Multi-modal Large Language Model

Methodology

Benchmark

Platform

Data

Presenter: Jing SHAO

© 2023 Jing SHAO. All Rights Reserved

Multi-modal Large Language Model

PaLM 2

Meta

文心一言

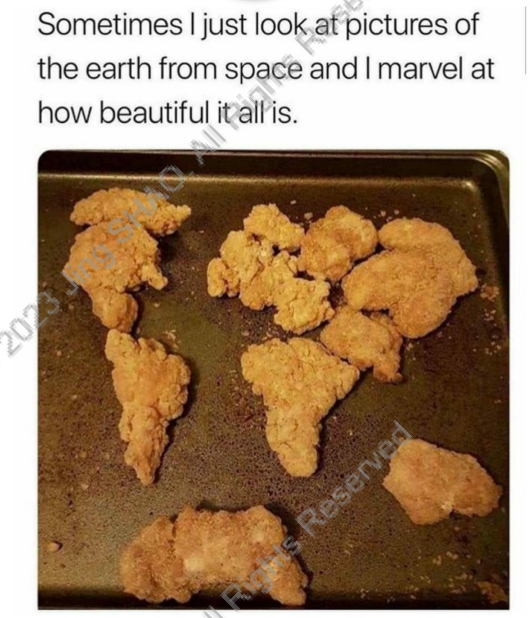
商量 - 语言大模型

通义千问

360智脑 × 360搜索

© 2023 Jing SHAO. All Rights Reserved

Vision ~ Large Language Model



Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.

Q: Can you explain this meme?

A: This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets.

The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world.

The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

© 2023 Jing SHAO. All Rights Reserved

How to

Adapt LLM to Multi-Modal LLM?

Balance Multi-Modal Conflict?

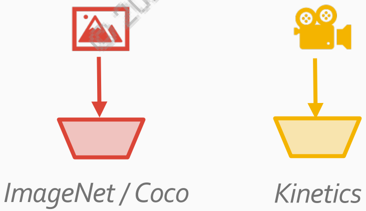
Construct Multi-Modal Instructions?

Decode Various Downstream Tasks?

Evaluate Multi-Modal LLM?

© 2023 Jing SHAO. All Rights Reserved

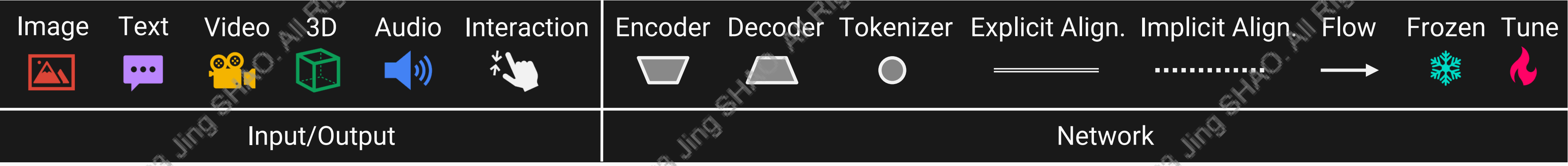
Image	Text	Video	3D	Audio	Interaction	Encoder	Decoder	Tokenizer	Explicit Align.	Implicit Align.	Flow	Frozen	Tune
								●	═	┈	→	❄	🔥
Input/Output						Network							

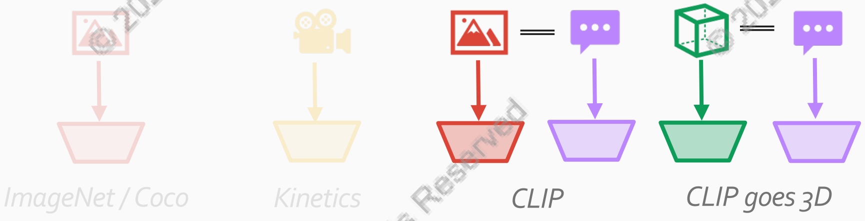


- Each modality is pre-trained **separately**
- **No** connection between different modalities

<Vision, Label> / <Vision, Vision>

© 2023 Jing SHAO. All Rights Reserved



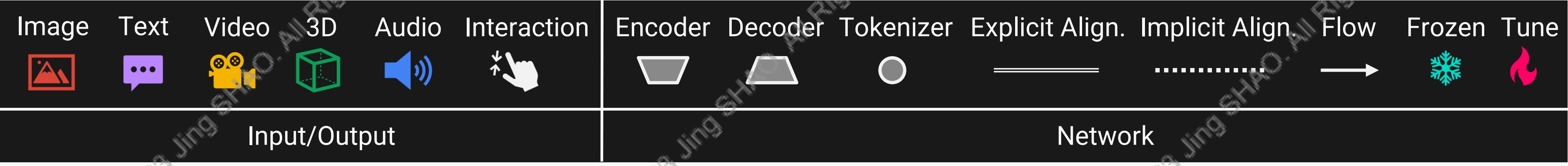


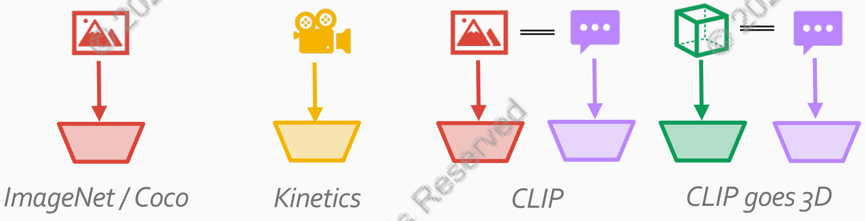
- Link vision with language
- Need **explicit** connection between different modalities, including large-scale paired data and explicit alignment -> Hard to extend

<Vision, Label> / <Vision, Vision>

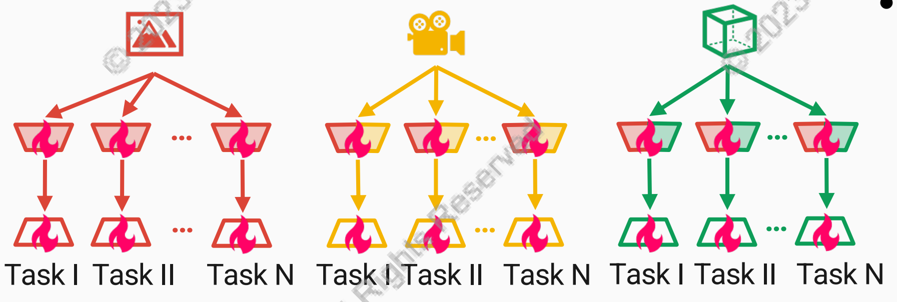
<Vision, Text>

© 2023 Jing SHAO. All Rights Reserved



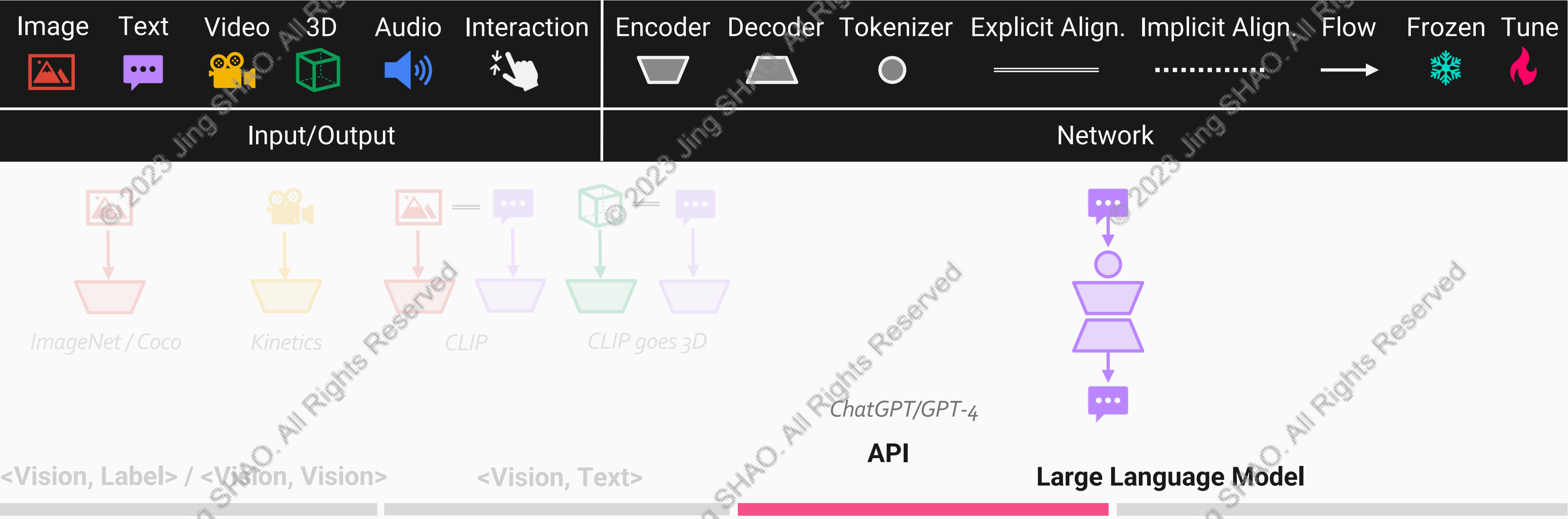


<Vision, Label> / <Vision, Vision> **<Vision, Text>**

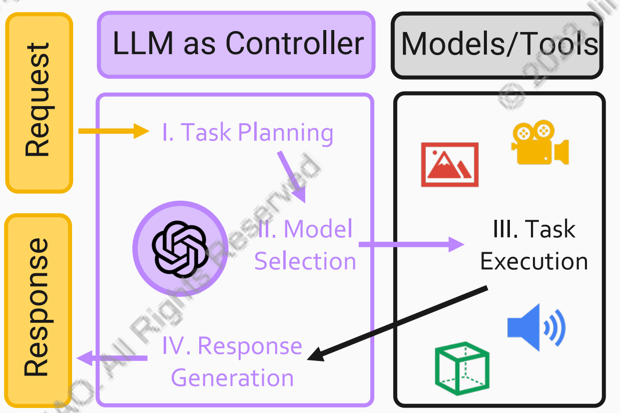


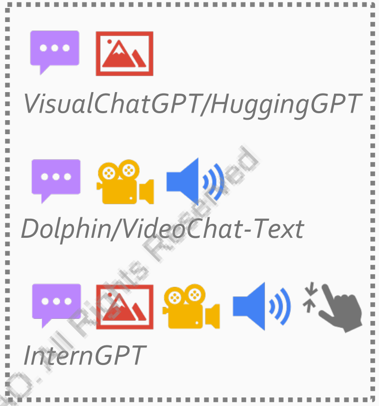
- Separate models for each task
  - **Separate** decoders
  - **Separate** finetuned encoders
  - **Limited** task number

© 2023 Jing SHAO. All Rights Reserved

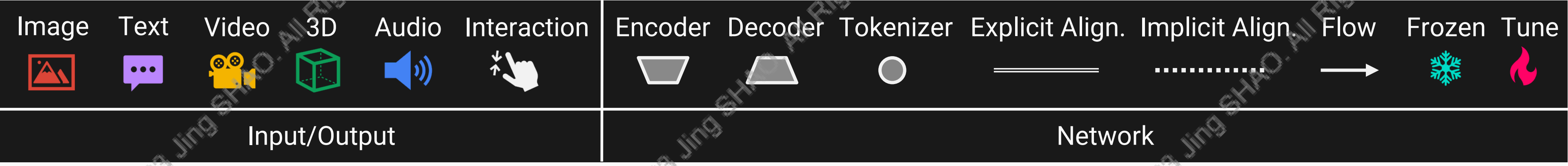


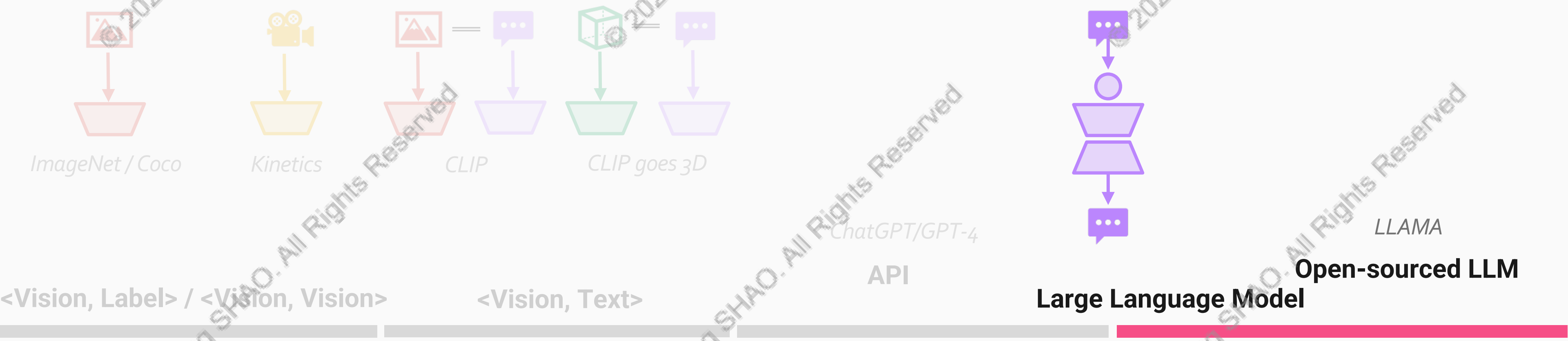
- MLLM based on **API**
  - LLM as **controller**
  - **N** models, one framework
  - Prompt engineering (e.g. flow design, prompt definition)
  - **No** need to finetune models



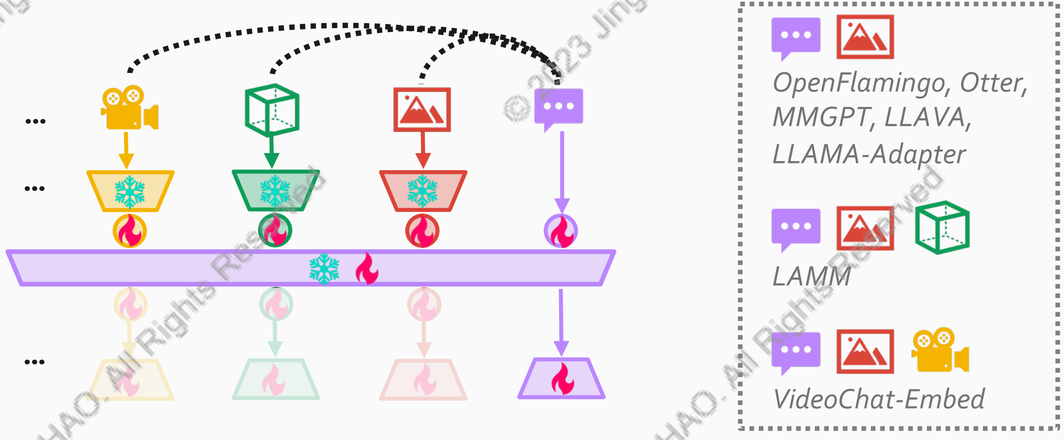


© 2023 Jing SHAO. All Rights Reserved



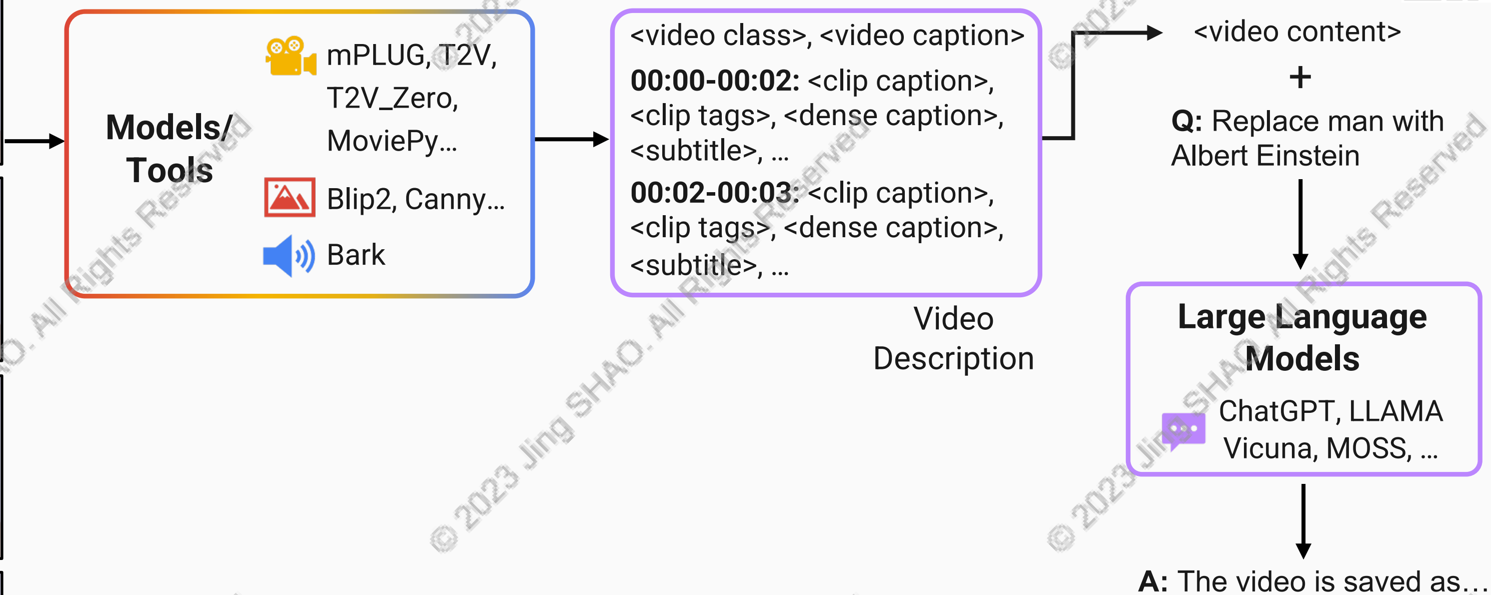


- MLLM based on **trainable models**
  - LLM as **base**
  - **One** model, one framework
  - Instruct tuning data (implicit alignment between different modalities)
  - **Trainable**, in-context learning (fewer data, parameter-efficient finetuning)



© 2023 Jing SHAO. All Rights Reserved

Dolphin | VideoChat | InternGPT



Models/ Tools

mPLUG, T2V, T2V_Zero, MoviePy…

Blip2, Canny…

Bark

<video class>, <video caption>

00:00-00:02: <clip caption>, <clip tags>, <dense caption>, <subtitle>, …

00:02-00:03: <clip caption>, <clip tags>, <dense caption>, <subtitle>, …

Video Description

<video content>

+

Q: Replace man with Albert Einstein

Large Language Models

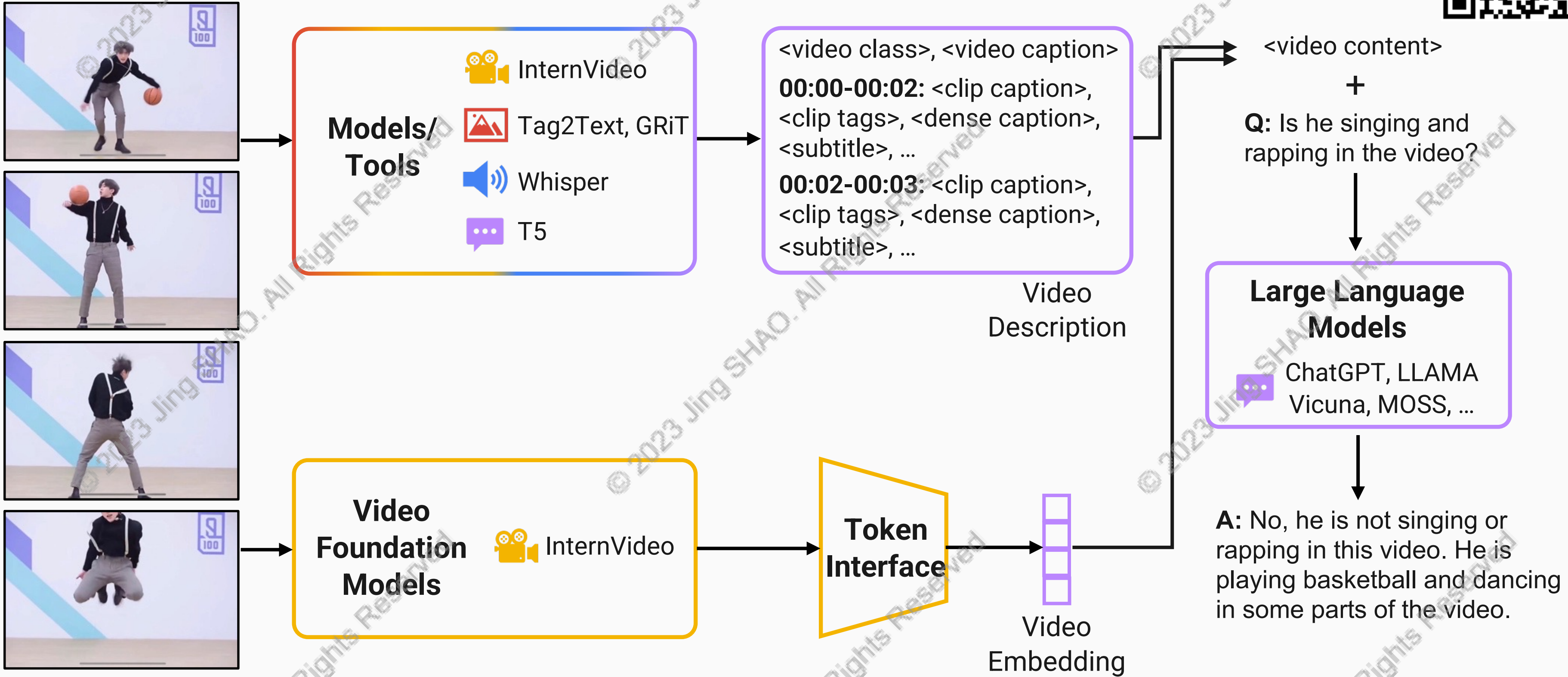
ChatGPT, LLAMA Vicuna, MOSS, …

A: The video is saved as…

Dolphin: A General Video Interaction Platform Based on LLMs, 2023

© 2023 Jing SHAO. All Rights Reserved

Dolphin | VideoChat | InternGPT



Models/ Tools

InternVideo

Tag2Text, GRiT

Whisper

T5

<video class>, <video caption>

00:00-00:02: <clip caption>, <clip tags>, <dense caption>, <subtitle>, …

00:02-00:03: <clip caption>, <clip tags>, <dense caption>, <subtitle>, …

Video Description

Video Foundation Models

InternVideo

Token Interface

Video Embedding

<video content>

+

Q: Is he singing and rapping in the video?

Large Language Models

ChatGPT, LLAMA Vicuna, MOSS, …

A: No, he is not singing or rapping in this video. He is playing basketball and dancing in some parts of the video.

© 2023 Jing SHAO. All Rights Reserved

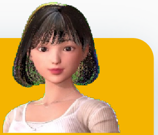
Dolphin | VideoChat | InternGPT

dolphin

General Video Interaction Platform
Based on LLMs

© 2023 Jing SHAO. All Rights Reserved

Dolphin | VideoChat | InternGPT



01

手语老师

Video Understanding & Video Generation

02

音乐家

Video Generation & Music Generation

03

演说肖像

Video Generation & Audio Generation

04

儿童教育

05

分镜制作

Q：请为我生成一段手语，内容是“你们好，很高兴认识你们，我是一名手语老师”。

A：好的，这个视频被保存为video/68g4.mp4。

Q：能帮我将视频中的手语翻译为中文嘛？

A：好的。这段手语对应的中文是：进入五月后，各地水果货物等农副产品大量上市，丰富着全国市民的菜篮子，也给农户们带来了收入。

© 2023 Jing SHAO. All Rights Reserved

Dolphin | VideoChat | InternGPT

01 手语老师

Video Understanding &
Video Generation

02 音乐家

Video Generation &
Music Generation

03 演说肖像

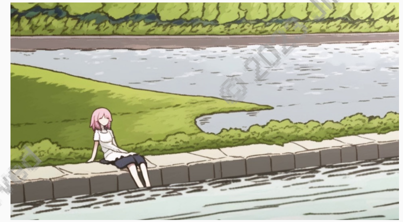
Video Generation &
Audio Generation

04 儿童教育

05 分镜制作

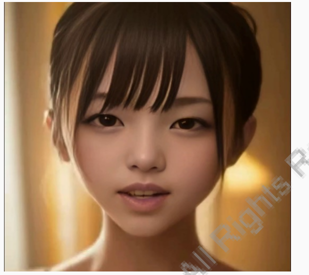
Q：请为我生成一段以”**悠闲的河畔**“为主题的音乐。

A：好的，这个视频被保存为video/pe7s.mp4。



Q：请为我生成一个小女孩的演说肖像，其演说内容为”大家好，我是一个演说肖像，我能够辅助直播，演讲，带货等各种工作，很高兴认识大家“。

A：好的，这个视频被保存为video/9qc7.mp4。



© 2023 Jing SHAO. All Rights Reserved

Dolphin | VideoChat | InternGPT

Video Chat 来啦

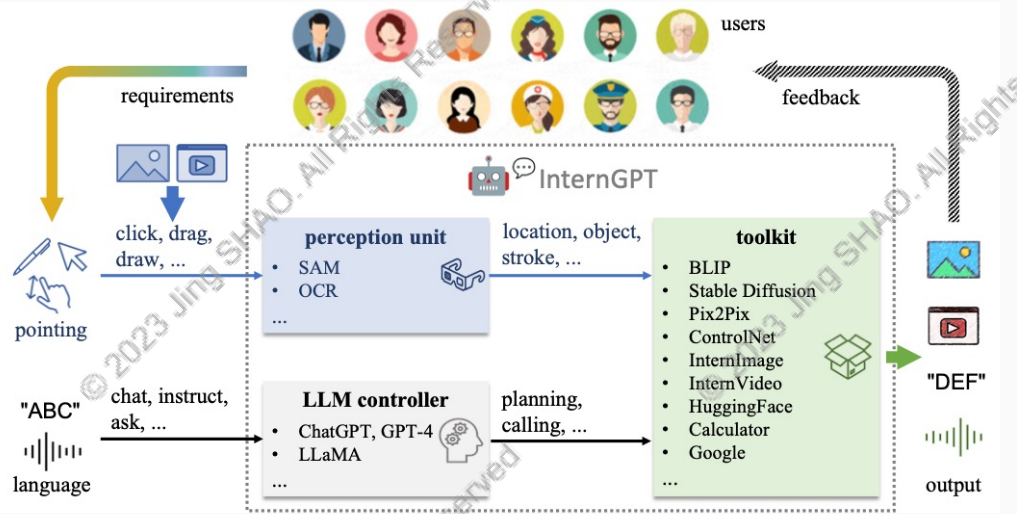
扫描右方二维码 加入讨论群

https://github.com/OpenGVLab/Ask-Anything

© 2023 Jing SHAO. All Rights Reserved

Dolphin | VideoChat | InternGPT

InternGPT is a multimodal interaction system that first proposed the cursor instruction



5 Modalities

It allows users to input visual signals to the language model through simple actions such as clicking and drawing lines

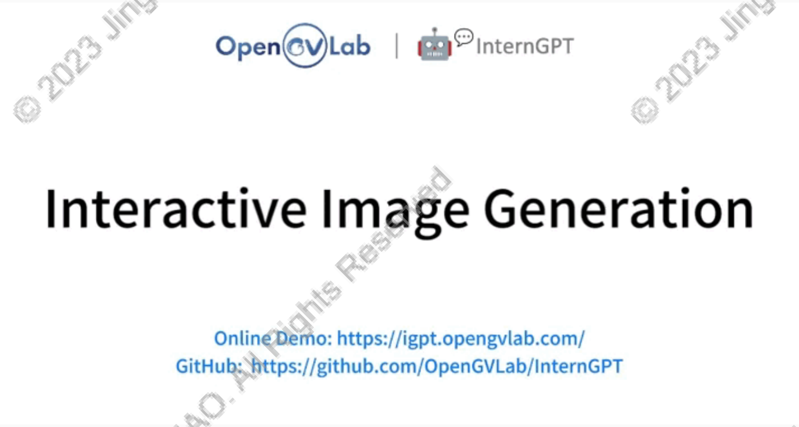
9 Tasks

Image cutouts, replacements, style transfer, image-based Q&A, drag-and-drop editing, sound-to-image conversion …

InternGPT: Solving Vision-Centric Tasks by Interacting with ChatGPT Beyond Language, 2023

© 2023 Jing SHAO. All Rights Reserved

Dolphin | VideoChat | InternGPT



OpenGVLab | InternGPT

Interactive Image Generation

Online Demo: https://igpt.opengvlab.com/
GitHub: https://github.com/OpenGVLab/InternGPT

OpenGVLab | InternGPT

InternGPT + DragGAN

Online Demo: https://igpt.opengvlab.com/
GitHub: https://github.com/OpenGVLab/InternGPT

OpenGVLab | InternGPT

Audio ⇨ Image

Online Demo: https://igpt.opengvlab.com/
GitHub: https://github.com/OpenGVLab/InternGPT

© 2023 Jing SHAO. All Rights Reserved

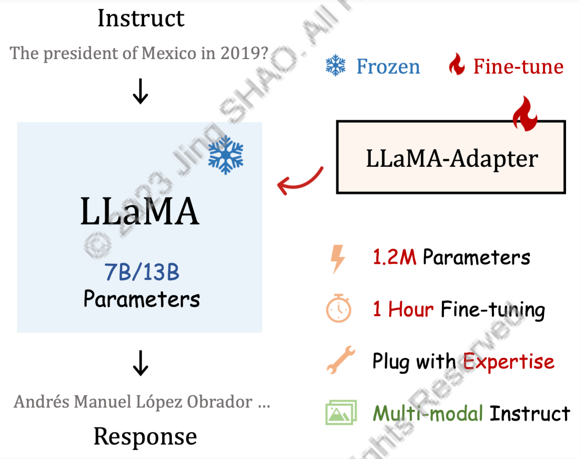
LLAMA-Adapter | LAMM

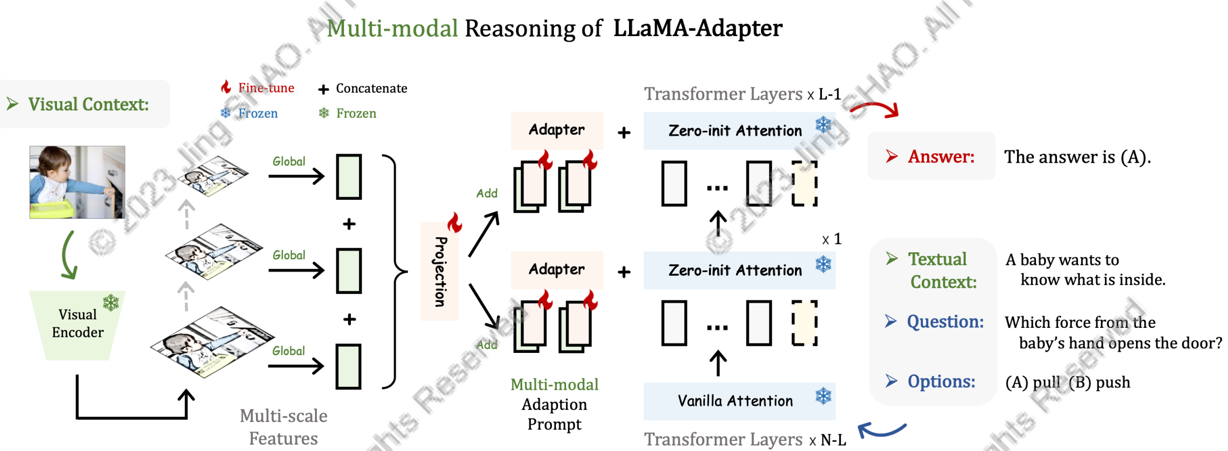
LLAMA-Adapter finetunes language models **efficiently** to support **multiple** modalities

4 Modalities
Image/Video/Sound/PointCloud

4 LLMs
StableLM/Pythia/INCITE/Falcon

High Impact
[Github] 4k Stars/[Twitter] 9M Views



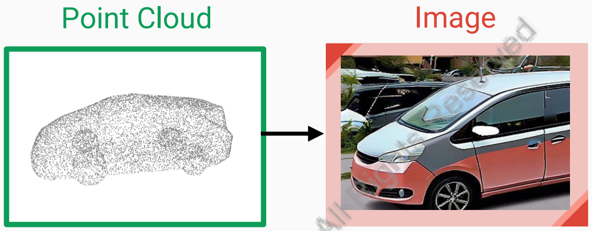


LLaMA-Adapter/LLaMA-Adapter v2, 2023

© 2023 Jing SHAO. All Rights Reserved

LLAMA-Adapter | LAMM

LLAMA-Adapter finetunes language models **efficiently** to support **multiple** modalities

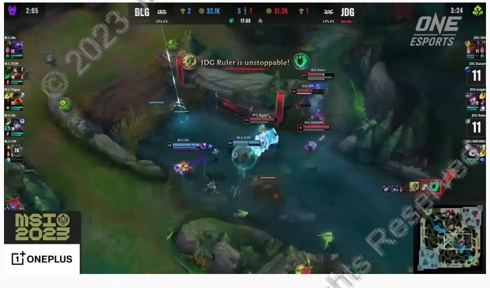


LLM+3D

Q: Describe the 3D object in detail.

A: The 3D object is a black car, specifically a minivan, parked in a parking lot. The car is positioned in the foreground of the image, with its front facing the viewer. The car appears to be a popular model, as there are several other cars parked in the lot, including a few cars in the background. The parking lot is located near a building, which could be a school or another type of establishment. The scene captures a typical day at a parking lot, with people going about their daily activities or running errands.

Video



LLM+Video

Q: What is the name of this game? What are the main characteristics?

A: The name of this game is “League of Legends,” and it is a multiplayer online battle arena (MOBA) game. The main characteristics of League of Legends include a team-based format, where two teams of five players each compete to destroy the opposing team‘s Nexus, a central structure that houses the game’s boss character, the “Nefarian.” Players use various champions, each with unique abilities and roles, to gain advantages and contribute to their team‘s success. The game is played on a map with multiple objectives, such as destroying enemy structures, capturing and defending areas, and engaging in teamfights. League of Legends is known for its strategic gameplay, skill-based mechanics, and competitive nature, making it a popular and engaging game for bot.

© 2023 Jing SHAO. All Rights Reserved

LLAMA-Adapter | LAMM

LAMM-InstructData

180k
<Image, Instruction, Response>

10k
<PC, Instruction, Response>

14
Datasets

LAMM-Benchmark

12+
Tasks

70k
<Image/PC, Instruction, Response>

6
4*A100 GPU Hours

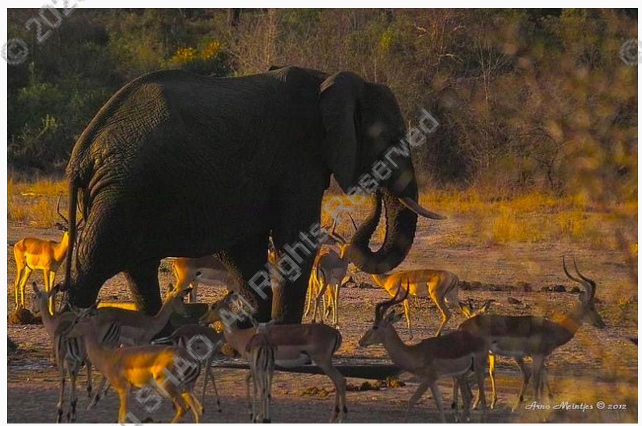
LAMM-Framework

LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark, 2023 June

© 2023 Jing SHAO. All Rights Reserved

LLAMA-Adapter | LAMM-InstructData

n-round Daily Dialogue



Instruction: <image> Is there only one elephant in the image?
Response: Yes, there is only one elephant in the image, surrounded by other animals.

Instruction: What other animals are present in the image with the elephant?
Response: The other animals in the image are deer, elk, and antelope. The presence of these animals suggests the picture captures a mix of wildlife species in their natural habitat.

Instruction: What is the size of the elephant in the image?
Response: The image shows a large elephant, bigger than the deer, elk, and antelope surrounding it.

Instruction: Is the elephant a baby or an adult?
Response: The elephant appears to be an adult, as it is described as "large" in the image.

Instruction: What are the animals doing in the image?
Response: The animals, including the elephant, are walking around in a field together. It appears to be a peaceful coexistence among them despite being different species. This signifies the uniqueness of the wildlife scene captured in the image.

LLaVA

LAMM

Instruction: <image> How many animals are depicted in the image?
Response: In the image, there are several animals depicted, including a large elephant and a group of antelope.

Instruction: What are the different animal species visible in the image?
Response: The image shows two distinct animal species - an elephant and several antelope.

Instruction: How can we identify the elephant in the image?
Response: The elephant in the image is identifiable by its large size, gray skin, long tusks, big ears, and a long trunk. These are the distinctive features of an elephant.

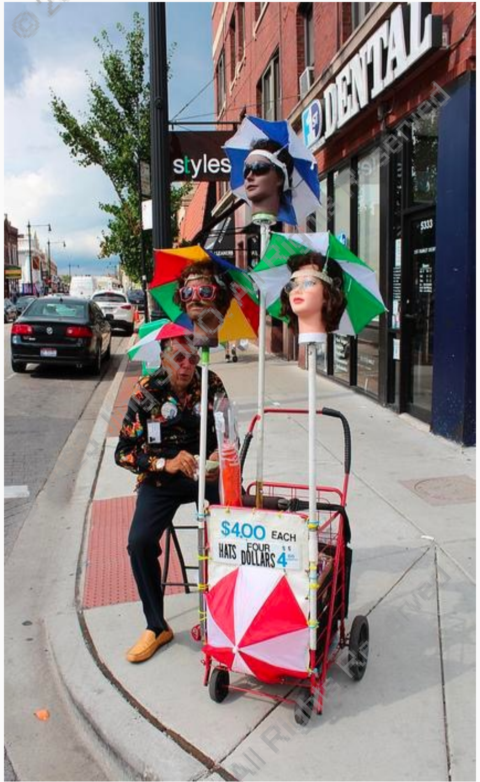
Instruction: How are the elephant and the antelope interacting in the image?
Response: The elephant appears to be walking peacefully among the antelope, suggesting a peaceful coexistence between these different species in their natural habitat.

Instruction: What can we infer from the image about the behavior of these animals?
Response: The image shows an elephant walking calmly among a group of antelope. This suggests that these animals are used to sharing the same habitat and living peacefully together. The absence of any visible signs of distress or aggression also indicates that there isn't any immediate threat present, and the animals feel safe in each other's presence..

© 2023 Jing SHAO. All Rights Reserved

LLAMA-Adapter | LAMM-InstructData

1-round Detailed Description



LLaVA

Instruction: <image> Write a detailed description of the given image.

Response: The image shows a man standing on a street corner with a cart, selling unique umbrella-style hats. There are colorful umbrellas attached to poles sticking out of a basket next to the man, drawing the attention of passersby. Five umbrellas of varying sizes can be seen displayed, with one large umbrella placed in the foreground.

Behind the man, two cars are parked along the side of the street, and another person is visible walking near the center of the scene. The man selling umbrella-style hats appears to be attracting interest or potentially waiting for customers to approach his makeshift street shop.

LAMM

Instruction: <image> Portray the image with a rich, descriptive narrative.

Response: The image depicts an intriguing outdoor scene on a city sidewalk. A man is seated comfortably on a chair, overseeing his eclectic array of merchandise for sale. A conspicuous part of his offerings include umbrella-style hats and sunglasses, displayed attractively on mannequin heads to attract potential buyers.

The man's temporary storefront is situated near a corner of the street, where he seems to be ready to engage with passers-by. A collection of green-leaved trees offers a pleasant shade and contributes to the street's overall ambiance.

The surroundings reveal an urban setting with multiple buildings constructed of red bricks and adorned with various windows. One building, in particular, has a sign indicating that it's a dental office, adding a touch of everyday life to the picture. A manhole cover can be seen on the sidewalk, suggesting a well-maintained urban infrastructure. Despite being an unusual sight, the man selling umbrella-style hats and sunglasses provides a touch of local color and personality to the city street corner. His presence adds a unique charm to the overall cityscape and hints at the diverse street culture in the area.

© 2023 Jing SHAO. All Rights Reserved

LLAMA-Adapter | LAMM-InstructData

© 2023 Jing SHAO. All Rights Reserved

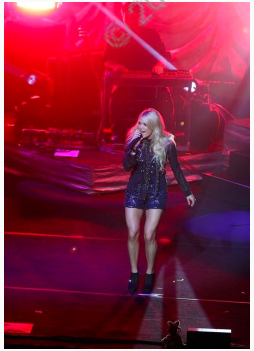
n-round Factual Knowledge Dialogue

Instruction: <image> What type of building is shown in the image?
Response: The building shown in the image is a monastery, which is the residence of a religious community.
Instruction: Can you describe the architectural features of the monastery in the image?
Response: While specific architectural features can vary widely between different monasteries, many share common elements such as a central courtyard or cloister, a chapel or church, dormitory-style living quarters, and communal spaces like a refectory for communal meals. The image shows a building that has a sturdy, typically stone-built structure designed for communal living and religious activities.
Instruction: What is the purpose of a monastery and how does its design reflect this?
Response: A monastery is a building or complex of buildings where monks or nuns live in a community. The primary purpose of a monastery is to support the religious work of its residents, whether that be prayer, meditation, or other forms of spiritual practice. As such, the design of a monastery often reflects these needs, with spaces dedicated to worship, study, work, and communal living. It's intended to be self-sufficient and promote a peaceful, contemplative environment.

LAMM

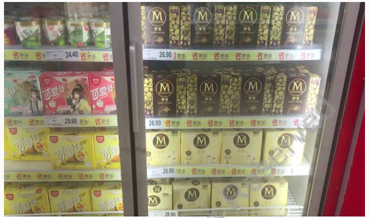
1-round Visual Task Dialogue

LAMM (Detection)



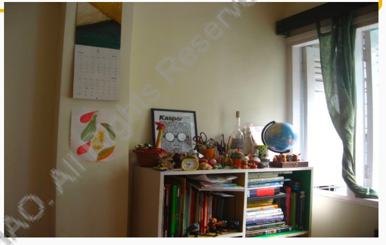
Instruction: <image> Can you accurately report the names and positions of all the objects visible in the image?
Response: The microphone object is present at the [0.49, 0.38, 0.56, 0.43] coordinate in the image. An object that can be classified as footwear is located at the [0.5, 0.77, 0.55, 0.84] position of the image. Upon close inspection of the image, it can be observed that there is an object positioned at [0.56, 0.78, 0.61, 0.85] that belongs to the footwear category. The image contains an object that can be classified as person and is located at the [0.48, 0.31, 0.78, 0.84] position.

LAMM (Counting)



Instruction: <image> Please identify the number of objects in the image that are labeled as a certain category.
Response: The system has detected 30 Ice cream present in the image.

LAMM (OCR)



Instruction: <image> what large name is on the framed picture on the shelf?
Response: A careful examination of the picture indicates that the correct answer is kasper.

LLAMA-Adapter | LAMM-InstructData

© 2023 Jing SHAO. All Rights Reserved

n-round Daily Dialogue

Instruction: <point cloud> Is there a TV cabinet in the room?
Response: Yes. The point cloud describes a full-sized corner cabinet with a large opening in the center for a television.

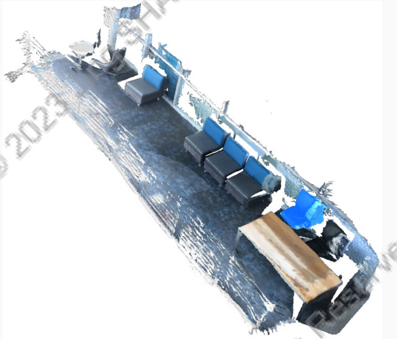
Instruction: <point cloud> Where is the desk located in the room?
Response: The desk is located in front of a black office chair with wheels.

1-round Detailed Description

Instruction: <point cloud> What is this point cloud about?
Response: The point cloud shows a living space with a large gray L-shaped couch positioned in the middle of the room. The couch is adjacent to the kitchen area and across from the bedroom area, providing a cozy and functional gathering space….

1-round Visual Task Dialogue

Instruction: <point cloud> Provide a detailed account of the objects observed in the point cloud, including their precise locations?
Response: An object within the table category can be observed at the [0.070, 3.632, 0.624, 1.110, 1.976, 1.083] position in the point cloud. Positioned at the [-1.705, 0.211, 0.417, 0.611, 9.417, 1.632] location, an object within the window category…



LLAMA-Adapter | LAMM-InstructData

© 2023 Jing SHAO. All Rights Reserved

Inputs

① System Message

You are an AI visual assistant …

C1. You need generate a conversation …

C2. You need to generate multiple rounds conversations based on Bamboo label and Wikipedia description …

C3. You need to describe the scene in detail …

C4. Give me 20 similar instruction-response prompts…

② In-context Learning Pairs

Human-labeled <instruction, response> pairs

③ Query

C1. <Caption> + <Boxes> + <Relationships>

C2. <Bamboo labels> + <Wikipedia description>

C3. <Caption> + <Boxes> + <Relationships>

C4. N/A

GPT API

Outputs

C1: n-round Daily Dialogue

C2: n-round Factual Knowledge Dialogue

C3. 1-round Detailed Description

C4: 1-round Visual Task Dialogue

2D & 3D Vision Tasks

- Object Detection
- Image Classification
- Keypoint Detection
- Object Counting
- OCR
- 3D indoor Detection

Instruction Pool & Response Template

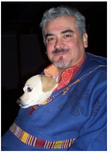
Instructions: [What are the names of the objects present in the image, and where are they positioned?, How many items in the image belong to a particular category? …]

Responses: [There is an object at the {P} position of the image, and its category is {C}. , There are {N} {C} visible in image …]

Dataset

LLAMA-Adapter | LAMM-Benchmark

Detection



You are performing an object detection task. Your goal is to locate all objects in the image, and give the corresponding coordinates.

Your output should be a single class name. Your answer should be the object name and the bounding box, represented as [x1, y1, x2, y2] with floating numbers ranging from 0 to 1.

Instruction: What is the category of the scene shown in the image?

Response: There is a dog [0.1, 0.4, 0.6, 0.7], and a person with the position from [0,0] to [1,1].

Counting

You are performing an object counting task. Your goal is to accurately count the number of objects in an image. Object counting is a computer vision task that involves detecting and counting the number of instances of specific objects within an image.

The answer needs to be an Arabic numeral.

Instruction: How many seashells are there in this image?

Response: There are eight seashells in this image.

{Task Definition | Output Structure | Instruction}

MLLM

Grounding

You are performing a point cloud grounding task, and our goal is to locate the position of objects in the question.

Your answer should be a bounding box, represented as [x1, y1, z1, x2, y2, z2] with floating numbers ranging from 0 to 1.

Instruction: There is a TV in the room. It is on the wall. Give me the bounding box of it in the point cloud.

Response: It locates at [0.790, 1.477, 0.219, 0.434, 0.687, 0.3667].

VQA

You are performing a visual question answering task, and your goal is to generate natural language answers that accurately solve the question.

Generate complex reasoning process and accurate answers to questions about visual content

Instruction: What is placed next to the fridge?

Response: Upon thorough examination of the point cloud, it can be inferred that the answer to this question is door.

Entity Extraction

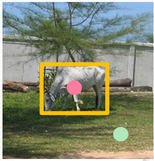
LAMM Evaluation

I. Traditional Metrics

Classification	Accuracy
Detection	mAP
Caption	BLEU
Counting	MAE
OCR	Word ACC

⋮

II. Binary-Loc. Metric

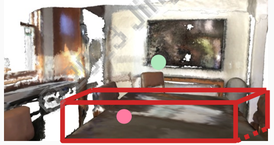


Inside GT bbox

● Positive

Outside GT bbox

● Negative



III. GPT Metric

Task Definition | GT | Eval-Instruct

Model A Response

Model B Response

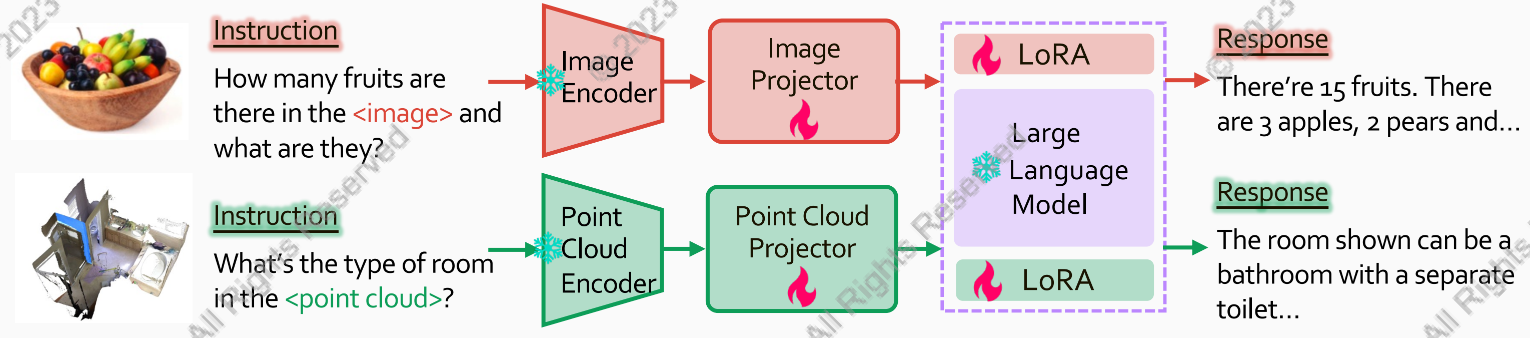
Model C Response

GPT Ranking

1. Model C
2. Model A
3. Model B

© 2023 Jing SHAO. All Rights Reserved

LLAMA-Adapter | LAMM-Benchmark



Instruction

How many fruits are there in the <image> and what are they?

Instruction

What's the type of room in the <point cloud>?

Response

There're 15 fruits. There are 3 apples, 2 pears and...

Response

The room shown can be a bathroom with a separate toilet...

IDs	Task	Dataset	Metric	LLaVA	MiniGPT4	mPLUG	LAMM
1	Classification	CIFAR10	Acc ↑	**60.83**	46.22	42.5	34.5
2	Detection	VOC2012	mAP ↑	1.42	0.92	0.158	**4.82**
3	VQA	SQAimage	Acc ↑	40.5	43.43	36.39	**47.15**
		AI2D	Acc ↑	18.13	Failed	19.31	**19.5**
4	Image Caption	flickr30k	BLEU4 ↑	**6.65**	5.1	2.74	0.70
5	F-g classification	UCMerced	Acc ↑	**47**	33.6	32.5	13
6	Counting	FSC147	MAE ↑	56.2	Failed	60.67	**53.97**
7	OCR	SVT	Word Acc ↑	**37.78**	16.97	30.39	4.2
8	Facial Classification	CelebA(Smile)	Acc ↑	Failed	**66.36**	Failed	51.3
		CelebA(Hair)	Acc ↑	**46.42**	43.47	40.93	30.48
9	Keypoints Detection	LSP	PCK ↑	Failed	Failed	Failed	Failed
10	Binary Localization	LAMM	Acc ↑	14.73	13.12	4.42	**31.2**
11	GPT Evaluation	LAMM	Score ↑	11	N/A	N/A	**89**

Capabilities:

- recognize open-vocabulary classes
- understand images
- answer questions

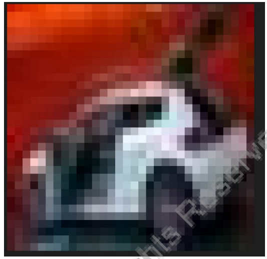
Limitations:

- localization-aware tasks
- human-centric tasks
- fine-grained specific domains

© 2023 Jing SHAO. All Rights Reserved

LLAMA-Adapter | LAMM-Benchmark

IDs	Task	Dataset	Metric	SOTA	LLaVA	MiniGPT4	mPLUG	LAMM	LAMM-FT
1	Classification	CIFAR10	Acc ↑	99.5	**60.83**	46.22	42.5	34.5	91.2
2	Detection	VOC2012	mAP ↑	97.2	1.42	0.92	0.158	**4.82**	13.48
3	VQA	SQAimage	Acc ↑	92.53	40.5	43.43	36.39	**47.15**	74.27
		AI2D	Acc ↑	N/A	18.13	Failed	19.31	**19.5**	
4	Image Caption	flickr30k	BLEU4 ↑	30.1	**6.65**	5.1	2.74	0.70	
5	F-g classification	UCMerced	Acc ↑	100	**47**	33.6	32.5	13	
6	Counting	FSC147	MAE ↑	10.79	56.2	Failed	60.67	**53.97**	
7	OCR	SVT	Word Acc ↑	97.9	**37.78**	16.97	30.39	4.2	
8	Facial Classification	CelebA(Smile)	Acc ↑	N/A	Failed	**66.36**	Failed	51.3	
		CelebA(Hair)	Acc ↑	N/A	**46.42**	43.47	40.93	30.48	
9	Keypoints Detection	LSP	PCK ↑	99.5	Failed	Failed	Failed	Failed	
10	Binary Localization	LAMM	Acc ↑	N/A	14.73	13.12	4.42	**31.2**	
11	GPT Evaluation	LAMM	Score ↑	N/A	11	N/A	N/A	**89**	


CIFAR10-Car

Fiat 500L 2012

Instruction: What is the category for this image?

Ground Truth: car

LAMM Response: Fiat 500L 2012

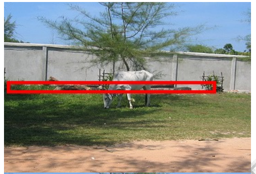
LLaVA Response: Car

Fine-Grained Recognition Capability

© 2023 Jing SHAO. All Rights Reserved

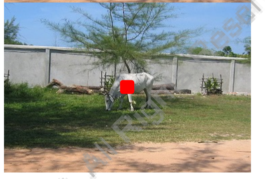
LLAMA-Adapter | LAMM-Benchmark

IDs	Task	Dataset	Metric	SOTA	LLaVA	MiniGPT4	mPLUG	LAMM	LAMM-FT
1	Classification	CIFAR10	Acc ↑	99.5	**60.83**	46.22	42.5	34.5	91.2
2	Detection	VOC2012	mAP ↑	97.2	1.42	0.92	0.158	**4.82**	13.48
3	VQA	SQAimage	Acc ↑	92.53	40.5	43.43	36.39	**47.15**	74.27
		AI2D	Acc ↑	N/A	18.13	Failed	19.31	**19.5**	
4	Image Caption	flickr30k	BLEU4 ↑	30.1	**6.65**	5.1	2.74	0.70	
5	F-g classification	UCMerced	Acc ↑	100	**47**	33.6	32.5	13	
6	Counting	FSC147	MAE ↑	10.79	56.2	Failed	60.67	**53.97**	
7	OCR	SVT	Word Acc ↑	97.9	**37.78**	16.97	30.39	4.2	
8	Facial Classification	CelebA(Smile)	Acc ↑	N/A	Failed	**66.36**	Failed	51.3	
		CelebA(Hair)	Acc ↑	N/A	**46.42**	43.47	40.93	30.48	
9	Keypoints Detection	LSP	PCK ↑	99.5	Failed	Failed	Failed	Failed	
10	Binary Localization	LAMM	Acc ↑	N/A	14.73	13.12	4.42	**31.2**	
11	GPT Evaluation	LAMM	Score ↑	N/A	11	N/A	N/A	**89**	



Instruction: Give the bounding box of all the objects in the image.

LLaVA Response: There is a white horse grazing in a fenced-in area with grass. The horse is positioned near the left side of the image. To provide the precise coordinates, the horse's bounding box would be (0.156, 0.477, 0.767, 0.523).



Instruction: Give the position of all the objects in the image.

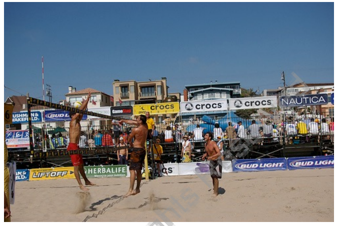
LLaVA Response: In the image, there is a white horse grazing on grass near a wall. The horse's position is approximately [0.5, 0.5].

Localization Capability

© 2023 Jing SHAO. All Rights Reserved

LLAMA-Adapter | LAMM-Benchmark

IDs	Task	Dataset	Metric	SOTA	LLaVA	MiniGPT4	mPLUG	LAMM	LAMM-FT
1	Classification	CIFAR10	Acc ↑	99.5	**60.83**	46.22	42.5	34.5	91.2
2	Detection	VOC2012	mAP ↑	97.2	1.42	0.92	0.158	**4.82**	13.48
3	VQA	SQAimage	Acc ↑	92.53	40.5	43.43	36.39	**47.15**	74.27
		AI2D	Acc ↑	N/A	18.13	Failed	19.31	**19.5**	
4	Image Caption	flickr30k	BLEU4 ↑	30.1	**6.65**	5.1	2.74	0.70	
5	F-g classification	UCMerced	Acc ↑	100	**47**	33.6	32.5	13	
6	Counting	FSC147	MAE ↑	10.79	56.2	Failed	60.67	**53.97**	
7	OCR	SVT	Word Acc ↑	97.9	**37.78**	16.97	30.39	4.2	
8	Facial Classification	CelebA(Smile)	Acc ↑	N/A	Failed	**66.36**	Failed	51.3	
		CelebA(Hair)	Acc ↑	N/A	**46.42**	43.47	40.93	30.48	
9	Keypoints Detection	LSP	PCK ↑	99.5	Failed	Failed	Failed	Failed	
10	Binary Localization	LAMM	Acc ↑	N/A	14.73	13.12	4.42	**31.2**	
11	GPT Evaluation	LAMM	Score ↑	N/A	11	N/A	N/A	**89**	



LLaVA Response: The image shows a group of people playing a game of volleyball on a beach. They are actively engaged in the game, with one man jumping in the air to hit the ball. The scene is lively and energetic, capturing the spirit of outdoor sports and beach activities.

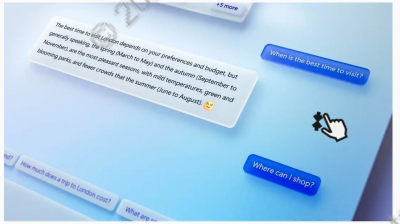
LAMM Response: The image captures an engaging scene from a professional sand volleyball game. Two men can be seen in the center of the court, fully immersed in the game. Both players are dressed in casual sportswear, with one wearing a white shirt and the other in a shirt of the same color. The sand court stretches across the image, providing ample space for the players to move around. There are multiple flags...

Description Capability

© 2023 Jing SHAO. All Rights Reserved

Insights for Future

Search Engine


+5 more

The best time to visit London depends on your preferences and budget, but generally speaking, the spring (March to May) and the autumn (September to November), are the most pleasant seasons, with mild temperatures, green and blooming parks, and fewer crowds that the summer (June to August). 😊

When is the best time to visit?

Where can I shop?

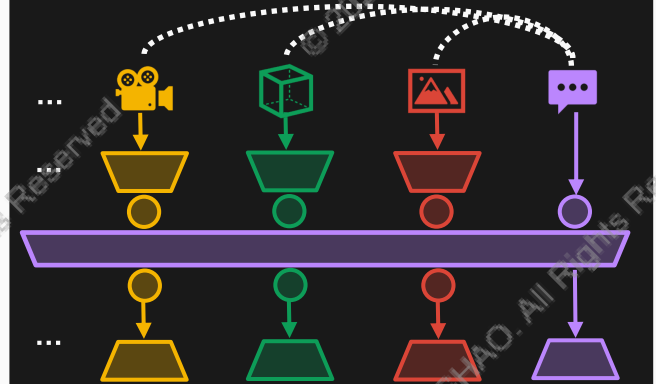
How much does a trip to London cost?

What are

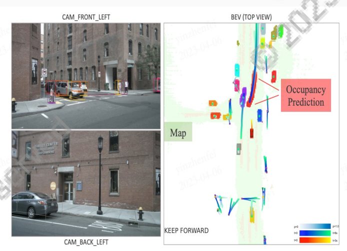
Instruction: <action> Based on my previous trip to Hawaii, how should I arrange my trip to Phuket?

Response: I recommend you to book a Marriott and then…

A Unified Multi-Modal Framework for AGI



Autonomous Driving



Instruction: <image> <pcloud> <action> Drive me efficiently to work and park underground.

Response: OK. We will take Rd. 66 and then… <action>

Robotics

Instruction: <image> <pcloud> <action> Cook a General Tso's Chicken for us before 6 p.m.

Response: OK. Since there're no chicken in the fridge, I need to… <action>

Human-Centric Applications

Instruction: <image> <mesh> Generate a motion of 3D digital humans while keeping the identification of the person in <image> but changing his clothes by <image> and acting as shooting the basketball.

Response: OK. Shooting the basketball should first jump and raise his right arm <mesh>….

Interactive Image Synthesis

Instruction: <action> <image> Draw a herd of running house…

Response: Here are a few images you might like… <image>

Video Applications

Instruction: Here is a real <video>, transfer its style to make an animation of 2mins, 30FPS…

Response: Here is the animation <video>, it tells a funny story that…

Low Level Applications

Instruction: <image> Denoise the image, remove the raindrop in the background, super resolution to 8K.

Response: Here are a few images…<image>

© 2023 Jing SHAO. All Rights Reserved

个人主页

OpenGVLab主页

OpenMMLab主页

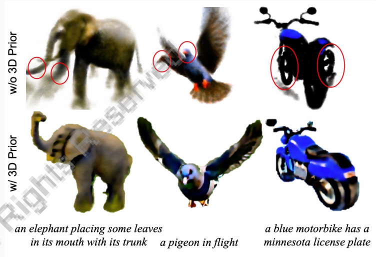
Hi3D

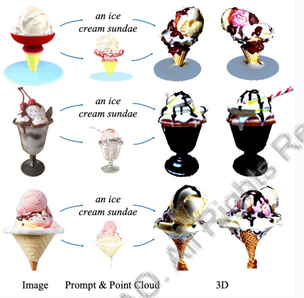
"From Geometry to Texture: A Hierarchical Framework for Efficient Text-to-3D Generation"

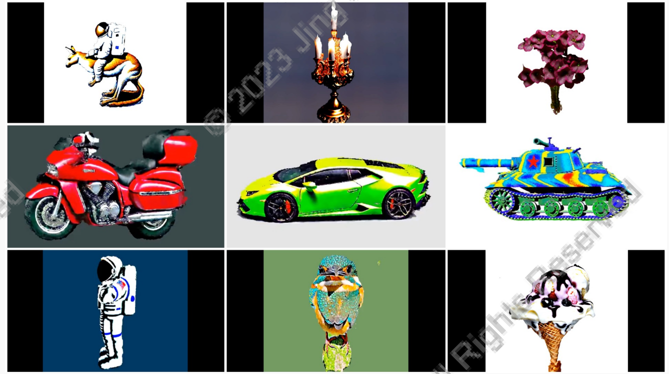
i. Efficient 10x 7x 4x speed up

DreamFusion/Magic3D/Fantasia3D

ii. Geometric Stable ***iii. Controllable***

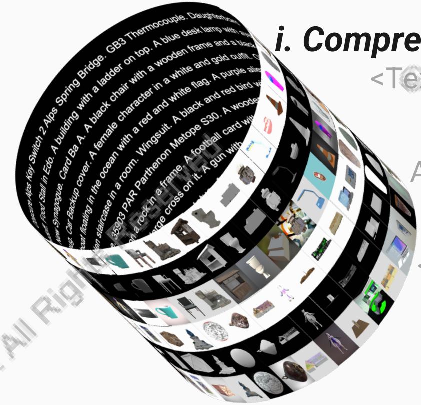

w/o 3D Prior
w/ 3D Prior
an elephant placing some leaves in its mouth with its trunk
a pigeon in flight
a blue motorbike has a minnesota license plate





UniG3D

"UniG3D: A Unified 3D Object Generation Dataset"



i. Comprehensive data format

<Text,3D-PCL,3D-Mesh,2D>

ii. Unified Pipeline

Adapt to any 3D dataset

iii. Large-scale

<550K,550K,5.5M,11M>

Otter+MIMIC-IT

"MIMIC-IT: Multi-Modal In-Context Instruction Tuning"

© 2023 Jing SHAO. All Rights Reserved

Thank You

© 2023 Jing SHAO. All Rights Reserved