examples

```
/* first in-context exemplar*/
Prompt: A building with a tower on top
that contains a clock.
Layout:
        clock {
        height: 37px;
        width: 38px;
        top: 104px;
        left: 60px;
        }
/* second in-context exemplar*/
```

Inference prompt: Prompt: a tower of a building that has a clock on it.

```
/* LLM Output*/
tower {
   height: 149px;
   width: 56px;
   top: 65px;
   left: 100px;
}...
```

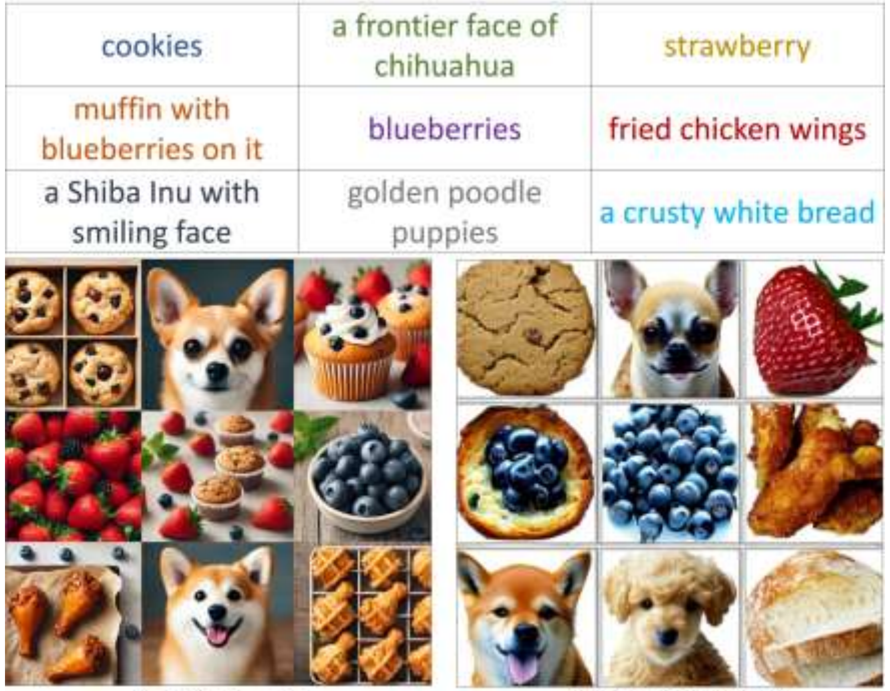
In CSS format



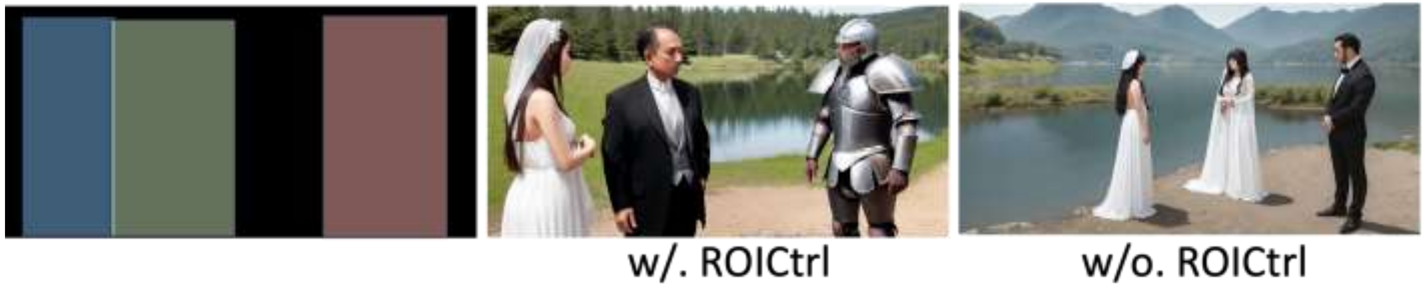
Layout-to-Image

[NeurIPS 2023] LayoutGPT: Compositional Visual Planning and Generation with Large Language Models

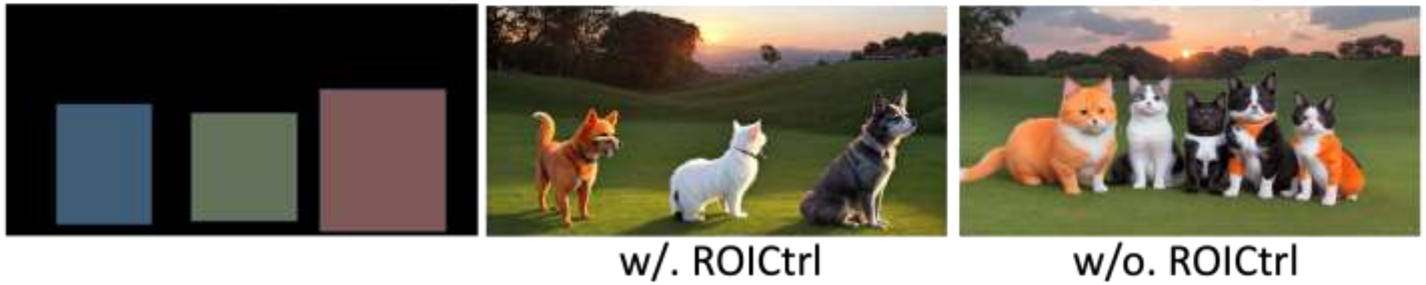
ROICtrl



(a) DALL-E 3 (b) ROICtrl (SD2)



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.

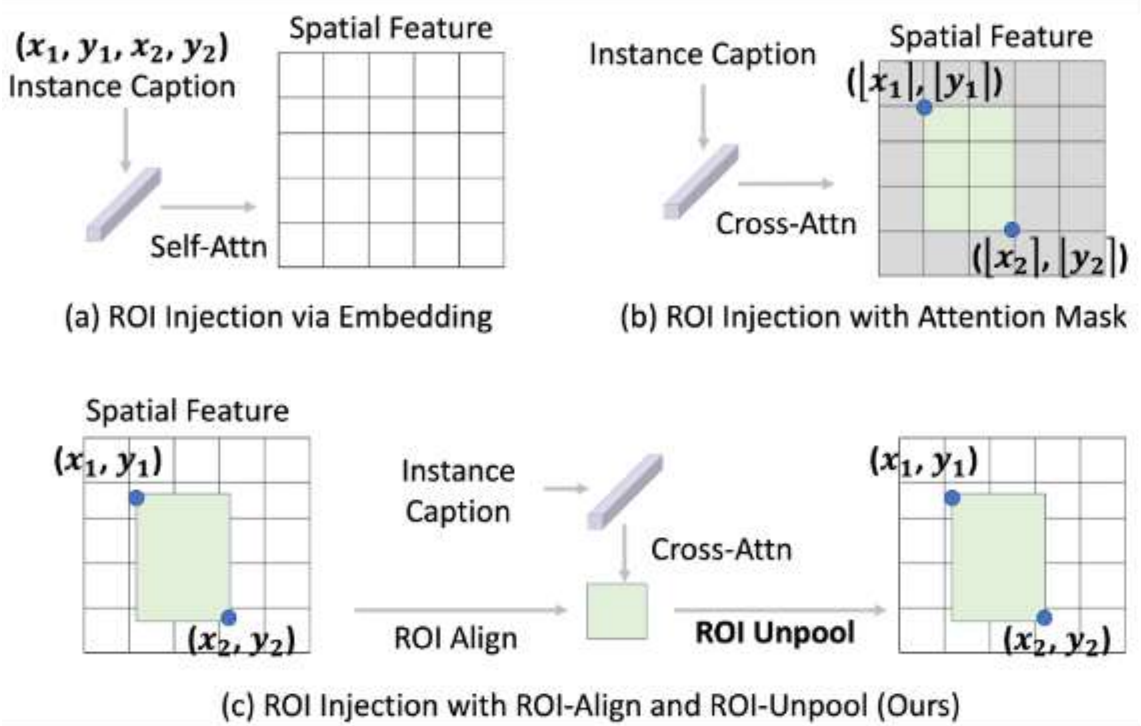


Global Caption: Two dogs and a cat on the grass.
ROI-1: A dog with orange fur. ROI-2: A cat with white fur. ROI-3: A dog with grey fur.

Управление пространственным размещением объектов при генерации

(11.2024) https://roictrl.github.io/

От ROI Align к ROI Unpool

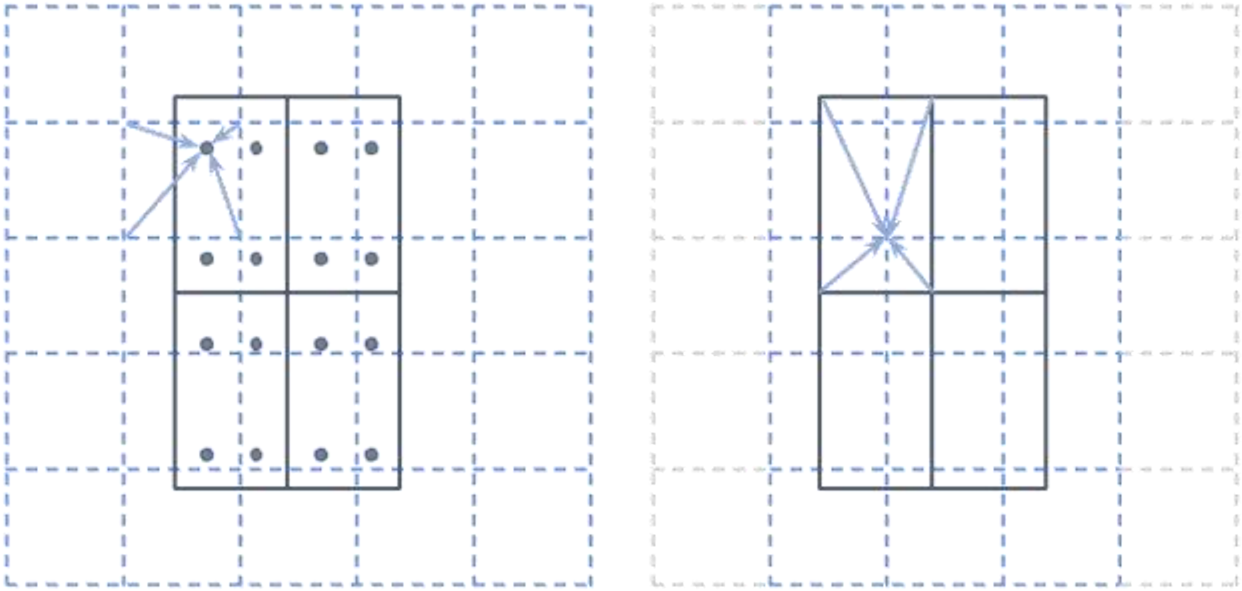


(a) ROI Injection via Embedding

(b) ROI Injection with Attention Mask

(c) ROI Injection with ROI-Align and ROI-Unpool (Ours)

Fig.1: Comparison with Various ROI Injections



(a) ROI-Align

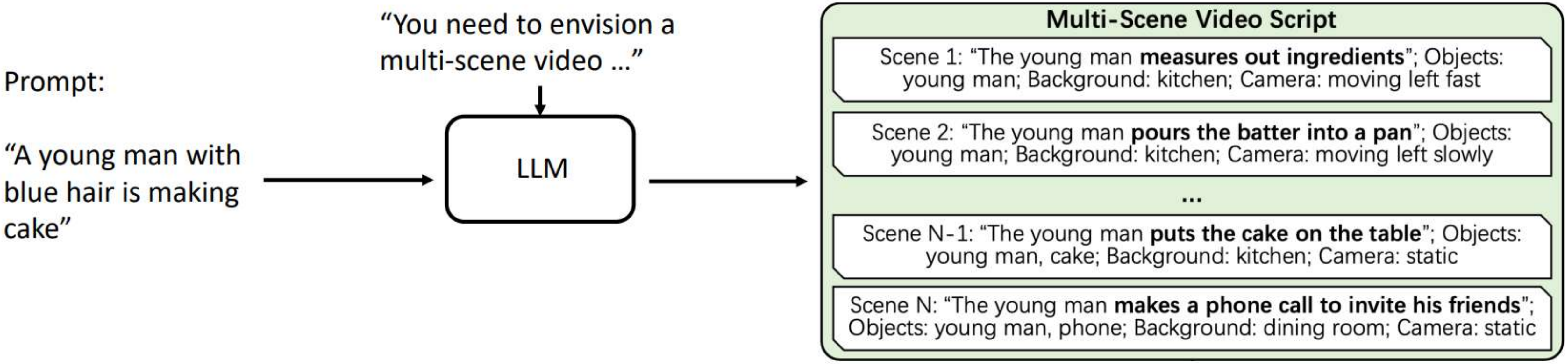
(b) ROI-Unpool

Fig. 2: ROI-Align and the complementary ROI-Unpool (Ours) for Visual Generation

[(11.2024) https://roictrl.github.io/](https://roictrl.github.io/)

Декомпозиция на сцены

- Декомпозиция промпта на несколько промптов



Prompt:

“A young man with blue hair is making cake”

“You need to envision a multi-scene video …”

Multi-Scene Video Script

Scene 1: “The young man **measures out ingredients**”; Objects: young man; Background: kitchen; Camera: moving left fast

Scene 2: “The young man **pours the batter into a pan**”; Objects: young man; Background: kitchen; Camera: moving left slowly

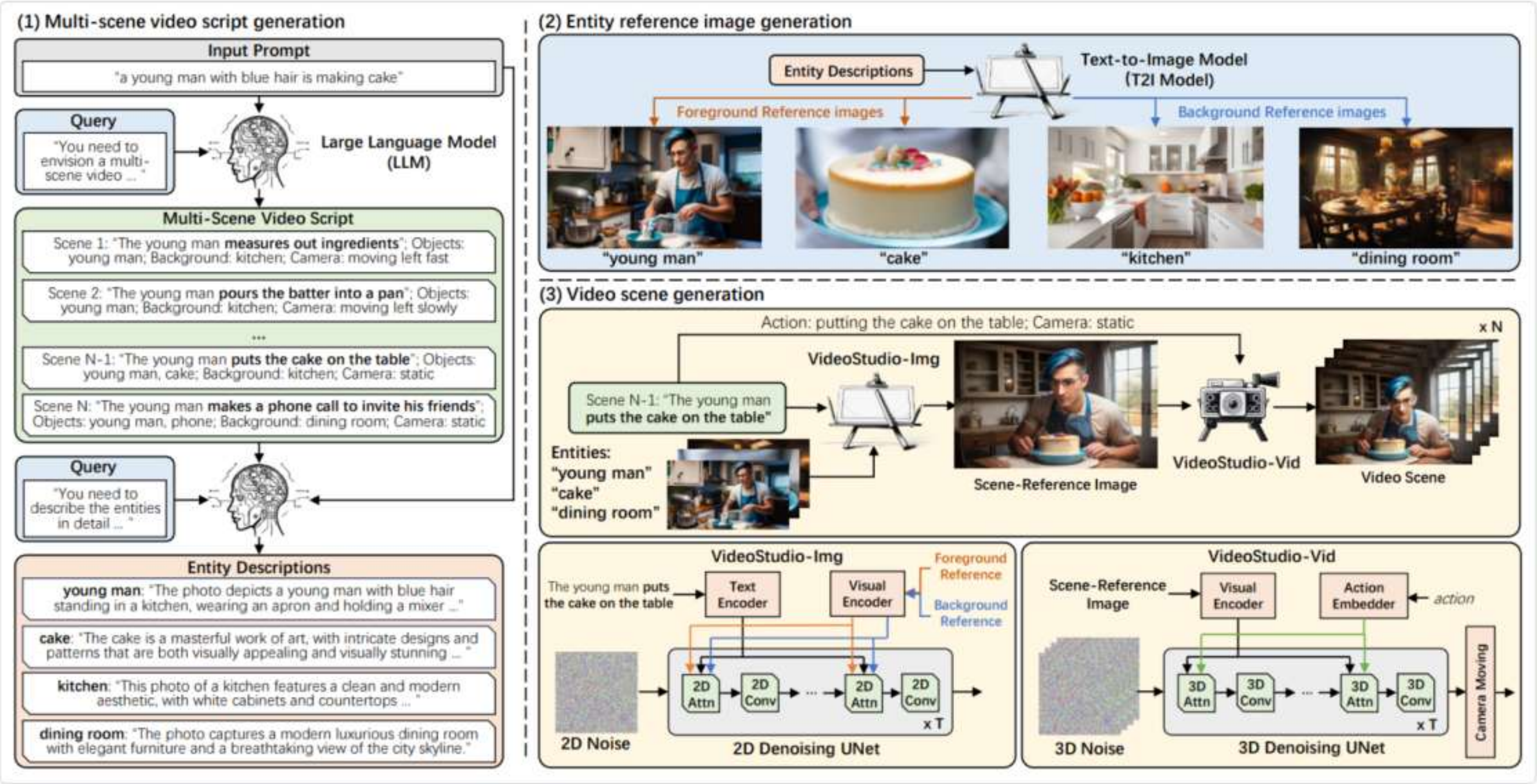
...

Scene N-1: “The young man **puts the cake on the table**”; Objects: young man, cake; Background: kitchen; Camera: static

Scene N: “The young man **makes a phone call to invite his friends**”; Objects: young man, phone; Background: dining room; Camera: static

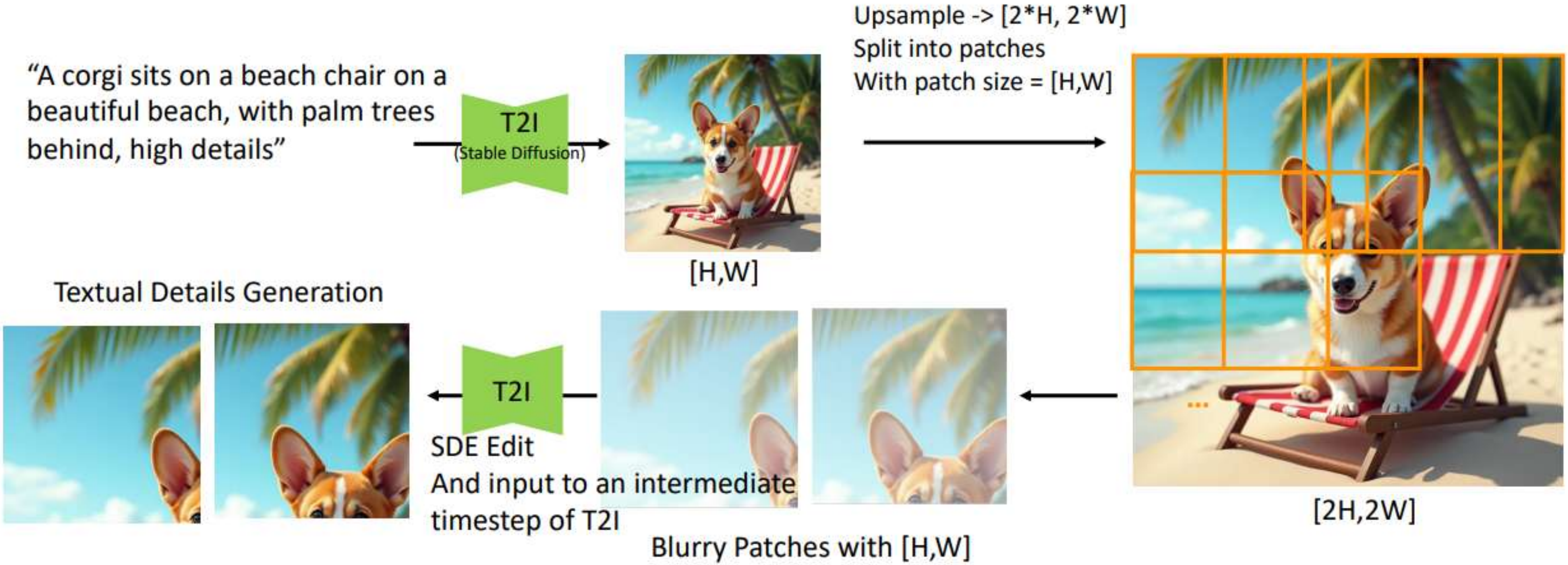
(2024-01) [VideoStudio: Generating Consistent-Content and Multi-Scene Videos]() [ECCV 2024]

Архитектура системы (уже фактические агент)


(1) Multi-scene video script generation
Input Prompt
"a young man with blue hair is making cake"
Query
"You need to envision a multi-scene video ..."
Large Language Model (LLM)
Multi-Scene Video Script
Scene 1: "The young man measures out ingredients"; Objects: young man; Background: kitchen; Camera: moving left fast
Scene 2: "The young man pours the batter into a pan"; Objects: young man; Background: kitchen; Camera: moving left slowly
...
Scene N-1: "The young man puts the cake on the table"; Objects: young man, cake; Background: kitchen; Camera: static
Scene N: "The young man makes a phone call to invite his friends"; Objects: young man, phone; Background: dining room; Camera: static
Query
"You need to describe the entities in detail ..."
Entity Descriptions
young man: "The photo depicts a young man with blue hair standing in a kitchen, wearing an apron and holding a mixer ..."
cake: "The cake is a masterful work of art, with intricate designs and patterns that are both visually appealing and visually stunning ..."
kitchen: "This photo of a kitchen features a clean and modern aesthetic, with white cabinets and countertops ..."
dining room: "The photo captures a modern luxurious dining room with elegant furniture and a breathtaking view of the city skyline."
(2) Entity reference image generation
Entity Descriptions
Text-to-Image Model (T2I Model)
Foreground Reference images
Background Reference images
"young man"
"cake"
"kitchen"
"dining room"
(3) Video scene generation
Action: putting the cake on the table; Camera: static
x N
VideoStudio-Img
Scene N-1: "The young man puts the cake on the table"
Entities:
"young man"
"cake"
"dining room"
Scene-Reference Image
VideoStudio-Vid
Video Scene
VideoStudio-Img
The young man puts the cake on the table
Text Encoder
Visual Encoder
Foreground Reference
Background Reference
2D Attn
2D Conv
...
2D Attn
2D Conv
x T
2D Noise
2D Denoising UNet
VideoStudio-Vid
Scene-Reference Image
Visual Encoder
Action Embedder
action
3D Attn
3D Conv
...
3D Attn
3D Conv
x T
Camera Moving
3D Noise
3D Denoising UNet

(2024-01) [VideoStudio: Generating Consistent-Content and Multi-Scene Videos] [ECCV 2024]

Иерархические промпты



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

T2I (Stable Diffusion)

[H,W]

Upsample -> [2*H, 2*W]
Split into patches
With patch size = [H,W]

[2H,2W]

Blurry Patches with [H,W]

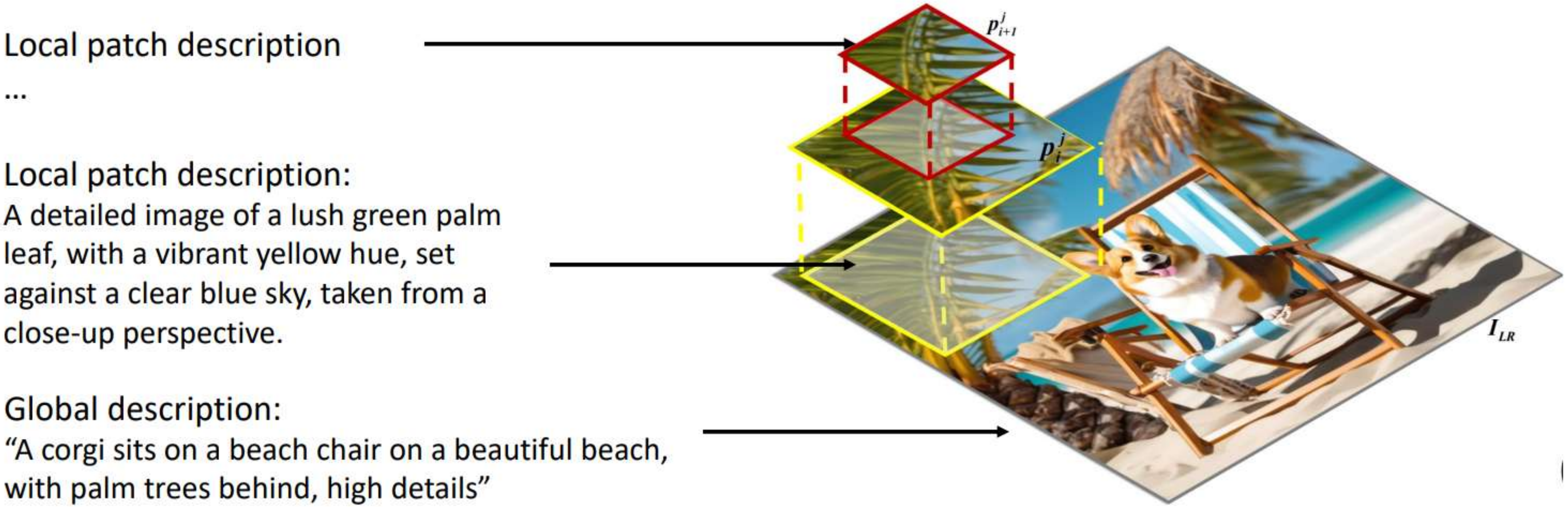
T2I

SDE Edit
And input to an intermediate timestep of T2I

Textual Details Generation

(2024-09) [HiPropmt : Tuning-free Higher-Resolution Generation with Hierarchical MLLM Prompts]()

Иерархические промпты



Local patch description

…

Local patch 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.

Global description:

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

(2024-09) HiPropmt : Tuning-free Higher-Resolution Generation with Hierarchical MLLM Prompts

Генерация текстовых описаний с БЯМ

Генерация текстовых описание

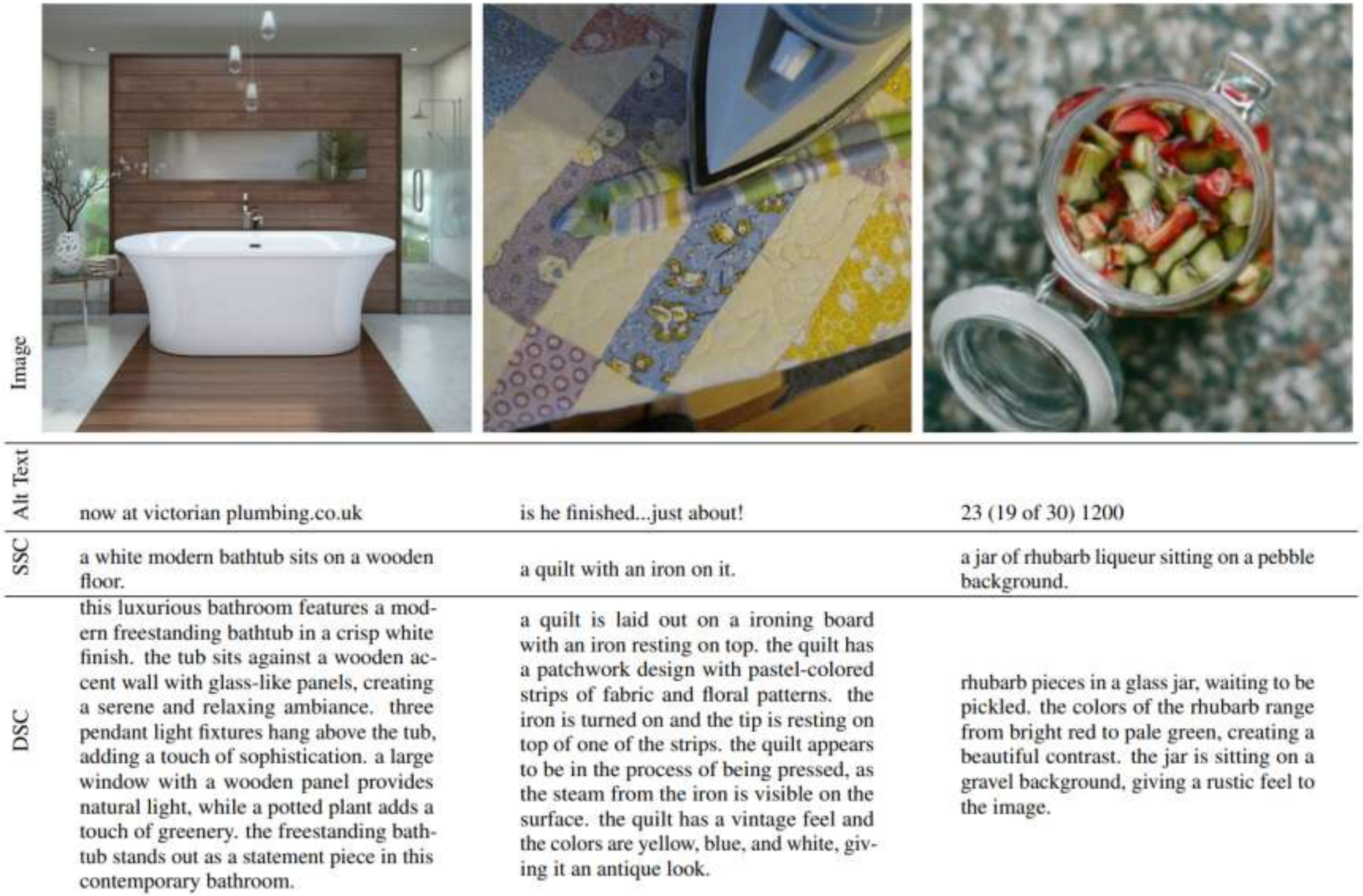


Image			
Alt Text	now at victorian plumbing.co.uk	is he finished...just about!	23 (19 of 30) 1200
SSC	a white modern bathtub sits on a wooden floor.	a quilt with an iron on it.	a jar of rhubarb liqueur sitting on a pebble background.
DSC	this luxurious bathroom features a modern freestanding bathtub 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.	a quilt is laid out on a 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.	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.

- Альтернативые и «человеческие» описания очень простые и шумные
- Люди обычно не описывают положение объектов, важные части (например, раковину на кухне), текстуры и т.д.
- VLLM можно обучить очень детализированным описаниям изображений

Improving Image Generation with Better Captions (DALL-E 3)

Профит от хороших описаний велик

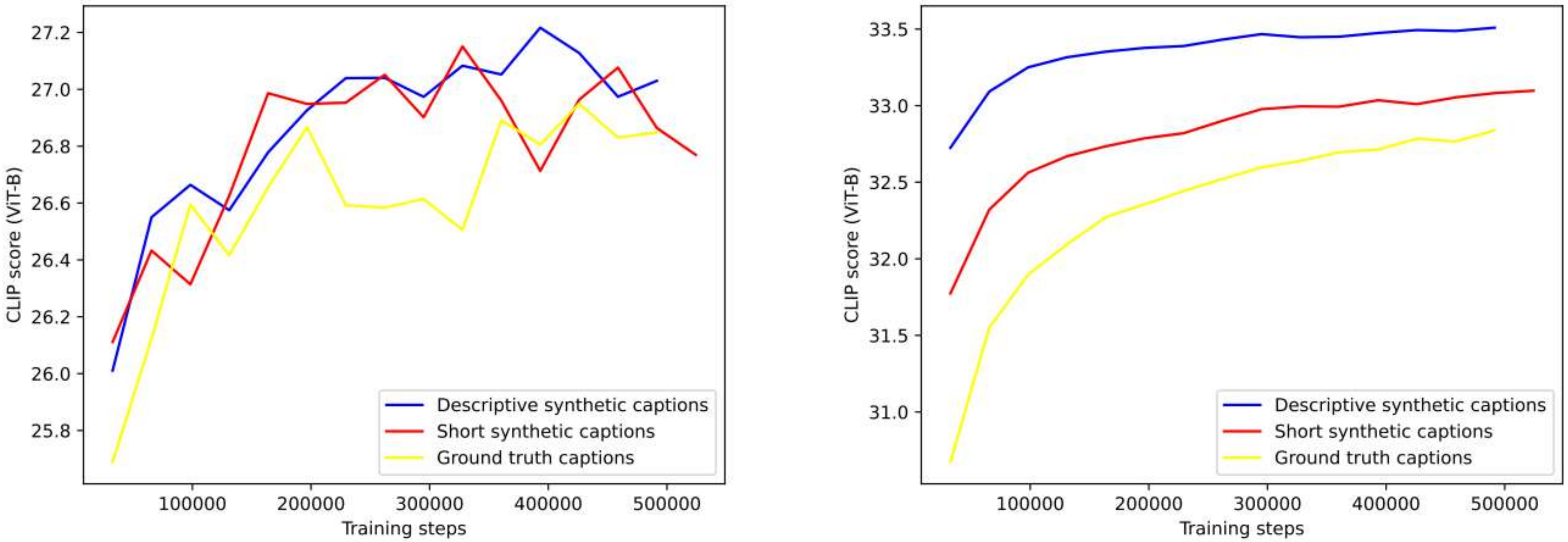
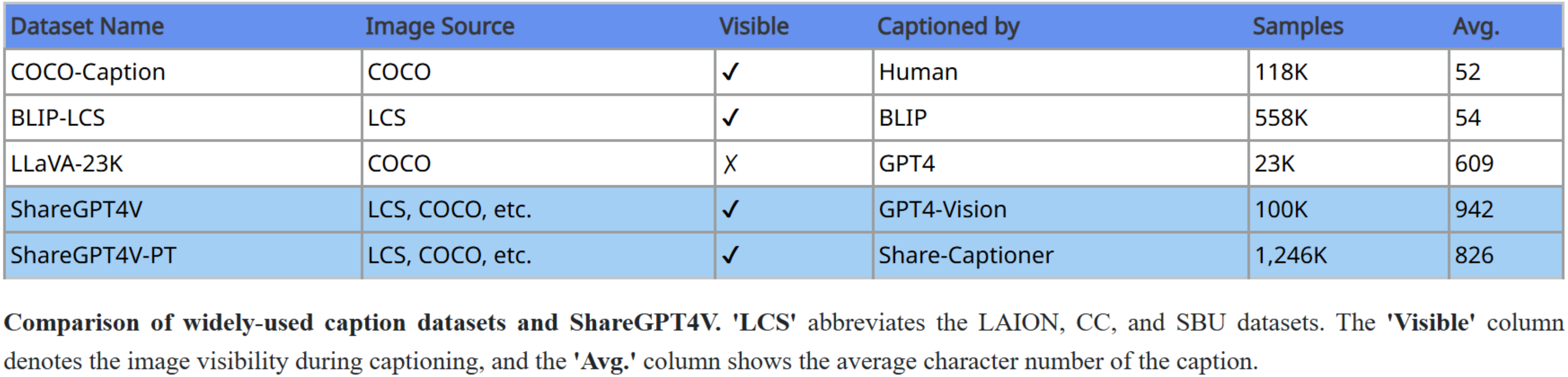


Figure 4 – CLIP scores for text-to-image models trained on different caption types. Left is evaluation results with ground truth captions on our evaluation dataset. Right uses the descriptive synthetic captions from the same dataset.

Improving Image Generation with Better Captions (DALL-E 3)

ShareGPT4V датасет

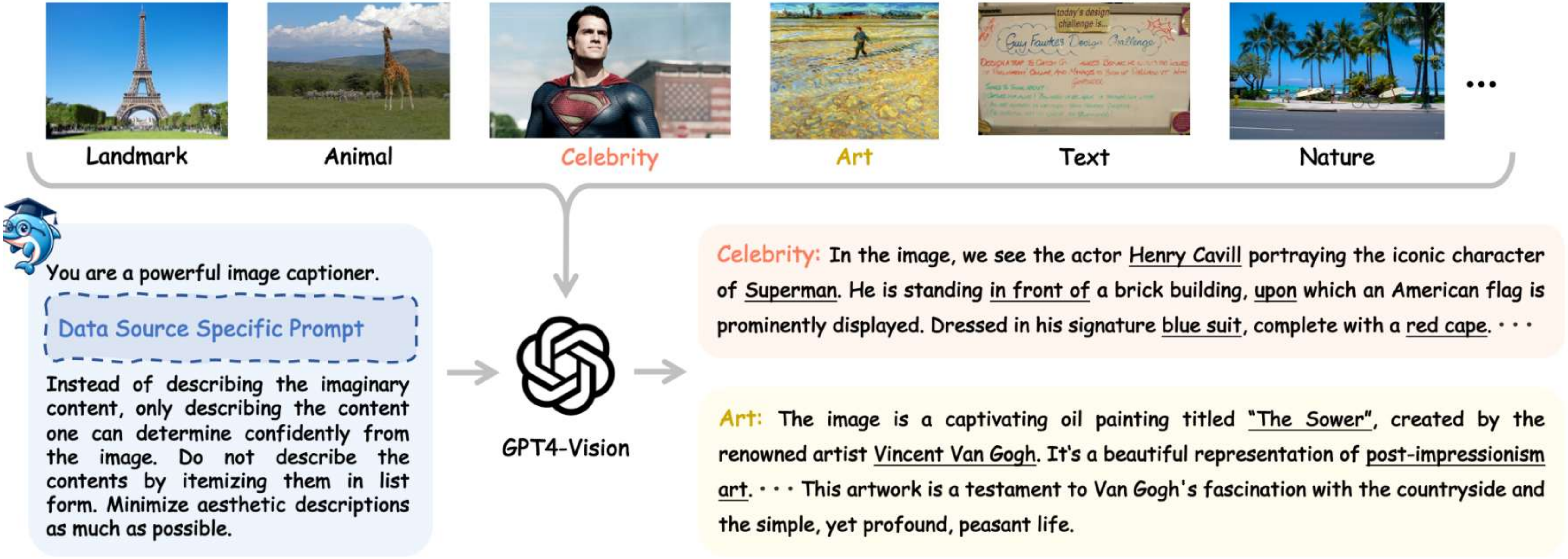


Dataset Name	Image Source	Visible	Captioned by	Samples	Avg.
COCO-Caption	COCO	✓	Human	118K	52
BLIP-LCS	LCS	✓	BLIP	558K	54
LLaVA-23K	COCO	✗	GPT4	23K	609
ShareGPT4V	LCS, COCO, etc.	✓	GPT4-Vision	100K	942
ShareGPT4V-PT	LCS, COCO, etc.	✓	Share-Captioner	1,246K	826

Comparison of widely-used caption datasets and ShareGPT4V. 'LCS' abbreviates the LAION, CC, and SBU datasets. The **'Visible'** column denotes the image visibility during captioning, and the **'Avg.'** column shows the average character number of the caption.

(2023-11) ShareGPT4V: Improving Large Multi-Modal Models with Better Captions

ShareGPT4V датасет



You are a powerful image captioner.

Data Source Specific Prompt

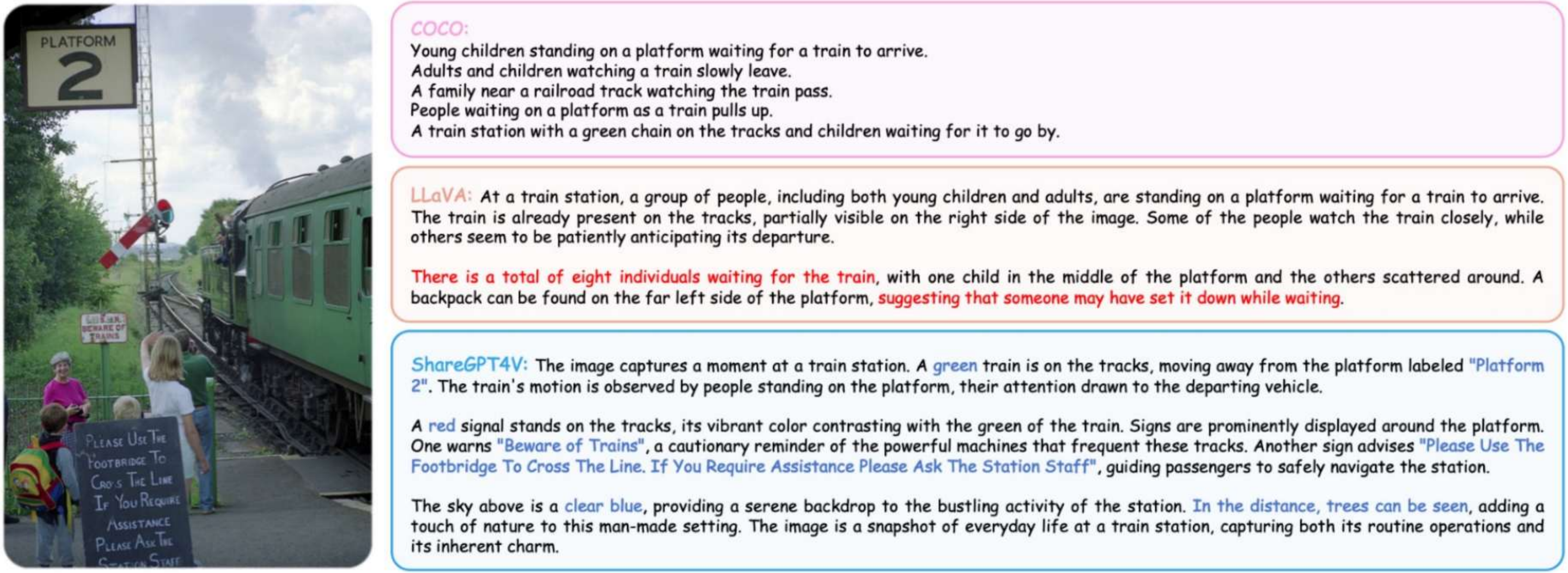
Instead of describing the imaginary content, only describing the content one can determine confidently from the image. Do not describe the contents by itemizing them in list form. Minimize aesthetic descriptions as much as possible.

Celebrity: In the image, we see the actor <u>Henry Cavill</u> portraying the iconic character of <u>Superman</u>. He is standing <u>in front of</u> a brick building, <u>upon</u> which an American flag is prominently displayed. Dressed in his signature <u>blue suit</u>, complete with a <u>red cape</u>. · · ·

Art: The image is a captivating oil painting titled <u>"The Sower"</u>, created by the renowned artist <u>Vincent Van Gogh</u>. It's a beautiful representation of <u>post-impressionism art</u>. · · · This artwork is a testament to Van Gogh's fascination with the countryside and the simple, yet profound, peasant life.

(2023-11) ShareGPT4V: Improving Large Multi-Modal Models with Better Captions

ShareGPT4V датасет


PLATFORM
2
BEWARE OF TRAINS
Please Use The Footbridge To Cross The Line If You Require Assistance Please Ask The Station Staff

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 chain 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.

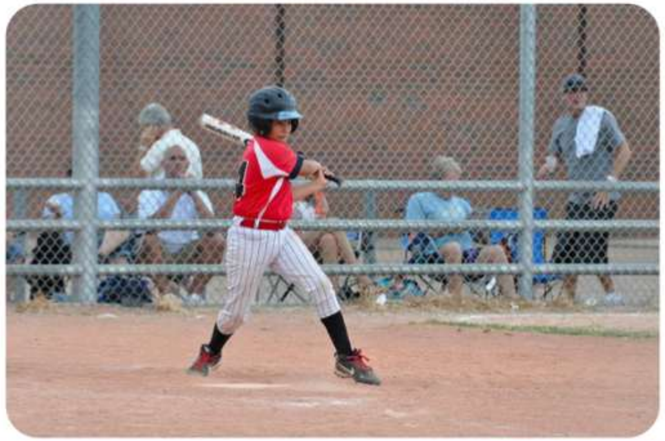
ShareGPT4V: The image captures a moment at a train station. A green train is on the tracks, moving away from the platform labeled "Platform 2". The train's motion is observed by people standing on the platform, their attention drawn to the departing vehicle.

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.

(2023-11) [ShareGPT4V: Improving Large Multi-Modal Models with Better Captions]

ShareGPT4V датасет



COCO: A young boy swinging a baseball bat during a baseball game.

BLIP: a young boy holding a baseball bat on a field.

LLaVA-1.5-13B: The image captures a young boy in a baseball uniform, standing at home plate and swinging a baseball bat. He is focused on the game and appears to be in the middle of a swing.
There are several people around the field, including a few sitting on chairs and benches, watching the game. Some of them are closer to the boy, while others are further away. A baseball glove can be seen on the ground, likely belonging to one of the players. The scene conveys the excitement and energy of a baseball game.

GPT4-Vision: In the image, a young baseball player is the central figure, poised in the batter's box, ready to swing at an incoming ball. Dressed in a vibrant red and white uniform, the player's stance exudes determination and focus. A black helmet, providing protection, sits snugly on the player's head, matching the black grip of the bat held firmly in their hands.

The batter's box, marked by white chalk lines, is clearly visible beneath the player's feet, setting the stage for the impending action. The chain-link fence in the background serves as a boundary, separating the playing field from the spectators.

Speaking of spectators, a group of them can be seen in the background, seated on bleachers. They appear to be engrossed in the game, their faces blurred, making it impossible to discern their expressions. Their presence, however, adds a sense of anticipation and excitement to the scene.

The image captures a moment frozen in time, just before the swing, the anticipation palpable. It's a snapshot of a typical baseball game, filled with tension, excitement, and the promise of action.

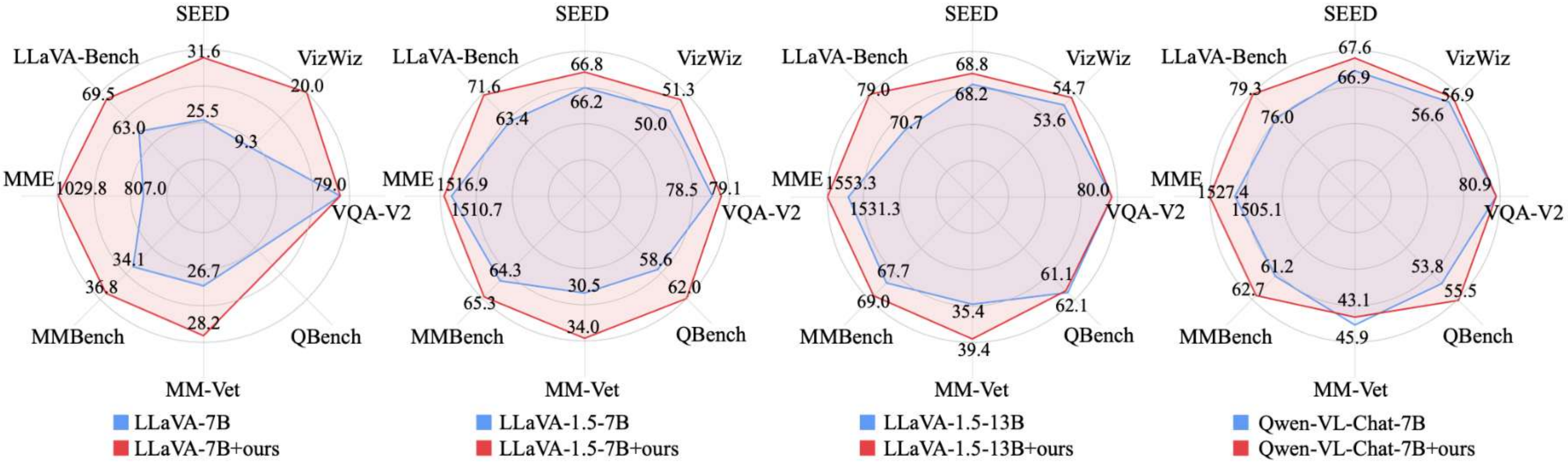
Share-Captioner: In the heart of the action, a young baseball player, donned in a vibrant red and white uniform, is captured mid-swing. The player's black helmet gleams under the sunlight, matching the intensity in his eyes as he focuses on the incoming ball. His black socks peek out from under his white pants, adding a touch of contrast to his uniform.

The player is positioned in the batter's box, his body coiled like a spring, ready to unleash the power of his swing. His bat, gripped firmly in his hands, is poised to connect with the ball.

Beyond the player, a chain-link fence stands guard, separating the field from the spectators. A handful of spectators can be seen in the background, their attention riveted on the unfolding drama at home plate. Their presence adds a sense of anticipation and excitement to the scene.

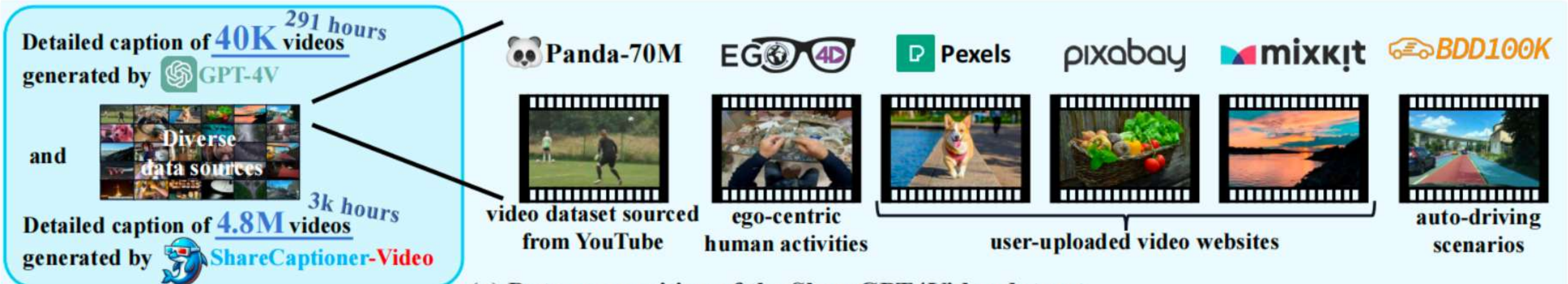
Every element in this image paints a vivid picture of a thrilling moment in a baseball game, from the player's intense focus to the spectators' eager anticipation.

Влияние датасета на качество ВБЯМ



(2023-11) ShareGPT4V: Improving Large Multi-Modal Models with Better Captions

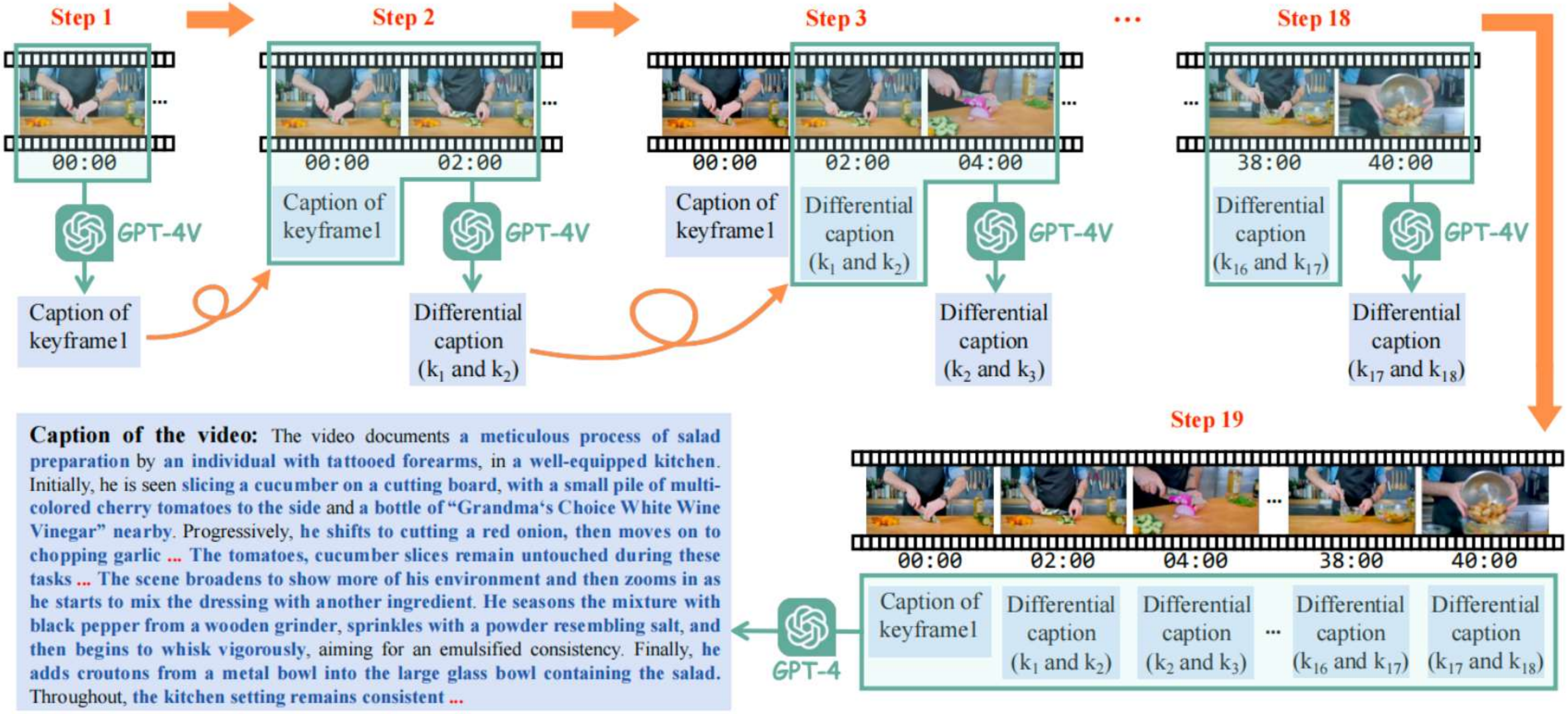
ShareGPT4Video


Detailed caption of 40K videos
291 hours
generated by GPT-4V
and
Diverse data sources
3k hours
Detailed caption of 4.8M videos
generated by ShareCaptioner-Video
Panda-70M
EGO4D
Pexels
pixabay
mixkit
BDD100K
video dataset sourced from YouTube
ego-centric human activities
user-uploaded video websites
auto-driving scenarios

(a) Data composition of the ShareGPT4Video dataset

(2024-06) [Improving Video Understanding and Generation with Better Captions]()

ShareGPT4Video

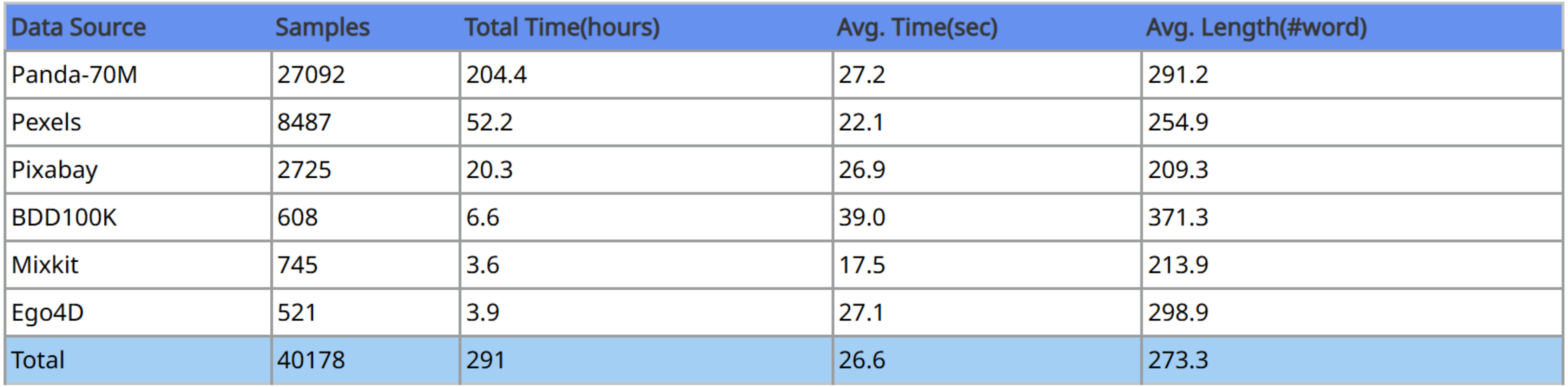

Step 1
00:00
GPT-4V
Caption of keyframe1
Step 2
00:00
02:00
Caption of keyframe1
GPT-4V
Differential caption (k_1 and k_2)
Step 3
00:00
02:00
04:00
Caption of keyframe1
Differential caption (k_1 and k_2)
GPT-4V
Differential caption (k_2 and k_3)
...
Step 18
38:00
40:00
Differential caption (k_{16} and k_{17})
GPT-4V
Differential caption (k_{17} and k_{18})
Step 19
00:00
02:00
04:00
38:00
40:00
Caption of keyframe1
Differential caption (k_1 and k_2)
Differential caption (k_2 and k_3)
Differential caption (k_{16} and k_{17})
Differential caption (k_{17} and k_{18})
GPT-4

Caption of the video: The video documents a meticulous process of salad preparation by an individual with tattooed forearms, in a well-equipped kitchen. Initially, he is seen slicing a cucumber on a cutting board, with a small pile of multi-colored cherry tomatoes to the side and a bottle of "Grandma's Choice White Wine Vinegar" nearby. Progressively, he shifts to cutting a red onion, then moves on to chopping garlic ... The tomatoes, cucumber slices remain untouched during these tasks ... The scene broadens to show more of his environment and then zooms in as he starts to mix the dressing with another ingredient. He seasons the mixture with black pepper from a wooden grinder, sprinkles with a powder resembling salt, and then begins to whisk vigorously, aiming for an emulsified consistency. Finally, he adds croutons from a metal bowl into the large glass bowl containing the salad. Throughout, the kitchen setting remains consistent ...

(b) Example of generating detailed video captions using GPT-4V

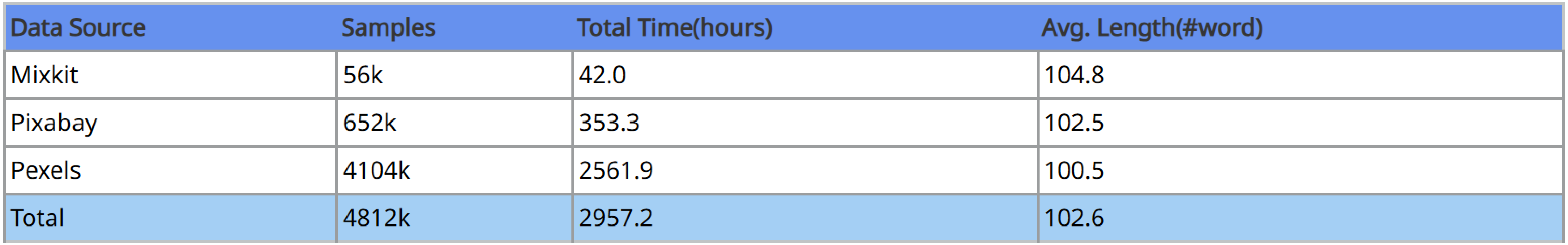
(2024-06) Improving Video Understanding and Generation with Better Captions

ShareCaptioner-Video



Data Source	Samples	Total Time(hours)	Avg. Time(sec)	Avg. Length(#word)
Panda-70M	27092	204.4	27.2	291.2
Pexels	8487	52.2	22.1	254.9
Pixabay	2725	20.3	26.9	209.3
BDD100K	608	6.6	39.0	371.3
Mixkit	745	3.6	17.5	213.9
Ego4D	521	3.9	27.1	298.9
Total	40178	291	26.6	273.3

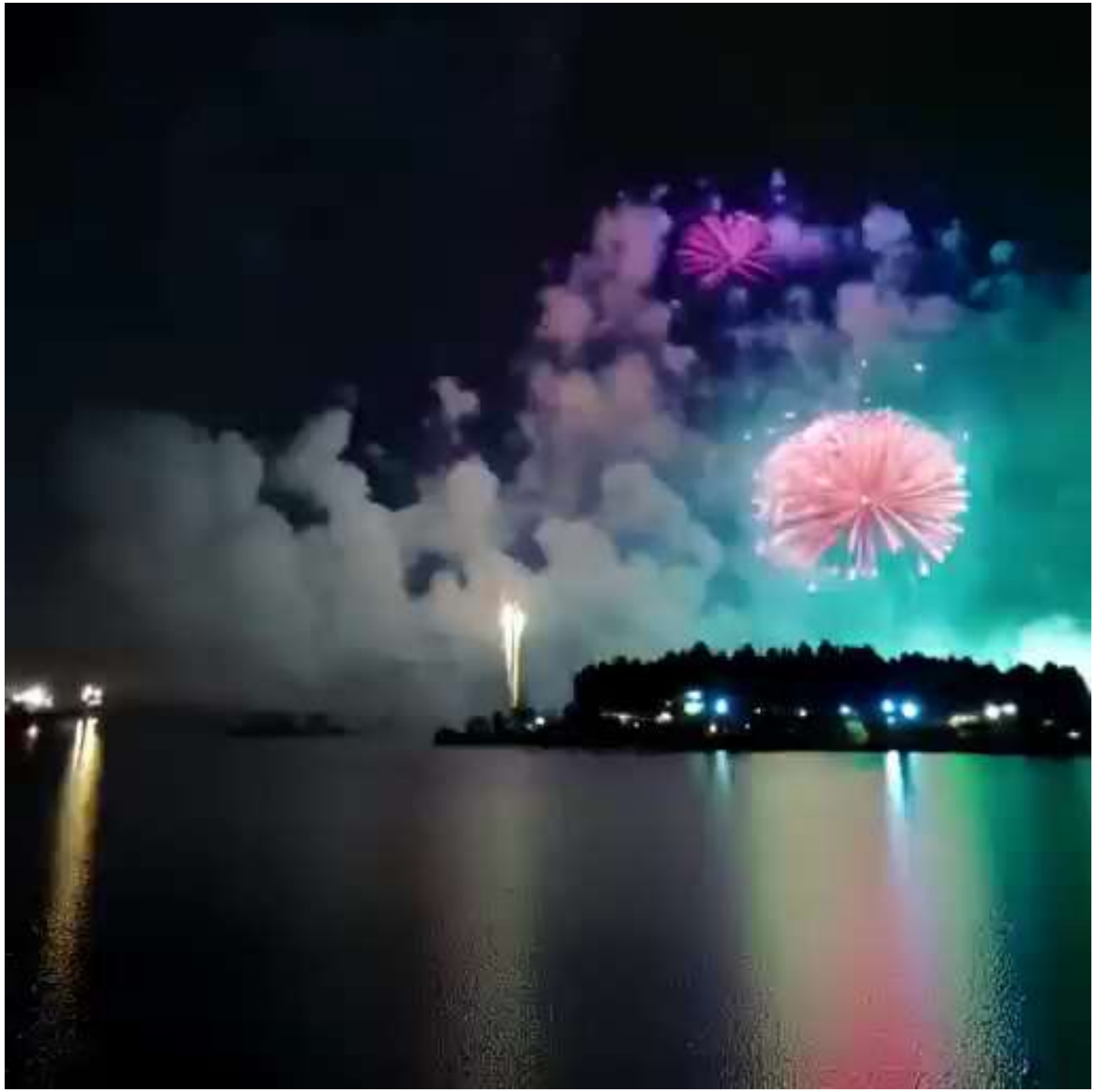
Обучаем свою модель ShareCaptioner-Video и размечаем ею другие датасеты



Data Source	Samples	Total Time(hours)	Avg. Length(#word)
Mixkit	56k	42.0	104.8
Pixabay	652k	353.3	102.5
Pexels	4104k	2561.9	100.5
Total	4812k	2957.2	102.6

(2024-06) [Improving Video Understanding and Generation with Better Captions]

Пример генерации видео



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.

Обуславливание (Conditioning) через БЯМ

Let’s Tango!

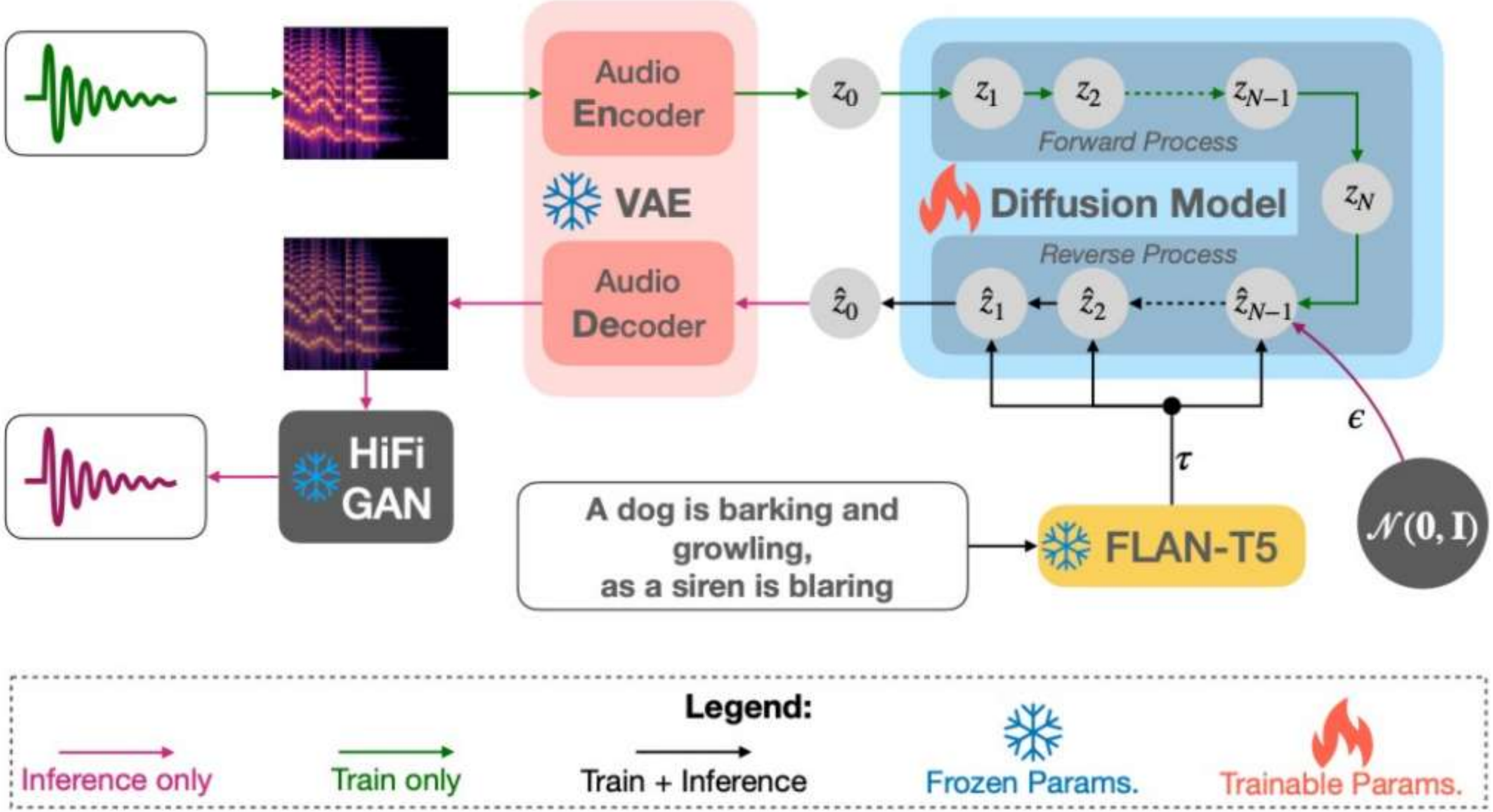
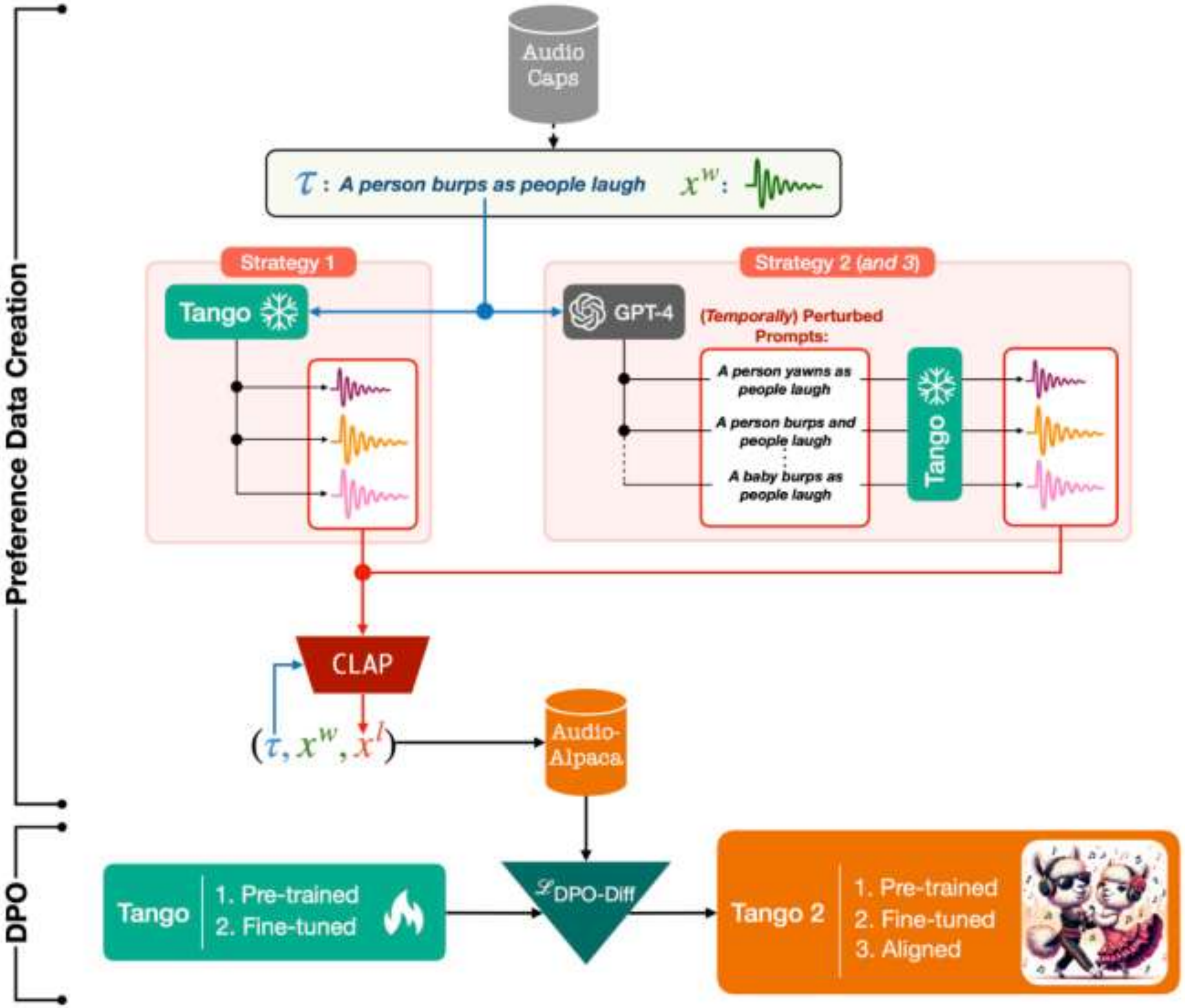


Figure 1: Overall architecture of TANGO.

(2023-04) Text-to-Audio Generation using Instruction-Tuned LLM and Latent Diffusion Model

Тюнинг Tango в Tango2



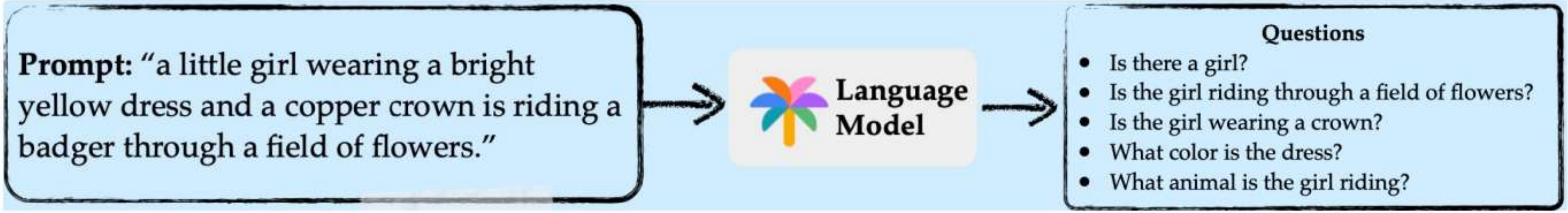
- Берём базовую TANGO
- Делает пертурбацию промптов
- Генерируем варианты аудио
- Сортируем их по качеству с помощью внешней модели
- Тюним Tango с помощью DPO

(2024-04) Aligning Diffusion-based Text-to-Audio Generations through Direct Preference Optimization

БЯМ как оценщик

БЯМ как оценщик

Используем БЯМ для генерации вопросов про изображение



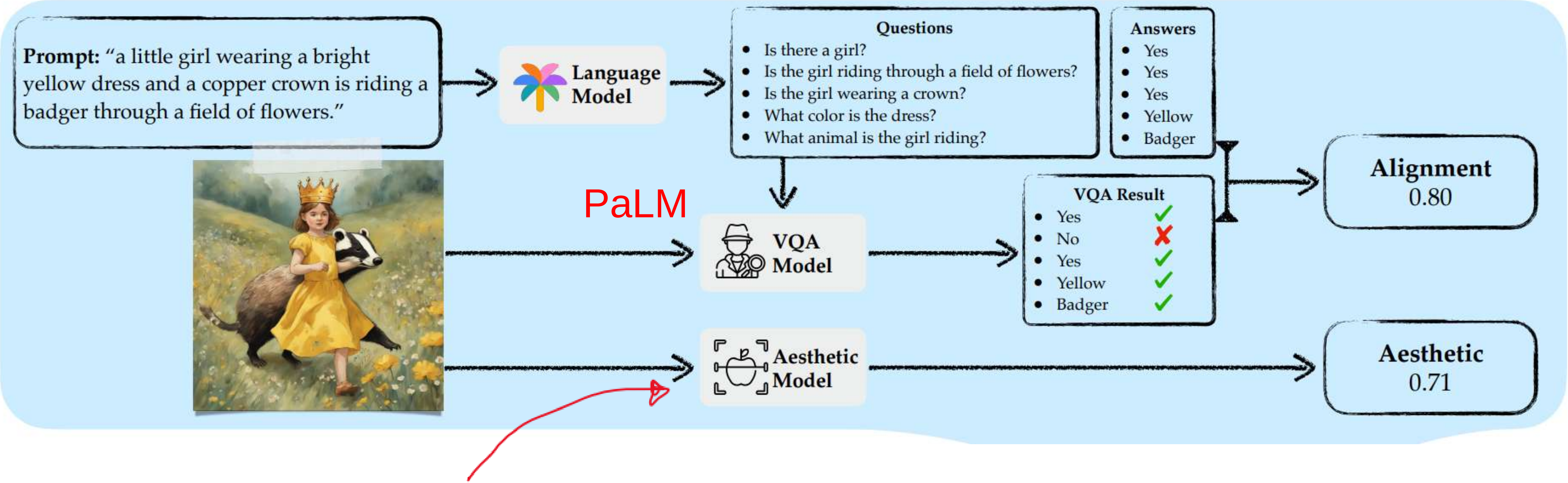
Prompt: "a little girl wearing a bright yellow dress and a copper crown is riding a badger through a field of flowers."

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?

(2023-11) DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback

Оценка изображения через ВБЯМ



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

Language Model

PaLM

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

VQA Model

VQA Result

- Yes ✓
- No ✗
- Yes ✓
- Yellow ✓
- Badger ✓

Alignment 0.80

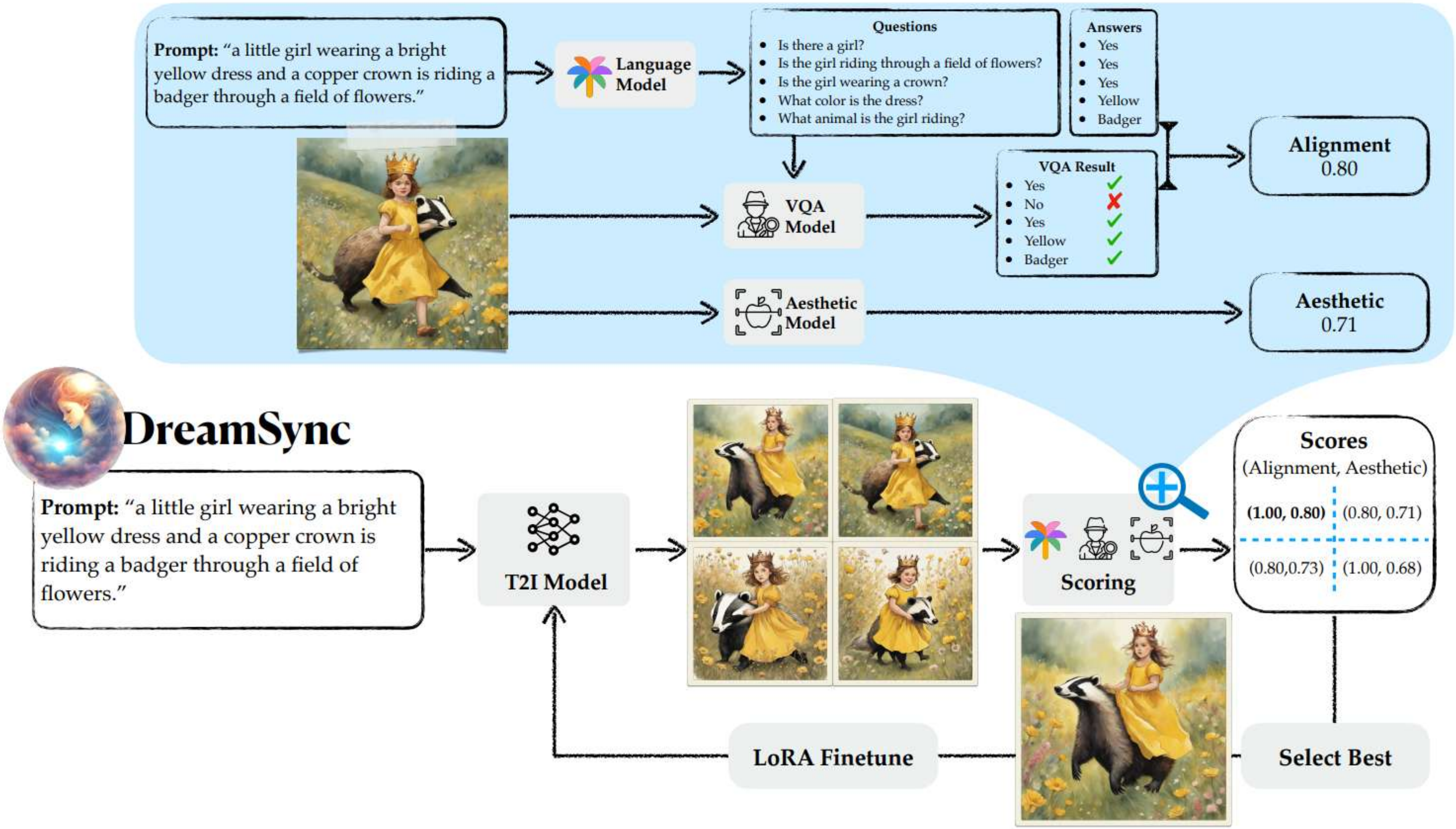
Aesthetic Model

Aesthetic 0.71

(2023-03) VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining – VLM, похожая на CLIP, обученная на поднаборе LAION и датасете AVA в 250К изображений, размеченных оценками и комментариями по качеству

(2023-11) [DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback]

Цикл обратной связи



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

VQA Model

VQA Result
- Yes ✓
- No ✗
- Yes ✓
- Yellow ✓
- Badger ✓

Alignment 0.80

Aesthetic Model

Aesthetic 0.71

DreamSync

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

T2I Model

Scoring

Scores
(Alignment, Aesthetic)

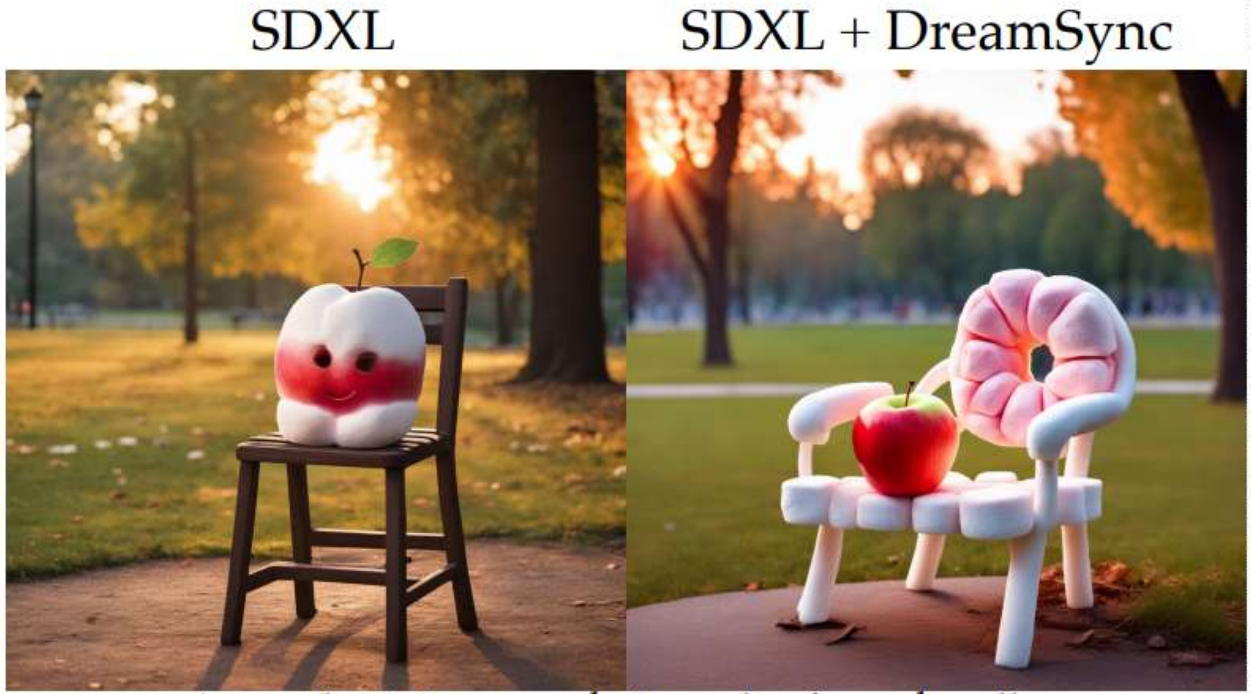
(1.00, 0.80)	(0.80, 0.71)
(0.80,0.73)	(1.00, 0.68)

Select Best

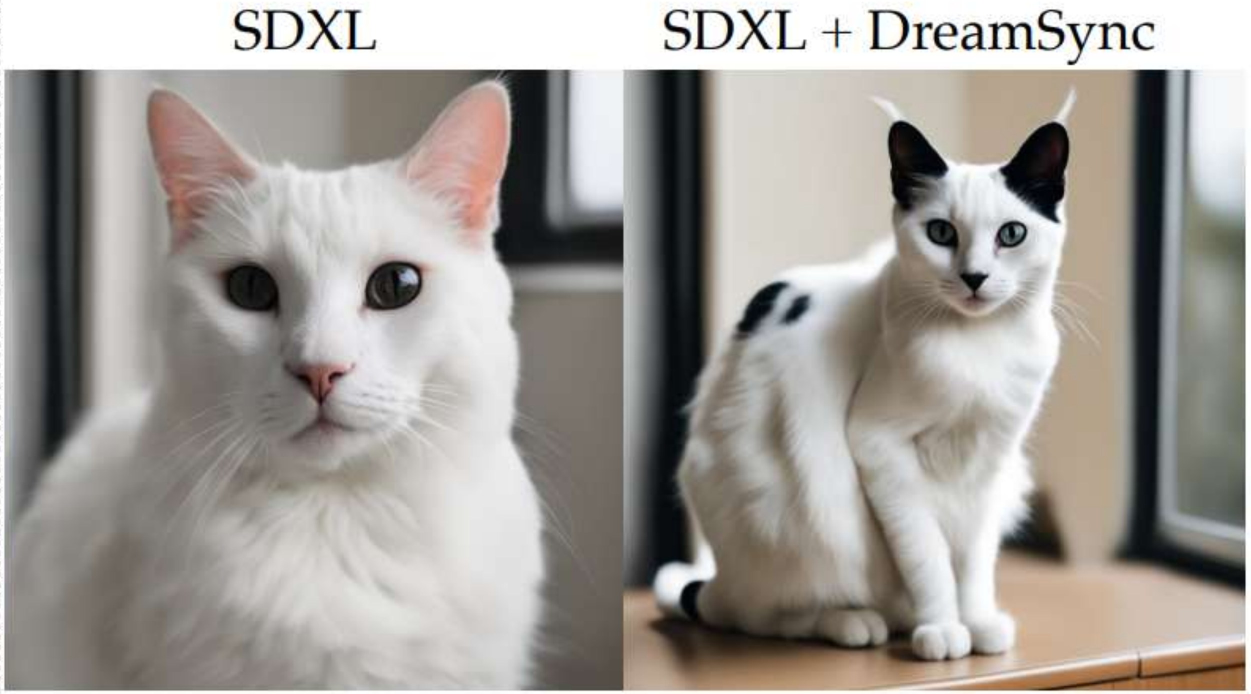
LoRA Finetune

(2023-11) DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback

Итог – улучшение качества генерации



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



A white cat with black ears and markings

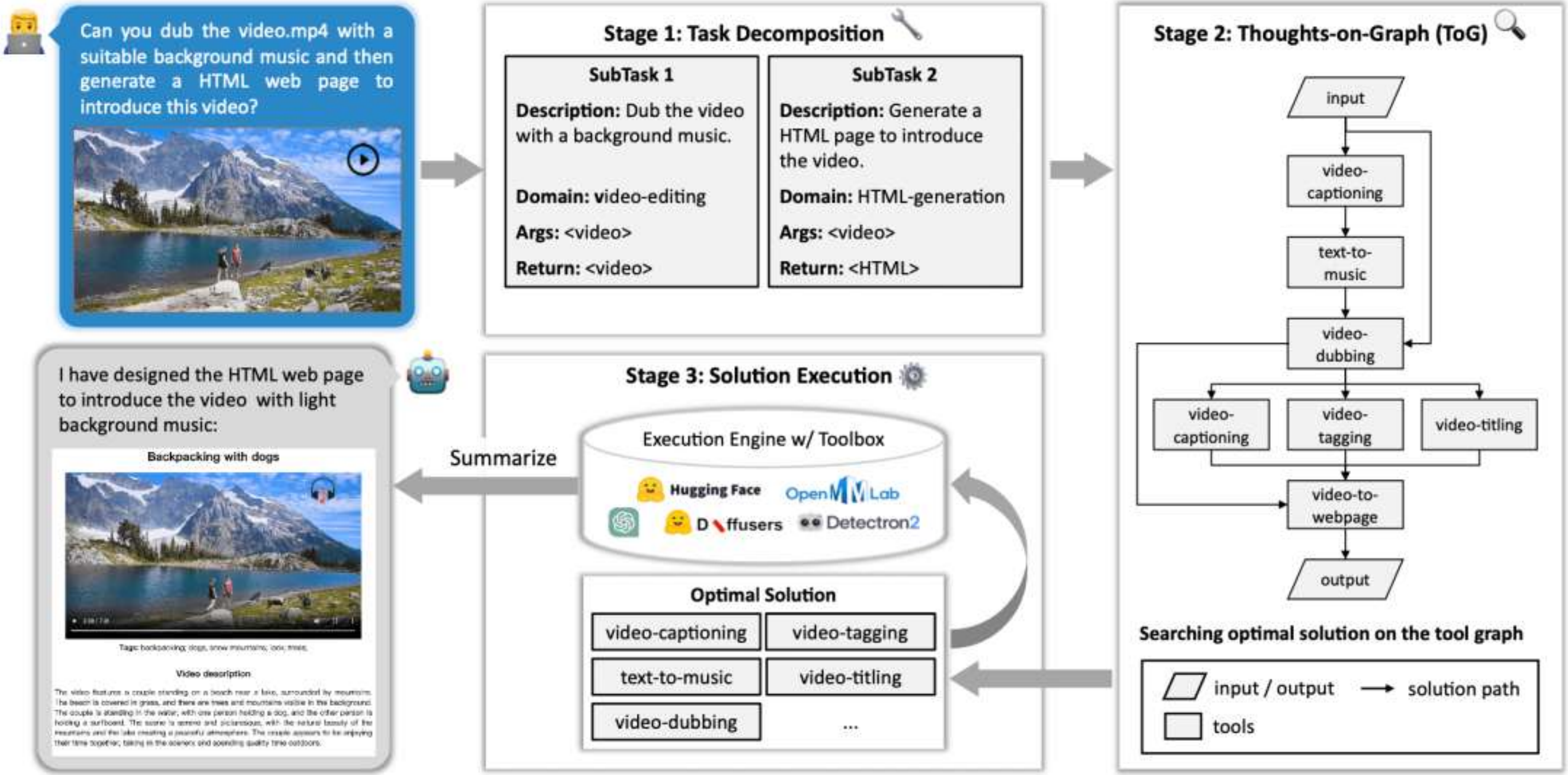
Агенты БЯМ

LLM как агенты для генерации

- Смысл агентов – справится со сложными задачами и сложными инструкциями пользователя, решая задачу «под ключ»

- Complex content:
  - “An oil painting, where a green vintage car, a blue scooter on the left of it and a black bicycle on the right of it, are parked on the road, with two birds in the sky.”
  - Multiple objects; Attribute binding; Position relationship; Object quantity;
- Multiple-task / Multi-modal system:
  - “Can you generate a video of a car running on the road” [Video Generation]
  - “Can you generate the corresponding sound” [Video-to-audio Generation]
  - “Can you modify the car into a big blue truck” [Video Editing]

ControlLLM



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

Stage 1: Task Decomposition

SubTask 1

Description: Dub the video with a background music.

Domain: video-editing

Args: <video>

Return: <video>

SubTask 2

Description: Generate a HTML page to introduce the video.

Domain: HTML-generation

Args: <video>

Return: <HTML>

Stage 2: Thoughts-on-Graph (ToG)

Searching optimal solution on the tool graph

Stage 3: Solution Execution

Optimal Solution

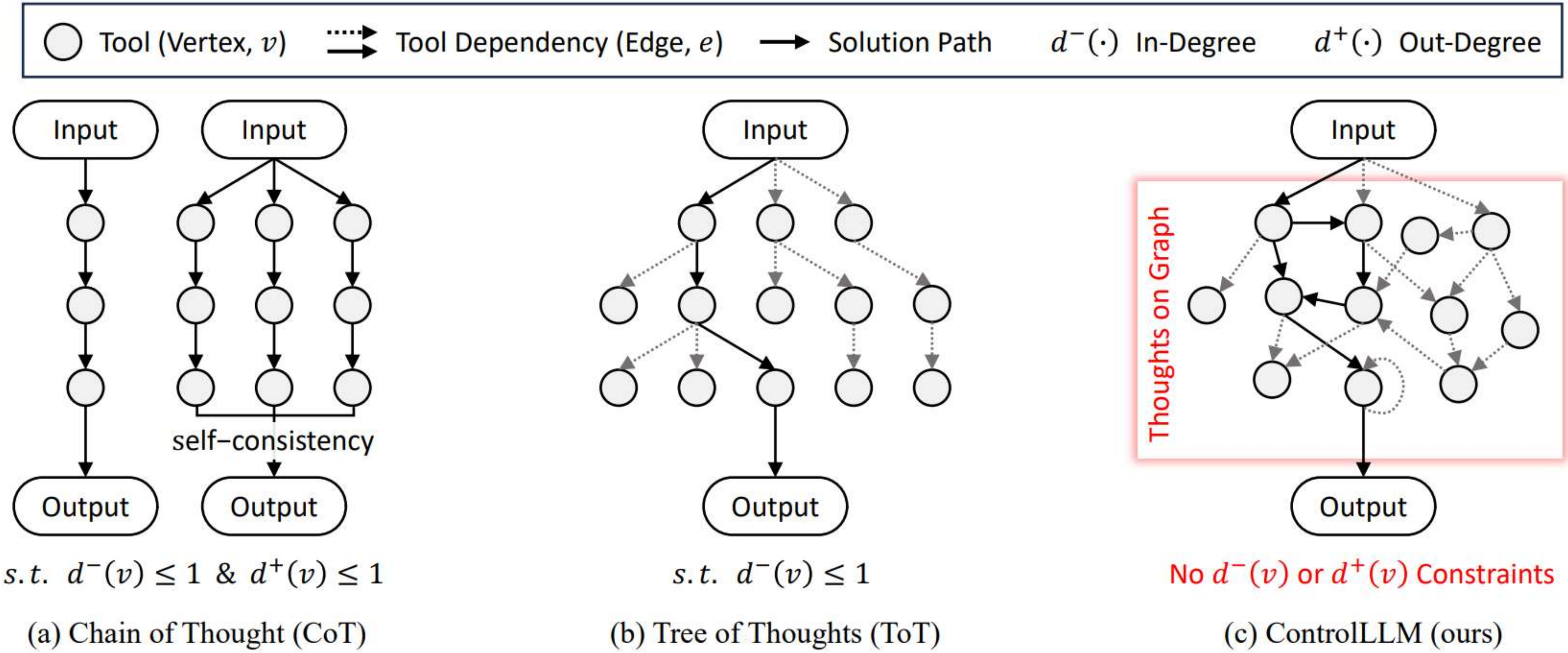
I have designed the HTML web page to introduce the video with light background music:

Backpacking with dogs

Video description

(2023-10) ControlLLM: Augment Language Models with Tools by Searching on Graphs

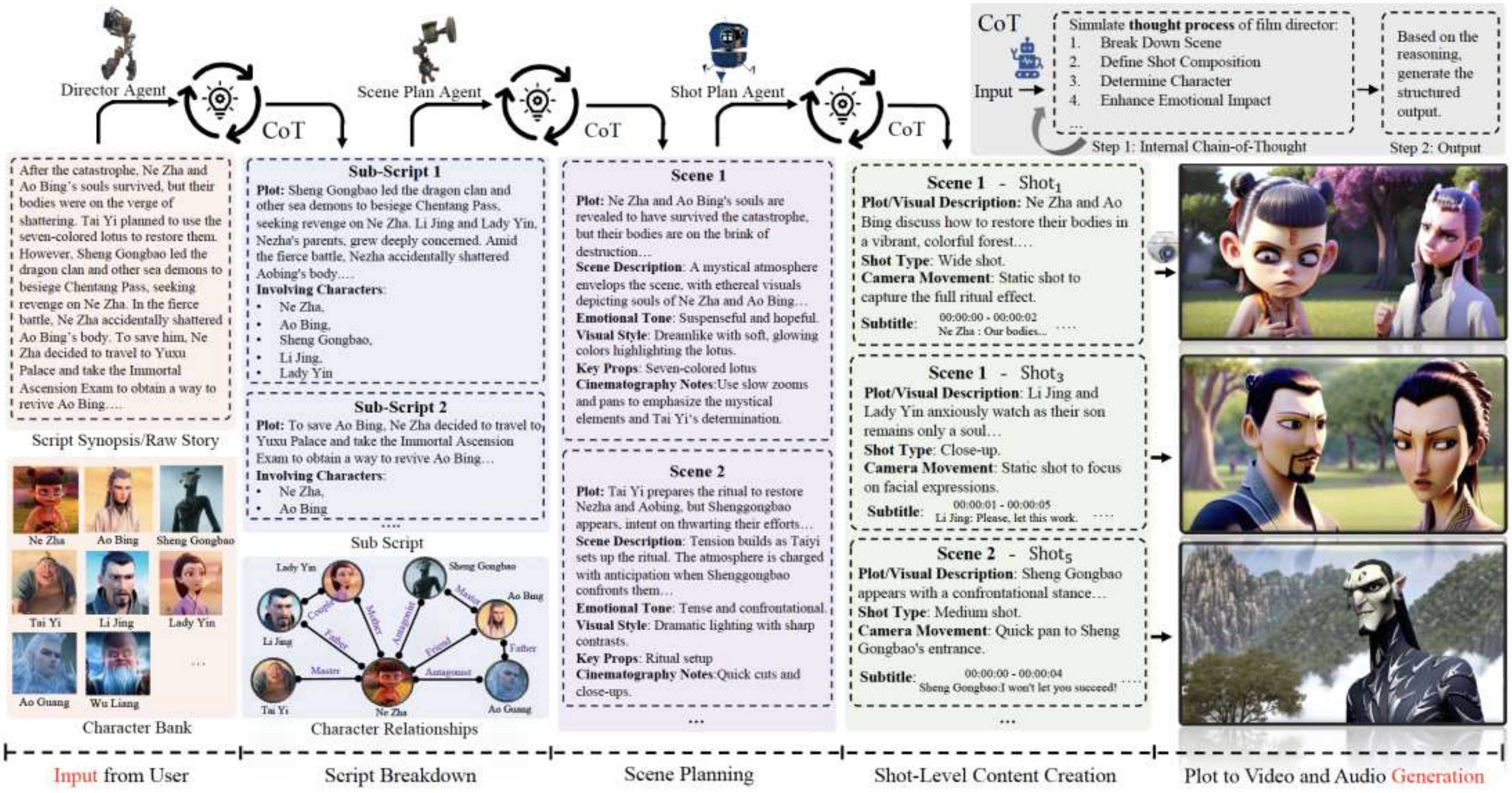
Растекаться мыслью по графу



(a) Chain of Thought (CoT) (b) Tree of Thoughts (ToT) (c) ControlLLM (ours)

Figure 1. **Comparisons of different paradigms for task planning.** (a) Chain of Thought (CoT) [41], CoT with self-consistency [40] and (b) Tree of Thoughts [47] (ToT) essentially rely on the LLMs to perform task planning, where the edge is actually formed by LLMs at run time. (c) The Thoughts-on-Graph (ToG) paradigm in our method searches for solutions on a pre-built graph that captures the dependencies of tools, which avoids the hallucination problem in tool invocation.

Movie Agent



Simulate **thought process** of film director:

1. Break Down Scene
2. Define Shot Composition
3. Determine Character
4. Enhance Emotional Impact

...

Based on the reasoning, generate the structured output.

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….

Sub-Script 1

Plot: Sheng Gongbao led the dragon clan and other sea demons to besiege Chentang Pass, seeking revenge on Ne Zha. Li Jing and Lady Yin, Nezha's parents, grew deeply concerned. Amid the fierce battle, Nezha accidentally shattered Aobing's body….

Involving Characters:

- Ne Zha,
- Ao Bing,
- Sheng Gongbao,
- Li Jing,
- Lady Yin

Sub-Script 2

Plot: To save Ao Bing, Ne Zha decided to travel to Yuxu Palace and take the Immortal Ascension Exam to obtain a way to revive Ao Bing…

Involving Characters:

- Ne Zha,
- Ao Bing

….

Scene 1

Plot: Ne Zha and Ao Bing's souls are revealed to have survived the catastrophe, but their bodies are on the brink of destruction…

Scene Description: A mystical atmosphere envelops the scene, with ethereal visuals depicting souls of Ne Zha and Ao Bing…

Emotional Tone: Suspenseful and hopeful.

Visual Style: Dreamlike with soft, glowing colors highlighting the lotus.

Key Props: Seven-colored lotus

Cinematography Notes:Use slow zooms and pans to emphasize the mystical elements and Tai Yi's determination.

Scene 2

Plot: Tai Yi prepares the ritual to restore Nezha and Aobing, but Shenggongbao appears, intent on thwarting their efforts…

Scene Description: Tension builds as Taiyi sets up the ritual. The atmosphere is charged with anticipation when Shenggongbao confronts them…

Emotional Tone: Tense and confrontational.

Visual Style: Dramatic lighting with sharp contrasts.

Key Props: Ritual setup

Cinematography Notes:Quick cuts and close-ups.

...

Scene 1 - Shot$_1$

Plot/Visual Description: Ne Zha and Ao Bing discuss how to restore their bodies in a vibrant, colorful forest….

Shot Type: Wide shot.

Camera Movement: Static shot to capture the full ritual effect.

Subtitle: 00:00:00 - 00:00:02 Ne Zha : Our bodies... ….

Scene 1 - Shot$_3$

Plot/Visual Description: Li Jing and Lady Yin anxiously watch as their son remains only a soul…

Shot Type: Close-up.

Camera Movement: Static shot to focus on facial expressions.

Subtitle: 00:00:01 - 00:00:05 Li Jing: Please, let this work. ….

Scene 2 - Shot$_5$

Plot/Visual Description: Sheng Gongbao appears with a confrontational stance…

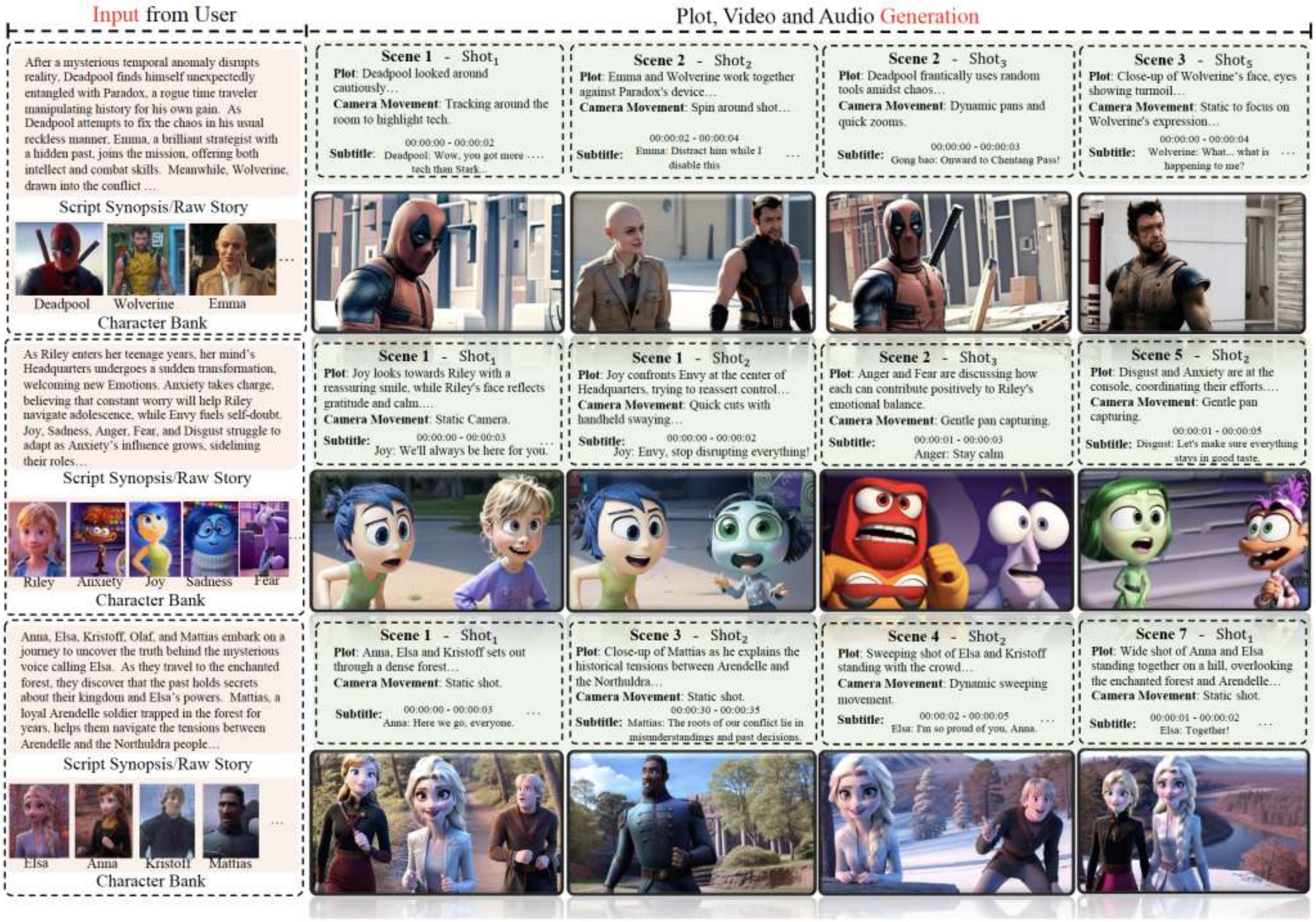
Shot Type: Medium shot.

Camera Movement: Quick pan to Sheng Gongbao's entrance.

Subtitle: 00:00:00 - 00:00:04 Sheng Gongbao:I won't let you succeed! ….

...

(2025-03) **MovieAgent: Automated Movie Generation via Multi-Agent CoT Planning** https://weijiawu.github.io/MovieAgent/



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 …

Scene 1 - Shot$_1$
Plot: Deadpool looked around cautiously…
Camera Movement: Tracking around the room to highlight tech.
00:00:00 - 00:00:02
Subtitle: Deadpool: Wow, you got more …. tech than Stark...

Scene 2 - Shot$_2$
Plot: Emma and Wolverine work together against Paradox's device…
Camera Movement: Spin around shot…
00:00:02 - 00:00:04
Subtitle: Emma: Distract him while I disable this …

Scene 2 - Shot$_3$
Plot: Deadpool frantically uses random tools amidst chaos…
Camera Movement: Dynamic pans and quick zooms.
00:00:00 - 00:00:03
Subtitle: Gong bao: Onward to Chentang Pass!

Scene 3 - Shot$_5$
Plot: Close-up of Wolverine's face, eyes showing turmoil…
Camera Movement: Static to focus on Wolverine's expression…
00:00:00 - 00:00:04
Subtitle: Wolverine: What... what is happening to me? …

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…

Scene 1 - Shot$_1$
Plot: Joy looks towards Riley with a reassuring smile, while Riley's face reflects gratitude and calm.…
Camera Movement: Static Camera.
00:00:00 - 00:00:03
Subtitle: Joy: We'll always be here for you. …

Scene 1 - Shot$_2$
Plot: Joy confronts Envy at the center of Headquarters, trying to reassert control…
Camera Movement: Quick cuts with handheld swaying…
00:00:00 - 00:00:02
Subtitle: Joy: Envy, stop disrupting everything!

Scene 2 - Shot$_3$
Plot: Anger and Fear are discussing how each can contribute positively to Riley's emotional balance.
Camera Movement: Gentle pan capturing.
00:00:01 - 00:00:03
Subtitle: Anger: Stay calm

Scene 5 - Shot$_2$
Plot: Disgust and Anxiety are at the console, coordinating their efforts.…
Camera Movement: Gentle pan capturing.
00:00:01 - 00:00:05
Subtitle: Disgust: Let's make sure everything stays in good taste.

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…

Scene 1 - Shot$_1$
Plot: Anna, Elsa and Kristoff sets out through a dense forest…
Camera Movement: Static shot.
00:00:00 - 00:00:03
Subtitle: Anna: Here we go, everyone. …

Scene 3 - Shot$_2$
Plot: Close-up of Mattias as he explains the historical tensions between Arendelle and the Northuldra…
Camera Movement: Static shot.
00:00:30 - 00:00:35
Subtitle: Mattias: The roots of our conflict lie in misunderstandings and past decisions.

Scene 4 - Shot$_2$
Plot: Sweeping shot of Elsa and Kristoff standing with the crowd…
Camera Movement: Dynamic sweeping movement.
00:00:02 - 00:00:05
Subtitle: Elsa: I'm so proud of you, Anna. …

Scene 7 - Shot$_1$
Plot: Wide shot of Anna and Elsa standing together on a hill, overlooking the enchanted forest and Arendelle…
Camera Movement: Static shot.
00:00:01 - 00:00:02
Subtitle: Elsa: Together! …

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.

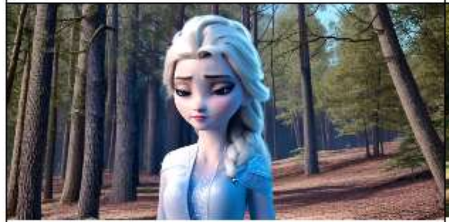
Involving Characters: Elsa, Anna, Kristoff, and Olaf.

Emotional Tone: Curiosity and determination.

Key Themes: Adventure, discovery, and mystery.

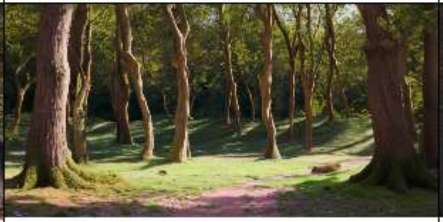
Scene 1 - Shot 1

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.}



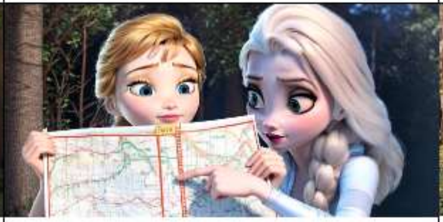
Scene 1 - Shot 3

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: {}



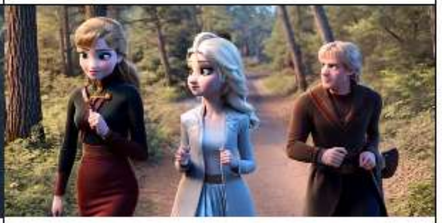
Scene 1 - Shot 4

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.}



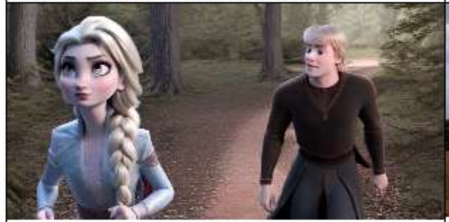
Scene 1 - Shot 6

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.}



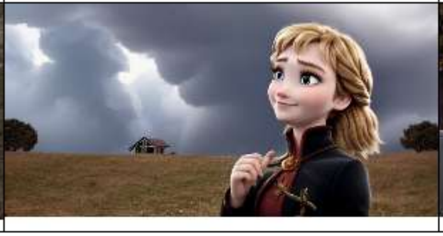
Scene 1 - Shot 7

Plot: A tracking shot of Elsa and Kristoff leading the way into the forest. The forest is dense and mysterious, with tall trees and a thick canopy. The sky is partly cloudy, casting dappled sunlight on the ground.
Involving Characters: Elsa, Kristoff.
Shot Type: Tracking shot.
Camera Movement: Dolly-in.
Subtitle: {Elsa: Let's go, everyone. Stay close. Kristoff: Right behind you, Elsa.}



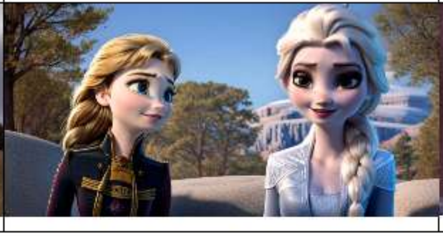
Scene 2 - Shot 1

Plot: Wind whips the hair of Anna across faces, prismatic storm light slicing diagonally through frame.
Involving Characters: Anna
Shot Type: Medium Shot.
Camera Movement: Handheld shaky pan following debris.
Subtitle: {}



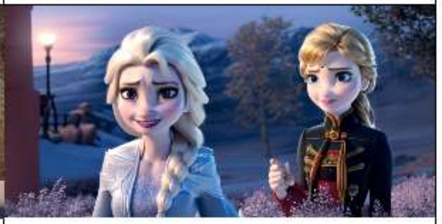
Scene 2 - Shot 4

Plot: Wide shot of Anna and Elsa surrounded by the ancient trees and historical artifacts.
Involving Characters: Elsa, Anna.
Shot Type: Wide Shot.
Camera Movement: Pull away.
Subtitle: {Anna: Understanding the past is key to moving forward. Elsa: We must learn from these stories.}



Scene 2 - Shot 5

Plot: Elsa and Anna stand in awe as mystical lights dance around them, casting ethereal glows and revealing ancient symbols.
Involving Characters: Elsa, Anna.
Shot Type: Close-up.
Camera Movement: Slow-motion.
Subtitle: {Elsa: Do you see that? It's like they're speaking to us. Anna: I've never seen anything like it.}

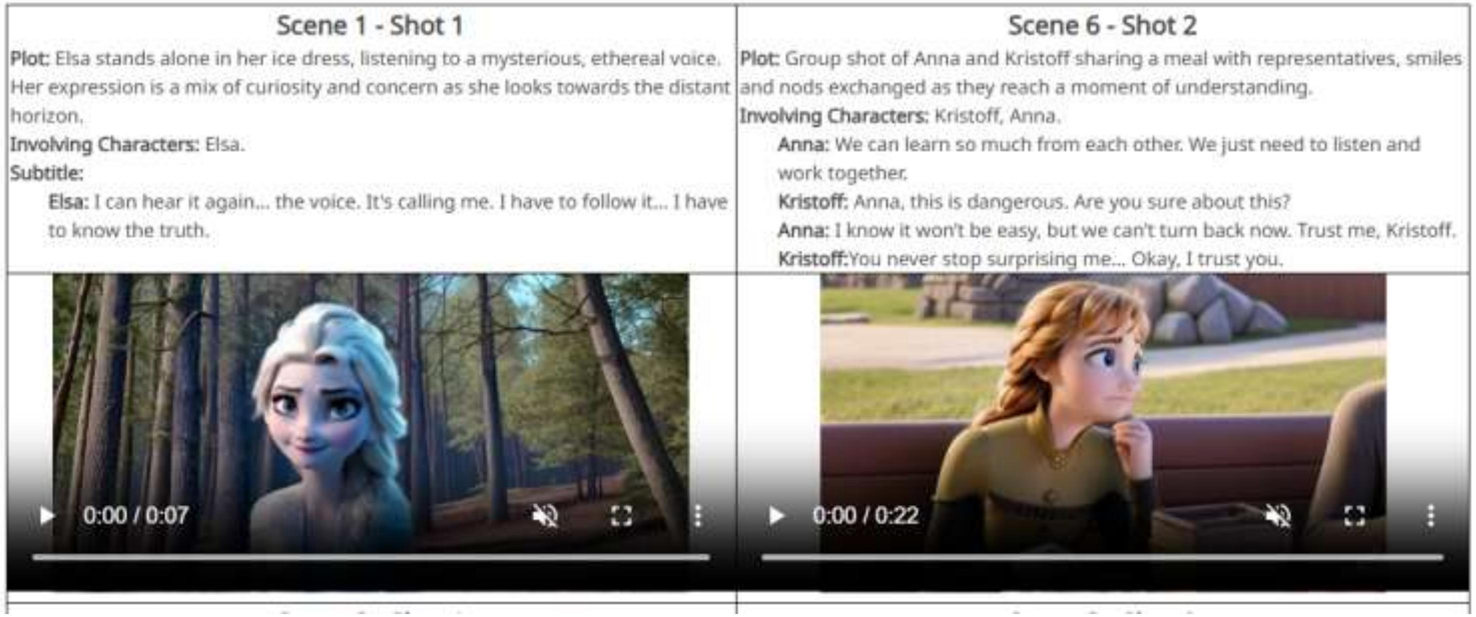


Озвучка

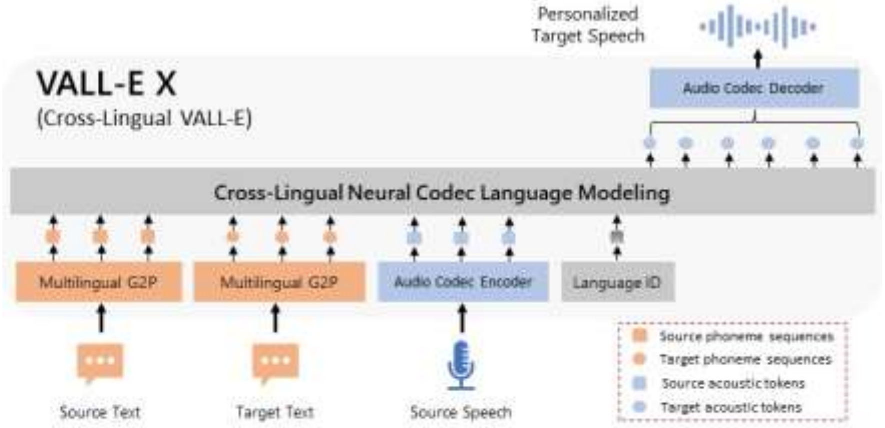
Input: Script Synopsis and Character Bank (Image,Audio Sample, Name)

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

Output: Multi-scene, multi-shot videos with a coherent narrative, synchronized audio.

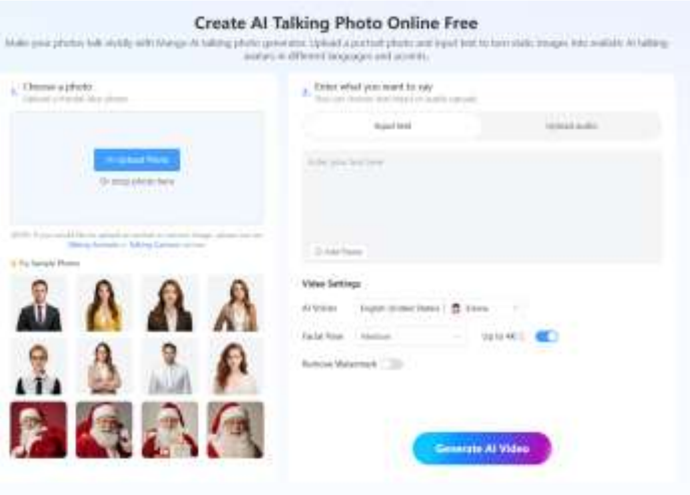


Scene 1 - Shot 1	Scene 6 - Shot 2
Plot: Elsa stands alone in her ice dress, listening to a mysterious, ethereal voice. Her expression is a mix of curiosity and concern as she looks towards the distant horizon. **Involving Characters:** Elsa. **Subtitle:** **Elsa:** I can hear it again... the voice. It's calling me. I have to follow it... I have to know the truth.	**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.



https://github.com/Plachtaa/VALL-E-X

(не-авторская реализация, т.к. MS не выпустила код)



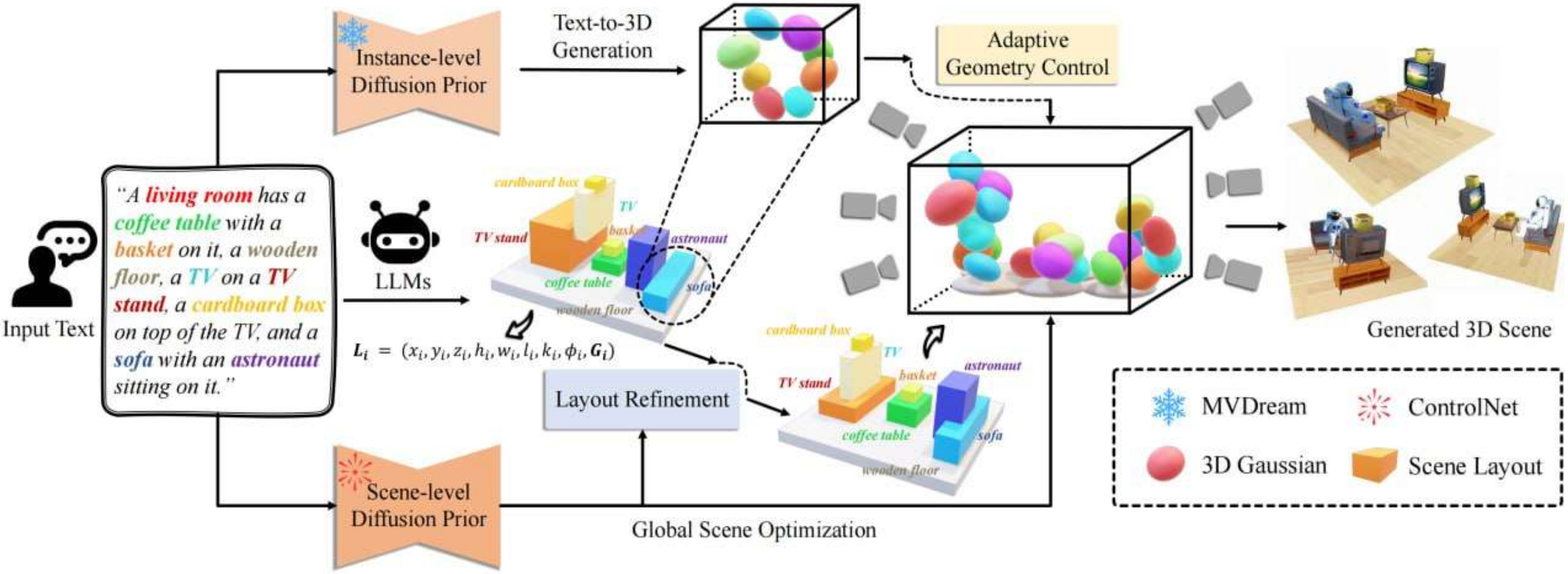
Create AI Talking Photo Online Free

https://mangoanimate.com/

(Китайский AI стартап)

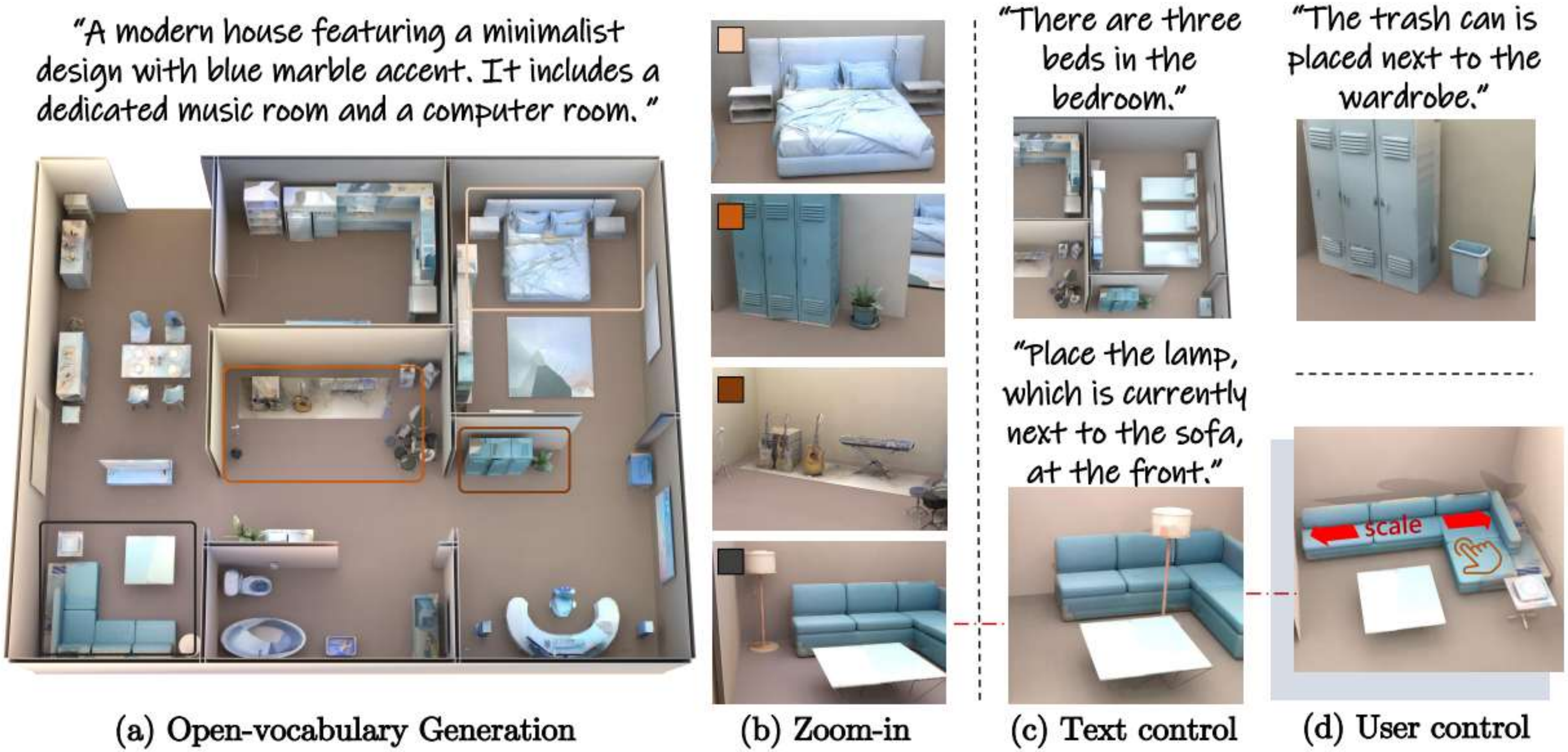
3D

Генерация 3D сцен


Instance-level Diffusion Prior
Text-to-3D Generation
Adaptive Geometry Control
Input Text
“A living room has a coffee table with a basket on it, a wooden floor, a TV on a TV stand, a cardboard box on top of the TV, and a sofa with an astronaut sitting on it.”
LLMs
cardboard box
TV
TV stand
basket
astronaut
coffee table
sofa
wooden floor
$L_i = (x_i, y_i, z_i, h_i, w_i, l_i, k_i, \phi_i, G_i)$
Layout Refinement
Scene-level Diffusion Prior
Global Scene Optimization
Generated 3D Scene
MVDream
ControlNet
3D Gaussian
Scene Layout

https://gala3d.github.io/

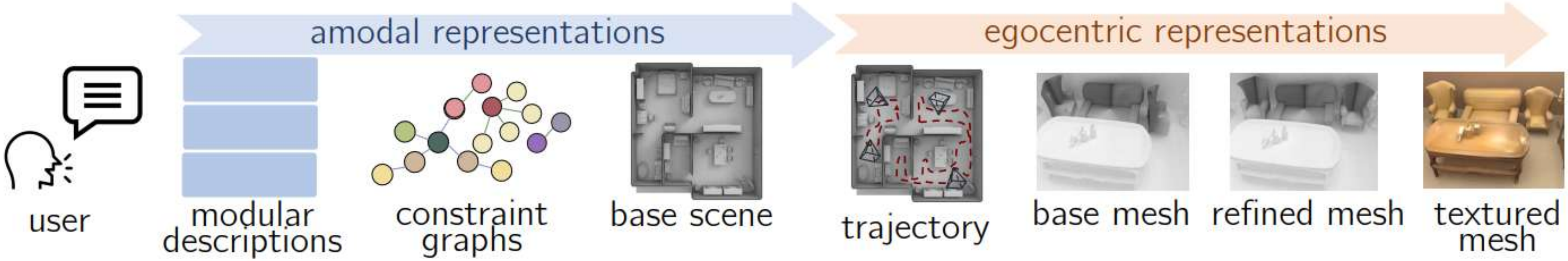
AnyHome (ECCV2024) – пример SoTA генерации сцен


"A modern house featuring a minimalist design with blue marble accent. It includes a dedicated music room and a computer room."
"There are three beds in the bedroom."
"The trash can is placed next to the wardrobe."
"Place the lamp, which is currently next to the sofa, at the front."
scale

(a) Open-vocabulary Generation (b) Zoom-in (c) Text control (d) User control

https://ivl.cs.brown.edu/research/anyhome

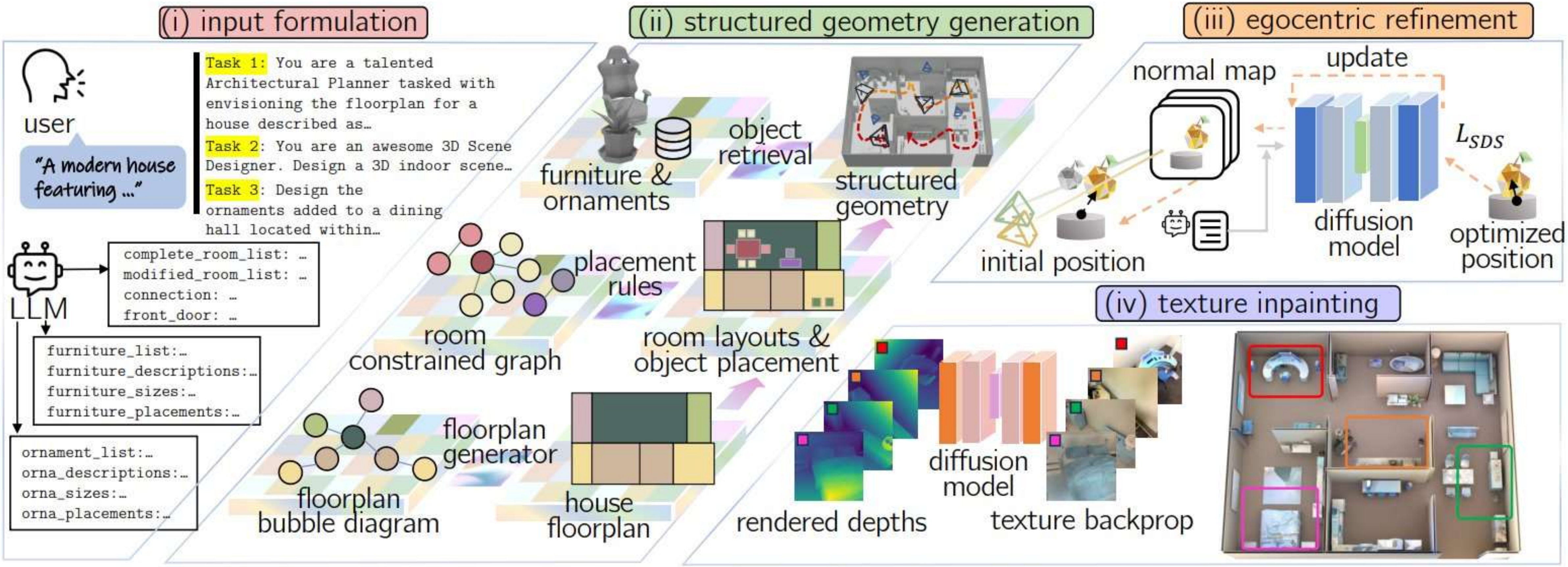
Декомпозиция задачи



- → Усложняем декомпозицию задачи, за счёт этого упрощая подзадачи и получая возможность активнее задействовать существующие модели
- → Эгоцентричный вид позволяет задействовать 2D модели активнее, а не только 3D модели
- → Больше возможностей для управления (control), уточнения (refinement), и редактирования (editing), т.к. больше ручек и прозрачнее воздействие ручки на результат

https://ivl.cs.brown.edu/research/anyhome

Схема работы подробнее



Task 1: You are a talented Architectural Planner tasked with envisioning the floorplan for a house described as…

Task 2: You are an awesome 3D Scene Designer. Design a 3D indoor scene…

Task 3: Design the ornaments added to a dining hall located within…

```
complete_room_list: …
modified_room_list: …
connection: …
front_door: …
```

```
furniture_list:…
furniture_descriptions:…
furniture_sizes:…
furniture_placements:…
```

```
ornament_list:…
orna_descriptions:…
orna_sizes:…
orna_placements:…
```

https://ivl.cs.brown.edu/research/anyhome

Примеры результатов и вариабельность

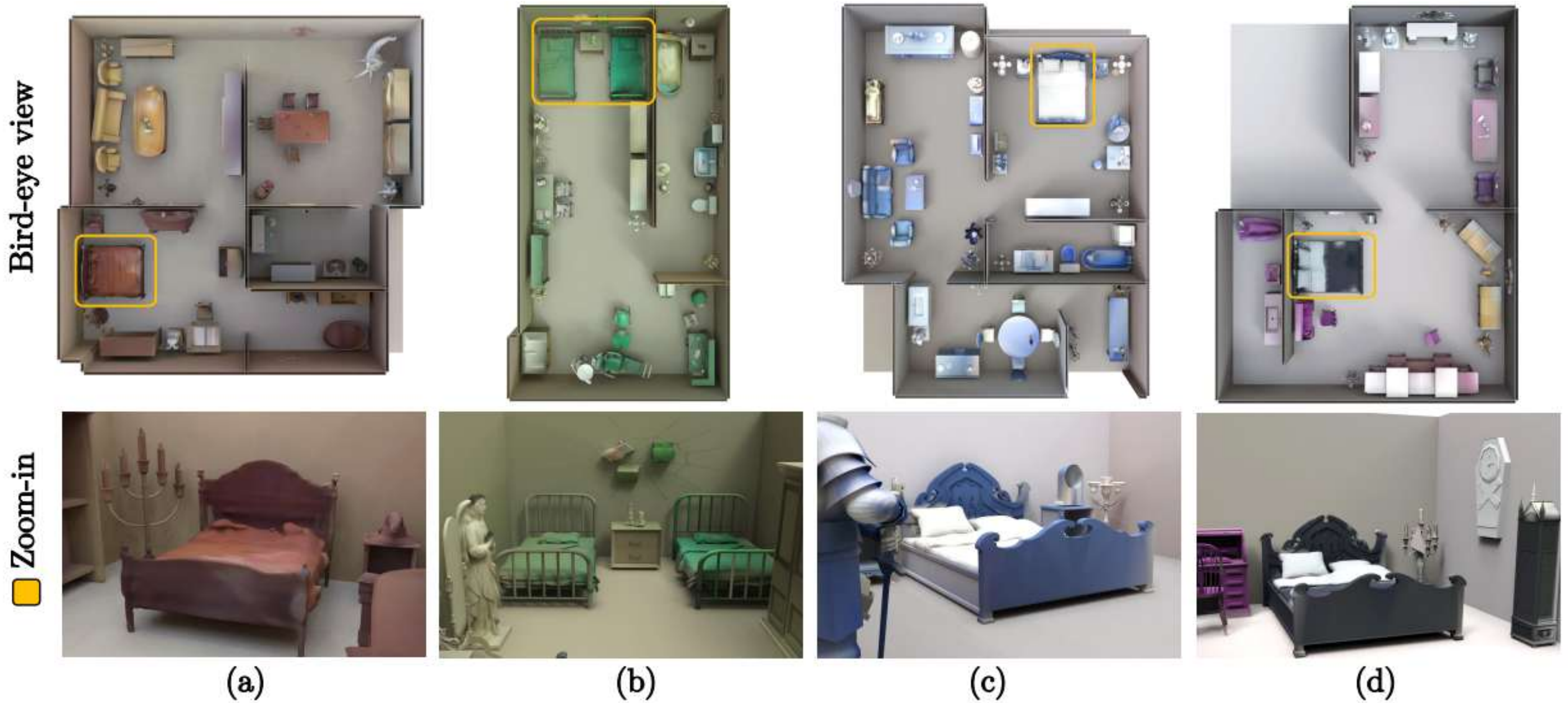


Fig. 6: Diverse Scene Results. Four distinct scenes generated for the prompt `"A one-bedroom, one-bathroom haunted house featuring dark wood and antique furnishings."` **AnyHome** produces houses with diverse floorplans, room types, room layouts, objects and textures.

Резюме

- Генеративные модели обязаны своим успехом возможности предобучения на больших объёмах неразмеченных данных
- Если множество способов использовать БЯМ для генерации данных
- БЯМ может быть универальной основой (universal backbone) для генерации мультимодальных данных, но есть сложности с разными требованиями кодировщиков/декодировщиков для понимания картинок и для их генерации
- БЯМ позволяют декомпозировать и раскрывать запросы пользователя, добавляя им нужную детализацию и подробность
- Качество генеративных моделей зависит от качества текстовых описаний данных, и ВБЯМ позволяют сгенерировать хорошие описания
- ВБЯМ также позволяют оценивать результат генерации и отбирать хорошие примеры для обучения и настройки