

LLM LE10 VT2025

Multi-Modal Large Language Models

Fredrik Heintz

Dept. of Computer Science

Linköping University

fredrik.heintz@liu.se

@FredrikHeintz

Outline:

- Introduction
- Architecture
- Modality and Functionality
- Multi-Modal Instruction Tuning
- Multi-modal Reasoning

Language modelling

- **Language modelling** is the task of predicting which word comes next in a sequence of words.
- More formally, given a sequence of words w_1, \dots, w_t we want to know the probability of the next word, w_{t+1} :

$$P(w_{t+1} | w_1, \dots, w_t)$$

- We are assuming that w_{t+1} comes from a finite vocabulary V .

language models = classifiers

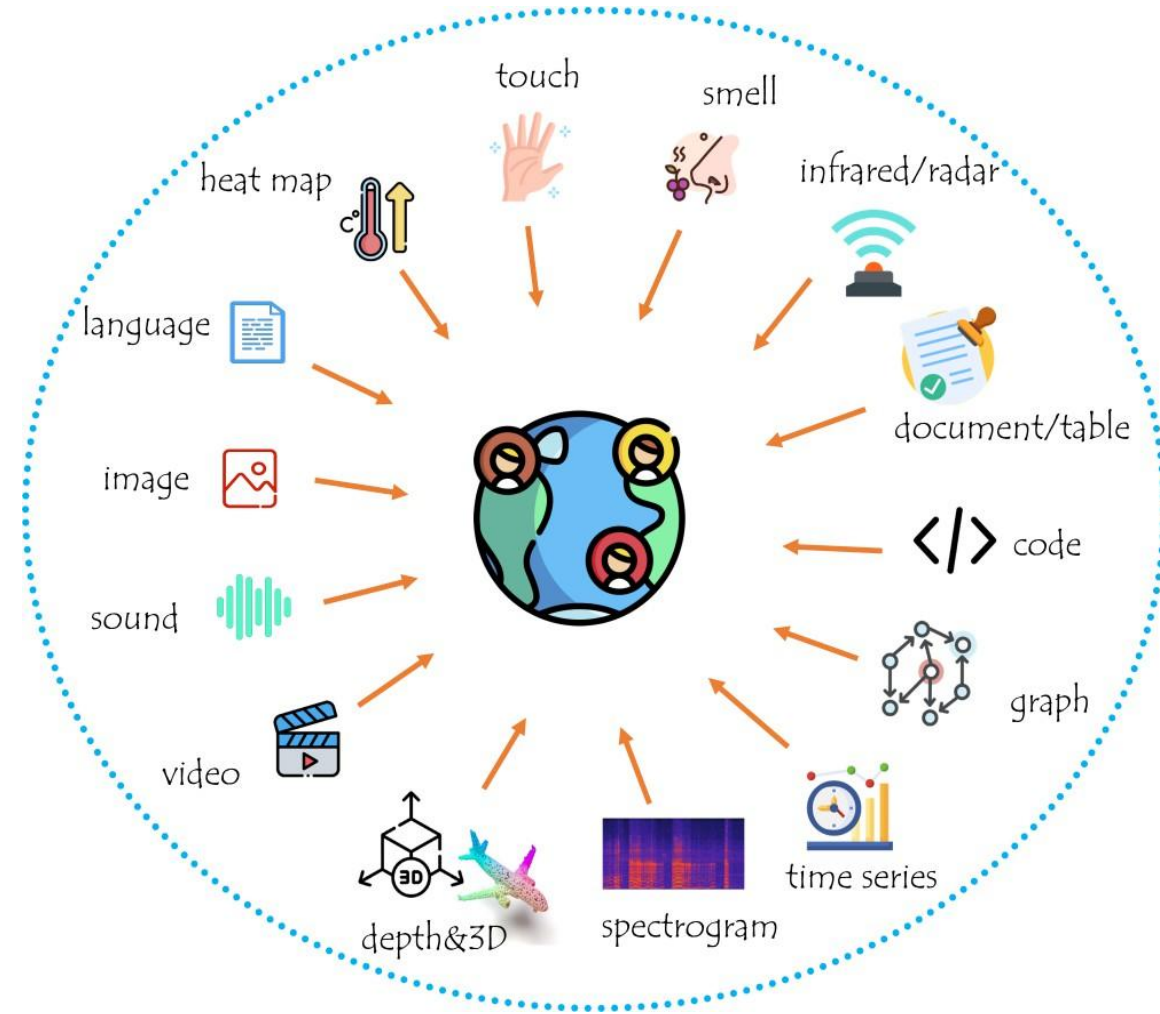


Intelligence in Multi-Sensory Data

- Harnessing Multimodality



This world we live in is replete with multimodal information & signals, **not just language.**



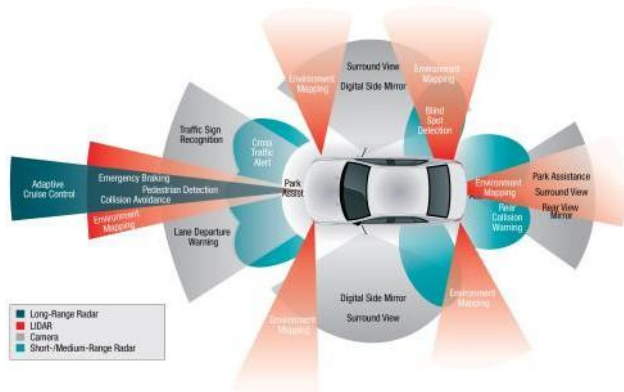
Intelligence in Multi-Sensory Data

● Harnessing Multimodality

👉 This world we live in is replete with multimodal information & signals, not just language.

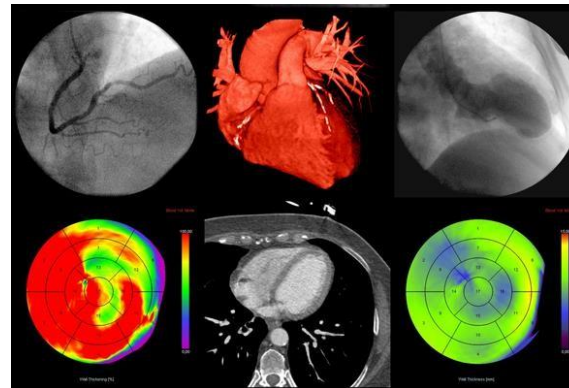
○ Autonomous Driving Systems

In this application, vehicles use a combination of visual data (cameras), spatial data (LiDAR), and auditory signals (sonar) to navigate safely.



○ Healthcare Diagnostics

*Medical **imaging** tools like MRI, CT scans, and X-rays, along with patient history and verbal symptoms, are used to diagnose diseases.*




○ Smart Home Assistants

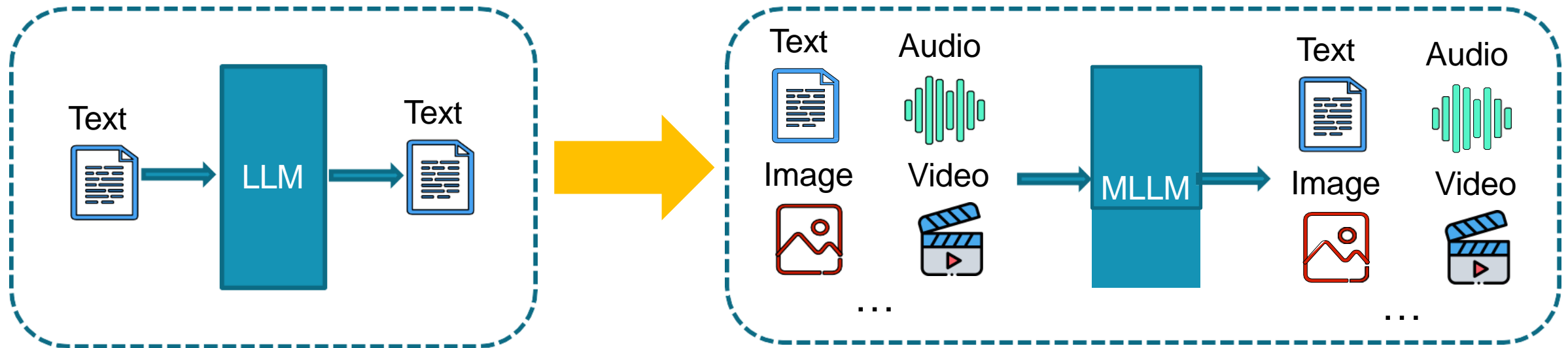
*Devices like Amazon Alexa and Google Home use voice commands (**audio**), physical interaction (**touch**), and sometimes **visual** cues to operate.*



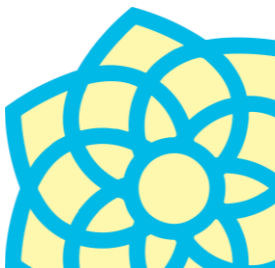
Intelligence in Multi-Sensory Data

- Building Multimodal LLMs (MLLMs)

 Can we transfer the success of **LLMs** to **MLLMs**, enabling LLMs to comprehend multimodal information as deeply as they understand language?

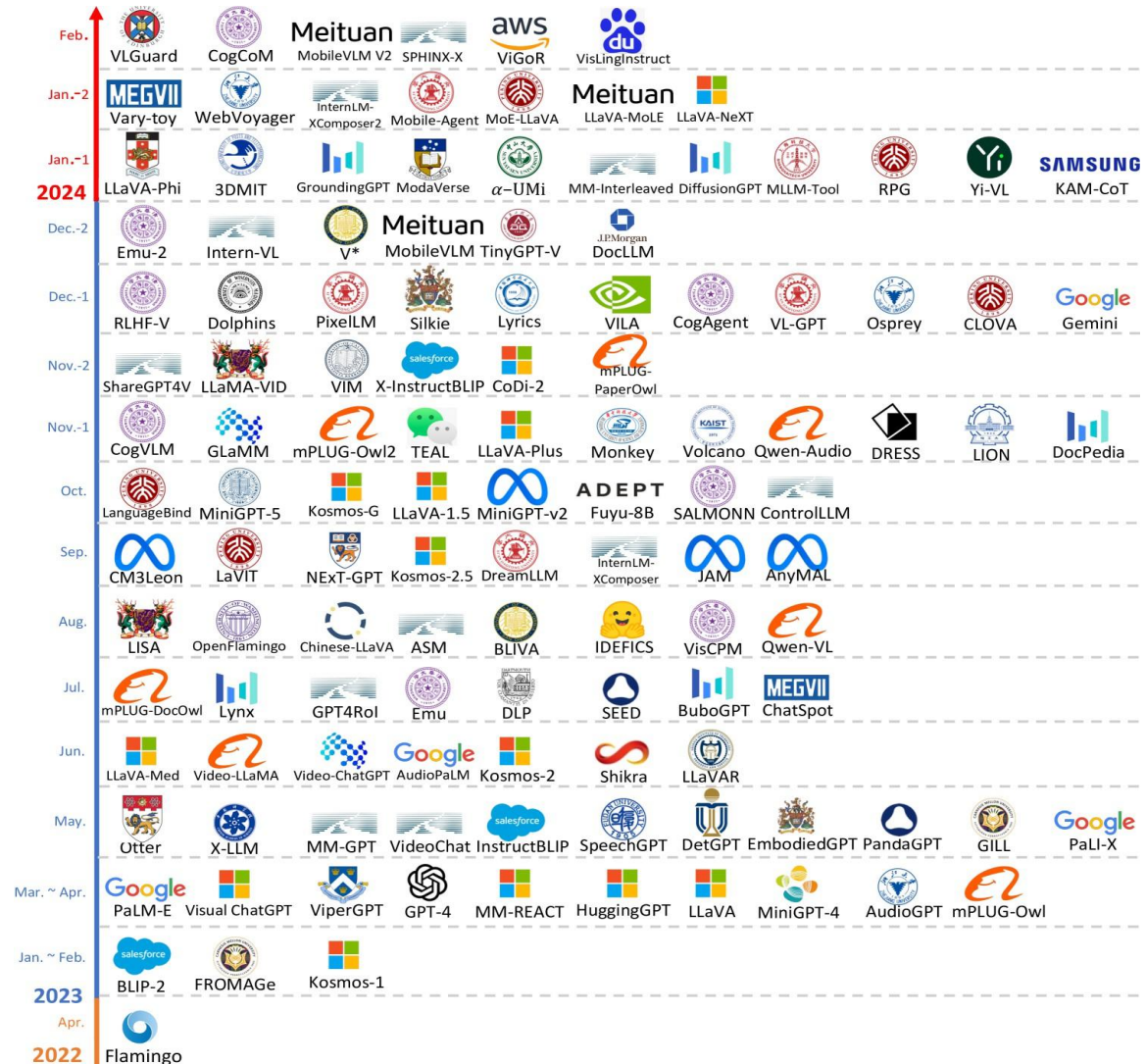


 Perceiving and interacting with the world as **HUMAN BEINGS** do, might be the key to achieving human-level AI.



Intelligence in Multi-Sensory Data

- Trends of MLLMs



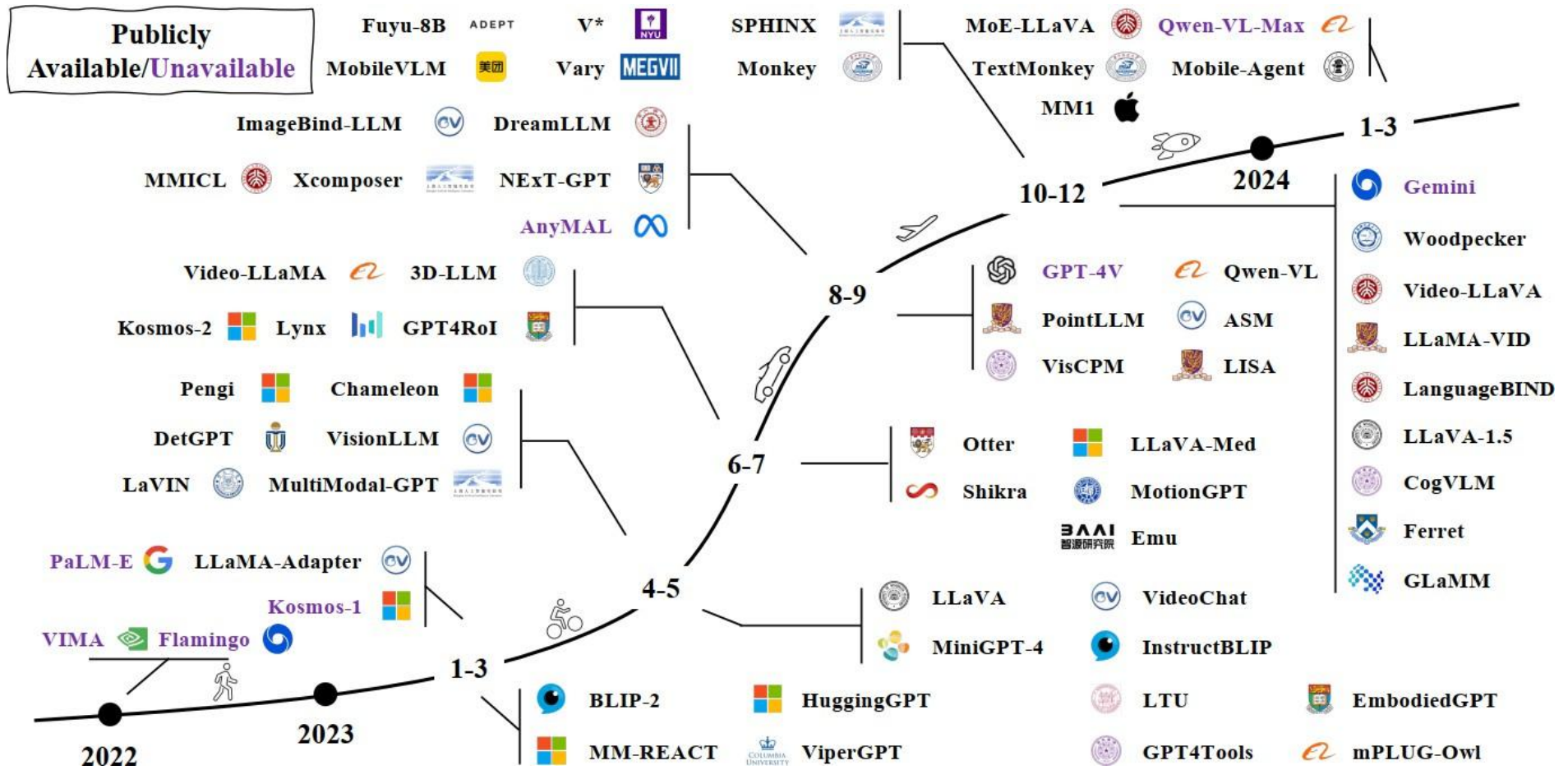
[1] MM-LLMs: Recent Advances in MultiModal

Large Language Models, 2023.



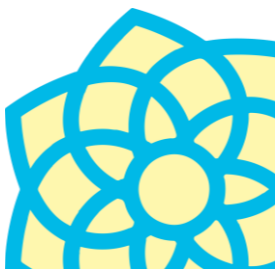
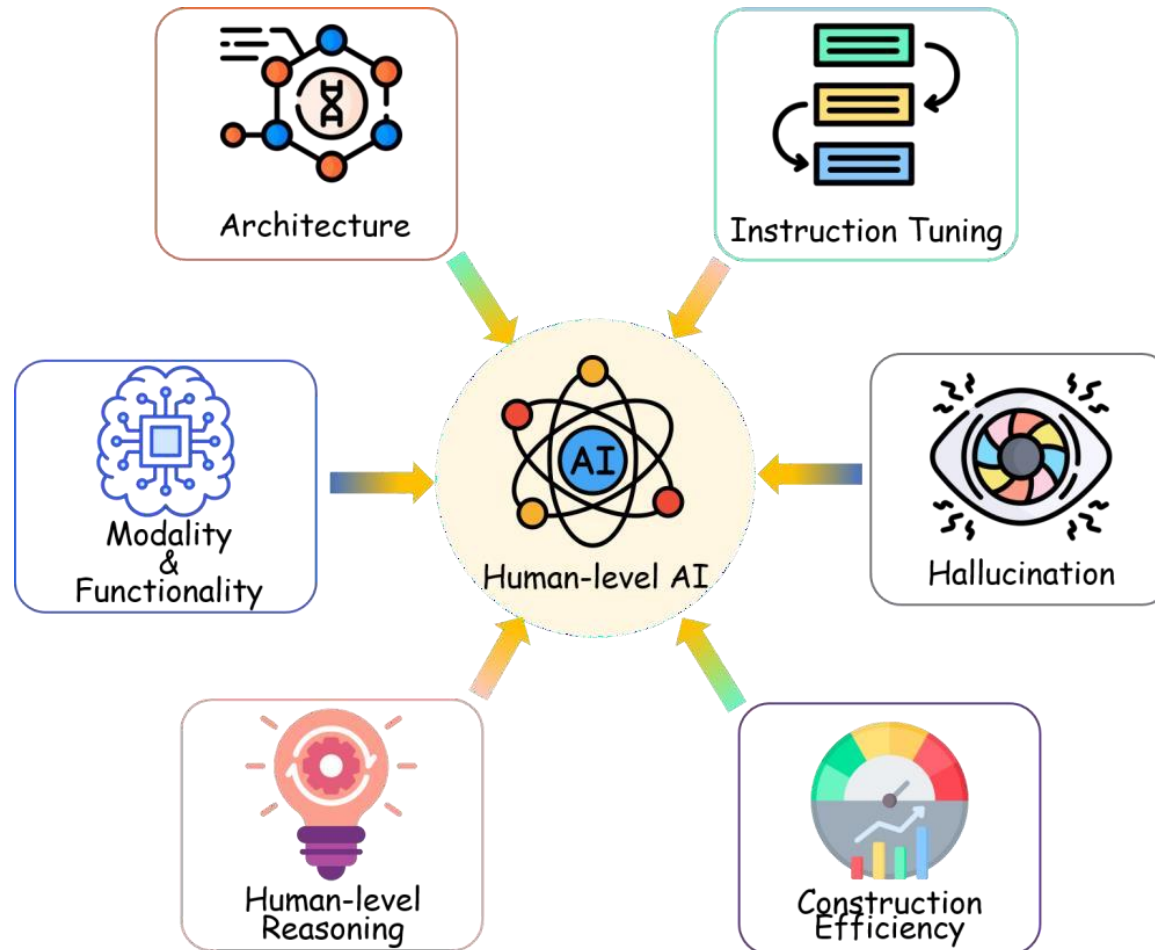
Intelligence in Multi-Sensory Data

Trends of MLLMs



From MLLMs to Human-level AI

- Key Aspects for Building Powerful MLLMs



* Part-II

MLLM Design: Architecture



Yuan Yao

Research Fellow

National University of Singapore

<https://yaoyuanthu.github.io/>

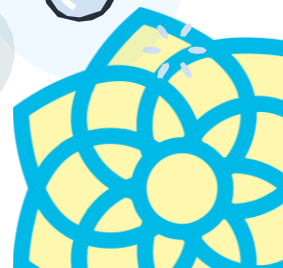


Table of Content

0 **Architecture**

- × Overview: Basic Architecture
- × Multimodal Encoding
- × Input-side Projection
- × Backbone LLMs
- × Decoding-side Projection
- × Multimodal Generation



1

Architecture of MLLM

How to design an MLLM?



Overview of MLLM Architecture

- Preliminary Idea: Intelligence over Language



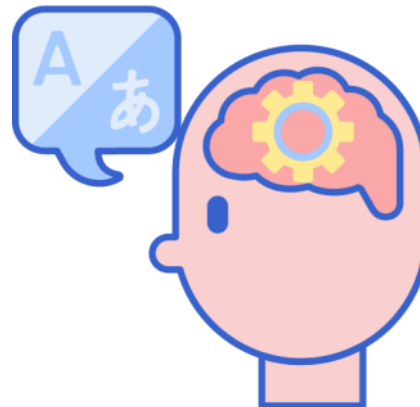
Emergent phenomena have extensively already occurred in language-based LLMs.



These LLMs now generally possess very powerful **semantic understanding capabilities**.



This also implies that **language is a crucial modality for carrying intelligence**.



language



Overview of MLLM Architecture

• Preliminary Idea: Language Intelligence as Pivot

- 👉 Given this premise, **nearly all CURRENT MLLMs are built based on language-based LLMs** as the core decision-making module (i.e., the brain or central processor).
- 👉 By adding additional external non-textual modality modules, LLMs are enabled with multimodal abilities.
 - Extend the capability boundary, next milestone towards more advanced intelligence
 - More applications



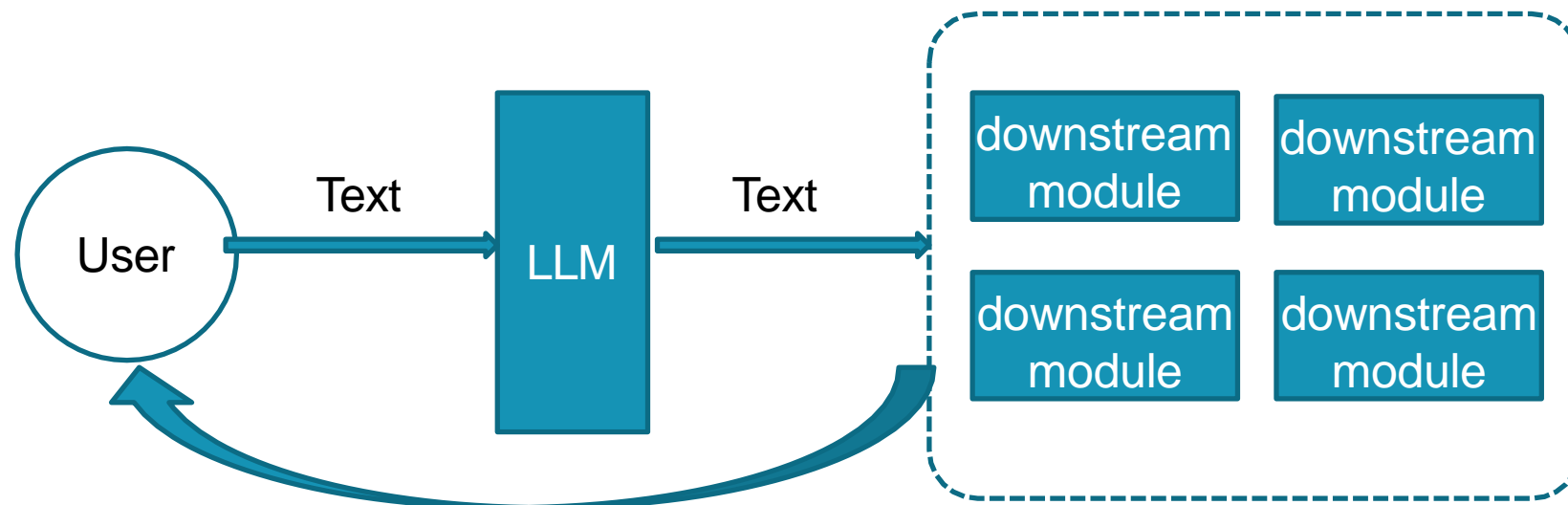
Overview of MLLM Architecture

- Architecture-I: LLM as Discrete Scheduler/Controller

 The role of the LLM is to **receive textual signals** and **instruct textual commands to call downstream modules**.

- Key feature:

All message passing within the system, such as “multimodal encoder to the LLM” or “LLM to downstream modules”, is facilitated through **pure textual** commands as the medium.

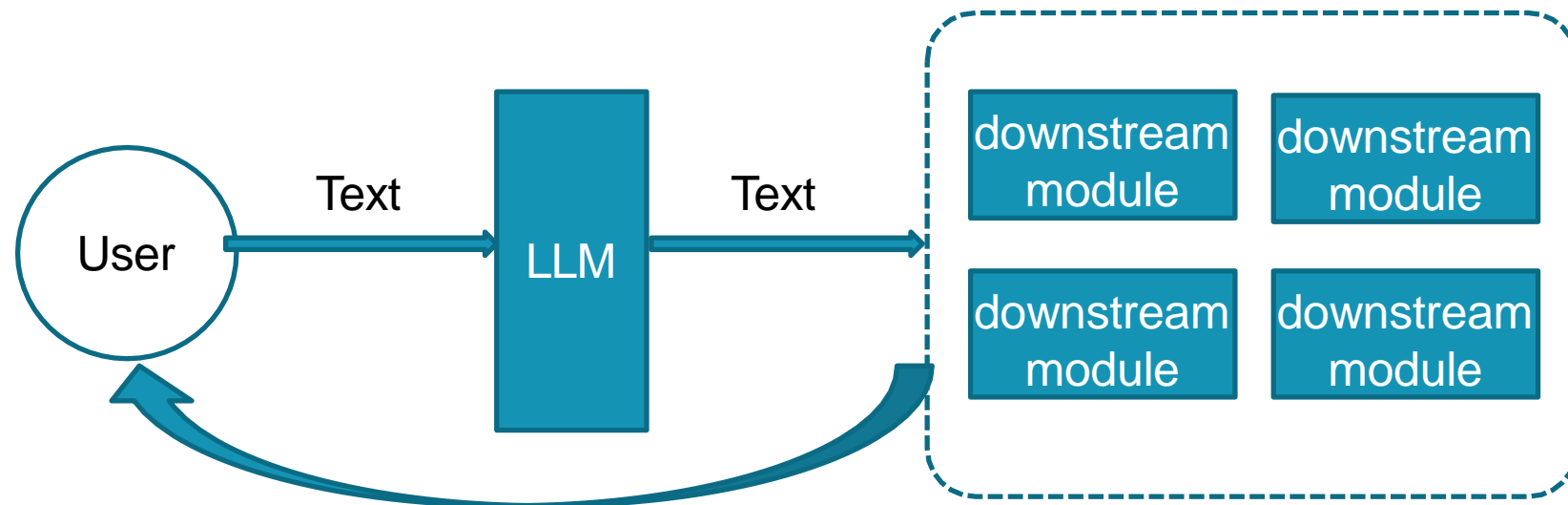


Overview of MLLM Architecture

- Architecture-I: LLM as Discrete Scheduler/Controller

- Representative MLLMs:

- Visual-ChatGPT
- HuggingGPT
- MM-REACT
- ViperGPT
- AudioGPT
- LLaVA-Plus
- ...

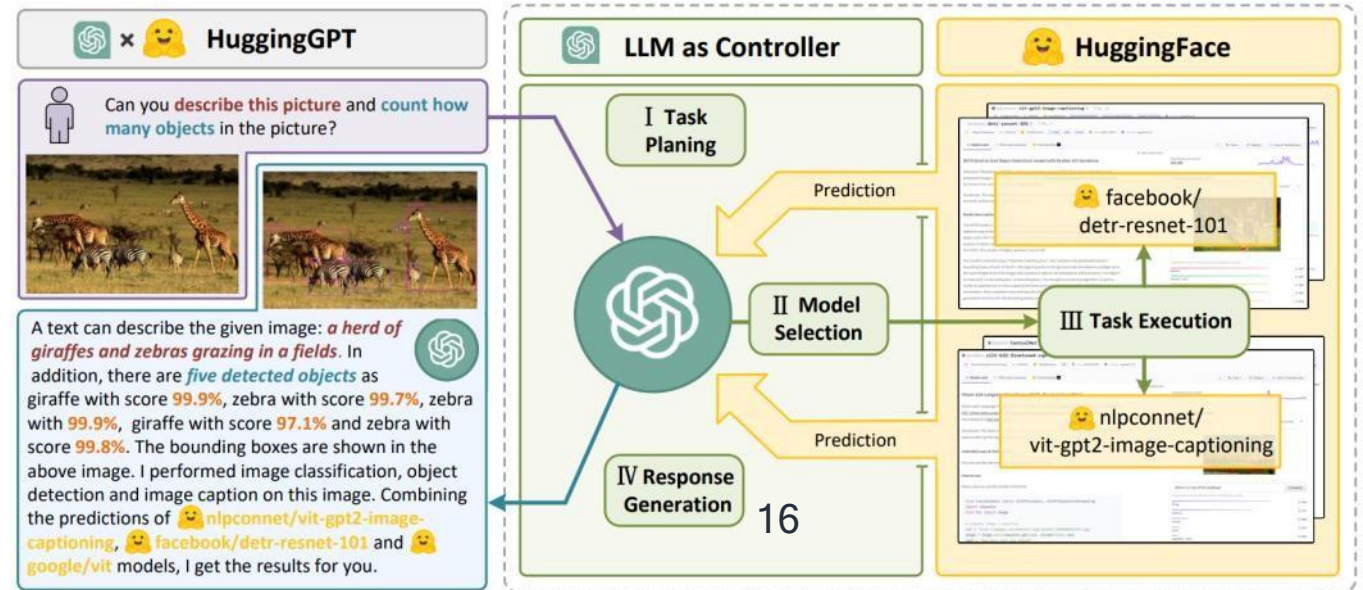
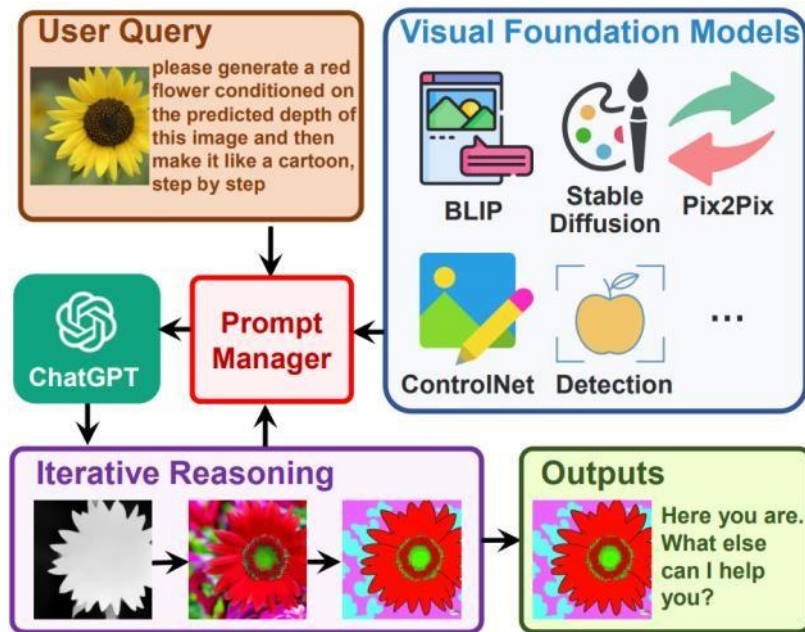


Overview of MLLM Architecture

- Architecture-I: LLM as Discrete Scheduler/Controller

- o Visual-ChatGPT

- o HuggingGPT

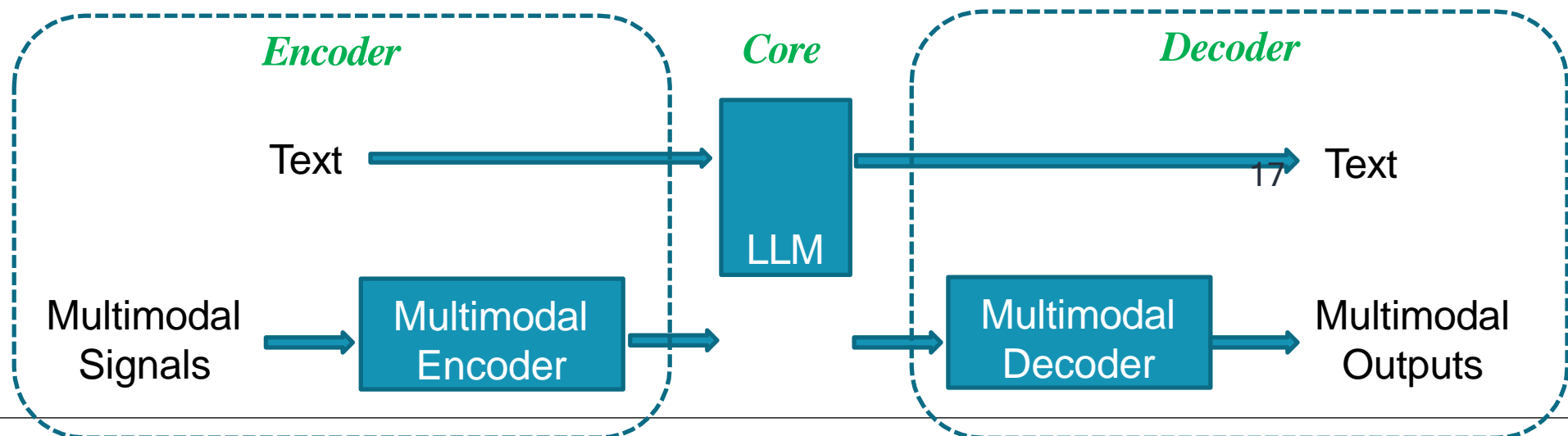


- o Quick to build (without training), flexible extension to many tool features
- o Information loss in text medium, the bottle-neck



Overview of MLLM Architecture

- Architecture-II: LLM as Joint Part of System
 - The role of the LLM is to perceive multimodal information, and **react by itself**, in a structure of **Encoder-LLM-Decoder**.
 - Key feature:
 - LLM is the key joint part of the system, **receiving multimodal information directly from outside**, and delegating instruction to decoders/generators in a more smooth manner.

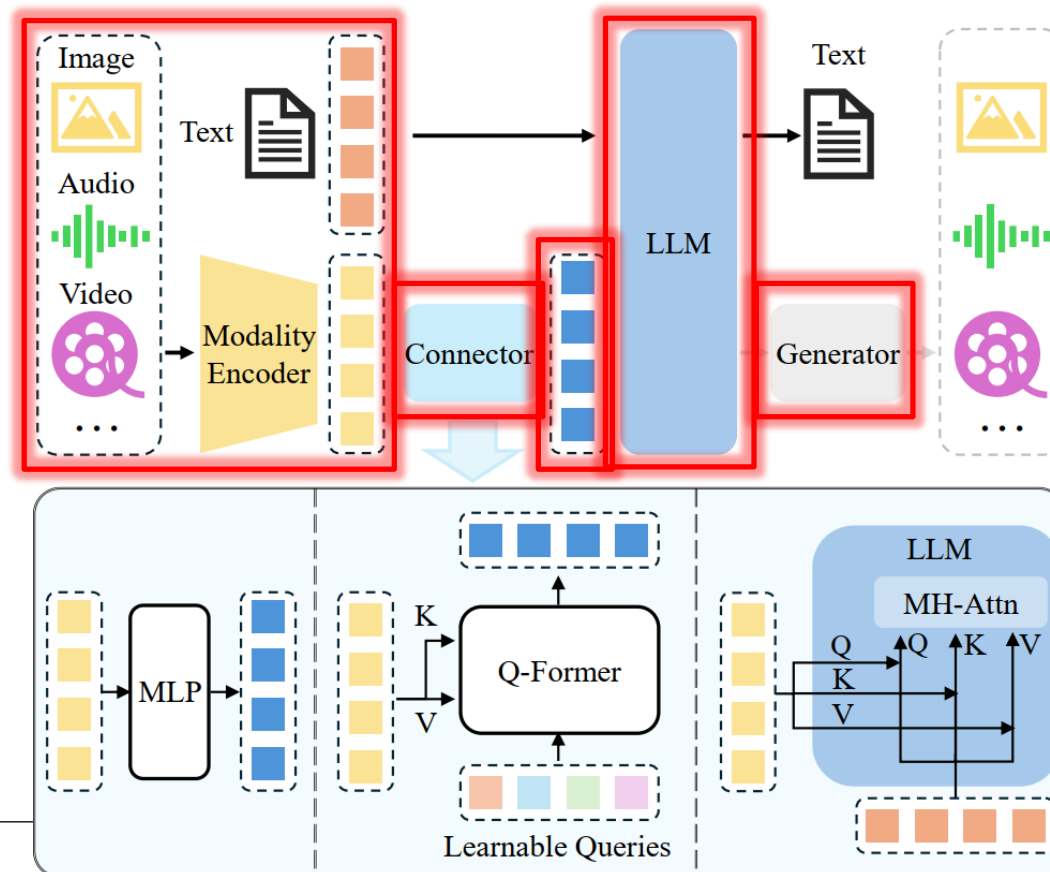


Overview of MLLM Architecture

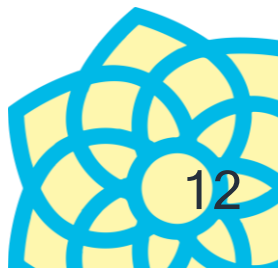
- Architecture-II: LLM as Joint Part of System

More promising

- > 90% MLLMs belong to this category.
- Higher upper-bound, better integrated into a unified model



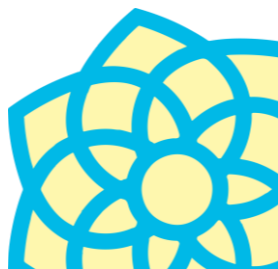
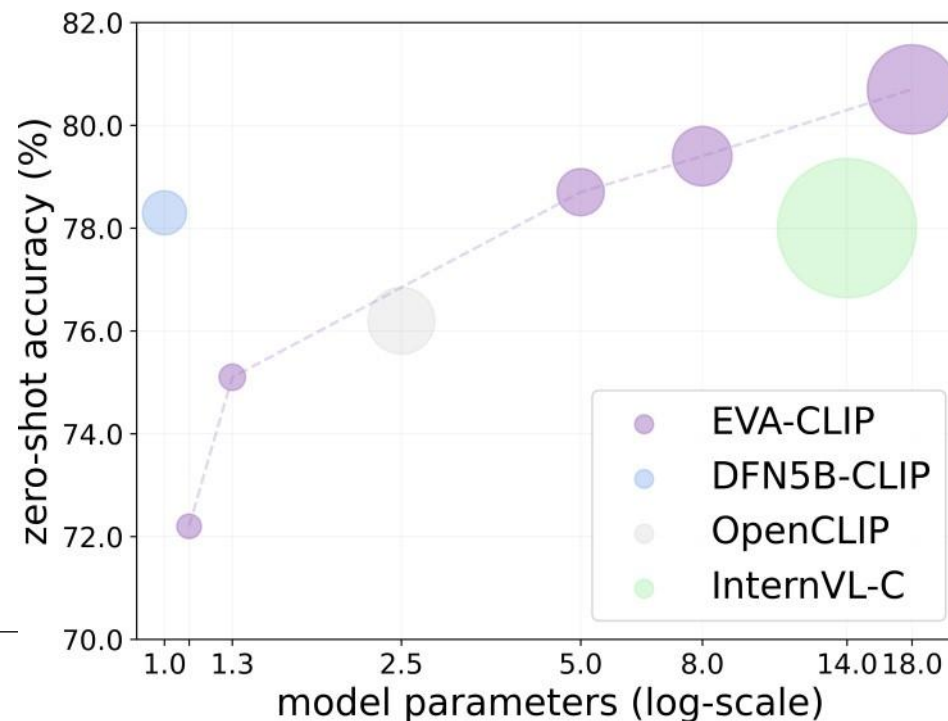
[1] A Survey on Multimodal Large Language Models.
<https://github.com/BradyFU/A-wesome-Multimodal-Large-Language-Models, 2023.>



Multimodal Encoding

• Visual Encoder

- **CLIP-ViT** is the most popular choice for vision-language models.
 - ✗ Providing image representations well aligned with text space.
 - ✗ Scale well with respect to parameters and data.
- SigLIP is gaining increasing popularity (smaller and stronger)



Multimodal Encoding

- Visual Encoder

- Limitations of existing pretrained ViTs:
 - × Fixed low-resolution (224x224 or 336x336) in square shape
- High-resolution perception is essential, especially for OCR capability!



Low resolution encoding misses fine-grained visual details!



Multimodal Encoding

- Visual Encoder

- High-resolution Multimodal LLMs

- ✗ Image slice-based: Split high-resolution images into slices

- ✗ Representatives:

- ◆ GPT-4V, LLaVA-NeXT, MiniCPM-V 2.0/2.5, LLaVA-UHD, mPLUG-DocOwl 1.5, SPHINX, InternLM-XComposer2-4KHD, Monkey

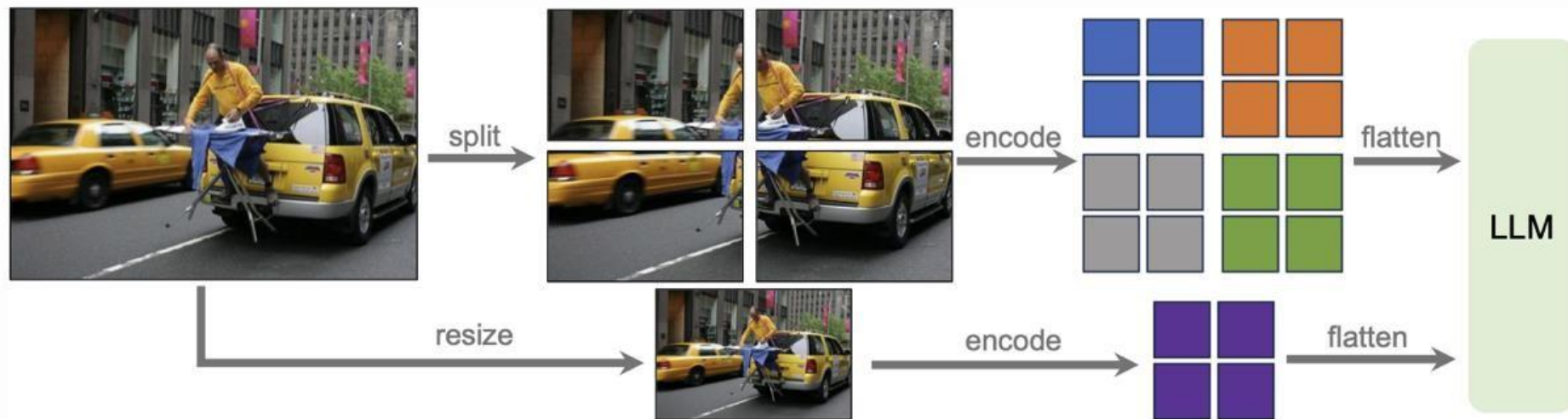


Illustration of dynamic high resolution scheme: a grid configuration of 2×2



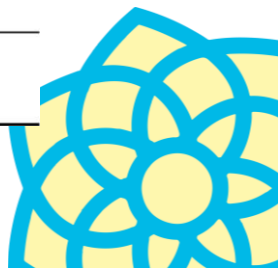
Multimodal Encoding

Visual Encoder

High-resolution Multimodal LLMs

- ✗ Image slice-based: Split high-resolution images into slices
- ✗ OCR capabilities improves significantly without new data

Model	#Data	MaxRes.	AR.	TFLOPs	VQA ^{v2}	GQA	VQA ^T	POPE	SQA	VizWiz	MME	MMB	MMB ^{CN}
BLIP-2 [21]	129M	224×224	Fix	1.0	41.0	41.0	42.5	85.3	61.0	19.6	1293.8	-	-
InstructBLIP [11]	130M	224×224	Fix	1.0	-	49.5	50.7	78.9	63.1	33.4	1212.8	-	-
Shikra [8]	6M	224×224	Fix	8.0	77.4	-	-	-	-	-	-	58.8	-
Qwen-VL [5]	1.4B	448×448	Fix	9.2	78.8	59.3	<u>63.8</u>	-	67.1	35.2	-	38.2	7.4
SPHINX [24]	1.0B	448×448	Fix	39.7	78.1	62.6	51.6	80.7	69.3	39.9	1476.1	66.9	56.2
SPHINX-2k [24]	1.0B	762×762	Fix	69.4	<u>80.7</u>	63.1	61.2	<u>87.2</u>	70.6	44.9	1470.7	65.9	57.9
MiniGPT-v2 [7]	326M	448×448	Fix	4.3	-	60.1	-	-	-	53.6	-	-	-
Fuyu-8B [6]	-	1024×1024	Any	21.3	74.2	-	-	74.1	-	-	728.6	10.7	-
OtterHD-8B [20]	-	1024×1024	Any	21.3	-	-	-	86.0	-	-	1223.4	58.3	-
mPLUG-Owl2 [43]	401M	448×448	Fix	1.7	79.4	56.1	58.2	86.2	68.7	54.5	1450.2	64.5	-
UReader [42]	86M	896×1120	Enum	26.0	-	-	57.6	-	-	-	-	-	-
Monkey [23]	1.0B	896×1344	Enum	65.3	80.3	60.7	-	67.6	69.4	61.2	-	-	-
LLaVA-1.5 [27]	1.2M	336×336	Fix	15.5	80.0	<u>63.3</u>	61.3	85.9	<u>71.6</u>	53.6	<u>1531.3</u>	<u>67.7</u>	<u>63.6</u>
LLaVA-UHD (ours)	1.2M	672×1008	Any	14.6	81.7	65.2	67.7	89.1	72.0	<u>56.1</u>	1535.0	68.0	64.8
Δ	-	×6 times	-	-0.9	+1.7	+1.9	+6.4	+3.2	+0.4	+2.5	+3.7	+0.3	+1.2



Multimodal Encoding

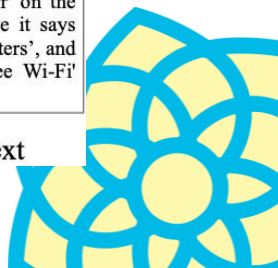
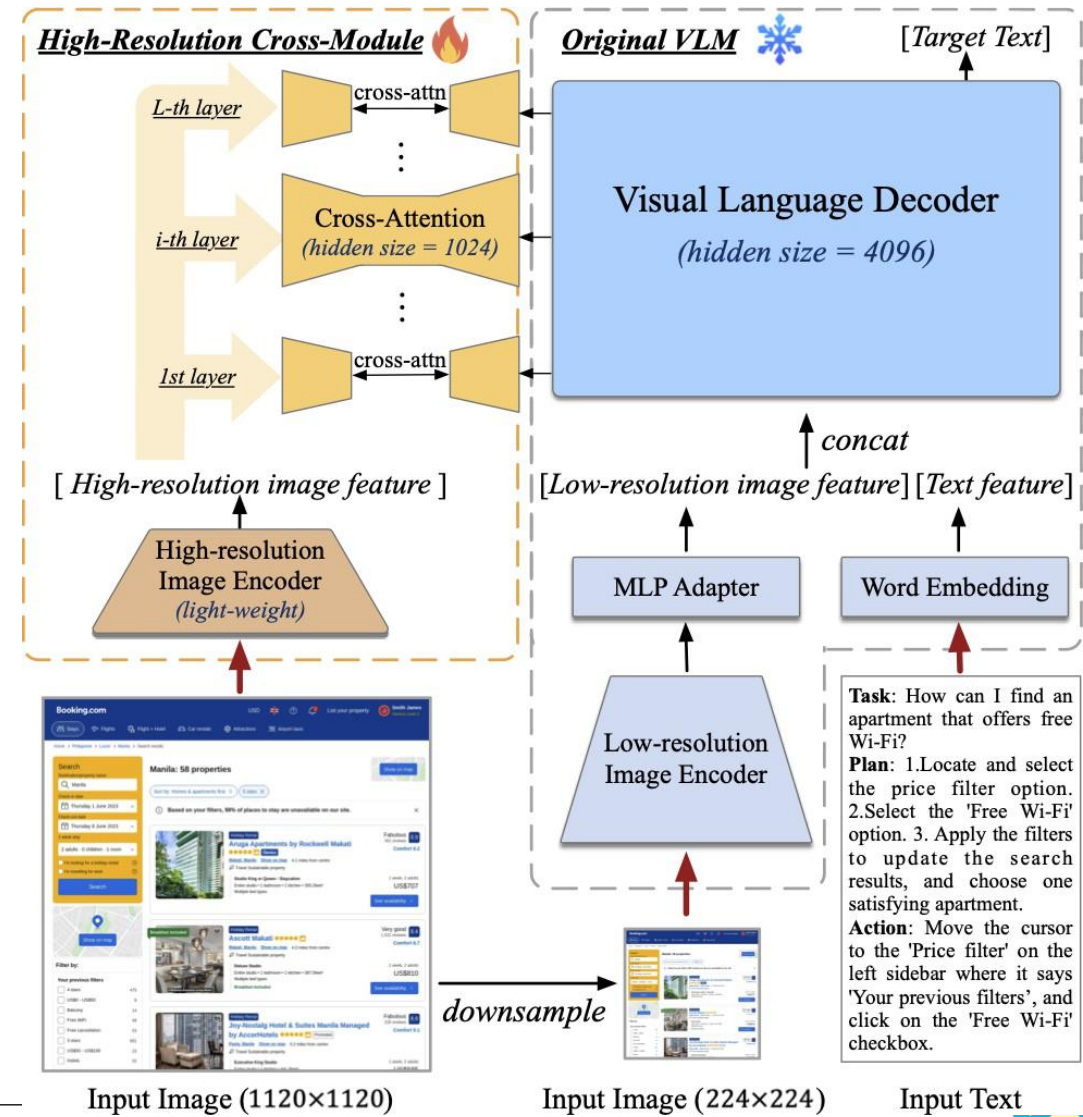
Visual Encoder

High-resolution Multimodal LLMs

✗ Dual branch encoders

✗ Representatives

- ◆ CogAgent
- ◆ Mini-Gemini
- ◆ DeepSeek-VL
- ◆ LLaVA-HR

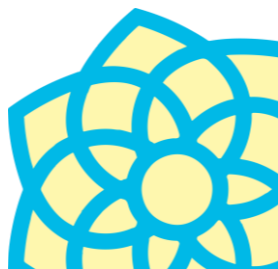
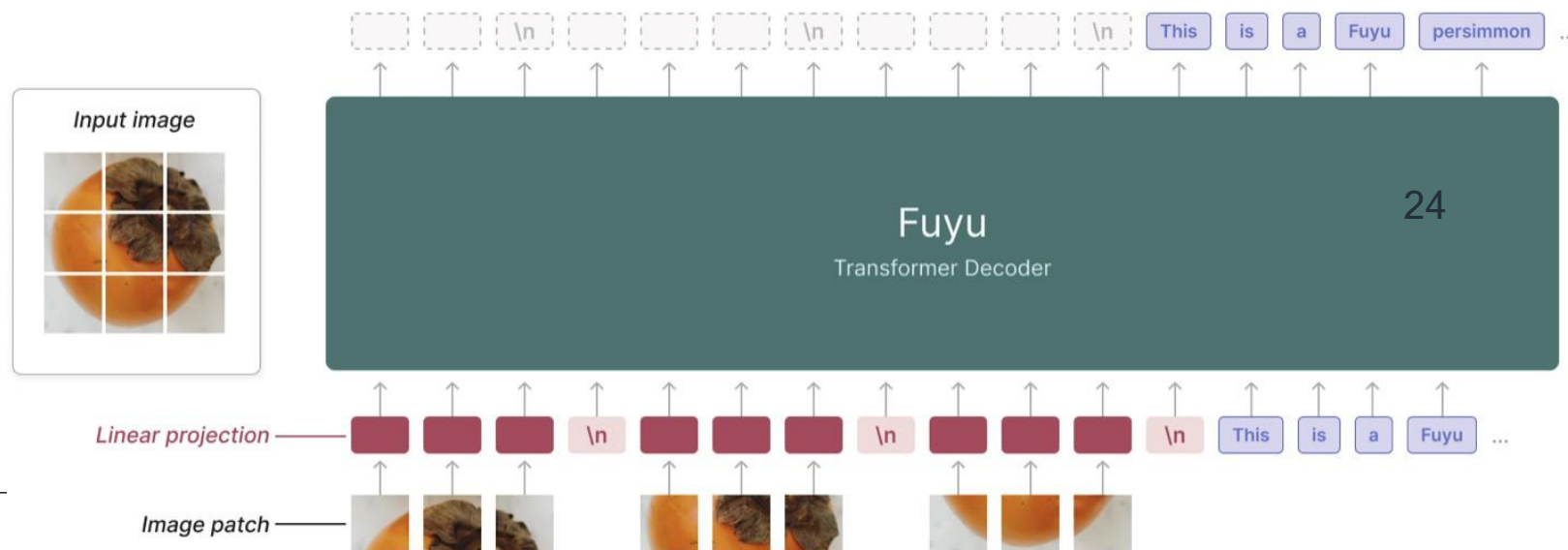


Multimodal Encoding

- Visual Encoder

- High-resolution Multimodal LLMs

- × ViT-free: linear project pixel-patches into tokens
 - × Representatives: Fuyu, OtterHD
 - × A potential unified way for MLLMs, getting rid of ViTs
 - × More costly to train, produce lengthy visual tokens



Multimodal Encoding

• Non-Visual Encoder

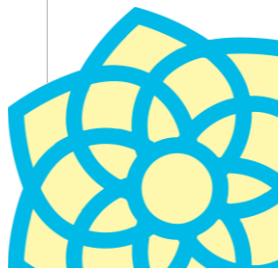
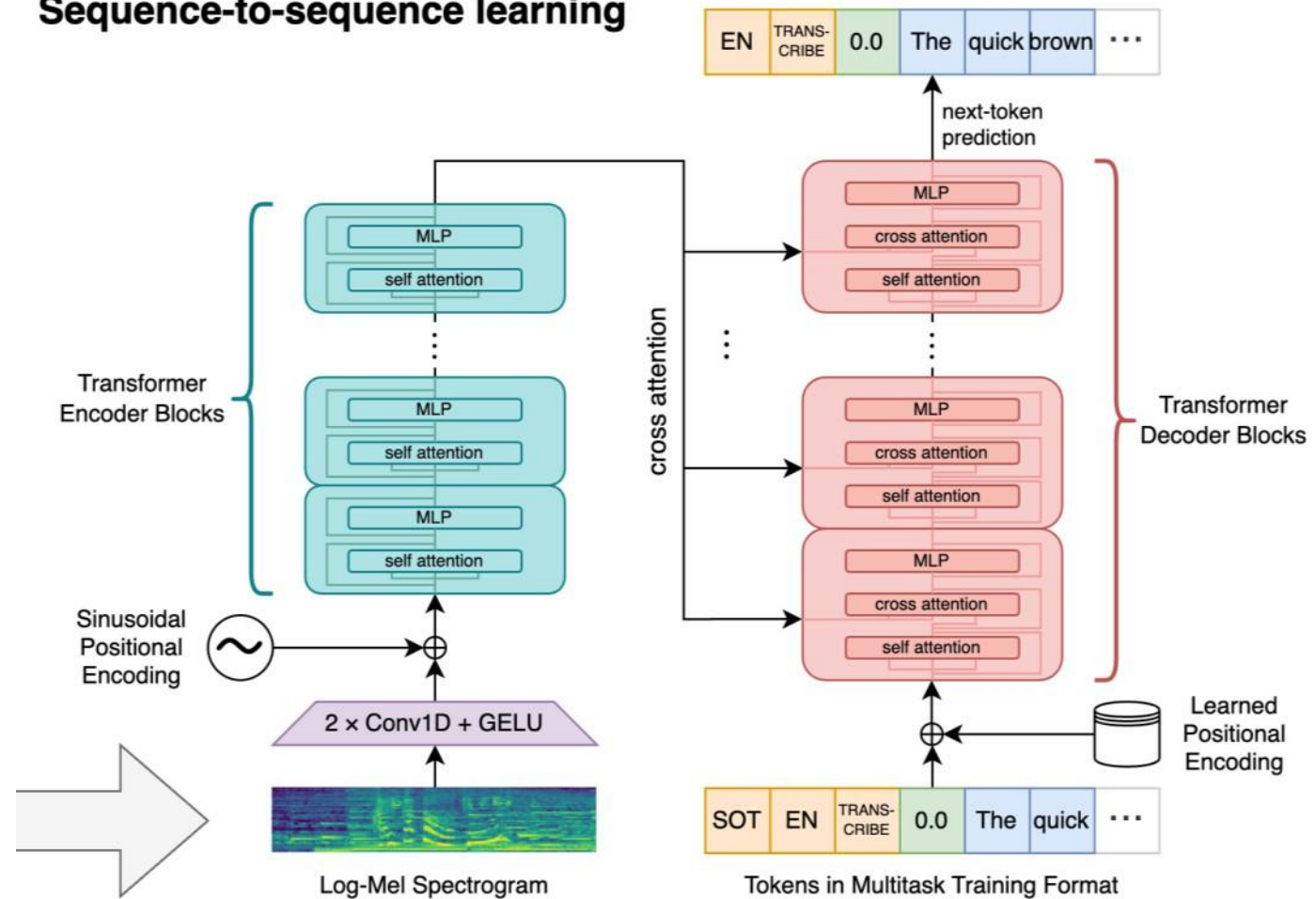
o Audio:

- ✗ Whisper
- ✗ AudioCLIP
- ✗ HuBERT
- ✗ BEATs

o 3D Point:

- ✗ Point-BERT

Sequence-to-sequence learning

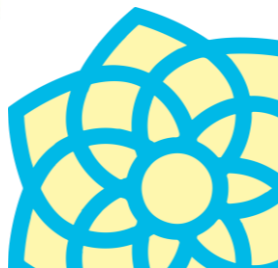
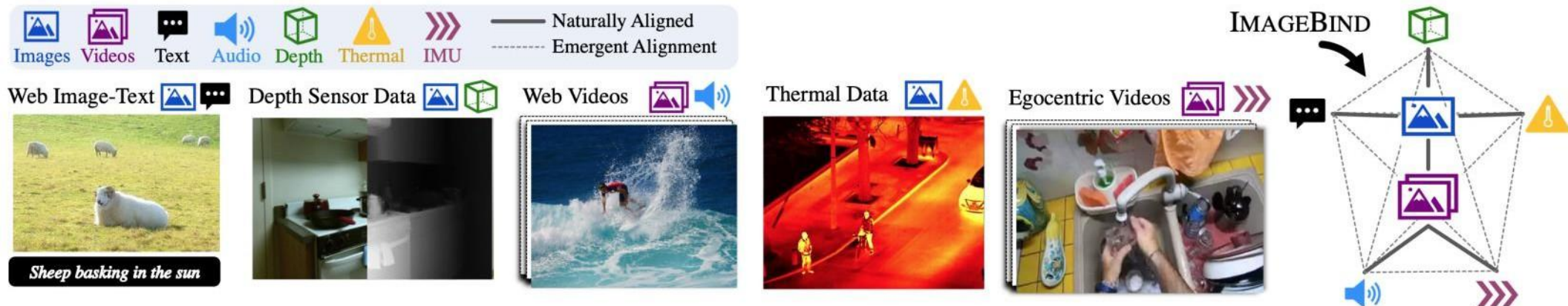


Multimodal Encoding

- Unified Multimodal Encoder

- ImageBind:

- × Embedding all modalities into a joint representation space of **Image**.
 - × Well aligned modality representations can benefit LLM understanding

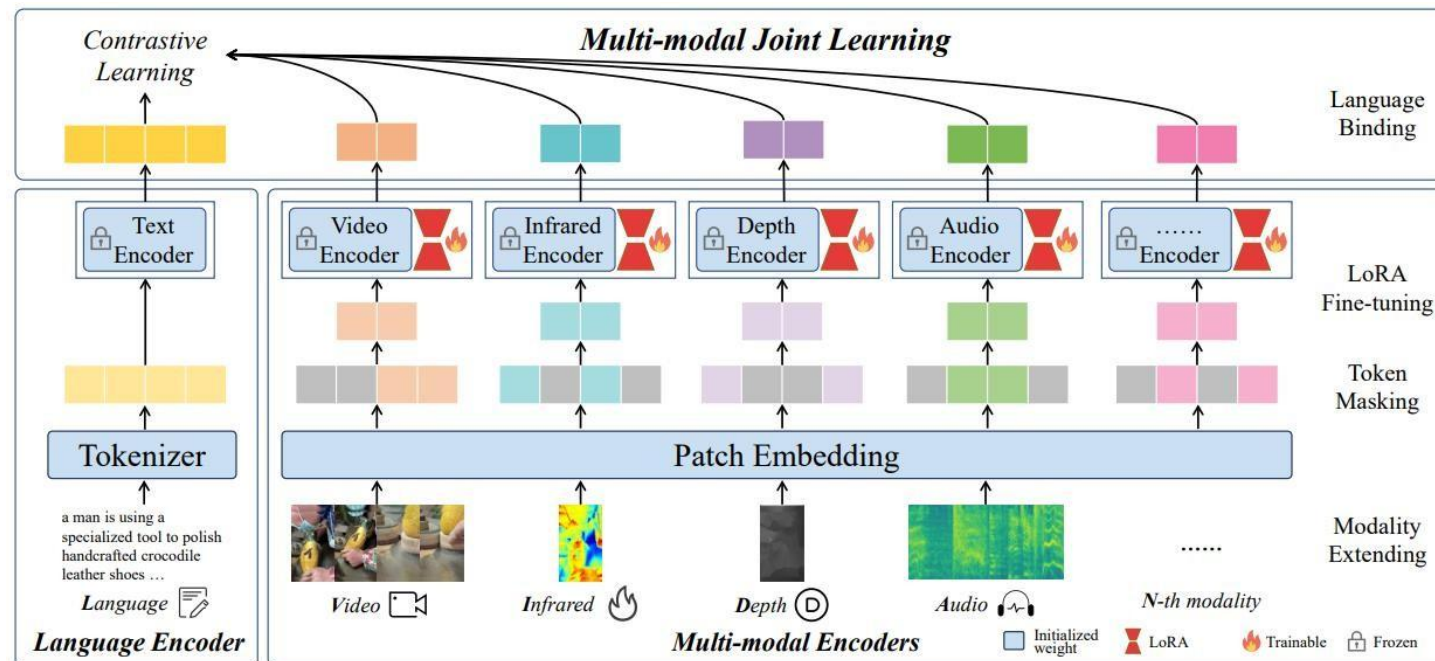


Multimodal Encoding

- Unified Multimodal Encoder

- o LanguageBind:

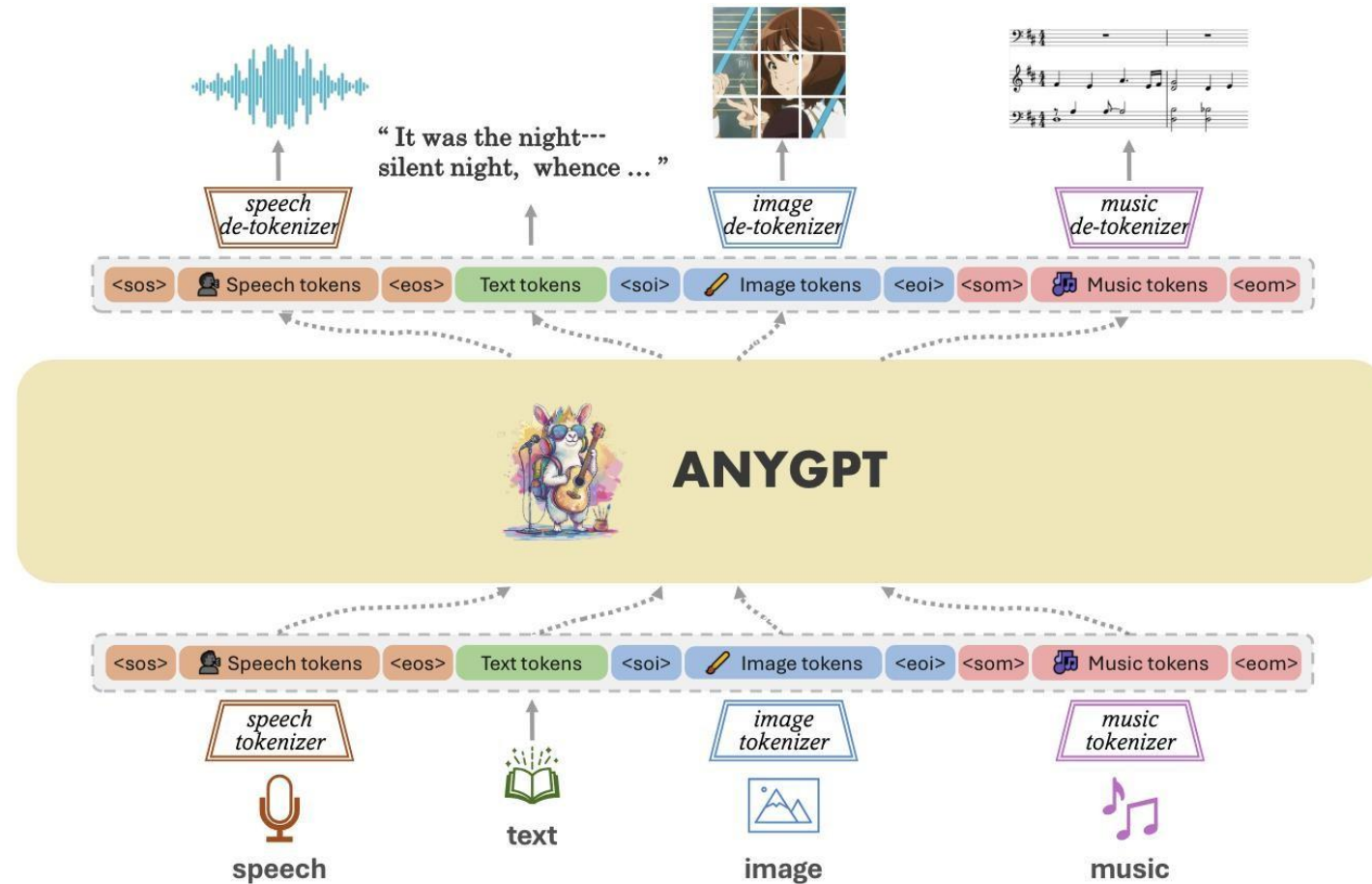
- ✗ Embedding all modalities into a joint representation space of **Language**.
- ✗ Well aligned modality representations can benefit LLM understanding



Multimodal Signal Tokenization

- Tokenization

- AnyGPT



Multimodal Signal Tokenization

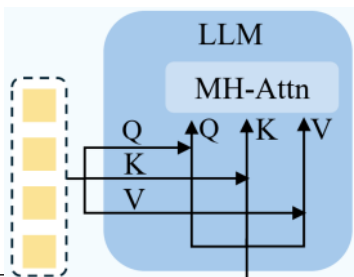
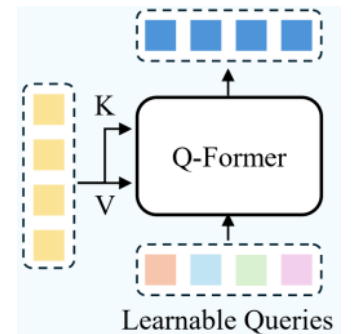
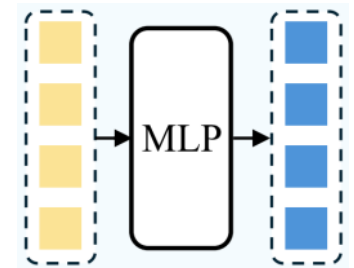
- Tokenization in Codebook
 - Represent multimodal signals as discrete tokens in a codebook
 - × Advantages: support **unified** multimodal signal **understanding** and **generation** in an auto-regressive next-token prediction framework
 - × More commonly used in image synthesizer
 - ◆ Parti
 - ◆ Muse (parallel)
 - ◆ MaskGIT (parallel)
 - × Representative Multimodal LLMs
 - ◆ Gemini
 - ◆ CM3
 - ◆ VideoPoet



Input-side Projection

• Methods to Connect Multimodal Representation with LLM

- Projecting multimodal (e.g., image) representations into LLM semantic space
 - × Q-Former: BLIP-2, InstructBLIP, VisCPM, VisualGLM
 - × Linear projection: LLaVA, MiniGPT-4, NExT-GPT
 - × Two-layer MLP: LLaVA-1.5/NeXT, CogVLM, DeepSeek-VL, Yi-VL
 - × Perceiver Resampler: Flamingo, Qwen-VL, MiniCPM-V, LLaVA-UHD
 - × C-Abstractor: HoneyBee, MM1



Input-side Projection

Some Insights

- Different papers have different conclusions about projection methods
 - ✗ Two-layer MLP is better than linear projection. (LLaVA 1.5)
 - ✗ Resampler is comparable to C-Abstractor (MM1) and MLP (LLaVA-UHD)

Method	LLM	Res.	GQA	MME	MM-Vet
InstructBLIP	14B	224	49.5	1212.8	25.6
<i>Only using a subset of InstructBLIP training data</i>					
0 LLaVA	7B	224	–	502.8	23.8
1 +VQA-v2	7B	224	47.0	1197.0	27.7
2 +Format prompt	7B	224	46.8	1323.8	26.3
3 +MLP VL connector	7B	224	47.3	1355.2	27.8
4 +OKVQA/OCR	7B	224	50.0	1377.6	29.6

Model	#TFLOPs	VQA ^{v2}	GQA	VQA ^T
LLaVA-1.5	15.50	74.6 (-5.4)	57.9 (-5.4)	58.4 (-3.9)
w/ adaptive enc.	15.50	74.9 (-5.2)	62.5 (-1.6)	60.7 (-1.1)
LLaVA-UHD	14.63	81.4 (-0.3)	61.8 (-3.4)	64.5 (-3.2)
w/ MLP	113.65	81.3 (-0.3)	62.0 (-3.4)	63.9 (-3.0)
w/ MLP & FP. [24]	80.10	79.6 (-1.6)	61.9 (-2.4)	58.5 (-7.6)



Input-side Projection

Some Insights

- Agreement: Number of visual token matters! Especially for efficiency
 - ✗ Resampler/Q-Former/C-Abstractor yield less visual tokens than MLP/Linear
 - ✗ Favorable in high-resolution image understanding

Model	#Data	MaxRes.	AR.	TFLOPs	VQA ^{v2}	GQA	VQA ^T	POPE	SQA	VizWiz	MME	MMB	MMB ^{CN}
BLIP-2 [21]	129M	224×224	Fix	1.0	41.0	41.0	42.5	85.3	61.0	19.6	1293.8	-	-
InstructBLIP [11]	130M	224×224	Fix	1.0	-	49.5	50.7	78.9	63.1	33.4	1212.8	-	-
Shikra [8]	6M	224×224	Fix	8.0	77.4	-	-	-	-	-	-	58.8	-
Qwen-VL [5]	1.4B	448×448	Fix	9.2	78.8	59.3	63.8	-	67.1	35.2	-	38.2	7.4
SPHINX [24]	1.0B	448×448	Fix	39.7	78.1	62.6	51.6	80.7	69.3	39.9	1476.1	66.9	56.2
SPHINX-2k [24]	1.0B	762×762	Fix	69.4	80.7	63.1	61.2	87.2	70.6	44.9	1470.7	65.9	57.9
MiniGPT-v2 [7]	326M	448×448	Fix	4.3	-	60.1	-	-	-	53.6	-	-	-
Fuyu-8B [6]	-	1024×1024	Any	21.3	74.2	-	-	74.1	-	-	728.6	10.7	-
OtterHD-8B [20]	-	1024×1024	Any	21.3	-	-	-	86.0	-	-	1223.4	58.3	-
mPLUG-Owl2 [43]	401M	448×448	Fix	1.7	79.4	56.1	58.2	86.2	68.7	54.5	1450.2	64.5	-
UReader [42]	86M	896×1120	Enum	26.0	-	-	57.6	-	-	-	-	-	-
Monkey [23]	1.0B	896×1344	Enum	65.3	80.3	60.7	-	67.6	69.4	61.2	-	-	-
LLaVA-1.5 [27]	1.2M	336×336	Fix	15.5	80.0	63.3	61.3	85.9	71.6	53.6	1531.3	67.7	63.6
LLaVA-UHD (ours)	1.2M	672×1008	Any	14.6	81.7	65.2	67.7	89.1	72.0	<u>56.1</u>	1535.0	68.0	64.8
Δ	-	×6 times	-	-0.9	+1.7	+1.9	+6.4	+3.2	+0.4	+2.5	+3.7	+0.3	+1.2

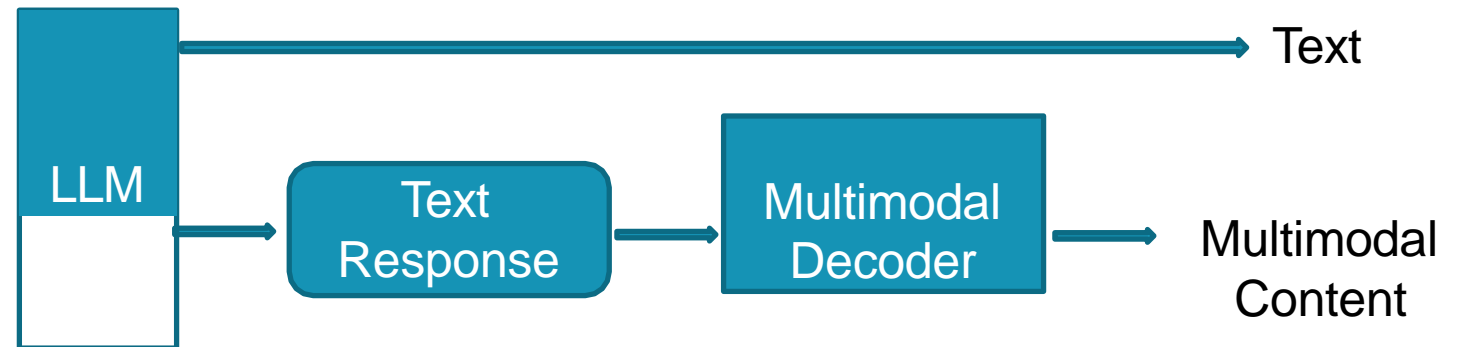


Decoding-side Connection

- Message passing via 1) text tokens

- Representative MLLMs:

- Visual-ChatGPT
- HuggingGPT
- GPT4Video
- MM-REACT
- ViperGPT
- ModaVerse
- Vitron
- ...



- Pros:

- High performance lower-bound
- More efficient, i.e., without tuning

- Cons:

- Loss of end-to-end tuning capabilities.
- Performance upper-bound is limited, i.e., some multimodal signals cannot be optimally conveyed through text.

1 *Visual-ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. 2023*

2 *HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. 2023*

3 *ModaVerse: Efficiently Transforming Modalities with LLMs. 2024*

4 *VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing. 2024*



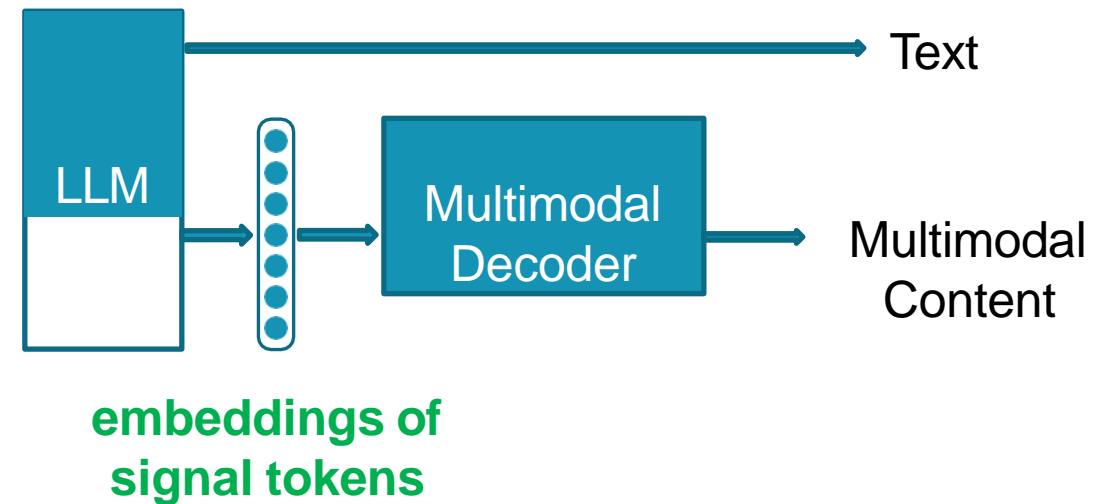
Decoding-side Connection

- Message passing via 2) continuous embedding

Passing the message from LLM to downstream decoders via soft embeddings, i.e., **signal tokens**.

- Merits

- Capable of end-to-end tuning, resulting in more efficient instruction transmission
- More able to convey various multimodal signals that text alone cannot express, e.g.,
 - *the numeration of vision*
 - *the visual-spatial relational semantics*



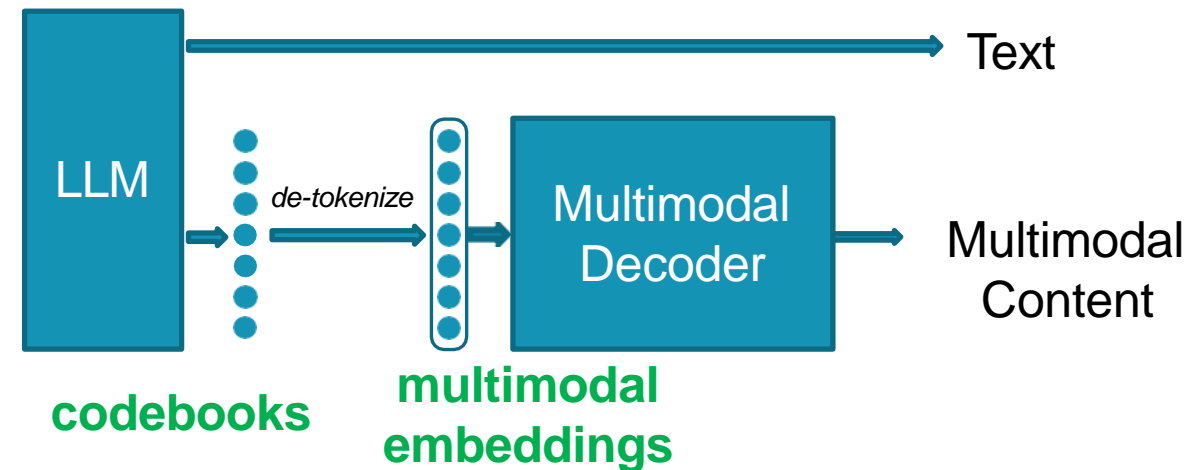
Decoding-side Connection

- Message passing via 3) codebooks

*LLM generates special tokens id, i.e., **codebooks**, to downstream (visual) decoders.*

- Merits

- Capable of end-to-end tuning for higher efficiency in command transmission
- Better at expressing various multimodal signals that cannot be captured by text alone
- Supports autoregressive multimodal token generation

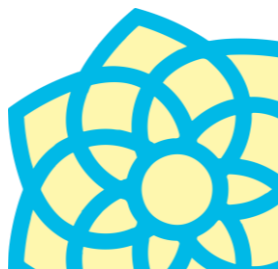


1 *Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action. 2023*

2 *LVM: Sequential Modeling Enables Scalable Learning for Large Vision Models. 2023*

3 *AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. 2024*

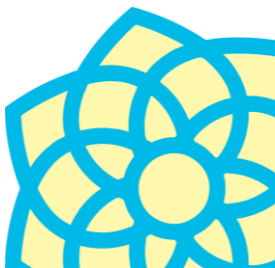
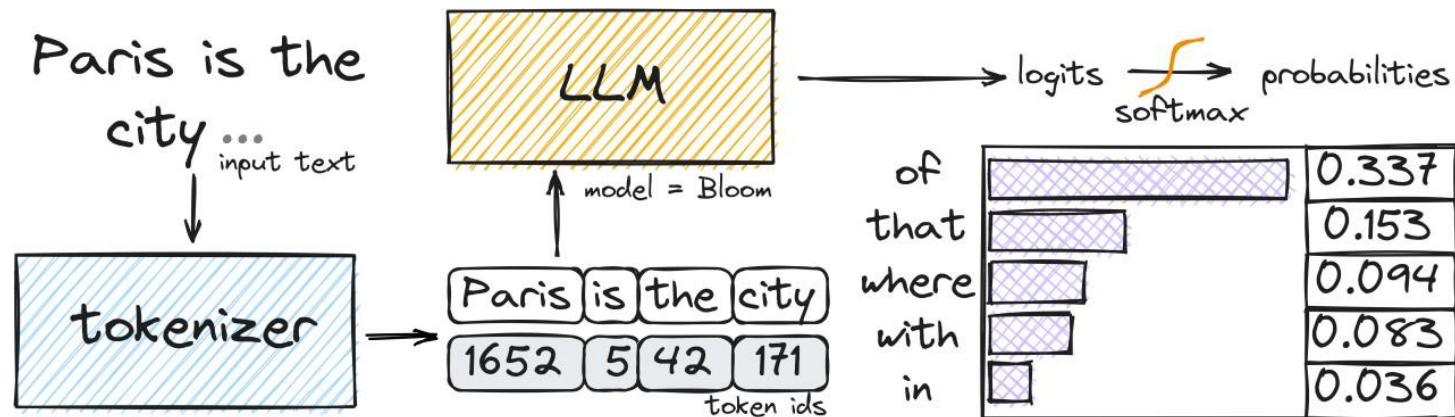
4 *VideoPoet: A Large Language Model for Zero-Shot Video Generation. 2024*



Multimodal Generation

- Text Generation
 - LLMs naturally support direct text generation

via e.g., BPE decoding, Beam search, ...



Multimodal Generation

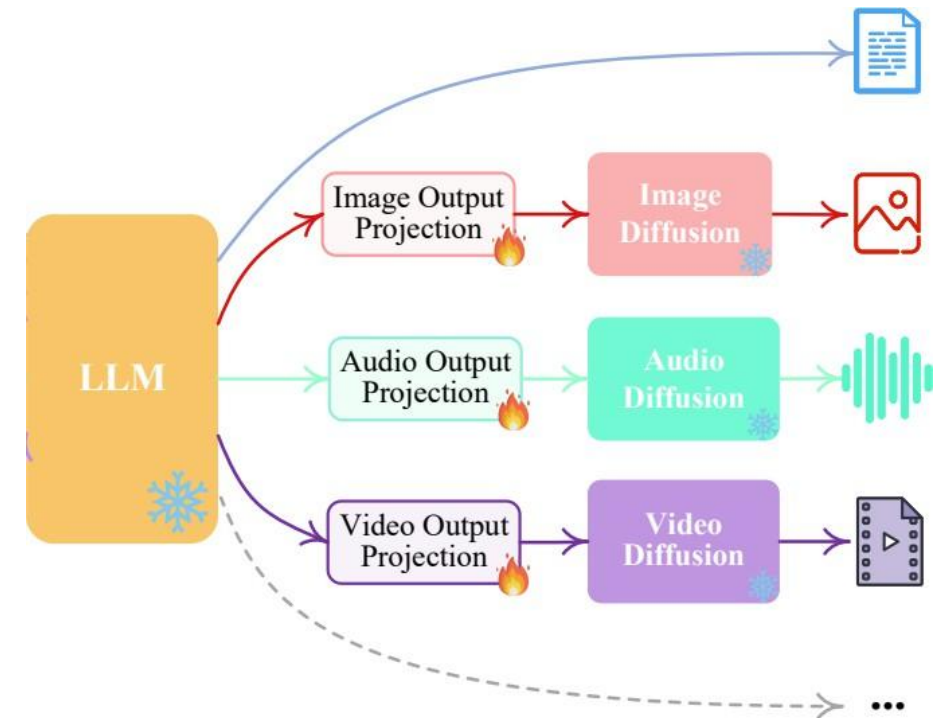
- Generation via Diffusion Models

- Visual (Image/Video) Generator

- Image Diffusion
- Video Diffusion

- Audio Generator

- Speech Diffusion
- Audio Diffusion



Multimodal Generation

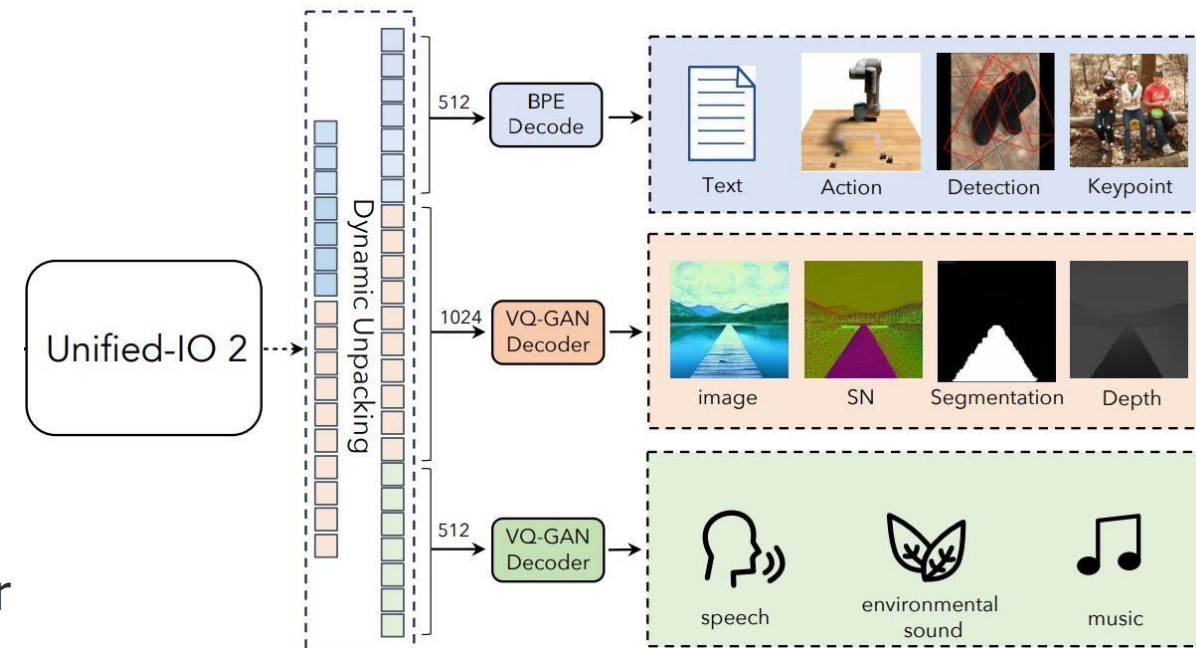
• Generation via Codebooks

o Visual (Image/Video) Generator

- o VQ-VAE + Codebooks
- o VQ-GAN + Codebooks

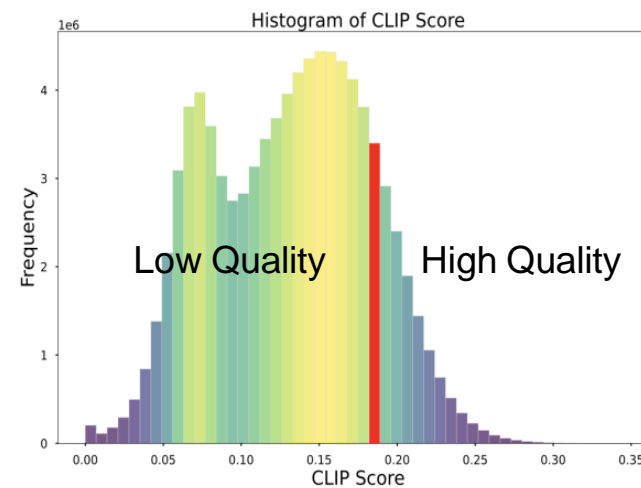
o Audio Generator

- o SpeechTokenizer + Residual Vector Quantizer
- o SoundStream + Residual Vector Quantizer



Multilingual Multimodal LLMs

- MLLMs are mostly strong in English, weak in other languages
 - Limited **scale** and **quality** of multimodal data in non-English languages
 - **Huge computation cost** for each language even if sufficient data available
 - Why not machine translation pipeline ?
 - ✗ Another LLM for translation: double computation cost and delay
 - ✗ Missing visual context can lead to incorrect translation
 - ✗ Not an elegant way to AGI

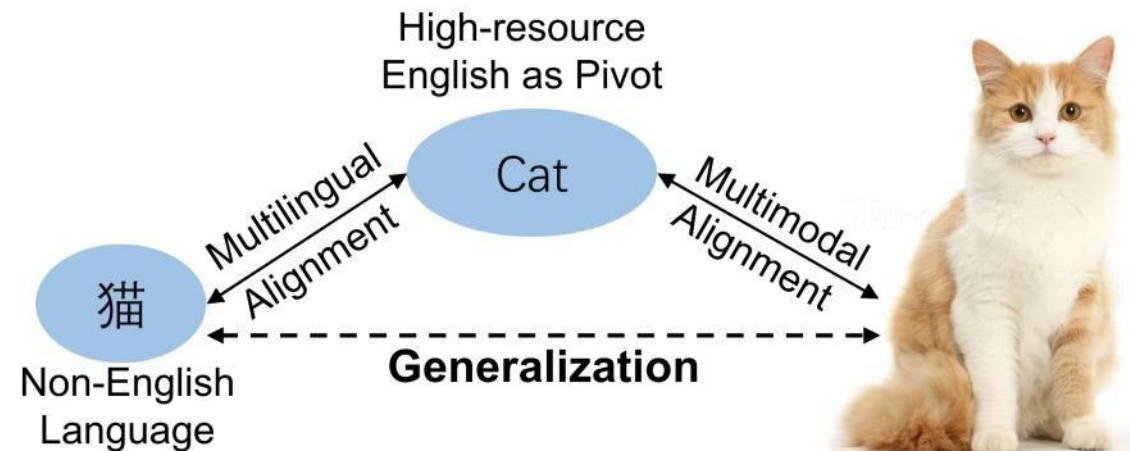


Chinese Image-Text Data Quality Distribution



Multilingual Multimodal LLMs

- MLLMs are mostly strong in English, weak in other languages
 - **Generalizing** multimodal capabilities to non-English languages via **multilingual LLM as pivots**
 - Basic idea
 - × Multilingual alignment: Reuse multilingual LLMs
 - × Multimodal alignment: Multimodal pretraining on English image-text pairs



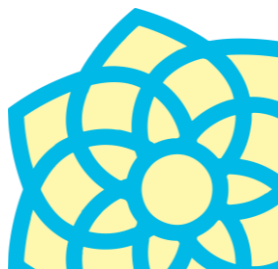
Multilingual Multimodal LLMs

- MLLMs are mostly strong in English, weak in other languages
 - **Generalizing** multimodal capabilities to non-English languages via **multilingual LLM as pivots**
 - Foundation capability can be well generalized already!

Was ist an diesem Bild ungewöhnlich?




An unusual aspect of the image is the man sitting on a folding chair in the back of a parked yellow taxi cab. This is not a typical scene, as one would expect passengers to be seated inside the cab or at the very least, be standing near the taxi. The man's presence in the back of the taxi, using a folding chair, adds an unconventional and unexpected element to the scene, making it stand out from regular taxi activities. Another notable aspect of the image is the presence of multiple people and vehicles, including other cars and pedestrians in the background. This urban setting contributes to the overall busy atmosphere, making the man on the chair in the taxi stand out even more.



Multilingual Multimodal LLMs

- MLLMs are mostly strong in English, weak in other languages
 - **Generalizing** multimodal capabilities to non-English languages via **multilingual LLM as pivots**
 - Lightweight SFT is sufficient to switch to the target language
 - Generalize to cultures. etc.



你能用苏轼的诗句描述这幅图像吗?
(Can you describe this image with Su Shi's poems?)

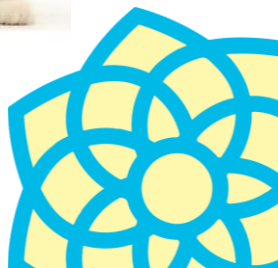
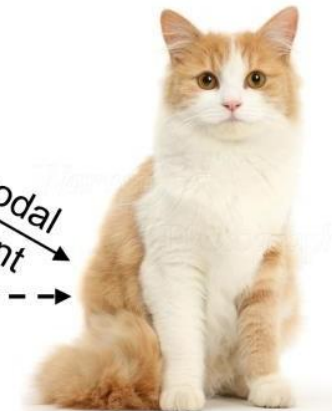
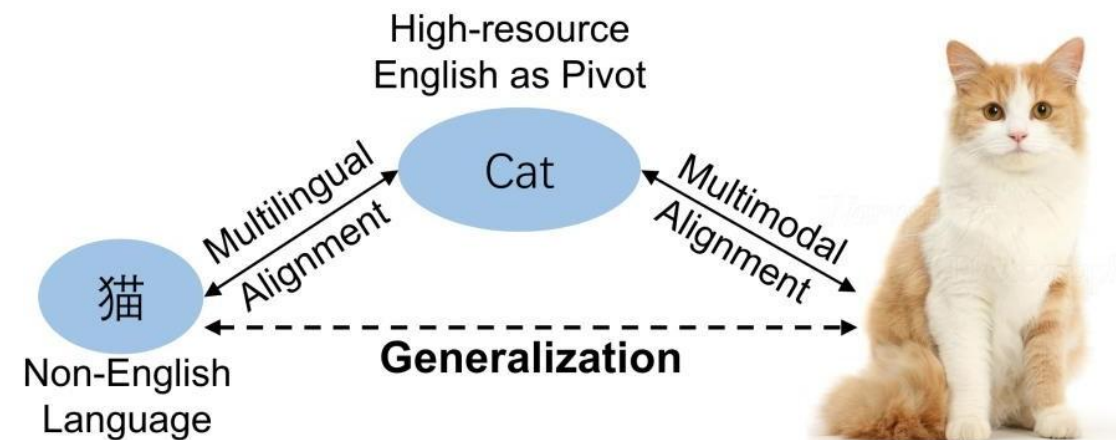
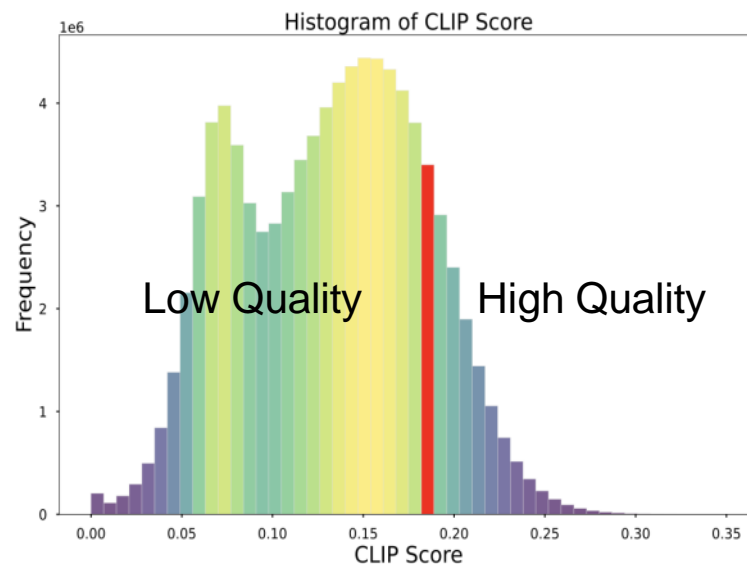
在这张照片中，月亮高高地挂在天空上，照亮了古老的中国建筑和它周围的环境。这创造出一种美丽而宁静的气氛，让人想起苏轼著名的诗句：“明月几时有？把酒问青天。不知天上宫阙，今夕是何年？”
In this photo, the moon hangs high in the sky, illuminating the ancient Chinese building and its surroundings. This creates a beautiful and serene atmosphere, reminiscent of Su Shi's famous poem: "When is the bright moon? Ask the blue sky for wine. I don't know the palace in the sky. What year is it?"

DoDo



Multilingual Multimodal LLMs

- MLLMs are mostly strong in English, weak in other languages
 - **Generalizing** multimodal capabilities to non-English languages via **multilingual LLM as pivots**
 - Common practice nowadays
 - ✗ VisCPM, LLaVA-NeXT, Yi-VL 34B, CogVLM2: English & Chinese
 - ✗ MiniCPM-Llama3-V 2.5: 30+ Languages:



* Part-III

Modality and Functionality



Hao Fei

Research Fellow

National University of Singapore

<http://haofei.vip/>

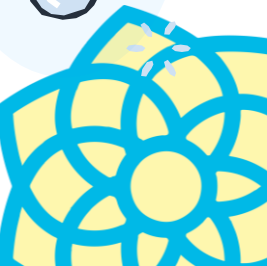
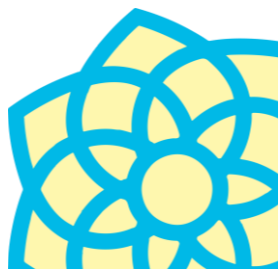


Table of Content

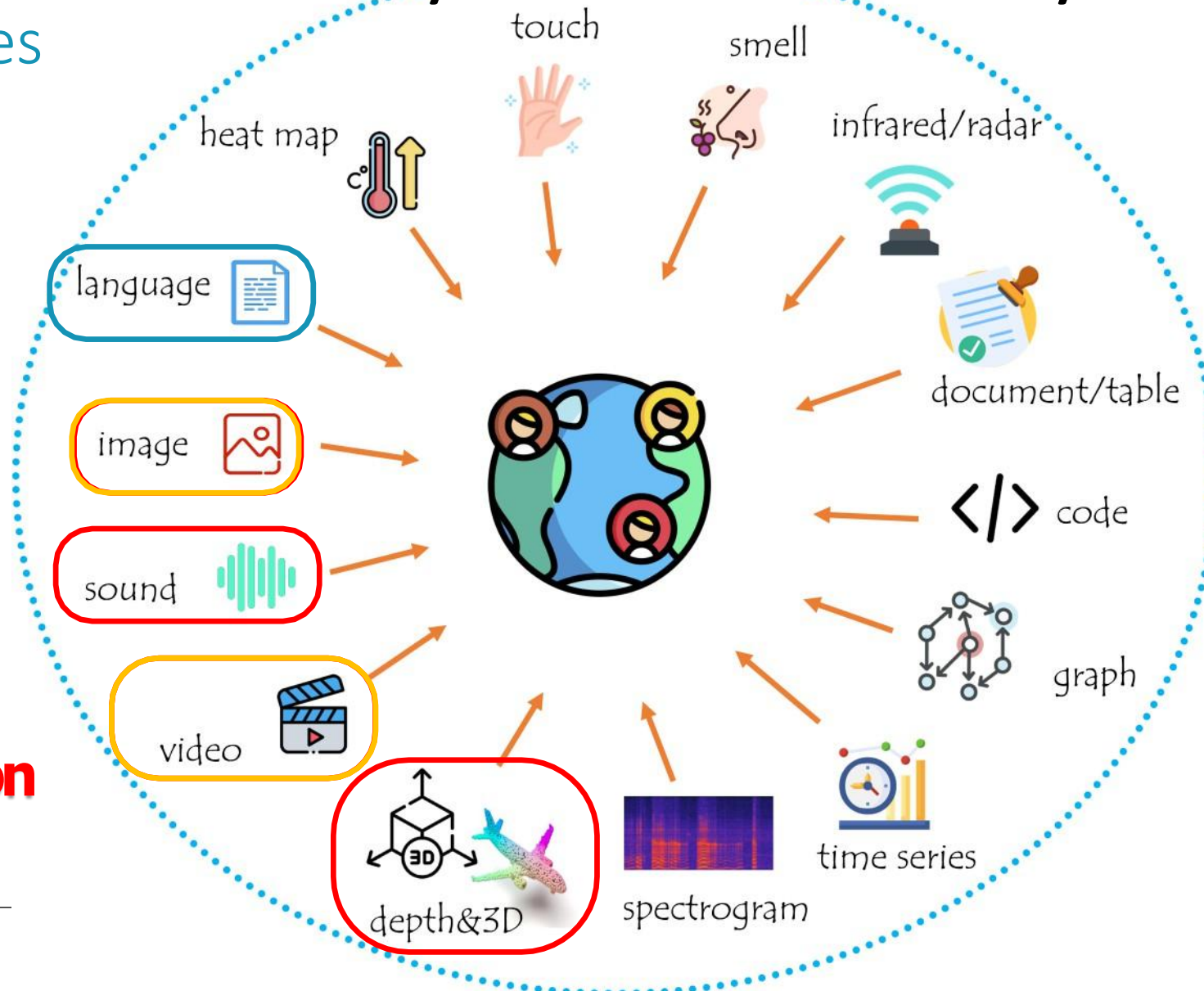
o **Modality & Functionality**

- × Overview
- × Multimodal Perceiving
- × Multimodal Generation
- × Unified MLLM
- × Fine-grained MLLM
- × What's Next



Overview of Modality and Functionality

- Modalities



Language + Vision



Overview of Modality and Functionality

	Modality (w/ Language)			
	Image	Video	Audio	3D
Input-side Perceiving	Flamingo, Kosmos-1, Blip2, mPLUG-Owl, Mini-GPT4, LLaVA, InstructBLIP, VPGTrans, CogVLM, Monkey, Chameleon, Otter, Qwen-VL, GPT-4v, SPHINX, Yi-VL, Fuyu, ...	VideoChat, VideoChatGPT, Video-LLaMA, PandaGPT, MovieChat, Video-LLaVA, LLaMA-VID, Momentor, ...	AudioGPT, SpeechGPT, VIOLA, AudioPaLM, SALMONN, MULLaMA, ...	3D-LLM, 3D-GPT, LL3DA, SpatialVLM, PointLLM, Point-Bind, ...
	[Pixel-wise] GPT4RoI, LION, MiniGPT-v2, NExT-Chat, Kosmos-2, GLaMM, LISA, DetGPT, Osprey, PixelLM, ...	[Pixel-wise] PG-Video-LLaVA, Merlin, MotionEpic, ...	-	-
	Video-LLaVA, Chat-UniVi, LLaMA-VID		-	-
	Panda-GPT, Video-LLaMA, AnyMAL, Macaw-LLM, Gemini, VideoPoet, ImageBind-LLM, LLMBind, LLaMA-Adapter, ...			-
Perceiving + Generating	GILL, EMU, MiniGPT-5, DreamLLM, LLaVA-Plus, InternLM-XComposer2, SEED-LLaMA, LaVIT, Mini-Gemini, ...	GPT4Video, Video-LaVIT, VideoPoet, ...	AudioGPT, SpeechGPT, VIOLA, AudioPaLM, ...	-
	[Pixel-wise] Vitron		-	-



Multimodal Perceiving

• Image-perceiving MLLM

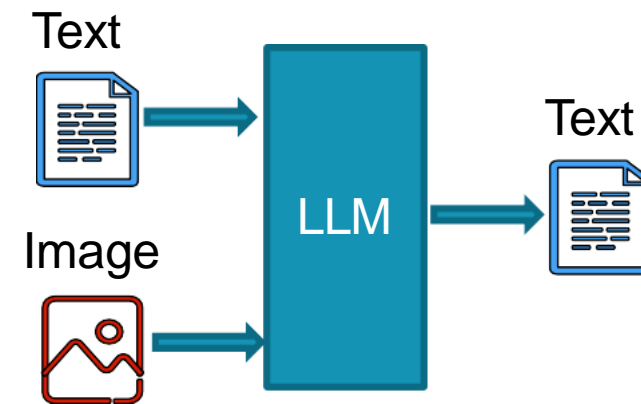
- Flamingo,
- Kosmos-1,
- Blip2, mPLUG-Owl,
- Mini-GPT4, LLaVA,
- InstructBLIP, Otter,
- VPGTrans
- Chameleon,
- Qwen-VL, GPT-4v,
- SPHINX,
- ...

1 *Flamingo: a Visual Language Model for Few-Shot Learning. 2022*

2 *Language Is Not All You Need: Aligning Perception with Language Models. 2023*

3 *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. 2023*

4 *MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. 2024*



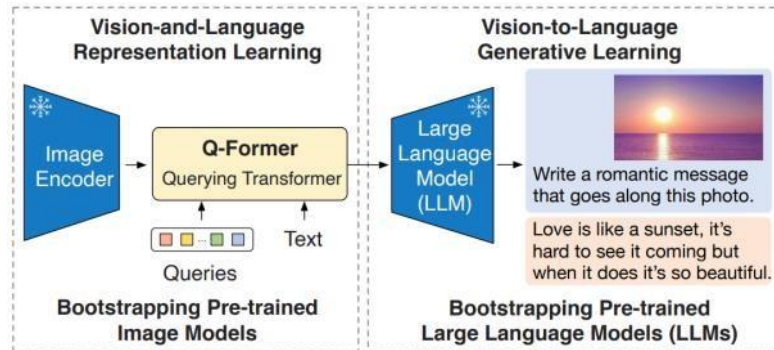
Encode input images with external image encoders, generating LLM-understandable visual feature, which is then fed into the LLM. LLM then interprets the input images based on the input text instructions and produces a textual response.



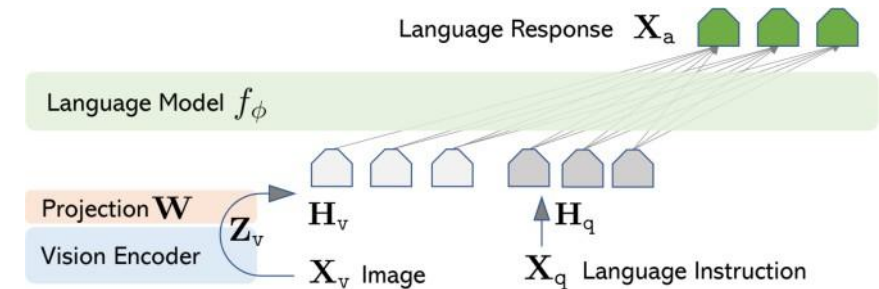
Multimodal Perceiving

Image-perceiving MLLM

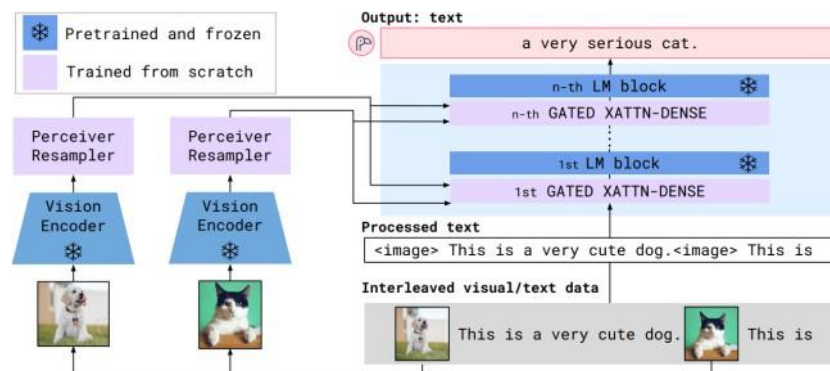
Blip2



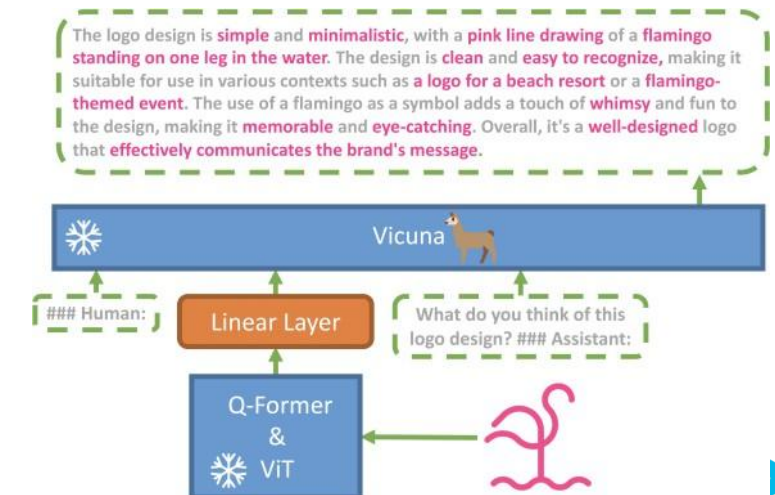
LLaVA



Flamingo



Mini-GPT4



1 Flamingo: a Visual Language Model for Few-Shot Learning. 2022

2 BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. 2023

3 Visual Instruction Tuning. 2023

4 A Survey on Multimodal Large Language Models.

<https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models>, 2023.



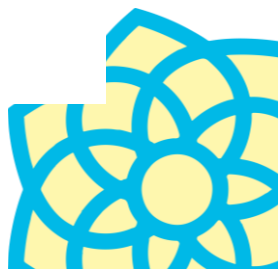
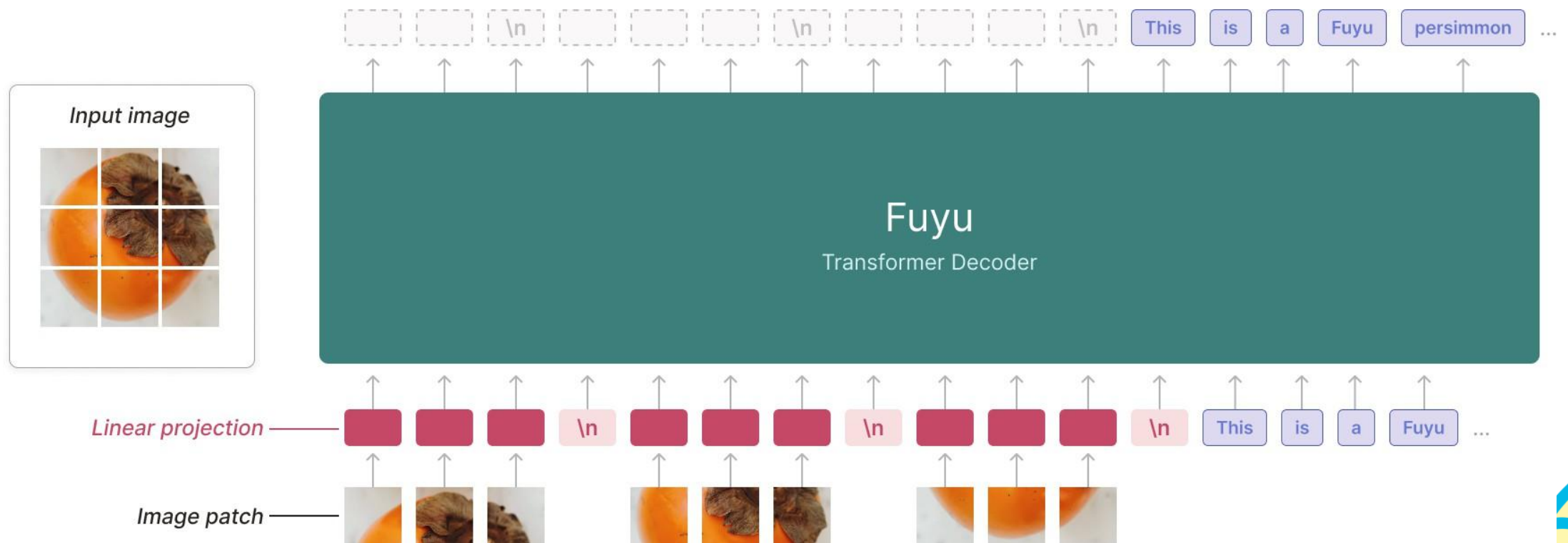
Multimodal Perceiving

- Image-perceiving MLLM

- Fuyu



Unlike all other existing image-oriented MLLMs, Fuyu processes image information without a frontend image encoder, and instead **directly inputs image patches into the LLM for interpretation.**

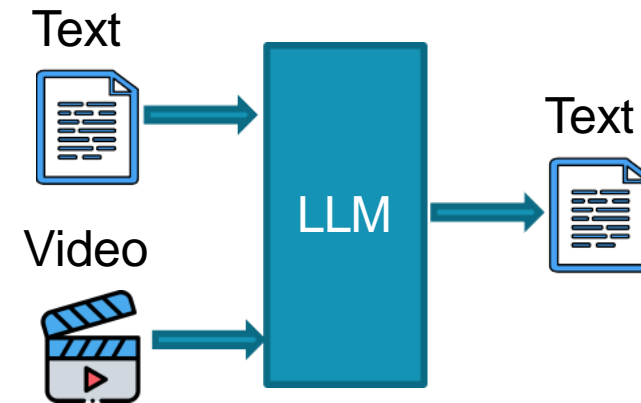


Multimodal Perceiving

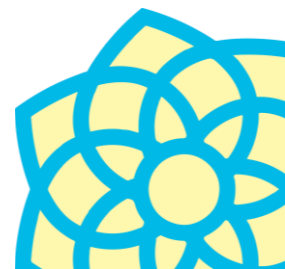
• Video-perceiving MLLM

- VideoChat,
- Video-ChatGPT,
- Video-LLaMA,
- PandaGPT,
- MovieChat,
- Video-LLaVA,
- LLaMA-VID,
- Momentor
- ...

- 1 *VideoChat: Chat-Centric Video Understanding. 2023*
- 2 *Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. 2023*
- 3 *Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. 2023*
- 4 *Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. 2023*
- 5 *Momentor: Advancing Video Large Language Model with Fine-Grained Temporal Reasoning. 2024*



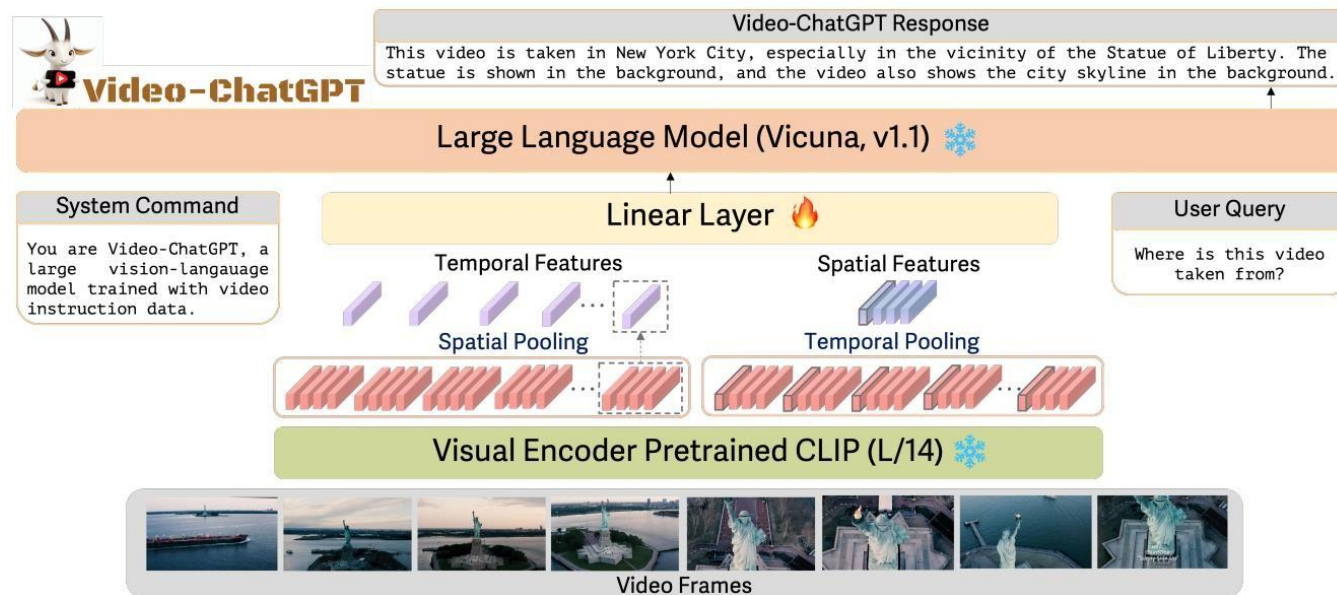
Encode input videos with external video encoders, generating LLM-understandable visual feature, feeding into LLM, which then interprets the input videos based on the input text instructions and produces a textual response.



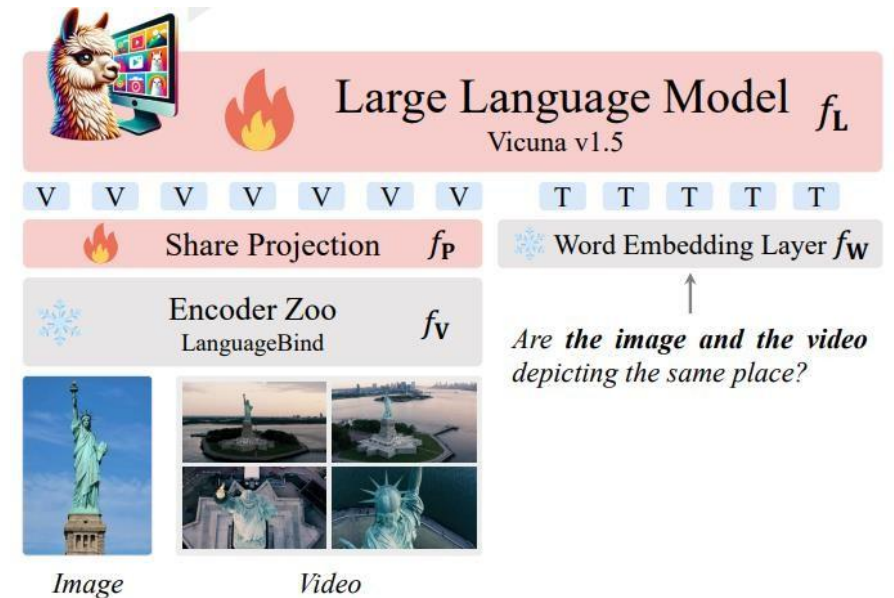
Multimodal Perceiving

• Video-perceiving MLLM

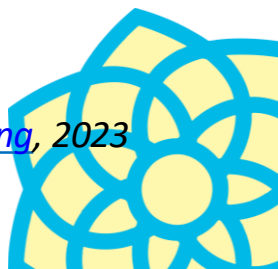
o Video-ChatGPT



o Video-LLaVA



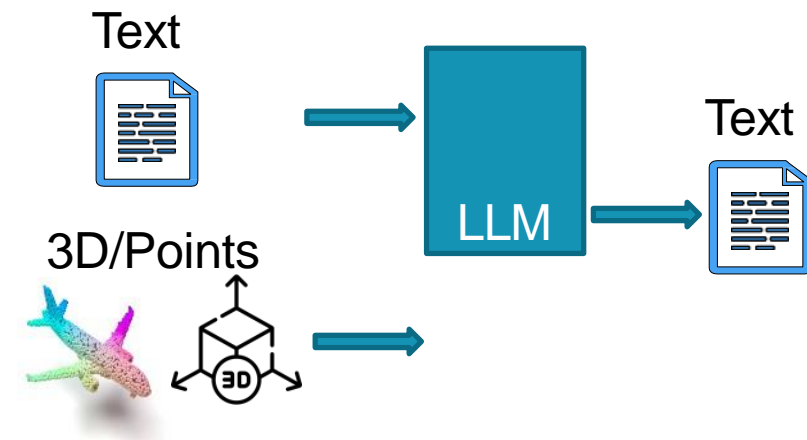
- 1 *Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. 2023*
- 2 *Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. 2023*
- 3 *Video Understanding with Large Language Models: A Survey. <https://github.com/yunlong10/Awesome-LLMs-for-Video-Understanding>, 2023*



Multimodal Perceiving

- 3D-perceiving MLLM

- 3D-LLM,
- 3D-GPT,
- LL3DA,
- SpatialVLM
- PointLLM
- Point-Bind
- ...



Encode input 3D information with external encoders, generating LLM-understandable 3D feature, feeding into LLM, which then interprets the input 3D/points based on the input text instructions and produces a textual response.

- 1 *3D-LLM: Injecting the 3D World into Large Language Models. 2023*
- 2 *3D-GPT: Procedural 3D Modeling with Large Language Models. 2023*
- 3 *LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning. 2023*
- 4 *PointLLM: Empowering Large Language Models to Understand Point Clouds. 2023*
- 5 *SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities. 2024*

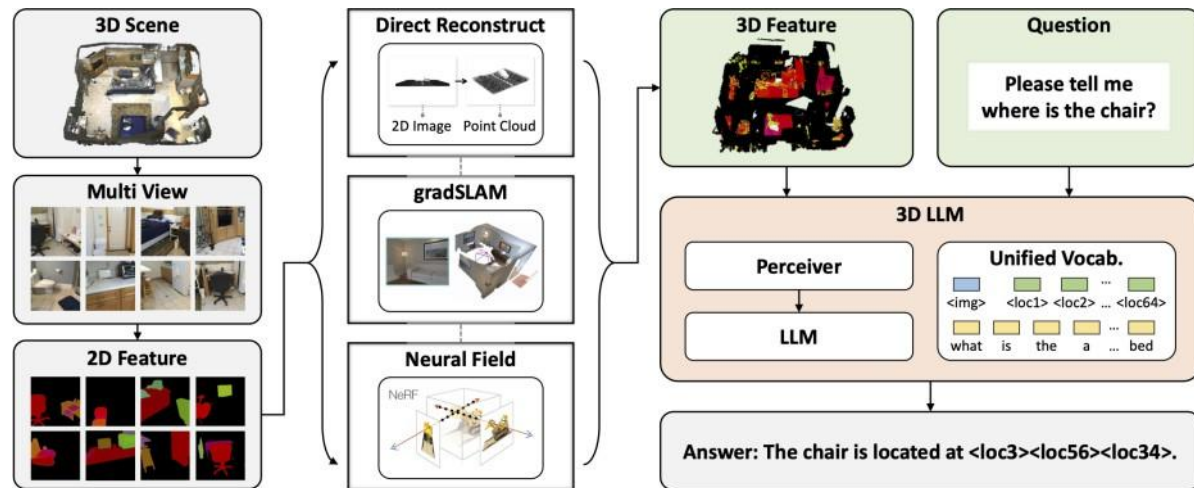
...



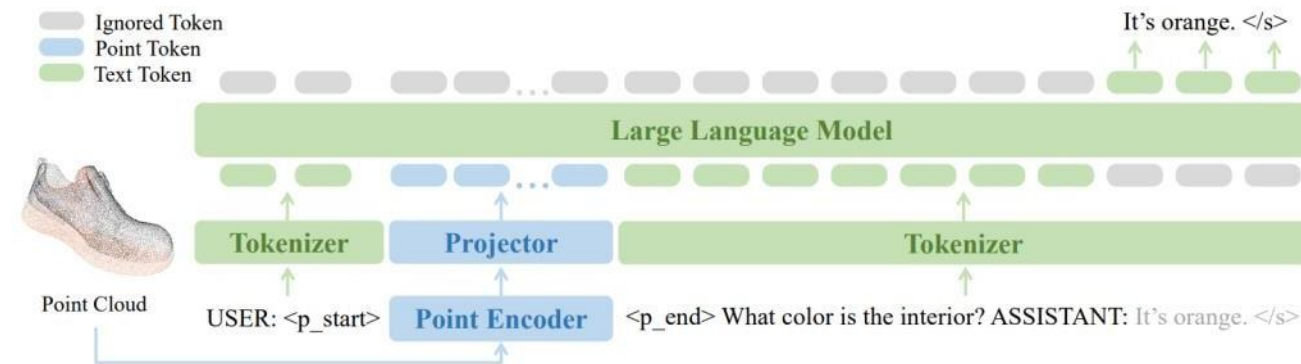
Multimodal Perceiving

- 3D-perceiving MLLM

- 0 3D-LLM



- 0 PointLLM



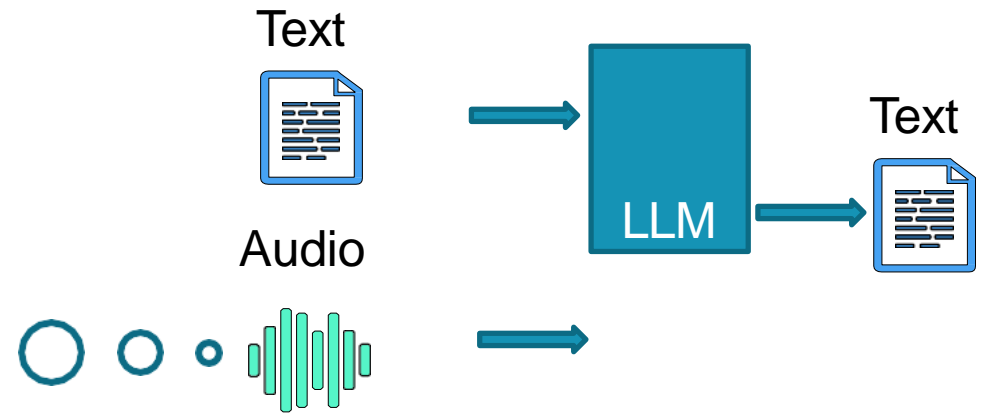
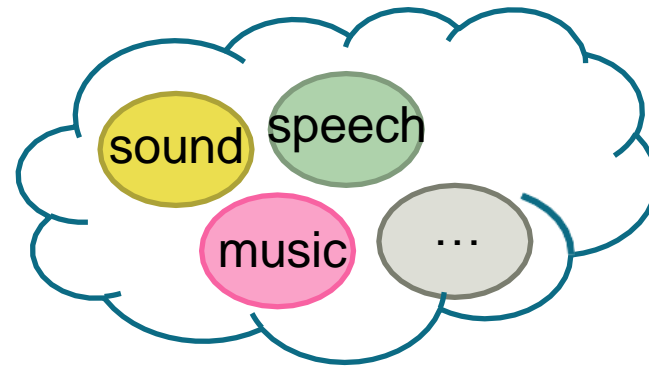
- 1 *3D-LLM: Injecting the 3D World into Large Language Models. 2023*
- 2 *PointLLM: Empowering Large Language Models to Understand Point Clouds. 2023*



Multimodal Perceiving

• Audio-perceiving MLLM

- 0 AudioGPT,
- 0 SpeechGPT,
- 0 VIOLA,
- 0 AudioPaLM
- 0 SALMONN
- 0 MU-LLaMA
- 0 ...



Encode input audio signals with external encoders, generating LLM-understandable signal features, feeding into LLM, which then interprets the audio based on the input text instructions and produces a textual response.

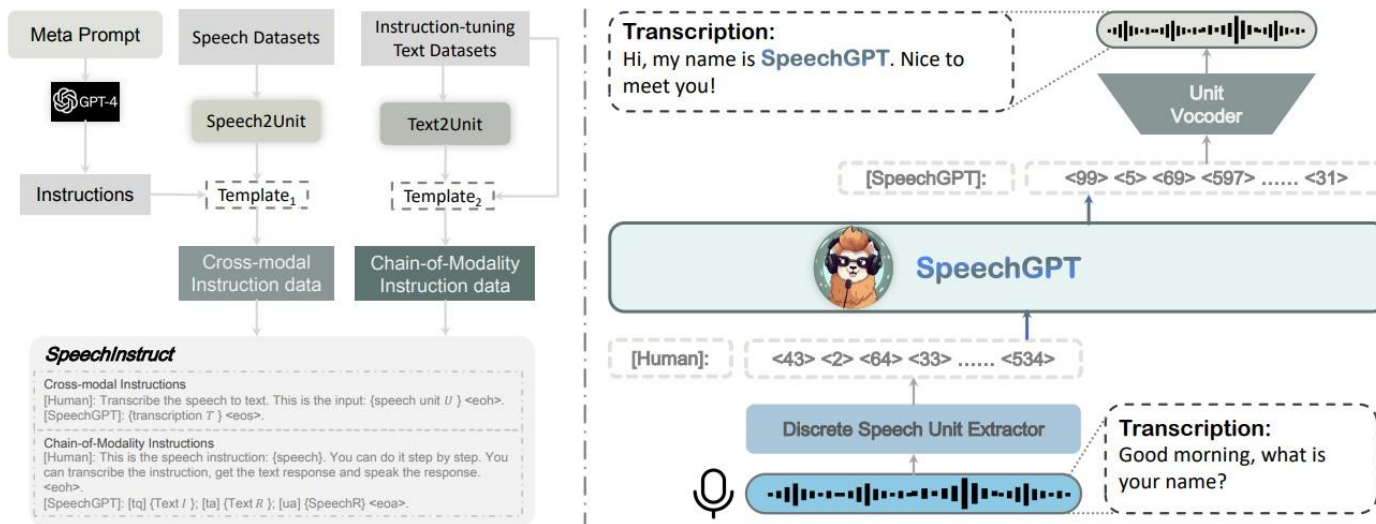
- 1 *AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. 2023*
- 2 *SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. 2023*
- 3 *VioLA: Unified Codec Language Models for Speech Recognition, Synthesis, and Translation. 2023*
- 4 *AudioPaLM: A Large Language Model That Can Speak and Listen. 2023*
- 5 *SALMONN: Towards Generic Hearing Abilities for Large Language Models. 2023*
- ...



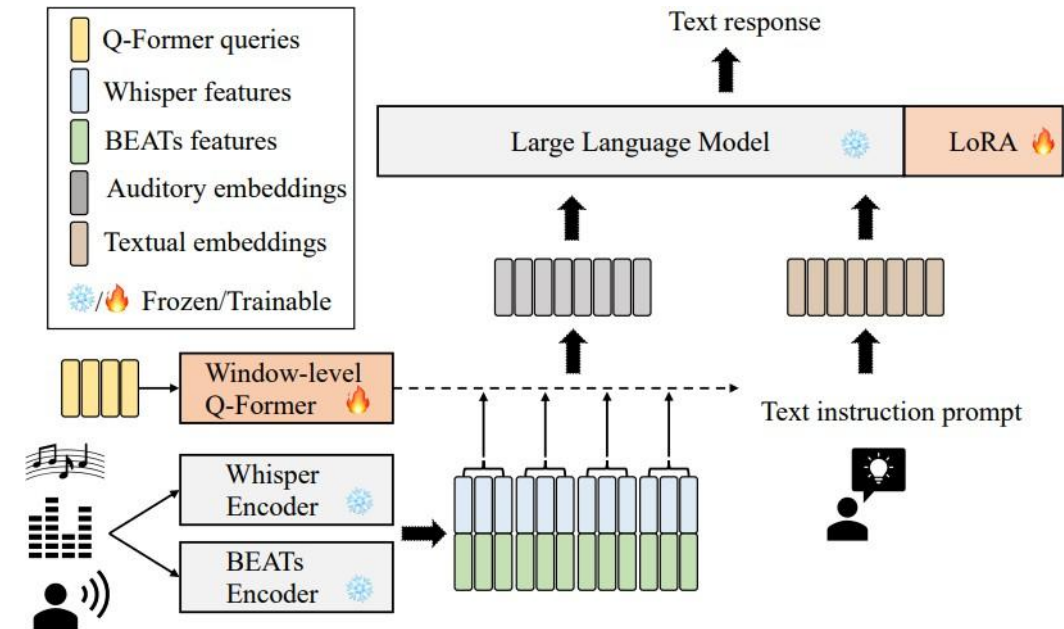
Multimodal Perceiving

• Audio-perceiving MLLM

0 SpeechGPT



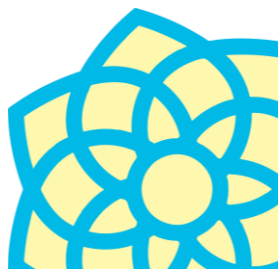
0 SALMONN



1 *SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities.* 2023

2 *SALMONN: Towards Generic Hearing Abilities for Large Language Models.* 2023

3 *Sparks of Large Audio Models: A Survey and Outlook.* <https://github.com/EmulationAI/awesome-large-audio-models>, 2023



Multimodal Perceiving

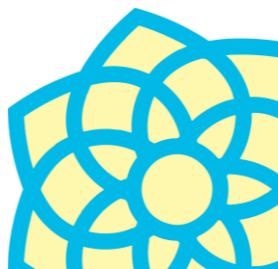
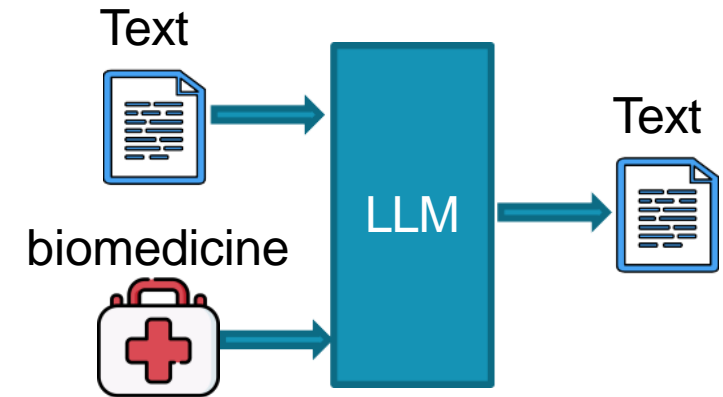
- X-perceiving MLLM

- **Bio-/Medical & Healthcare**

- | | | |
|----------------|---------------|-------------|
| ○ BioGPT | ○ DoctorGLM | ○ MedAlpaca |
| ○ DrugGPT | ○ BianQue | ○ AlpaCare |
| ○ BioMedLM | ○ ClinicalGPT | ○ Zhongjing |
| ○ OphGLM | ○ Qilin-Med | ○ PMC-LLaMA |
| ○ GatorTron | ○ ChatDoctor | ○ CPLLM |
| ○ GatorTronGPT | ○ BenTsao | ○ MedPaLM 2 |
| ○ MEDITRON | ○ HuatuoGPT | ○ BioMedGPT |

- 1 *BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. 2022*
- 2 *DrugGPT: A GPT-based Strategy for Designing Potential Ligands Targeting Specific Proteins. 2023*
- 3 *MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. 2023*
- 4 *HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge. 2023*
- 5 *AlpaCare: Instruction-tuned Large Language Models for Medical Application. 2023*
- 6 *A Survey of Large Language Models in Medicine: Progress, Application, and Challenge,*

<https://github.com/AI-in-Health/MedLLMsPracticalGuide>. 2023.



Multimodal Perceiving

- X-perceiving MLLM

- **Molecule & Chemistry**

+

○ ChemGPT

○ SPT

○ T5 Chem

○ ChemLLM

○ MolCA

○ MolXPTM

○ GIMLET

○ ...

- **Graph**

○ StructGPT

○ GPT4Graph

○ GraphGPT

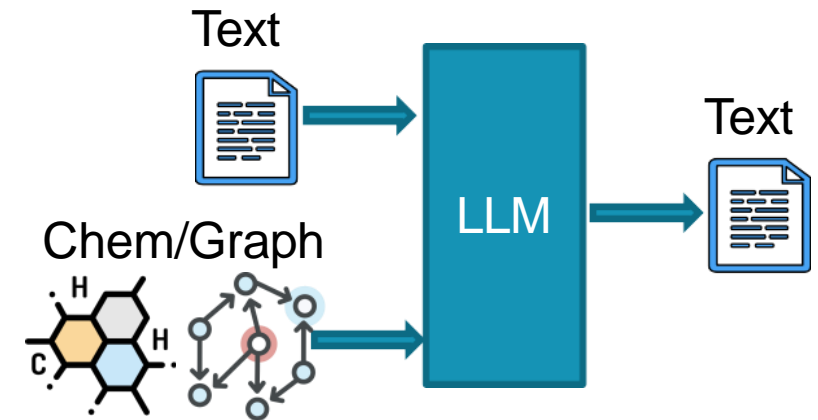
○ LLaGA

○ HiGPT

○ ...

- **Geographical Information System (GIS)**

○ GeoGPT



1 *Neural Scaling of Deep Chemical Models.* 2022

2 *ChemLLM: A Chemical Large Language Model.* 2023

3 *MolCA: Molecular Graph-Language Modeling with Cross-Modal Projector and Uni-Modal Adapter.* 2023

4 *StructGPT: A General Framework for Large Language Model to Reason on Structured Data.* 2023

5 *LLaGA: Large Language and Graph Assistant.* 2023

6 *Awesome-Graph-LLM, <https://github.com/XiaoxinHe/Awesome-Graph-LLM>.* 2023



Unified MLLM: Perceiving + Generation

● Scenarios



*Often, MLLMs need to not only **understand** the input multimodal information, but also to **generate** information in that modality.*

- Image Captioning
- Visual Question Answering
- Text-to-Vision Synthesis
- Vision-to-Vision Translation
- Scene Text Recognition
- Scene Text Inpainting
- ...



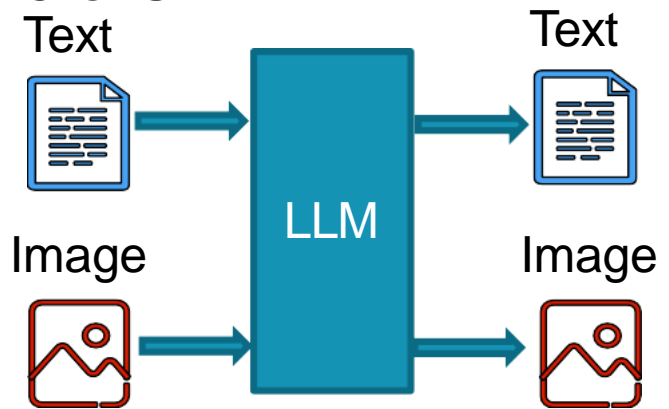
Overview of Modality and Functionality

	Modality (w/ Language)			
	Image	Video	Audio	3D
Input-side Perceiving	Flamingo, Kosmos-1, Blip2, mPLUG-Owl, Mini-GPT4, LLaVA, InstructBLIP, VPGTrans, CogVLM, Monkey, Chameleon, Otter, Qwen-VL, GPT-4v, SPHINX, Yi-VL, Fuyu, ...	VideoChat, Video-ChatGPT, Video-LLaMA, PandaGPT, MovieChat, Video-LLaVA, LLaMA-VID, Momentor, ...	AudioGPT, SpeechGPT, VIOLA, AudioPaLM, SALMONN, MU-LLaMA, ...	3D-LLM, 3D-GPT, LL3DA, SpatialVLM, PointLLM, Point-Bind, ...
	[Pixel-wise] GPT4RoI, LION, MiniGPT-v2, NExT-Chat, Kosmos-2, GLaMM, LISA, DetGPT, Osprey, PixelLM, ...	[Pixel-wise] PG-Video-LLaVA, Merlin, MotionEpic, ...	-	-
	Video-LLaVA, Chat-UniVi, LLaMA-VID		-	-
	Panda-GPT, Video-LLaMA, AnyMAL, Macaw-LLM, Gemini, VideoPoet, ImageBind-LLM, LLMBind, LLaMA-Adapter, ...			-
Perceiving + Generating	GILL, EMU, MiniGPT-5, DreamLLM, LLaVA-Plus, InternLM-XComposer2, SEED-LLaMA, LaVIT, Mini-Gemini, ...	GPT4Video, Video-LaVIT, VideoPoet, ...	AudioGPT, SpeechGPT, VIOLA, AudioPaLM, ...	-
	[Pixel-wise] Vitron		-	-
	NExT-GPT, Unified-IO 2, AnyGPT, CoDi-2, Modaverse, ViT-Lens, ...			-

Unified MLLM: Perceiving + Generation

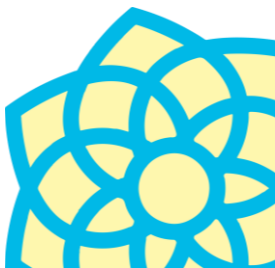
- Image

- GILL
- EMU
- MiniGPT-5
- DreamLLM
- LLaVA-Plus
- LaVIT
- ...



Central LLMs take as input both texts and images, after semantics comprehension, and generate both texts and images.

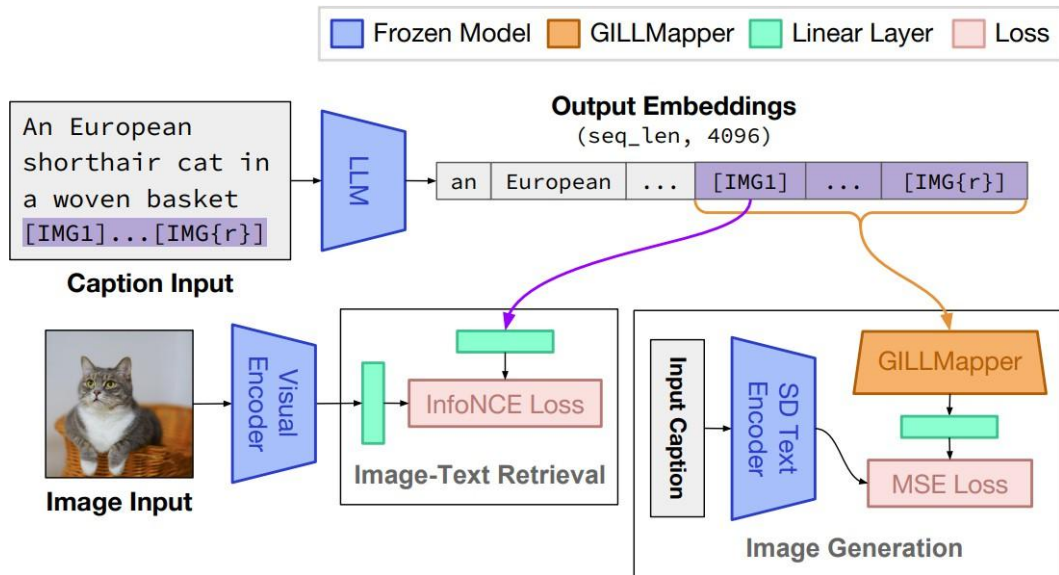
- 1 *Generating Images with Multimodal Language Models. 2023*
- 2 *Generative Pretraining in Multimodality. 2023*
- 3 *MiniGPT-5: Interleaved Vision-and-Language Generation via Generative Vokens. 2023*
- 4 *DreamLLM: Synergistic Multimodal Comprehension and Creation. 2023*
- 5 *LLaVA-Plus: Learning to Use Tools for Creating Multimodal Agents. 2023*
- ...



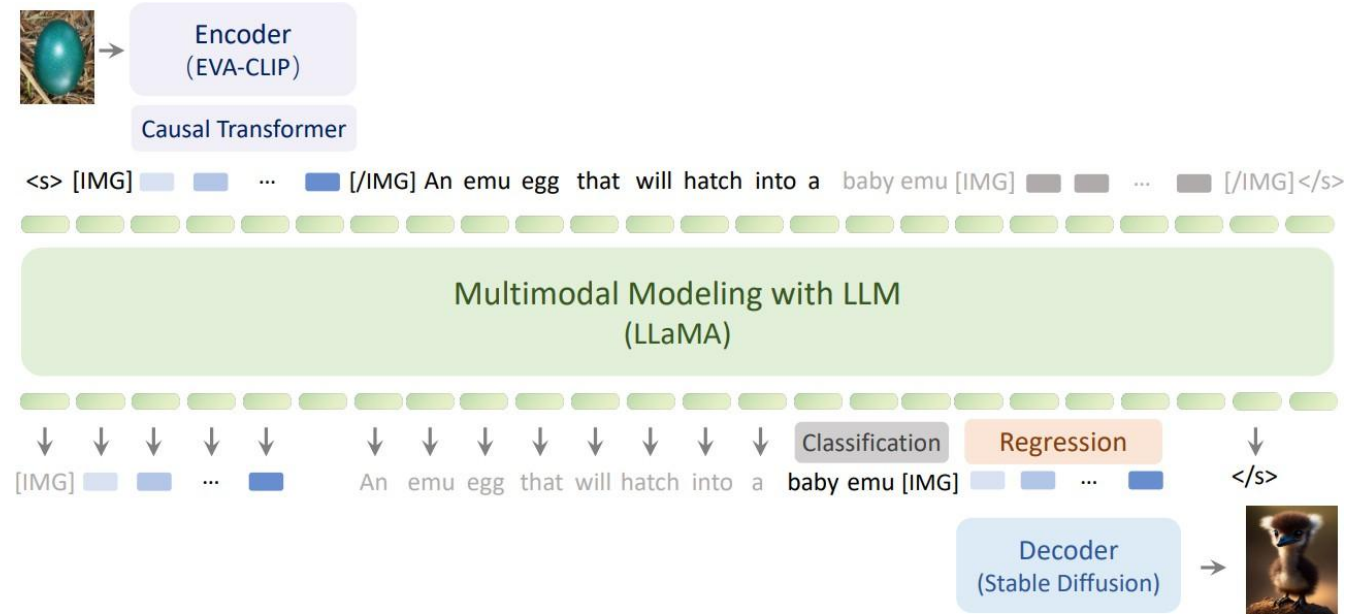
Unified MLLM: Perceiving + Generation

- Image

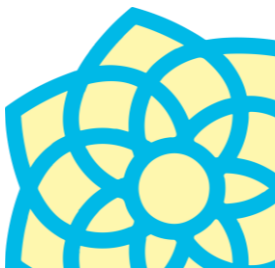
- o GILL



- o EMU



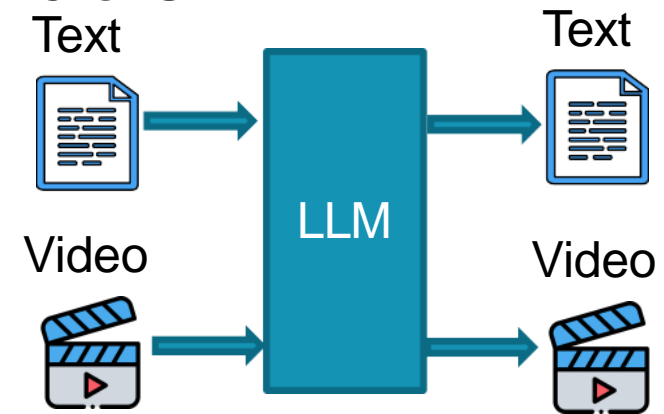
- 1 *Generating Images with Multimodal Language Models. 2023*
- 2 *Generative Pretraining in Multimodality. 2023*



Unified MLLM: Perceiving + Generation

- Video

- GPT4Video
- VideoPoet
- Video-LaVIT
- ...



Central LLMs take as input both texts and videos, after semantics comprehension, and generate both texts and videos.

1 *GPT4Video: A Unified Multimodal Large Language Model for Instruction-Followed Understanding and Safety-Aware Generation. 2023*

2 *VideoPoet: A Large Language Model for Zero-Shot Video Generation. 2023*

3 *Video-LaVIT: Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization. 2024*

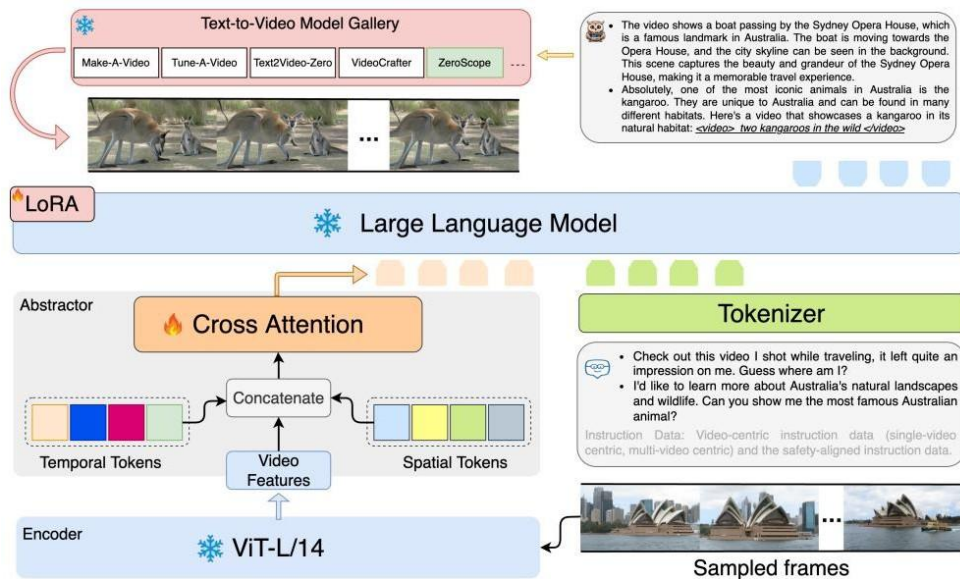
...



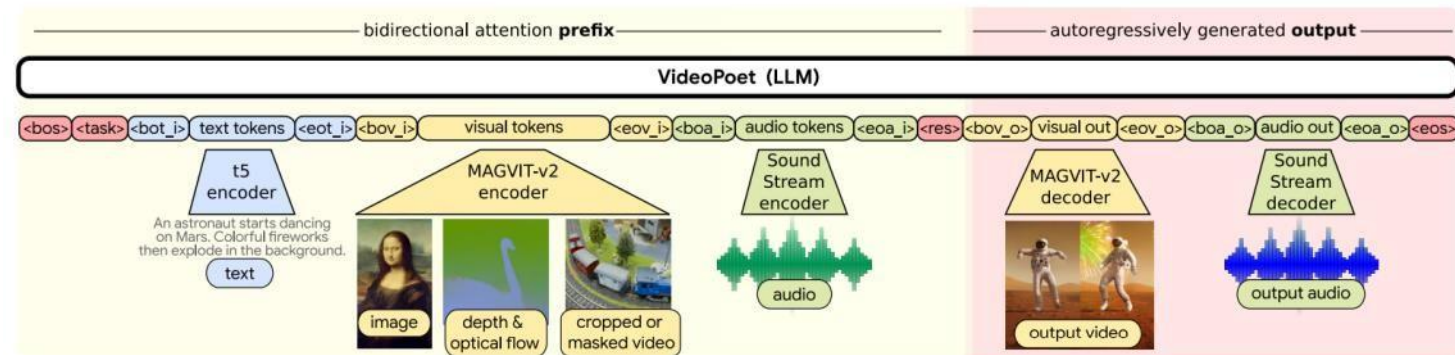
Unified MLLM: Perceiving + Generation

Video

GPT4Video



VideoPoet



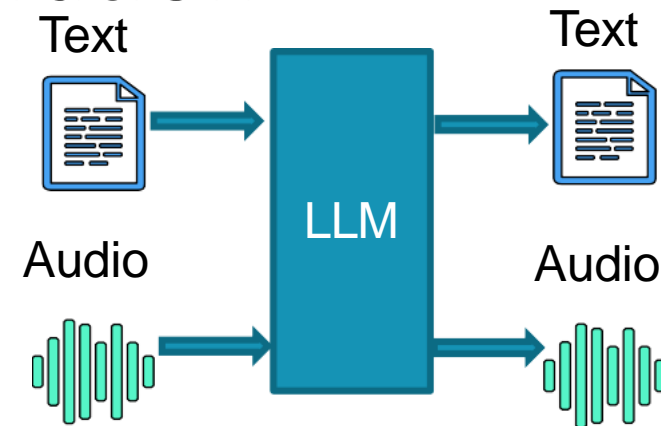
- 1 *GPT4Video: A Unified Multimodal Large Language Model for Instruction-Followed Understanding and Safety-Aware Generation. 2023*
- 2 *VideoPoet: A Large Language Model for Zero-Shot Video Generation. 2023*



Unified MLLM: Perceiving + Generation

• Audio

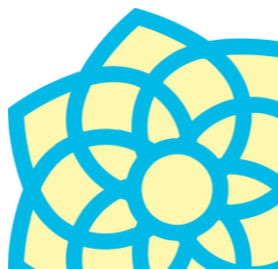
- AudioGPT,
- SpeechGPT,
- VIOLA,
- AudioPaLM,
- ...



Central LLMs take as input both texts and audio, after semantics comprehension, and generate both texts and audio.

- 1 *AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. 2023*
- 2 *SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. 2023*
- 3 *VioLA: Unified Codec Language Models for Speech Recognition, Synthesis, and Translation. 2023*
- 4 *AudioPaLM: A Large Language Model That Can Speak and Listen. 2023*

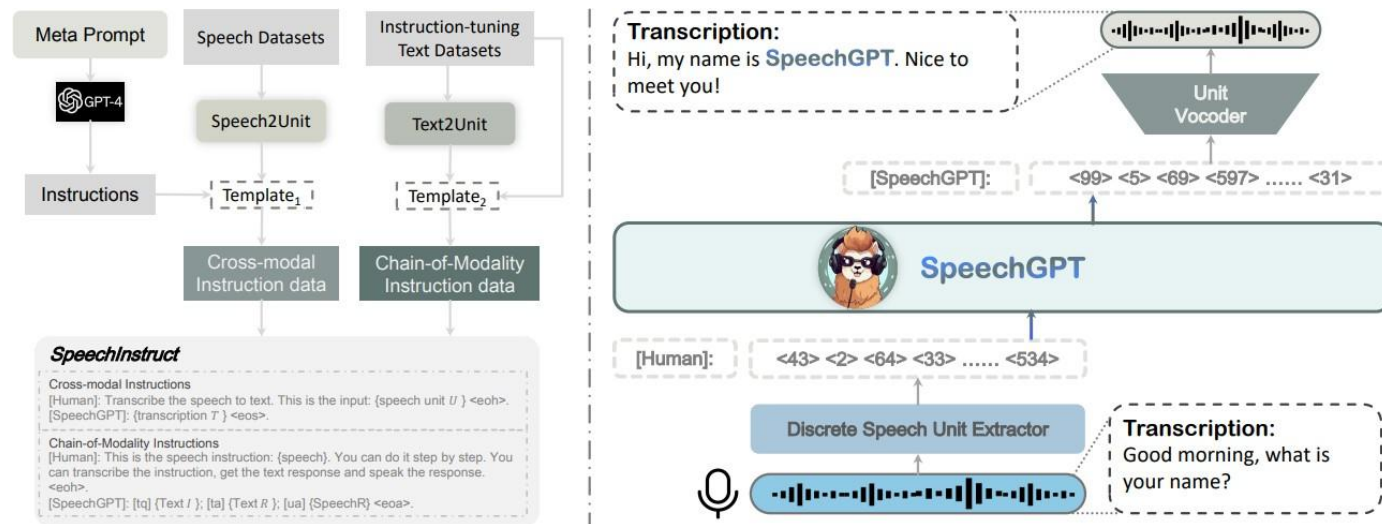
...



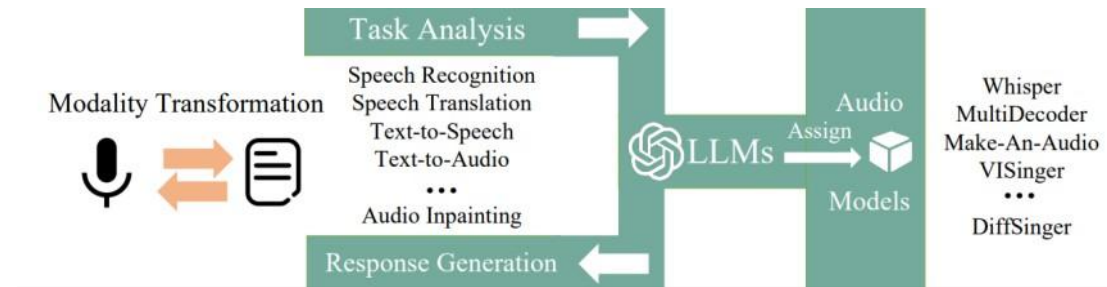
Unified MLLM: Perceiving + Generation

- Audio

- SpeechGPT

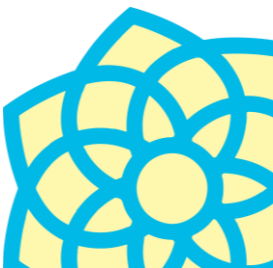


- AudioGPT



1 *SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities.* 2023

2 *AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head.* 2023



Unified MLLM: Harnessing Multi-Modalities

- Scenarios:



*In reality, modalities often have strong interconnections simultaneously. Thus, it is frequently necessary for MLLMs to handle the understanding of **multiple non-textual modalities at once**, rather than just one single (non-textual) modality.*

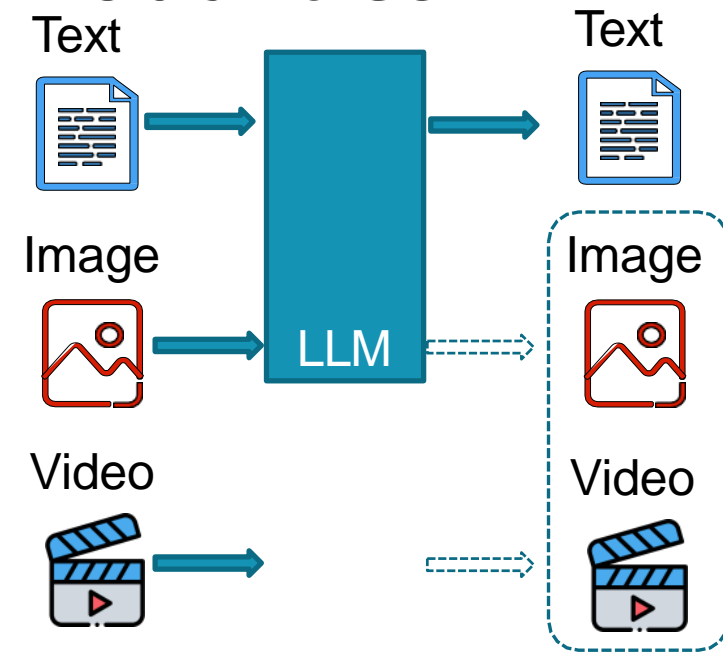
- Image+Video
- Audio+Video
- Image+Video+Audio
- Any-to-Any
- ...



Unified MLLM: Harnessing Multi-Modalities

- Text+Image+Video

- Video-LLaVA
- Chat-UniVi
- LLaMA-VID
- ...



Central LLMs take as input texts, image and video, after semantics comprehension, and generate texts (maybe also image and video, or combination).

- 1 *Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. 2023*
- 2 *Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. 2023*
- 3 *LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models. 2023*

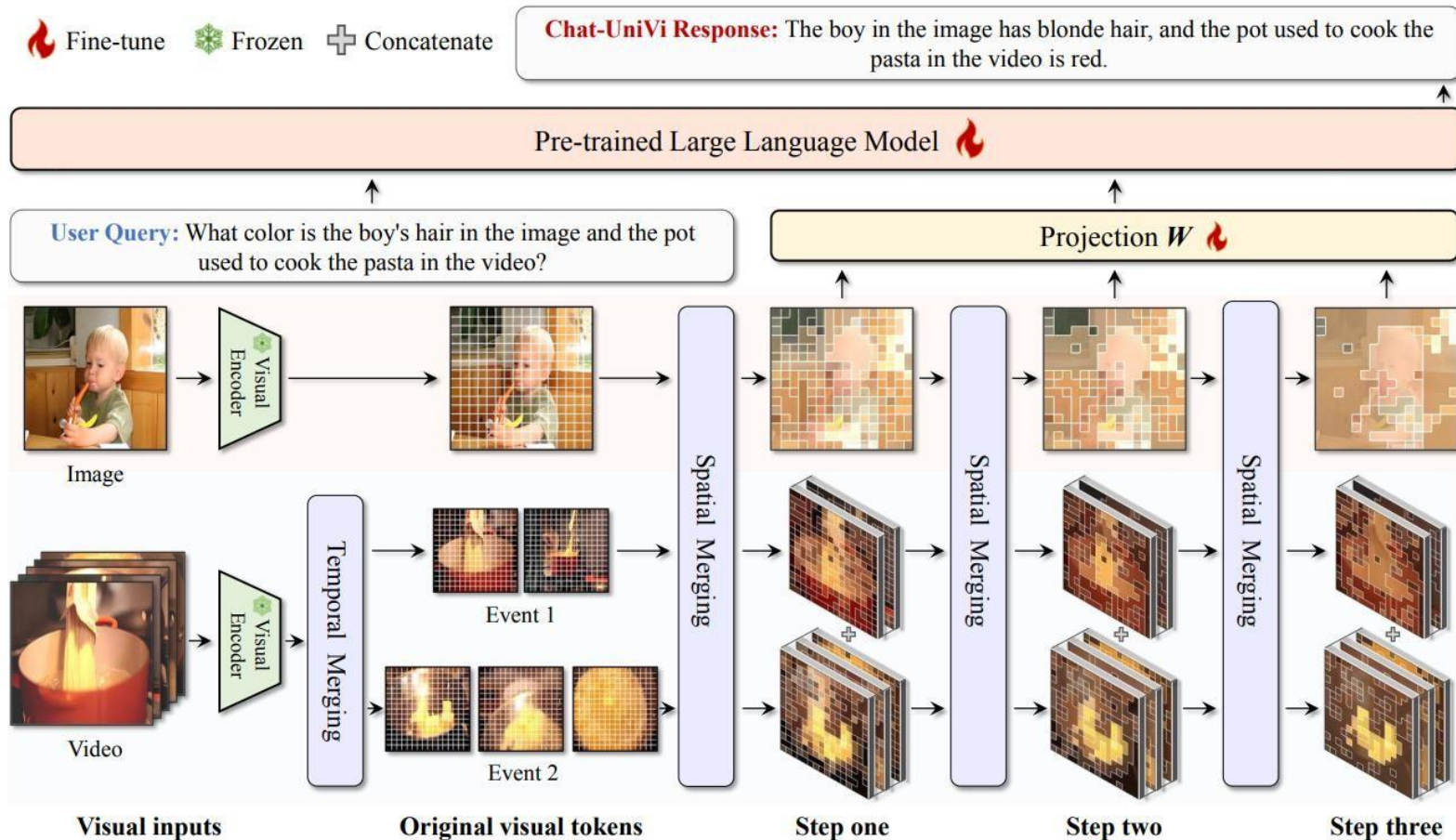
...



Unified MLLM: Harnessing Multi-Modalities

Text+Image+Video

Chat-UniVi



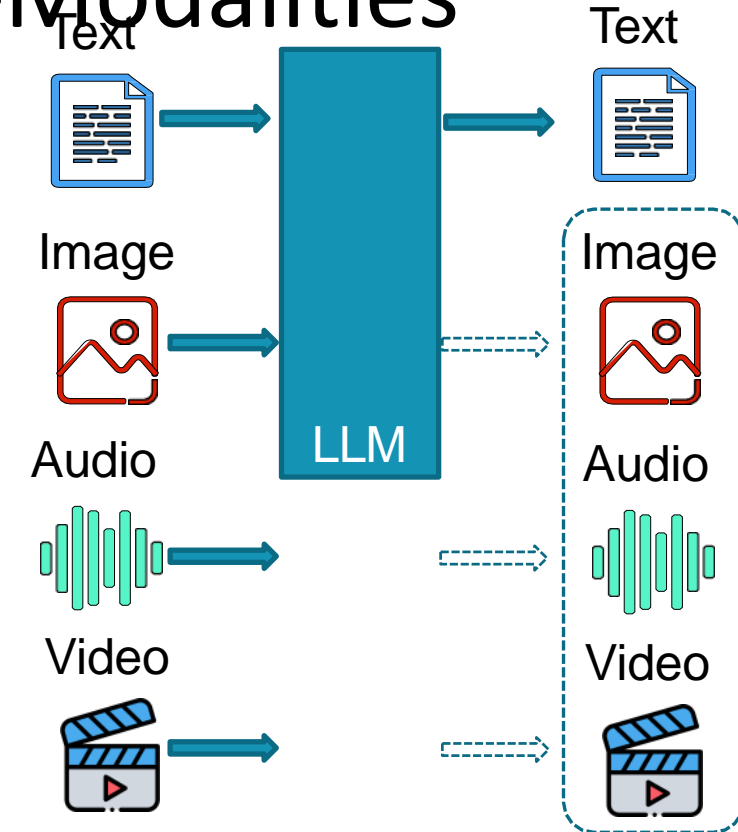
[1] Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. 2023



Unified MLLM: Harnessing Multi-Modalities

- Text+Image+Video+Audio

- Panda-GPT
- Video-LLaMA
- AnyMAL
- Macaw-LLM
- VideoPoet
- ImageBind-LLM
- LLMBind
- LLaMA-Adapter
- ...



Central LLMs take as input texts, audio, image and video, and generate texts (maybe also audio, image and video, or combination).

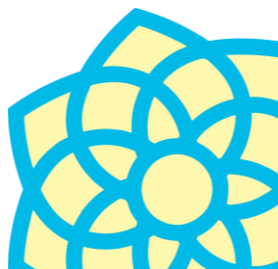
1 *PandaGPT: One Model to Instruction-Follow Them All. 2023*

2 *Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. 2023*

3 *AnyMAL: An Efficient and Scalable Any-Modality Augmented Language Model. 2023*

4 *Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. 2023*

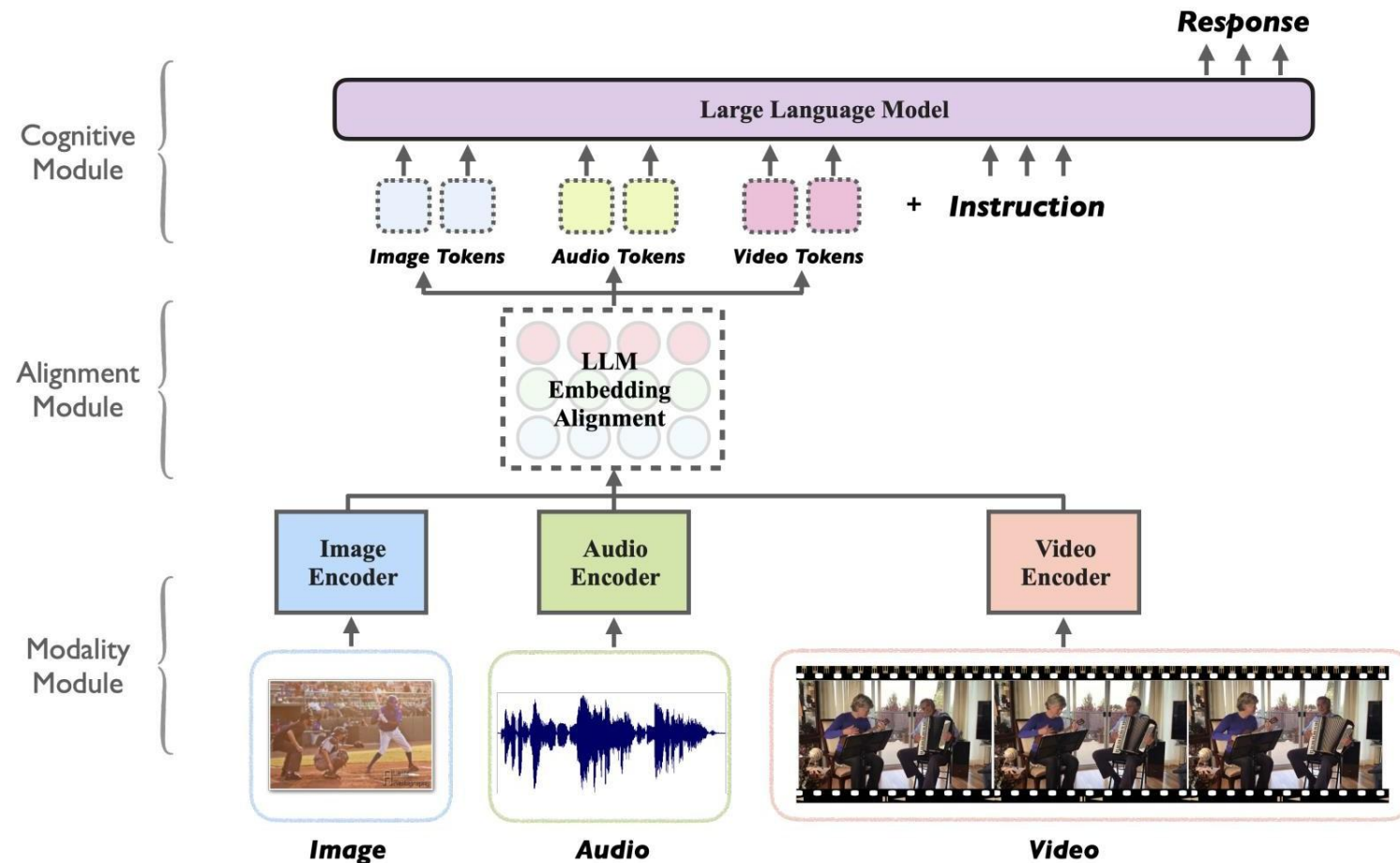
...



Unified MLLM: Harnessing Multi-Modalities

- Text+Image+Video+Audio

- o Macaw-LLM



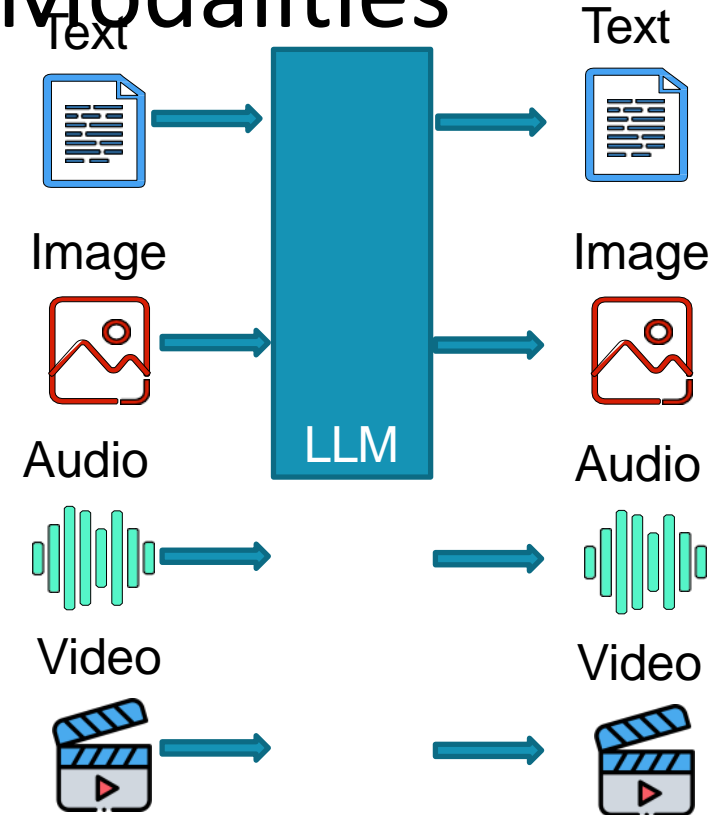
[1] Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. 2023



Unified MLLM: Harnessing Multi-Modalities

- Any-to-Any MLLM

- NExT-GPT
- Unified-IO 2 (w/o video)
- AnyGPT (w/o video)
- CoDi-2
- Modaverse
- ...



Central LLMs take as input texts, audio, image and video, and freely generate texts, audio, image and video, or combination.

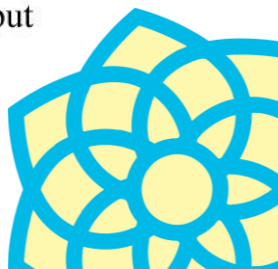
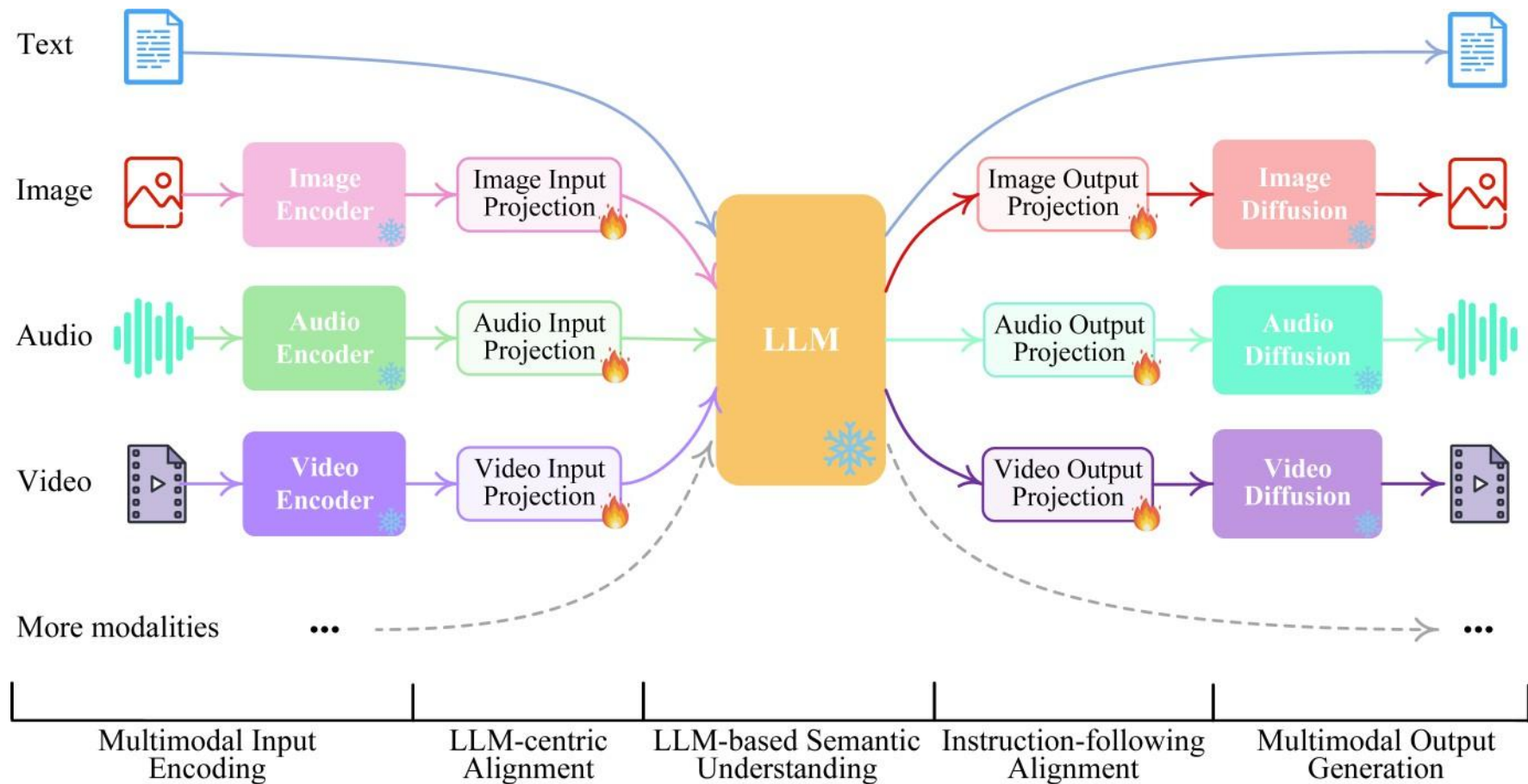
- 1 NExT-GPT: Any-to-Any Multimodal LLM. 2023
- 2 AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. 2023
- 3 CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation. 2023
- 4 ModaVerse: Efficiently Transforming Modalities with LLMs. 2023



Unified MLLM: Harnessing Multi-Modalities

- Any-to-Any MLLM

- o NExT-GPT



Unified MLLM: Harnessing Multi-Modalities

- Any-to-Any MLLM  NExT-GPT

- NExT-GPT



Text + Audio
↓
Text + Image + Video

Project: <https://next-gpt.github.io>

Paper: <https://arxiv.org/pdf/2309.05519>

Code: <https://github.com/NExT-GPT/NExT-GPT>



Fine-grained Capability of MLLM

Pixel-level Vision MLLM



The vision MLLMs described above generally only support coarse-grained, instance-level visual understanding. This can lead to **imprecise visual interpretations**. Also due to the lack of visual grounding, these MLLMs will potentially **produce hallucinations**.

- 0 Visual Grounding
- 0 Visual Segmentation
- 0 Visual Editing
- 0 Visual Inpainting
- 0 ...



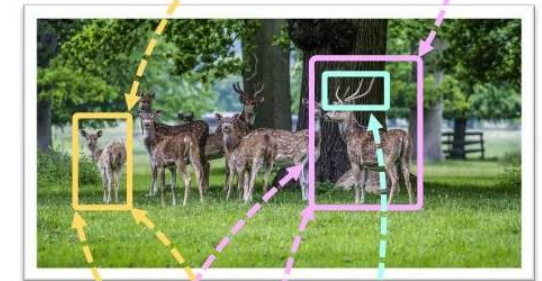
Describe the image in detail. Please output interleaved segmentation mask



The image showcases a large, white building with a red roof, surrounded by a well-manicured lawn and palm trees. The sky is visible over the building, the pavement, and the grass. The grass is also seen extending to the pavement.



What is the difference between this deer and another deer?



The main difference between the two deer is their size and antlers. The adult deer is larger with more prominent antlers, while the younger deer is smaller with no visible antlers.

Shikra



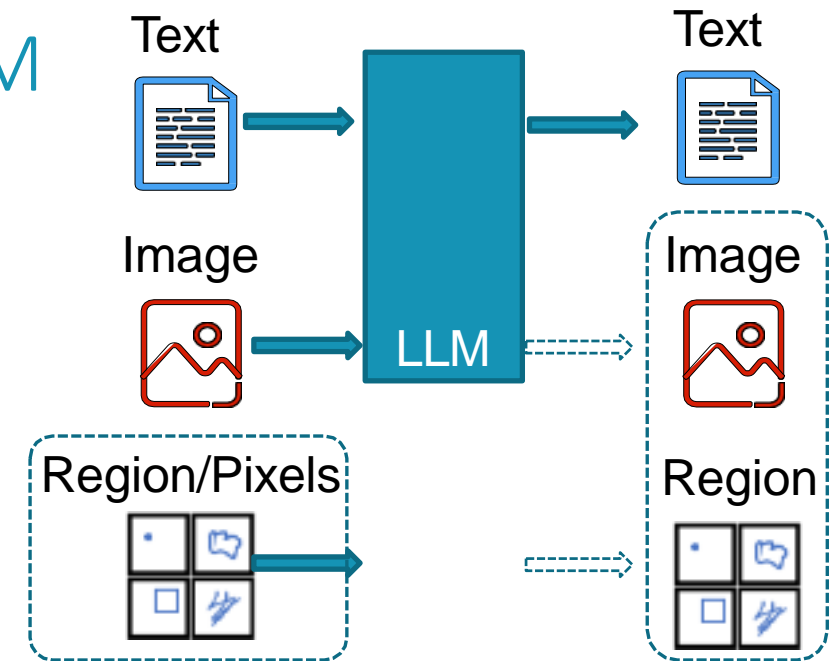
Fine-grained Capability of MLLM

Image-oriented Pixel-wise Regional MLLM

- GPT4RoI
- NExT-Chat
- MiniGPT-v2
- Shikra
- Kosmos-2
- GLaMM
- LISA
- DetGPT
- Osprey
- PixelLM
- LION
- ...



Users input an image (potentially specifying a region), and the LLM outputs content based on its understanding, grounding the visual content to specific pixel-level regions of the image.



- 1 *GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. 2023*
- 2 *NExT-Chat: An LMM for Chat, Detection and Segmentation. 2023*
- 3 *MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. 2023*
- 4 *Osprey: Pixel Understanding with Visual Instruction Tuning. 2023*
- 5 *GLaMM: Pixel Grounding Large Multimodal Model. 2023*
- 6 *Kosmos-2: Grounding Multimodal Large Language Models to the World. 2023*
- 7 *DetGPT: Detect What You Need via Reasoning. 2023*
- 8 *PixelLM: Pixel Reasoning with Large Multimodal Model. 2023*
- 9 *Lisa: Reasoning segmentation via large language model. 2023*
- 10 *Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. 2023*
- ...



Overview of Modality and Functionality

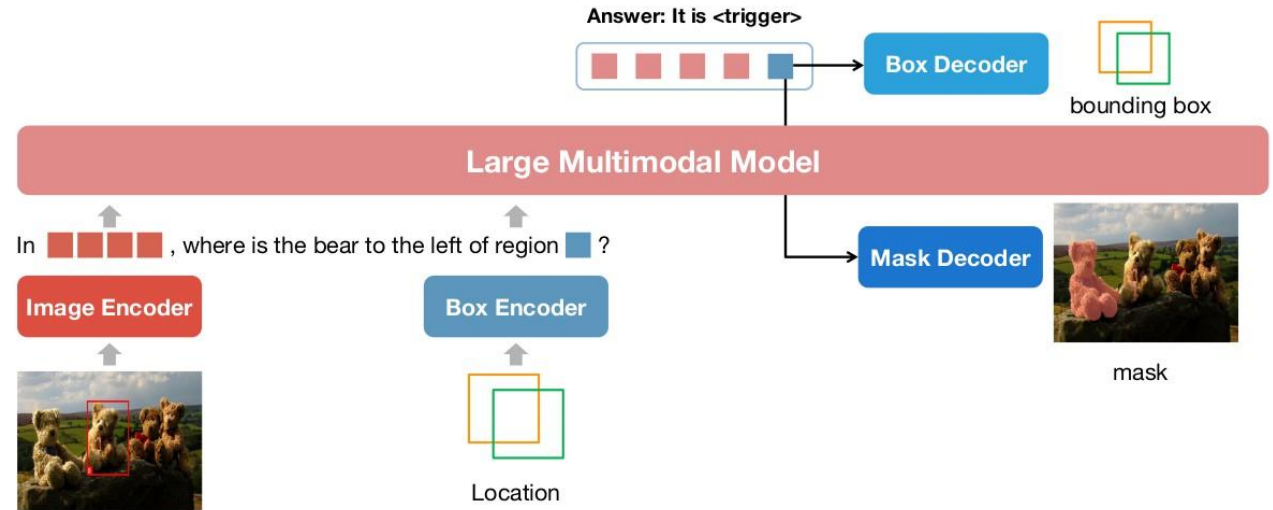
Modality (w/ Language)

	Image	Video	Audio	3D
Input-side Perceiving	Flamingo, Kosmos-1, Blip2, mPLUG-Owl, Mini-GPT4, LLaVA, InstructBLIP, VPGTrans, CogVLM, Monkey, Chameleon, Otter, Qwen-VL, GPT-4v, SPHINX, Yi-VL, Fuyu, ...	VideoChat, Video-ChatGPT, Video-LLaMA, PandaGPT, MovieChat, Video-LLaVA, LLaMA-VID, Momentor, ...	AudioGPT, SpeechGPT, VIOLA, AudioPaLM, SALMONN, MU-LLaMA, ...	3D-LLM, 3D-GPT, LL3DA, SpatialVLM, PointLLM, Point-Bind, ...
	[Pixel-wise] GPT4RoI, LION, MiniGPT-v2, NExT-Chat, Kosmos-2, GLaMM, LISA, DetGPT, Osprey, PixelLM, ...	[Pixel-wise] PG-Video-LLaVA, Merlin, MotionEpic, ...	-	-
	Video-LLaVA, Chat-UniVi, LLaMA-VID		-	-
	Panda-GPT, Video-LLaMA, AnyMAL, Macaw-LLM, Gemini, VideoPoet, ImageBind-LLM, LLMBind, LLaMA-Adapter, ...			-
Perceiving + Generating	GILL, EMU, MiniGPT-5, DreamLLM, LLaVA-Plus, InternLM-XComposer2, SEED-LLaMA, LaVIT, Mini-Gemini, ...	GPT4Video, Video-LaVIT, VideoPoet, ...	AudioGPT, SpeechGPT, VIOLA, AudioPaLM, ...	-
	[Pixel-wise] Vitron		-	-
	NExT-GPT, Unified-IO 2, AnyGPT, CoDi-2, Modaverse, ViT-Lens, ...			-

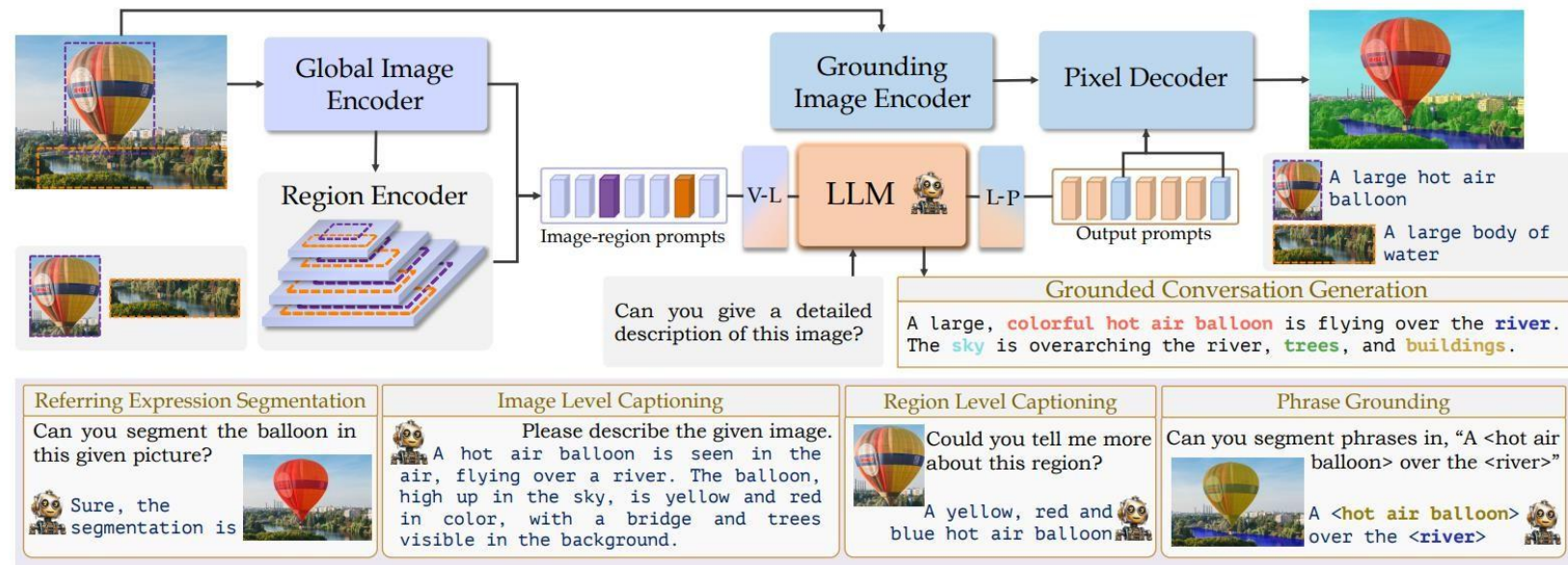
Fine-grained Capability of MLLM

- Image-oriented Pixel-wise

o NExT-Chat



o GLaMM



Fine-grained Capability of MLLM

- Image-oriented Pixel-wise Regional MLLM



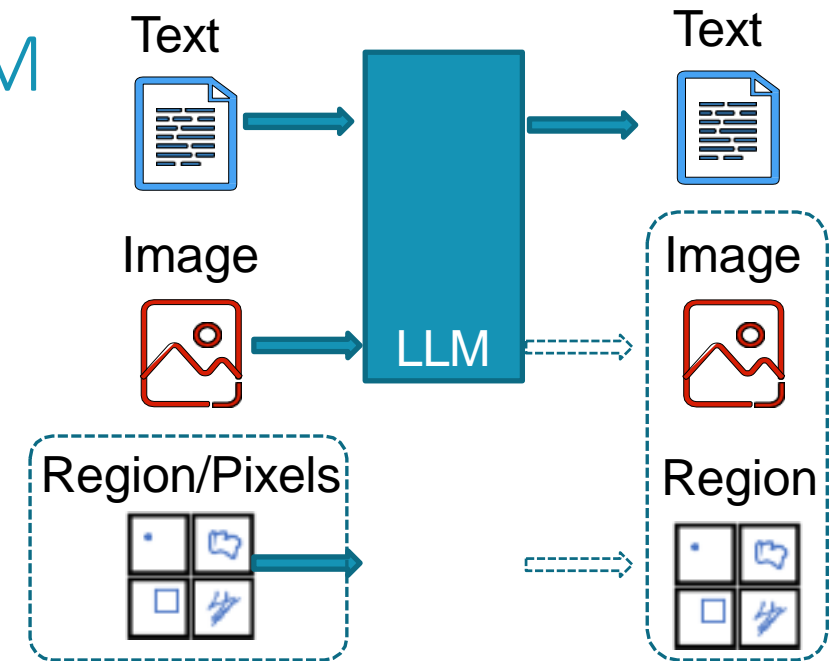
Pixel-level Awareness at Input/Output

- Output-side Only Pixel-wise Awareness

LISA, PixelLM, DetGPT, MiniGPT-v2, LION

- Input-&Output-side Pixel-wise Awareness

NExT-Chat, GPT4RoI, Shikra, KOSMOS-2, GLaMM, Osprey



Fine-grained Capability of MLLM

- Image-oriented Pixel-wise Regional MLLM

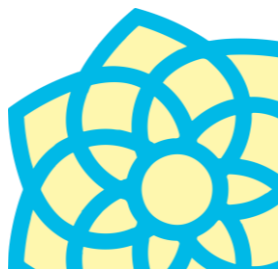
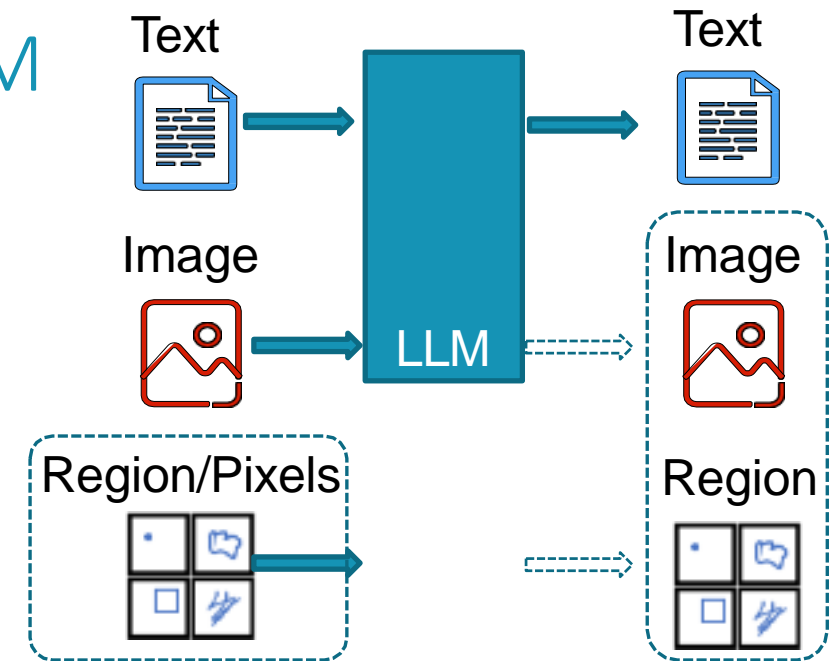
 Pixel Granularity

- Bounding-box Coordinates

NExT-Chat, GPT4RoI, Shikra, LION,
KOSMOS-2, DetGPT, MiniGPT-v2

- Finer-grained Mask-based Segments

NExT-Chat, LISA, PixellM,
GLaMM, Osprey



Fine-grained Capability of MLLM

- Image-oriented Pixel-wise Regional MLLM

 User Input Interaction

- No Image User Interaction

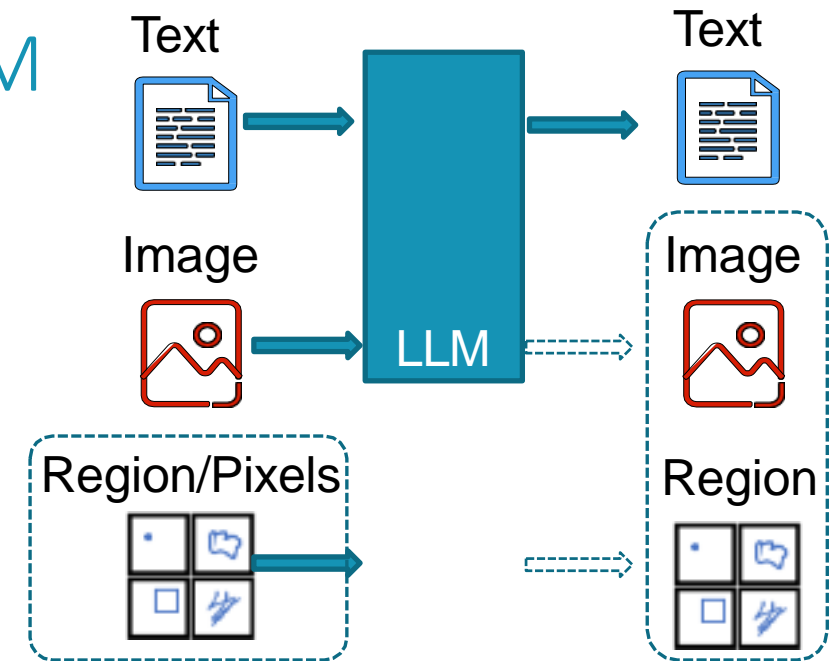
LISA, PixellLM, DetGPT, MiniGPT-v2, LION

- Bounding-box Coordinates

GPT4RoI, Shikra, KOSMOS-2, GLaMM

- User Sketches

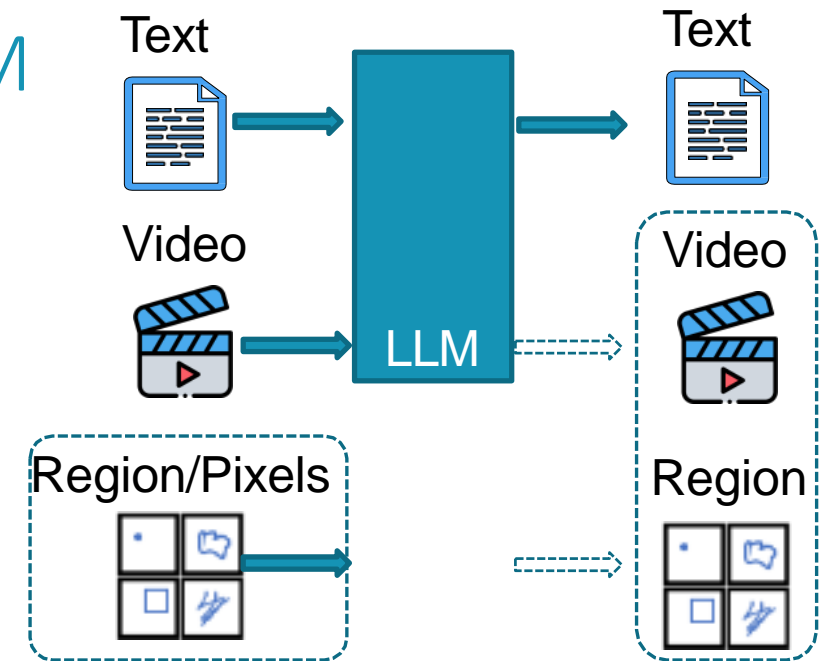
NExT-Chat, Osprey,



Fine-grained Capability of MLLM

- Video-oriented Pixel-wise Regional MLLM

- PG-Video-LLaVA
- Merlin
- MotionEpic
- ...



Users input an video (potentially specifying a region), and the LLM outputs content based on its understanding, grounding or tracking the content to specific pixel-level regions of the video.

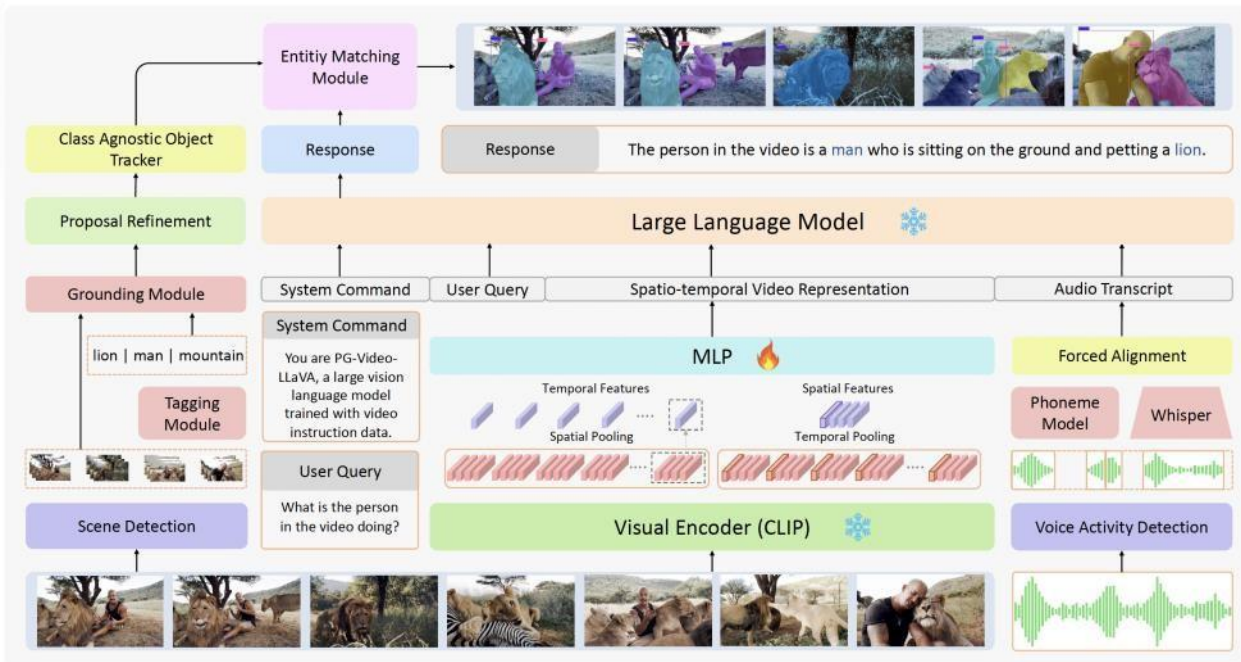
- 1 PG-Video-LLaVA: Pixel Grounding in Large Multimodal Video Models. 2023
- 2 Merlin: Empowering Multimodal LLMs with Foresight Minds. 2023
- 3 Video-of-Thought: Step-by-Step Video Reasoning from Perception to Cognition. 2024
- ...



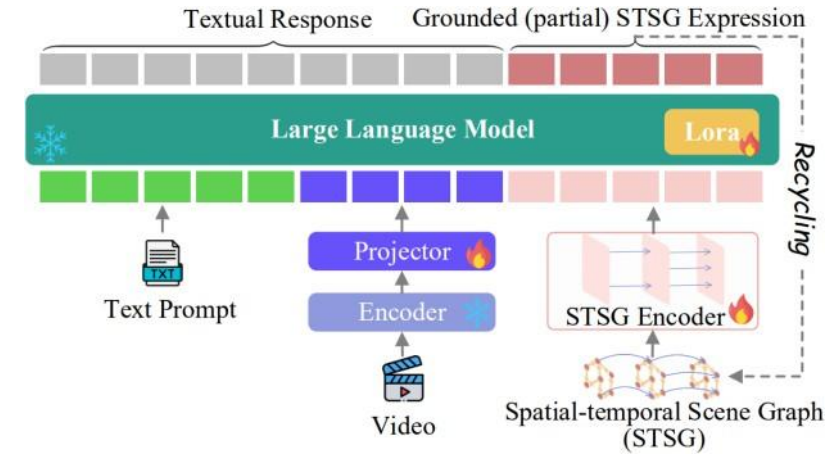
Fine-grained Capability of MLLM

Video-oriented Pixel-wise Regional MLLM

PG-Video-LLaVA



MotionEpic



Question: What is the least likely category for the animal in this video?
 A. Police Dog Competitive Animal Circus Performer D. Companion Pet E. Search and Rescue Dog



Step-1: The involved target is [dog].

Step-2: The partial STSG in tracking [dog] is:



Step-3: According to the video scene and STSG, the dog is crossing multiple hurdles with the dog being visible both before and after the hurdles. The accompanying man is observed providing instructions to guide the dog through the obstacles... Drawing on factual commonsense understanding, it might be inferred that the man is a trainer who is imparting various commands and training the dog on a grassy field.

Step-4:
 The video depicting professional training and complex actions suggests it might be a police dog performing daily training ... The rationality of the answer [A. Police Dog] is 2.
 The companion dog is to support companionship and emotional support to their owners rather than engaging in specialized tasks ... The answer [D. Companion Pet] has a coherence score of 8.

After ranking the rationale score, the final answer is [D. Companion Pet].

Step-5: Let's verify the [D. Companion Pet] based on visual perception ...
 1. Pixel Grounding Information Check: Based on the video scene, it depicts a training ground with a dog, so the answer is fitting.
 2. Commonsense Check: Observing the dog's energetic behavior during training aligns with the common understanding that companion pet are less likely to undergo such training, supporting the chosen answer.
 Conclusion: The answer [D. Companion Pet] is supported both by ...



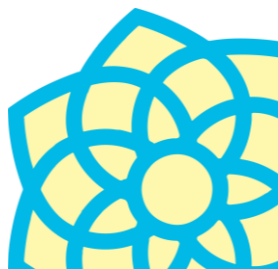
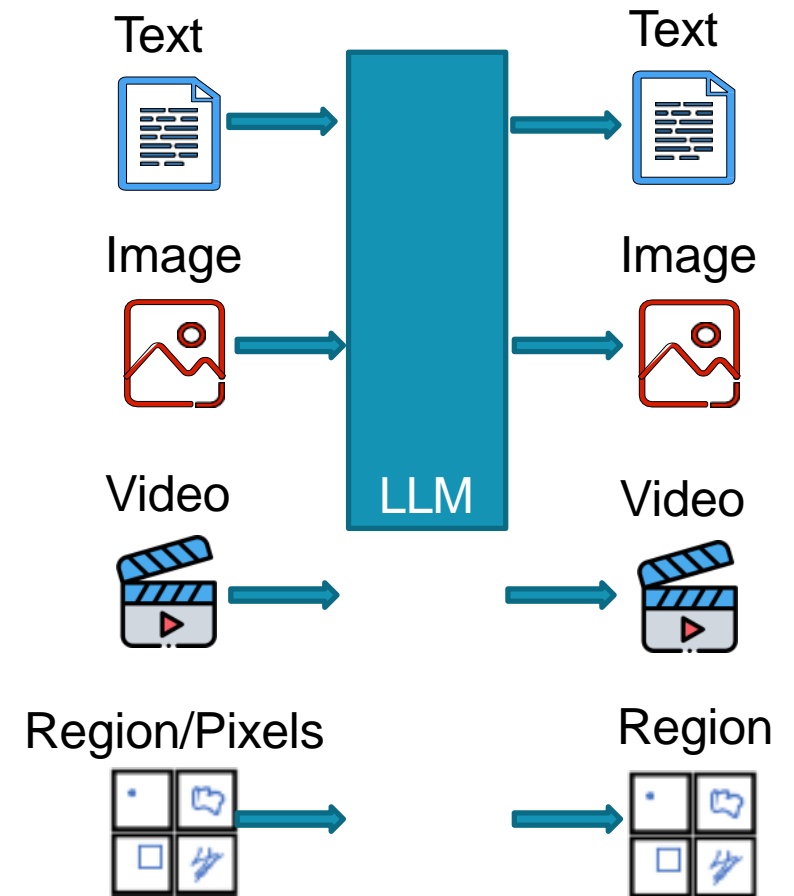
Fine-grained Capability of MLLM

- Unified Pixel-wise MLLM

- Vitron



Users input either an image or video (potentially specifying a region), and the LLM outputs content based on its understanding, generating, grounding or tracking the content to specific pixel-level regions of the image, video.



Fine-grained Capability of MLLM

Unified Pixel-wise MLLM

o Vitron

Model	Vision Supporting		Pixel/Regional Understanding	Segmenting/ Grounding	Generating	Editing
	Image	Video				
Flamingo [1]	✓	✗	✗	✗	✗	✗
BLIP-2 [45]	✓	✗	✗	✗	✗	✗
MiniGPT-4 [126]	✓	✗	✗	✗	✗	✗
LLaVA [57]	✓	✗	✗	✗	✗	✗
GILL [39]	✓	✗	✗	✗	✓	✗
Emu [90]	✓	✗	✗	✗	✓	✗
MiniGPT-5 [125]	✓	✗	✗	✗	✓	✗
DreamLLM [23]	✓	✗	✗	✗	✓	✗
GPT4RoI [122]	✓	✗	✓	✓	✗	✗
NExT-Chat [118]	✓	✗	✓	✓	✗	✗
MiniGPT-v2 [13]	✓	✗	✓	✓	✗	✗
Shikra [14]	✓	✗	✓	✓	✗	✗
Kosmos-2 [72]	✓	✗	✓	✓	✗	✗
GLaMM [78]	✓	✗	✓	✓	✗	✗
Osprey [117]	✓	✗	✓	✓	✗	✗
PixelLM [79]	✓	✗	✓	✓	✗	✗
LLaVA-Plus [58]	✓	✗	✗	✓	✓	✓
VideoChat [46]	✗	✓	✗	✗	✗	✗
Video-LLaMA [120]	✗	✓	✗	✗	✗	✗
Video-LLaVA [52]	✓	✓	✗	✗	✗	✗
Video-ChatGPT [61]	✗	✓	✗	✗	✗	✗
GPT4Video [99]	✗	✓	✗	✗	✓	✗
PG-Video-LLaVA [67]	✗	✓	✓	✓	✗	✗
NExT-GPT [104]	✓	✓	✗	✗	✓	✗
VITRON (Ours)	✓	✓	✓	✓	✓	✓



Fine-grained Capability of MLLM

- Unified Pixel-wise MLLM

- Vitron



Overview of Modality and Functionality

Modality (w/ Language)

	Image	Video	Audio	3D
Input-side Perceiving	Flamingo, Kosmos-1, Blip2, mPLUG-Owl, Mini-GPT4, LLaVA, InstructBLIP, VPGTrans, CogVLM, Monkey, Chameleon, Otter, Qwen-VL, GPT-4v, SPHINX, Yi-VL, Fuyu, ...	VideoChat, Video-ChatGPT, Video-LLaMA, PandaGPT, MovieChat, Video-LLaVA, LLaMA-VID, Momentor, ...	AudioGPT, SpeechGPT, VIOLA, AudioPaLM, SALMONN, MU-LLaMA, ...	3D-LLM, 3D-GPT, LL3DA, SpatialVLM, PointLLM, Point-Bind, ...
	[Pixel-wise] GPT4RoI, LION, MiniGPT-v2, NExT-Chat, Kosmos-2, GLaMM, LISA, DetGPT, Osprey, PixelLM, ...	[Pixel-wise] PG-Video-LLaVA, Merlin, MotionEpic, ...	-	-
	Video-LLaVA, Chat-UniVi, LLaMA-VID		-	-
	Panda-GPT, Video-LLaMA, AnyMAL, Macaw-LLM, Gemini, VideoPoet, ImageBind-LLM, LLMBind, LLaMA-Adapter, ...			-
Perceiving + Generating	GILL, EMU, MiniGPT-5, DreamLLM, LLaVA-Plus, InternLM-XComposer2, SEED-LLaMA, LaVIT, Mini-Gemini, ...	GPT4Video, Video-LaVIT, VideoPoet, ...	AudioGPT, SpeechGPT, VIOLA, AudioPaLM, ...	-
	[Pixel-wise] Vitron		-	-
	NExT-GPT, Unified-IO 2, AnyGPT, CoDi-2, Modaverse, ViT-Lens, ...			-

Fine-grained Capability of MLLM

- Unified Pixel-wise MLLM

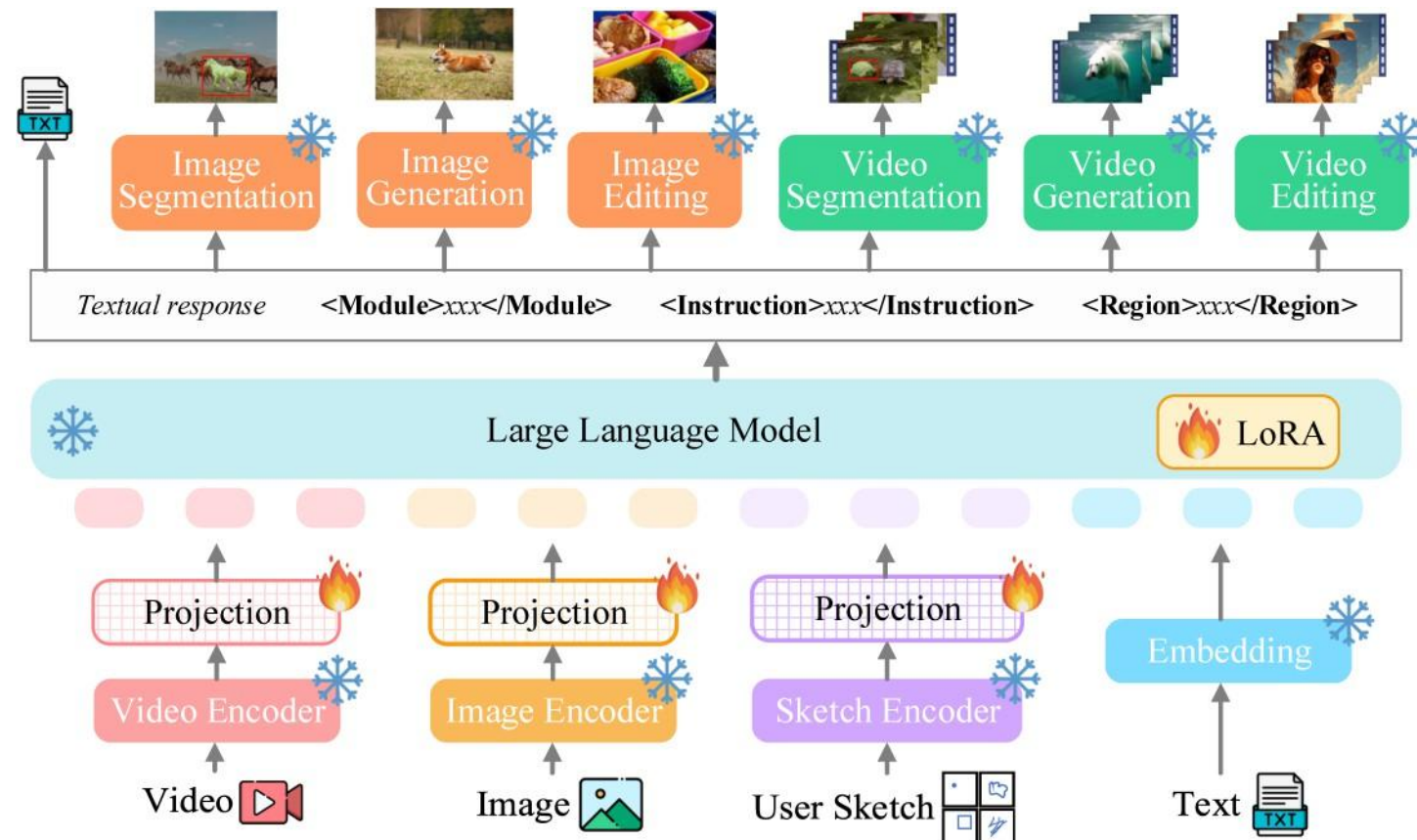
- o Vitron



Project: <https://vitron-llm.github.io/>

Paper: <https://is.gd/aGu0VV>

Code&Demo: <https://github.com/SkyworkAI/Vitron>



What's Next

- Angle-I: Unification of as Many Modalities & Tasks as Possible

- Modality Perspective: Going Broader

 *Currently, the majority of MLLM research focuses primarily on the integration of visual signals (e.g., **Image**, **Video**).*



What's Next from Multimodal LLM to AGI

- Angle-I: Unification of as Many Modalities & Tasks as Possible
 - Modality Perspective: Going Broader

➤ Modalities in current NExT-GPT:


language 


image 

sound 

video 


➤ More modalities to go:

heat map 

code 

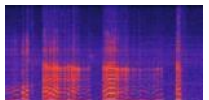
time series 

touch 

depth&3D 

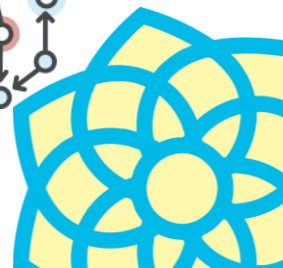
infrared/radar 

document/table 

spectrogram 

smell 

graph 



What's Next from Multimodal LLM to AGI

- Angle-I: Unification of as Many Modalities & Tasks as Possible

- Task Perspective: Going Deeper

👉 *Vision-based MLLM, **Vitron**, has focused on unifying image and video processing under the scope of pixel-wise tasks, ranging from low-level to high-level.*

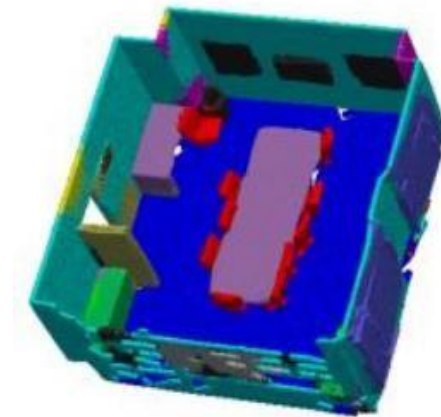
👉 *The next step could involve expanding MLLM support on the task level to more in-depth levels.*



Referring Segmentation



Panoptic Segmentation



3D Scene Segmentation



What's Next from Multimodal LLM to AGI

- Angle-II: Stronger Generation Ability via Better Tokenization

- Core Idea



*High-quality multimodal generation requires the system to **recover a sufficient amount of detailed multimodal information from the core LLM.***

- Remove the equivalence constraint between pre-LLM and post-LLM, as the roles of input and output multimodal tokens differ.
- Increase the information content of multimodal tokens to include more high-frequency details.

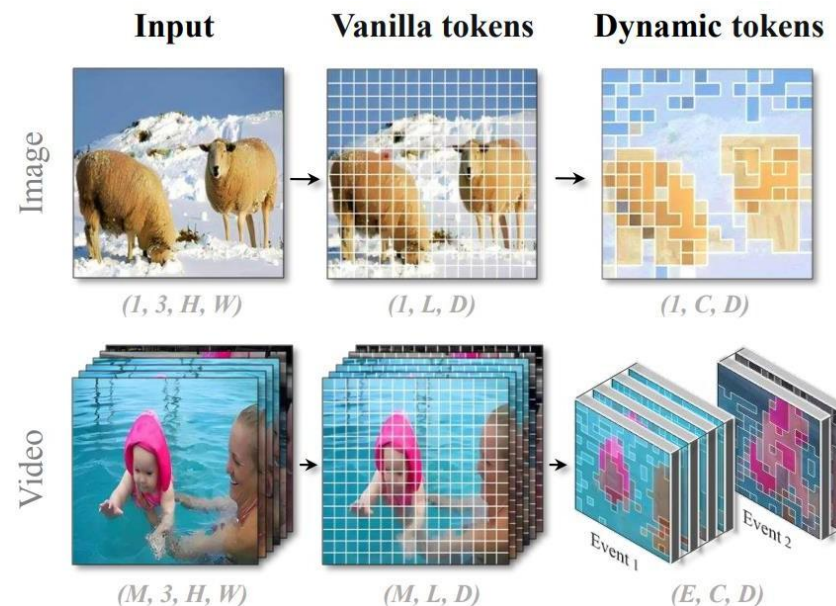


What's Next from Multimodal LLM to AGI

- Angle-II: Stronger Generation Ability via Better Tokenization

- A Hot Trend: Video tokenization

👉 Supporting both images and videos: more carefully model the spatial aspects of images and the temporal dynamics of videos.



1 LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models. 2024

2 Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. 2024

3 Video-LaViT: Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization. 2024



What's Next from Multimodal LLM to AGI

- Angle-III: More Multimodality & Multi-Task Synergy

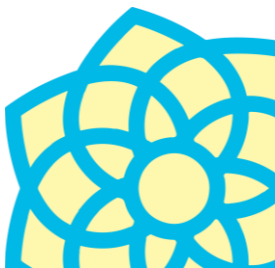
- Core Idea

- 👉 *Achieving a stronger MLLM, and potentially reaching AGI, necessitates enhanced Multimodality & Multi-Task Synergy for the MLLM generalist.*
- 👉 *Master **abductive reasoning** to facilitate **analogical thinking**, allowing different modalities and tasks, as well as the comprehension and generation processes, to mutually assist each other and create synergistic effects.*



1 *Abductive reasoning: Logic, visual thinking, and coherence. 1997.*

2 *Reasoning. <https://www.butte.edu/departments/cas/tipsheets/thinking/reasoning.html>*



* Part-IV

Multimodal Instruction Tuning

Haotian Liu

Ph.D.

University of Wisconsin, Madison

<https://hliu.cc>

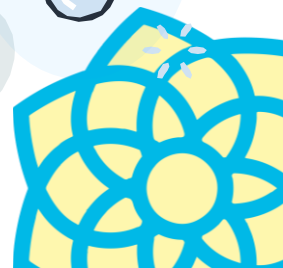
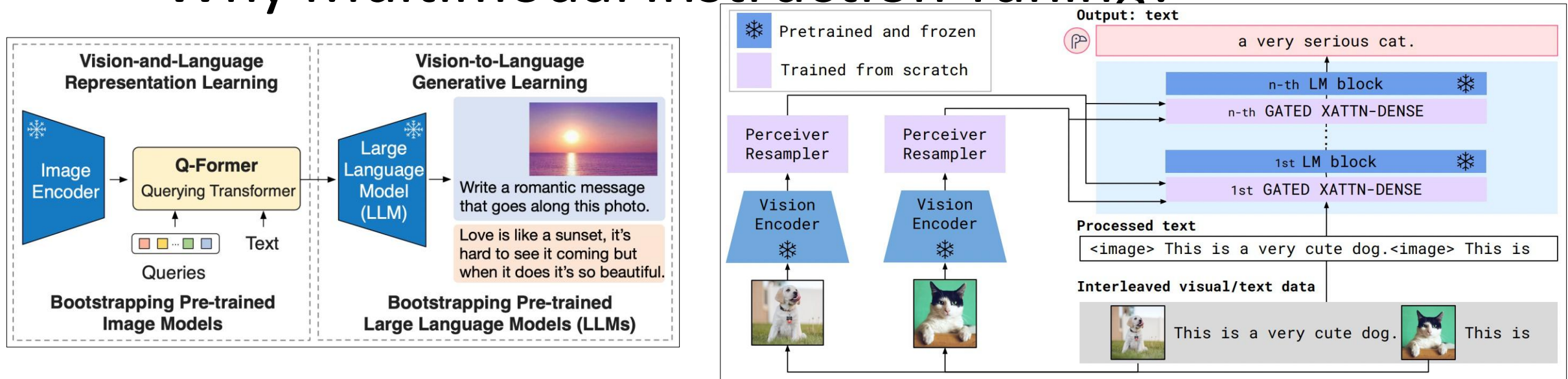


Table of Content

- **Motivations**
- **Multimodal Instruction Tuning Framework**
 - × Framework
 - × Training Paradigms
- **Multimodal Instruction Tuning Data Construction**
 - × Pretraining Data
 - × Instruction Tuning Data
 - × Existing Datasets



Why Multimodal Instruction Tuning?



Pretrained models aligns multiple modalities, can understand basic information from different modalities, and sometimes perform simple question-answering.



Cannot follow complex instructions, and often require **task-specific** fine-tuning for it to perform well on downstream tasks.

[Wang et al. 2022] GIT: A Generative Image-to-text Transformer for Vision and Language

[Li et al. 2023] Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models

[Alayrac et al. 2022] Flamingo: a visual language model for few-shot learning



Why Multimodal Instruction Tuning?

- From **Single-Purpose** to **General-Purpose**



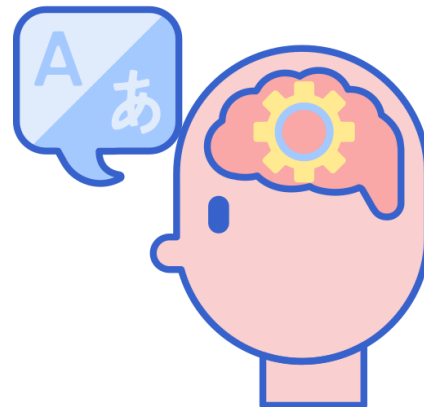
Traditional vision models are **task-specific**, which requires training and using multiple models for different tasks and **restrict the potential synergies from diverse tasks**;



These vision models typically have a pre-defined and fixed interface, leading to **limited interactivity and adaptability in following users' task instructions**.



Multimodal Instruction Tuning allows multimodal models to **generalize to unseen tasks by following new instructions**, thus boosting **zero-shot** performance.



Instruction Tuning is NOT multitask learning

- **Multitask learning (with task tokens)**

Training



INPUT: <image><tok_task_1=short_cap>
OUTPUT: <generated short descriptions>

INPUT: <image><tok_task_2=yes_no>
OUTPUT: yes/no

Testing

Only with <tok_task_1>, <tok_task_2>...

Does not work with <new_task=long_cap>

- **Instruction tuning (with natural language task instructions)**

Training



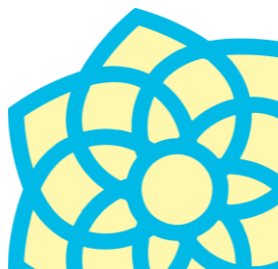
INPUT: <image>Describe this image briefly.
OUTPUT: <generated short descriptions>

INPUT: <image>Is this xxx?
OUTPUT: yes/no

Testing

INPUT: <image>Describe this image in detail.
OUTPUT: <long descriptions>

Generalizes to new instructions zero-shot.

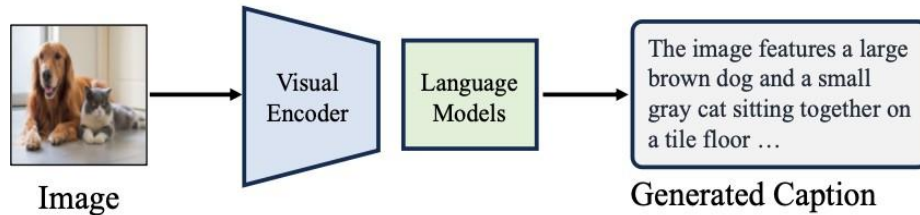


Why Multimodal Instruction Tuning?

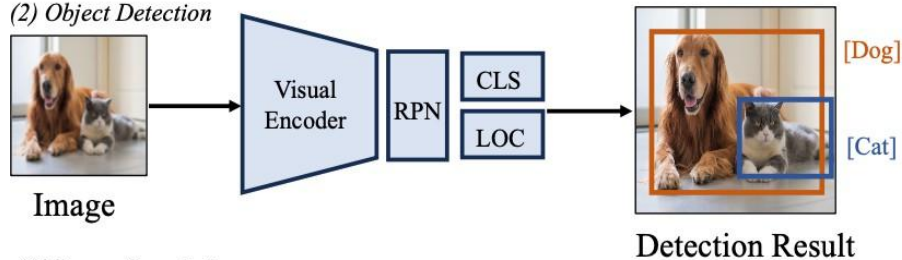
- From **Single-Purpose** to **General-Purpose**

(a). Traditional Task Paradigm for Computer Vision

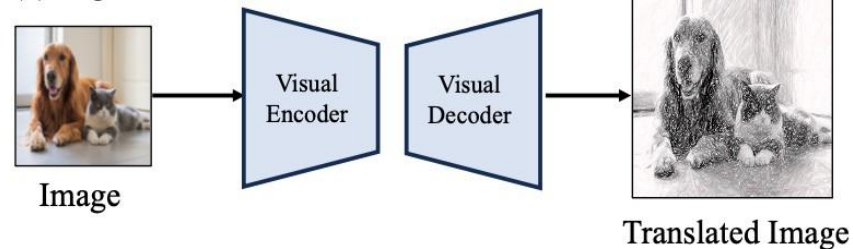
(1) Image Captioning



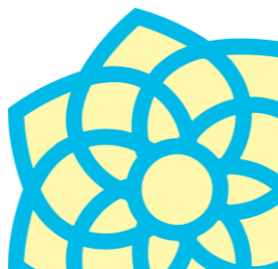
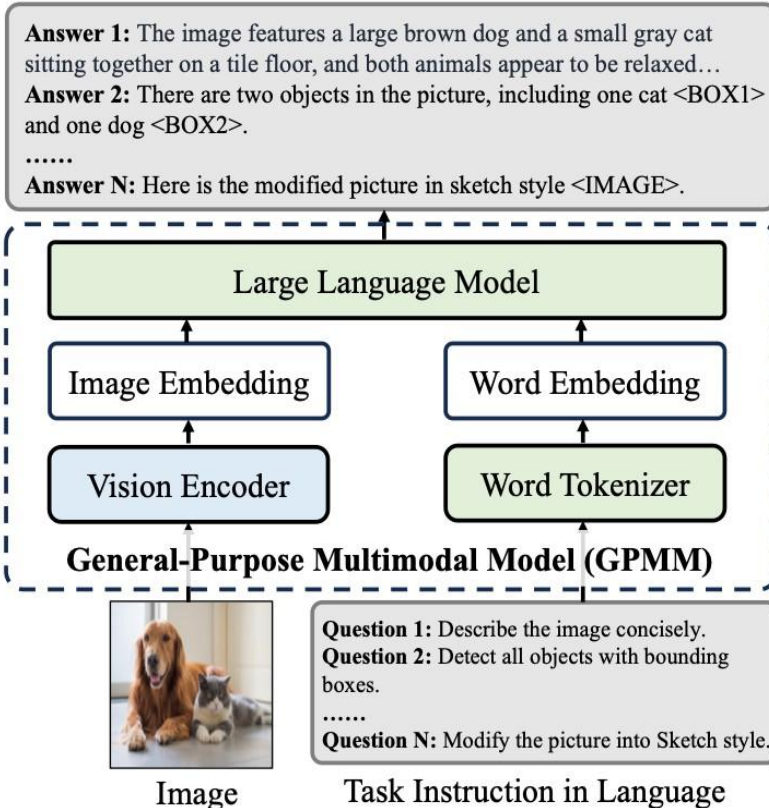
(2) Object Detection



(N) Image Translation



(b). Instruction-based Task Paradigm for Computer Vision



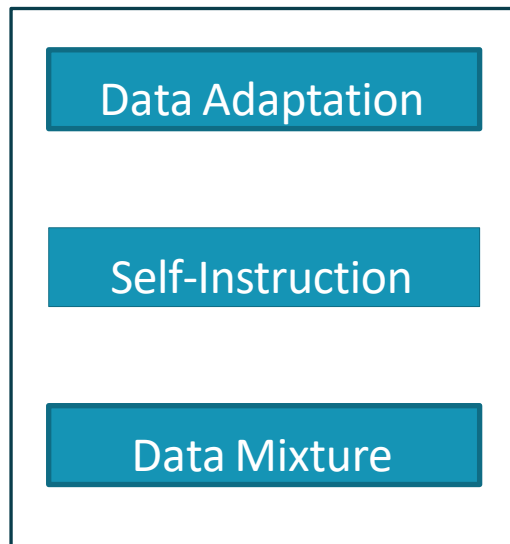
1

Multimodal Instruction Tuning Framework

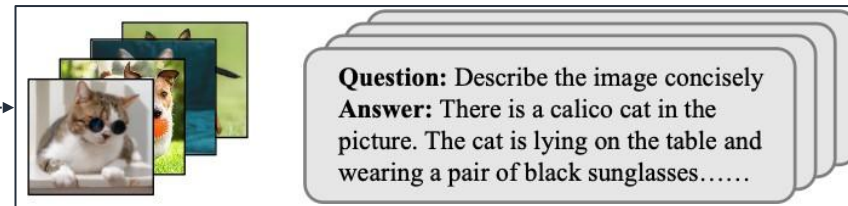


MLLM Instruction Tuning Framework

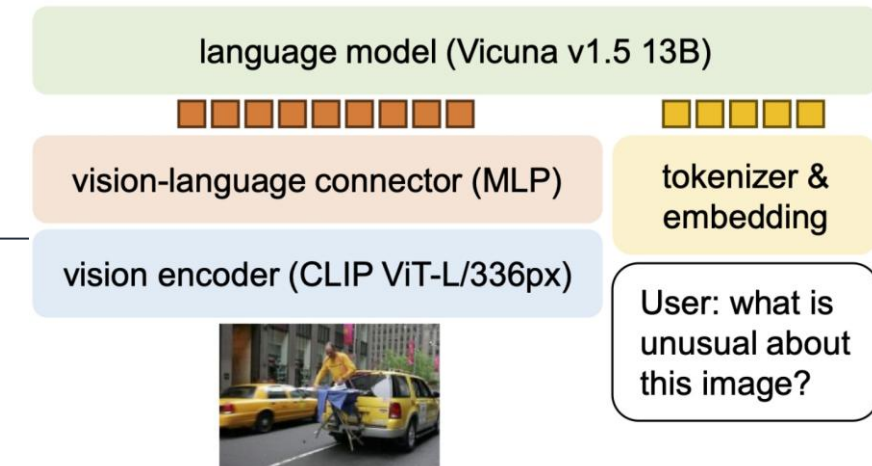
Data Construction



Visual Instruction-following data



Visual Instruction Tuning Framework Example: LLaVA-1.5



Popular MLLMs: *LLaVA, MiniGPT4, LLaVA-NeXT, ViP-LLaVA, LLaVA-UHD, MiniCPM, Qwen-VL, CogAgent, InternVL, mPLUG-OWL, Monkey, MiniGemini, LLaVA-HR, SPHINX, DeepSeek-VL, MoAI*

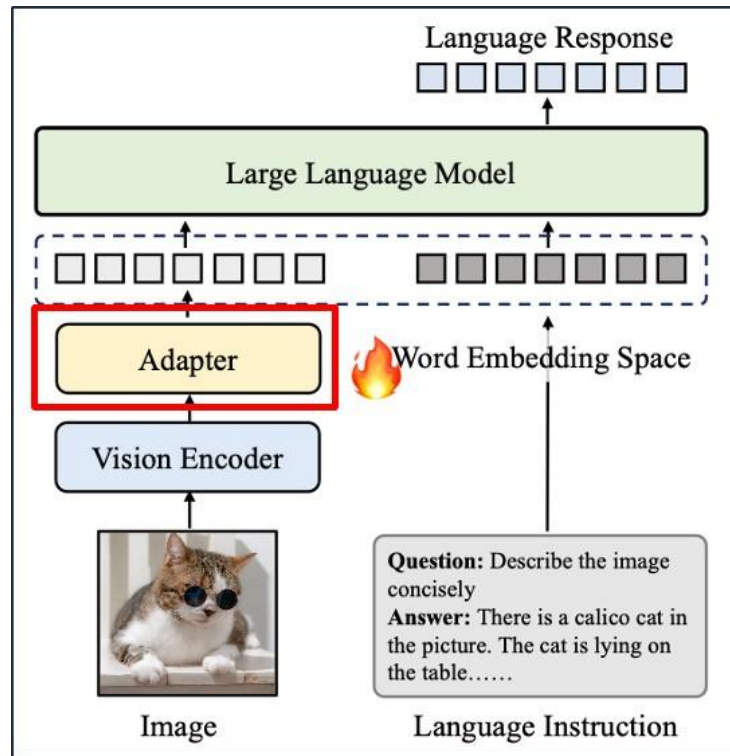


Training Paradigms



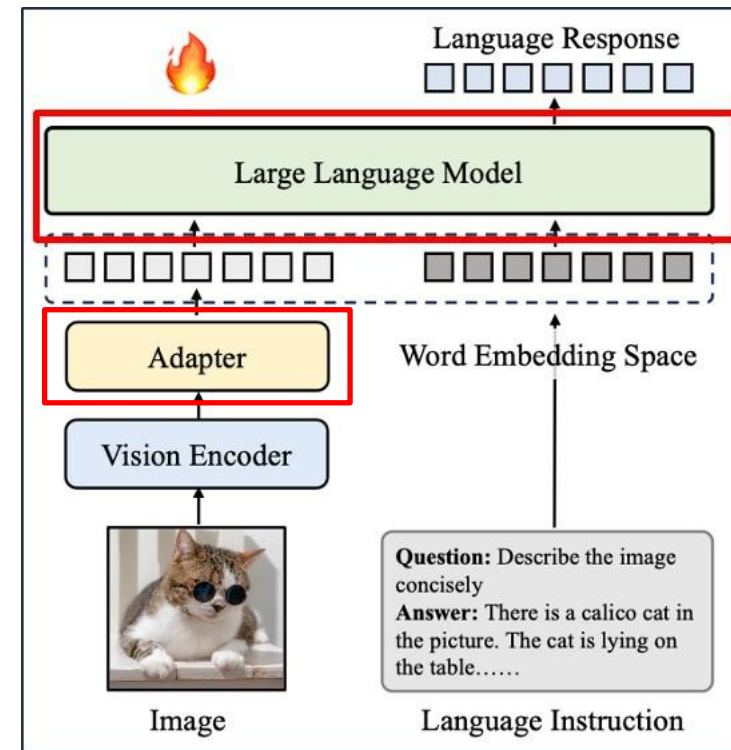
Stage1: Pretraining Stage

- Align different modalities, provide world knowledge




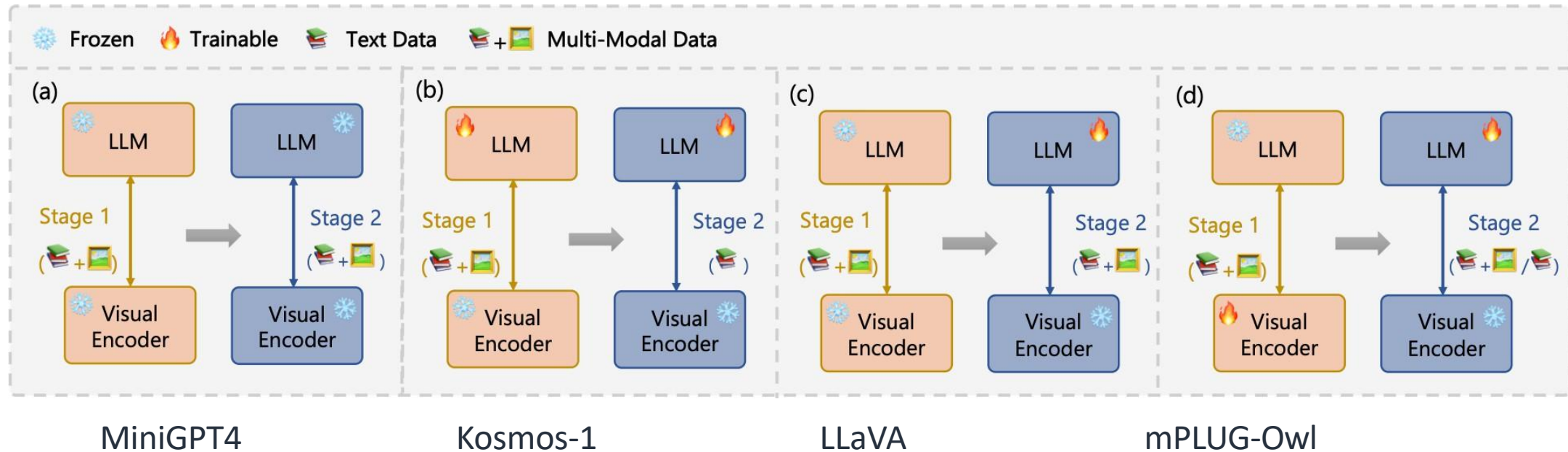
Stage2: Instruction Tuning Stage

- Teach models to better understand the instructions from users and fulfill the demanded tasks.



Training Paradigms

 Training paradigms of popular multimodal large language models.



1 *mPLUG-Owl: Language Models with Multimodality. 2023.*

2 *Visual Instruction Tuning. NeurIPS 2023.*

3 *MINIGPT-4: ENHANCING VISION-LANGUAGE UNDERSTANDING WITH ADVANCED LARGE LANGUAGE MODELS. 2023.*

4 *Language Is Not All You Need: Aligning Perception with Language Models. 2023.*



1.5

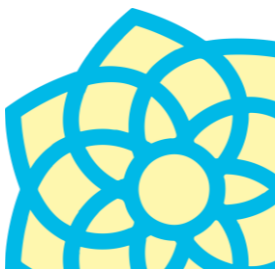
Another Perspective of Multimodal Instruction Tuning



How can we create such multimodal models that follow human's intent?



How can we create such multimodal models that follow human's intent efficiently?



How can we create such **multimodal** models that follow human's intent **efficiently**?



How can we create such **multimodal** models that **follow human's intent** **efficiently**?



How can we make an instruction-following LLM **multimodal** **efficiently**?

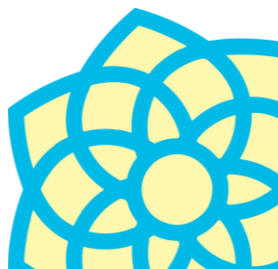
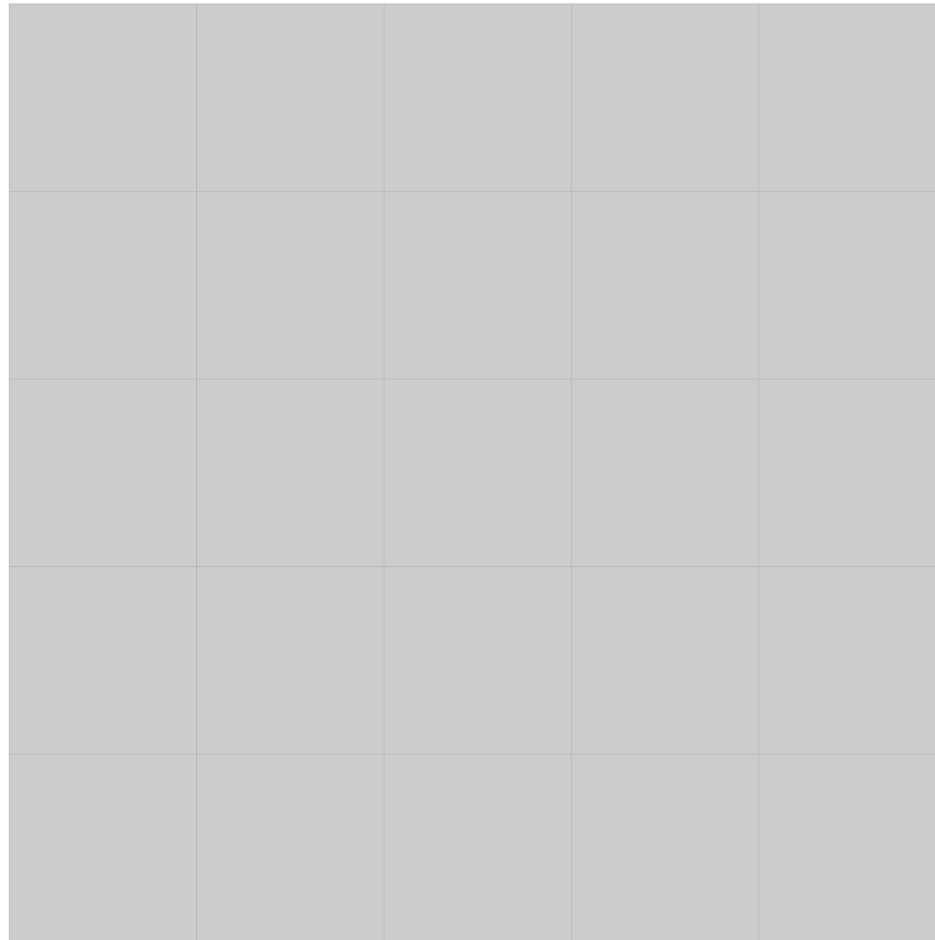


LLM “learns” a foreign language efficiently.

- LLaMA is almost trained on English tokens solely.
- LLaMA learns foreign languages with 52K conversations
 - E.g. Chinese, Japanese, etc.
 - ~1 hour training



Multimodal learning as a translation problem



Multimodal learning as a translation problem



		llama lava		
	llama lava glasses	llama lava glasses	Glasses	
		llama lava	llama lava	
		llama lava	llama lava	llama lava
		llama feet	llama feet	

Q: What's in the image?

A: A llama that's made of lava.

Q: What's special of this image?

A: The llama is wearing glasses.



Multimodal learning as a translation problem

Instruction

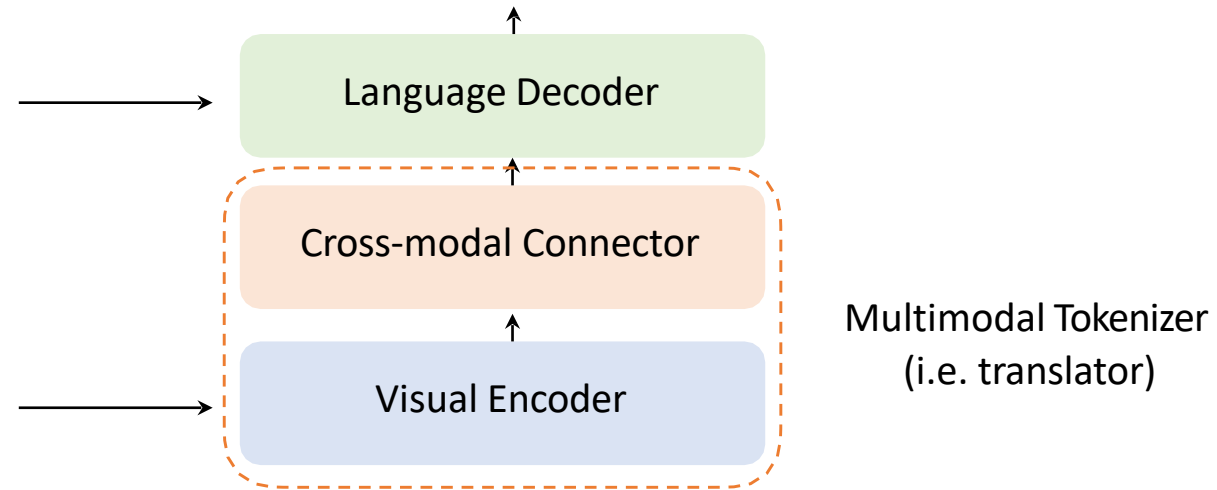
What's the color of the shirt that the man is wearing?

Image



Output

The color of the shirt the man's wearing is yellow.



LLM “learns” a **visual** foreign language efficiently.

Some questions are still hard



		llama lava		
	llama lava glasses	llama lava glasses	Glasses	
		llama lava	llama lava	
		llama lava	llama lava	llama lava
		llama feet	llama feet	

Q: Is the llama facing left or right?

A: Hmm...



Still struggles to follow complex visual instructions

Instruction

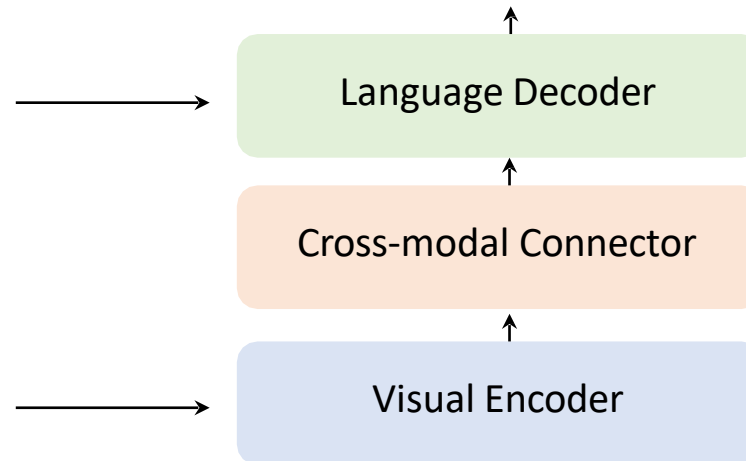
What is unusual about this image?

Image



Output

The unusual aspect of this image is ...



What do we need?

Tuning the model for following multimodal instructions



2

Multimodal Instruction Tuning Data Generation

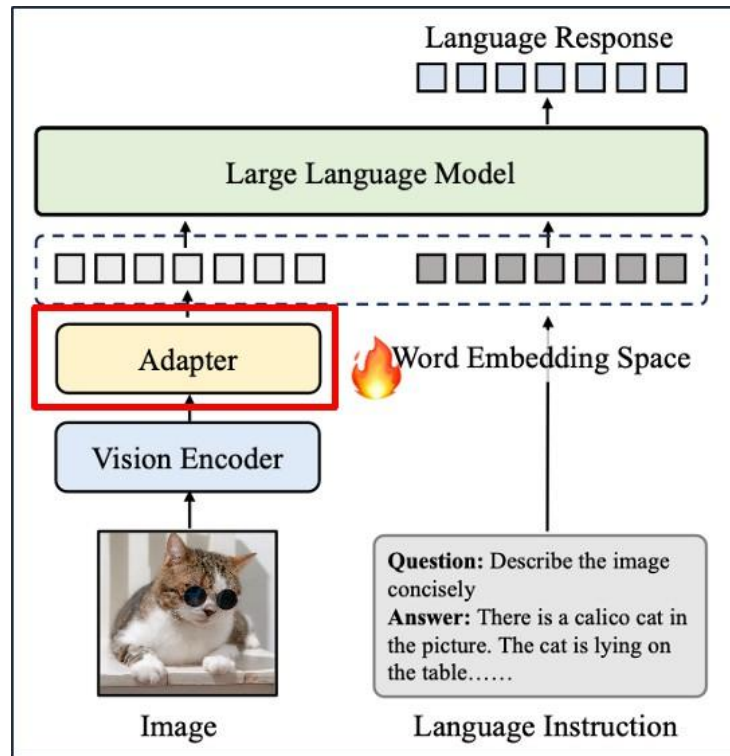


Pretraining Data



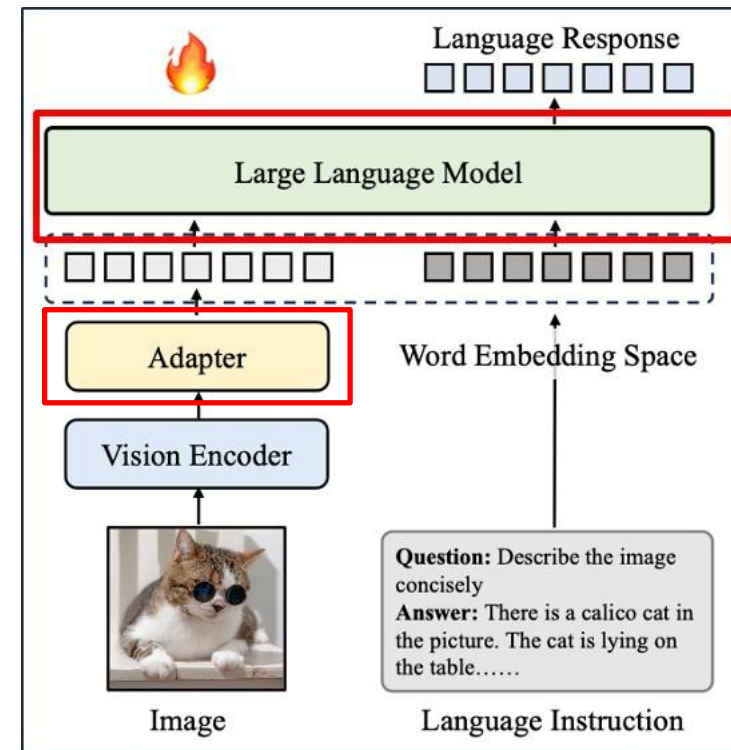
Stage1: Pretraining Stage

- Align different modalities, provide world knowledge



Stage2: Instruction Tuning Stage

- Teach models to better understand the instructions from users and fulfill the demanded tasks.



Pretraining Data (Paired)

❖ Coarse-gained Image-text

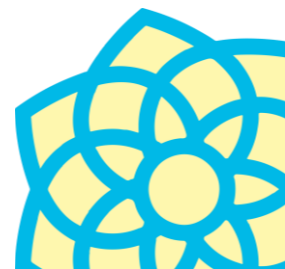
Data volume is **large**, the captions are **shorts** and **noisy**.

❖ Fine-gained Image-Text

High quality, **longer** and **more accurate descriptions**, fine-gained alignment between different modalities.


Dataset	Samples	Taxonomies
CC-3M	3.3M	Coarse-grained Image-Text
CC-12M	12.4M	Coarse-grained Image-Text
LAION-5B	5.9B	Coarse-grained Image-Text
SBU-Captions	1M	Coarse-grained Image-Text
ShareGPT4V-PT	1.2M	Fine-grained Image-Text
LVIS-Instruct4V	111k	Fine-grained Image-Text
<u>ALLaVA</u>	709k	Fine-grained Image-Text
MSR-VTT	200k	Video-Text
<u>WavCaps</u>	24k	Audio-Text
MMC-Instruction	600k	Chart-Text

[1] A Survey on Multimodal Large Language Models. 2024



Pretraining Data

❖ ShareGPT4V



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

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

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

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

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

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

Coarse-gained Image-text

Coarse-gained Image-text

Fine-gained Image-text

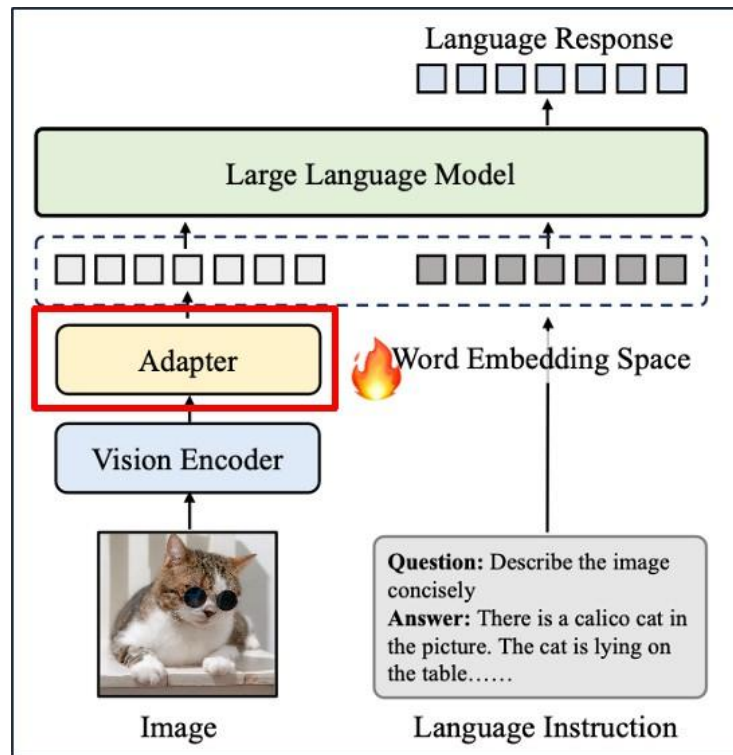


Instruction Data Generation



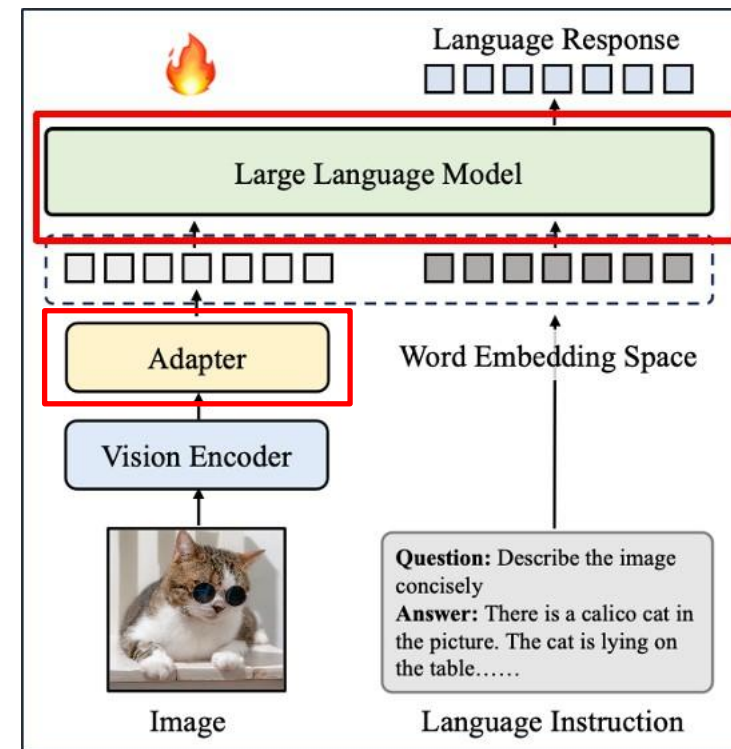
Stage1: Pretraining Stage

- Align different modalities, provide world knowledge



Stage2: Instruction Tuning Stage

- Teach models to better understand the instructions from users and fulfill the demanded tasks.



Instruction Data Generation

Image



Context (caption)

A group of people standing outside of a black vehicle with various luggage.

Context (bbox)




person: [0.68, 0.24, 0.77, 0.69], person: [0.63, 0.22, 0.68, 0.51],
person: [0.44, 0.23, 0.48, 0.34], backpack: [0.38, 0.69, 0.48, 0.91],
....



Instruction Data Generation

❖ Self Instruction

First, Translate images into **dense captions and bounding boxes**. Second, prompt **text-only GPT-4**.

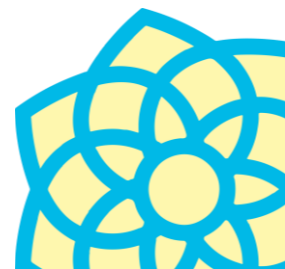
<p>Prompt: Give an image with following information: bounding box, positions that are the object left-top corner coordinates(X, Y), object sizes(Width, Height). Highly overlapping bounding boxes may refer to the same object.</p>	
<p>bounding box: elephant heard on rocks X: 73 Y: 80 Width: 418 Height: 418 woman wearing long dress X: 176 Y: 298 Width: 35 Height: 83 group of green chairs X: 153 Y: 326 Width: 95 Height: 126 an orange bucket on the ground X: 91 Y: 341 Width: 38 Height: 36 a group of white umbrellas X: 99 Y: 82 Width: 112 Height: 28 a man in an orange shirt X: 204 Y: 265 Width: 31 Height: 47 a woman wearing a yellow dress X: 169 Y: 298 Width: 47 Height: 76 ...</p>	
<p>Task: image captioning, Image Sentiment Analysis, Image Quality Assessment, Object Interaction Analysis, Object Attribute Detection, Multi-choice VQA ...</p>	
<p>Come up with 20 diverse instructions for all the tasks above with different language styles and accurate answers. The instructions should contain interrogative sentence and declarative sentences. The answers should be less than 30 words. Each task should have less than 3 instructions.</p>	
<p>GPT4 OUTPUT Example:</p>	
<p>Instruction: Craft a brief narrative about the baby elephant and adult elephant. Answer: A baby elephant is depicted behind an adult elephant, possibly seeking protection.</p>	

Bounding boxes,
dense Captions →

Task Descriptions →

Generation
Requirement →

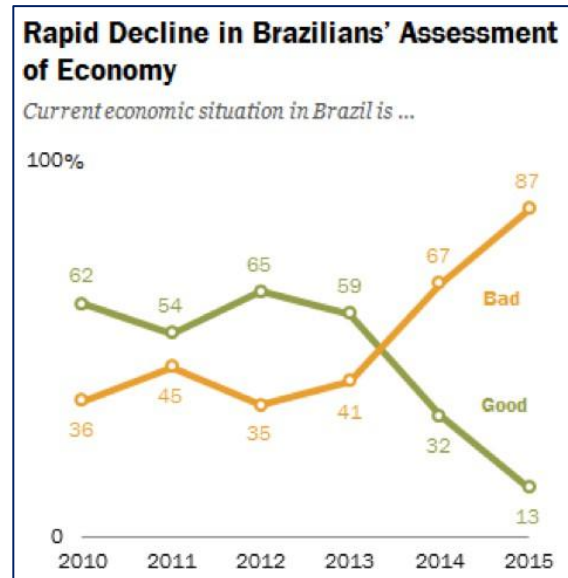
Output from GPT4 →



Instruction Data Generation

❖ Existing Data

The answers of existing VQA and caption datasets are usually **concise**, directly using these datasets for instruction tuning **may limit the output length of MLLMs**.



Question:

Which year has the most divergent opinions about Brazil's economy?

Answer:

2015

[1] ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. ACL 2022 Findings.



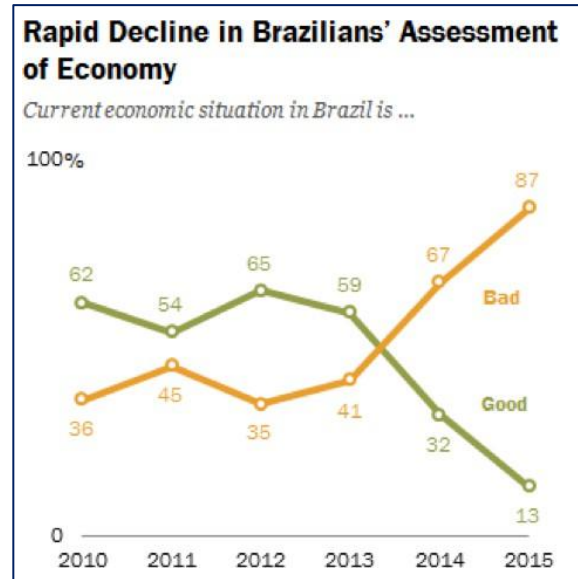
Instruction Data Generation

❖ Existing Data

The answers of existing VQA and caption datasets are usually **concise**, directly using these datasets for instruction tuning **may limit the output length of MLLMs**.

❖ Data Adaptation

Declares **short and brief** for short-answer data in the instruction.



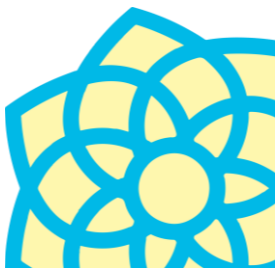
Question:

Which year has the most divergent opinions about Brazil's economy?
A short answer to the question is:

Answer:

2015

[1] ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. ACL 2022 Findings.



Instruction Data Generation



User

Can you tell me what I can cook with these?

Yes, with the variety of food items stored in the refrigerator, you can prepare several. For example, you can create a fruit salad using the strawberries, blueberries, and carrots. Additionally, ...



after

before



Directly training with VQA-v2 data (short answer) with our conversation data. The model refuses to provide natural answers.



Instruction Data Generation

Visual input example, Different Format Prompts:



Normal prompt

What is the color of the shirt that the man is wearing?



Yellow

Response

The man is wearing a yellow shirt.

Ambiguous prompt

Q: What is the color of the shirt that the man is wearing? A:



Yellow

Response

The man is wearing a yellow shirt.



Instruction Data Generation

Visual input example, Different Format Prompts:



Normal prompt

What is the color of the shirt that the man is wearing?

Response

The man is wearing a yellow shirt.

Ambiguous prompt

Q: What is the color of the shirt that the man is wearing? A:

Response

The man is wearing a yellow shirt.

Formatting prompt

What is the color of the shirt that the man is wearing? **Answer the question using a single word or phrase.**

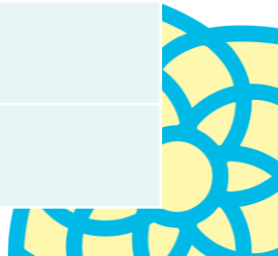
Response

Yellow.



Existing Instruction Tuning Dataset

Dataset	Size	Modalities	Constructions
LLaVA-Instruct-158k	158k	Image, Text	ChatGPT-generated
LRV-Instruction	400k	Image, Text	GPT4-generated
MMC-Instruction	600k	Chart, Text	GPT4-generated/adapted
Clotho-Detail	3.9k	Text, Audio	GPT4-generated
MACAW-LLM	119k	Image, Video, Text	GPT-3.5-turbo-generated
MIMIC-IT	2.8M	Image, Video, Text	ChatGPT-generated
StableLLaVA	126k	Image, Text	StableDiffusion & ChatGPT-generated
LAMM	196k	Image, PointCloud, Text	GPT4-generated
VIGC-LLaVA	1.8M	Image, Text	Model-generated
X-LLM	10k	Image, Video, Text	ChatGPT-generated



Summary

- How we teach multimodal models:

- Pretraining:



- A dictionary to teach LLM to understand (vocabularies from) a new modality

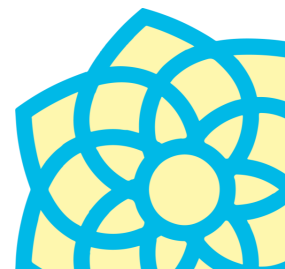
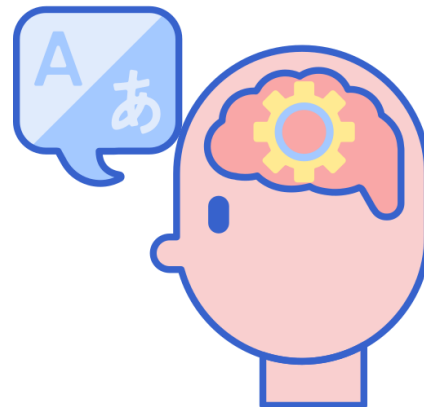


- Instruction tuning (short answer VQA):

- Small puzzles to effectively/efficiently injects new domain knowledge

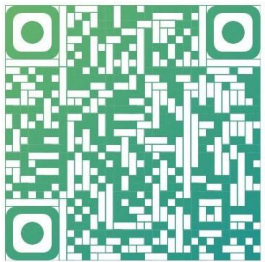


- Instruction tuning (natural conversation VQA):
Real-world applications to practice the skills



* Part-VI

Multimodal Reasoning



Zhuosheng Zhang

Tenure-Track Assistant Professor

Shanghai Jiao Tong University

<https://bcmi.sjtu.edu.cn/~zhangzs/>

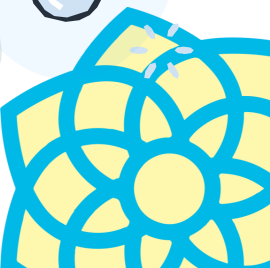
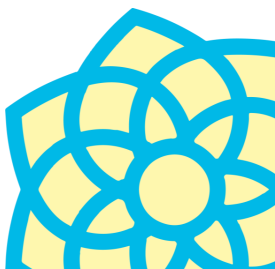


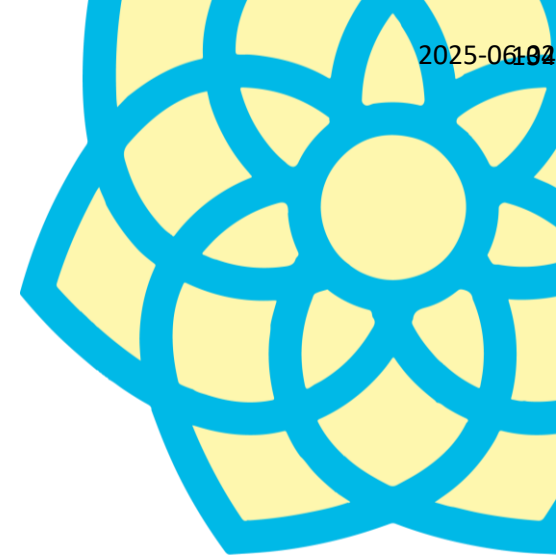
Table of Content

- o Basics of Multimodal Reasoning
 - × Background, Definition, and Development
- o Multimodal Chain-of-Thought Reasoning
 - × Paradigm Shift, the Role of Multimodal CoT
- o Towards Multimodal LLM Agents
 - × Taxonomy, Architecture, Applications
- o Challenges
 - × Evolutionary Reasoning, Interactive Reasoning, Reasoning Alignment

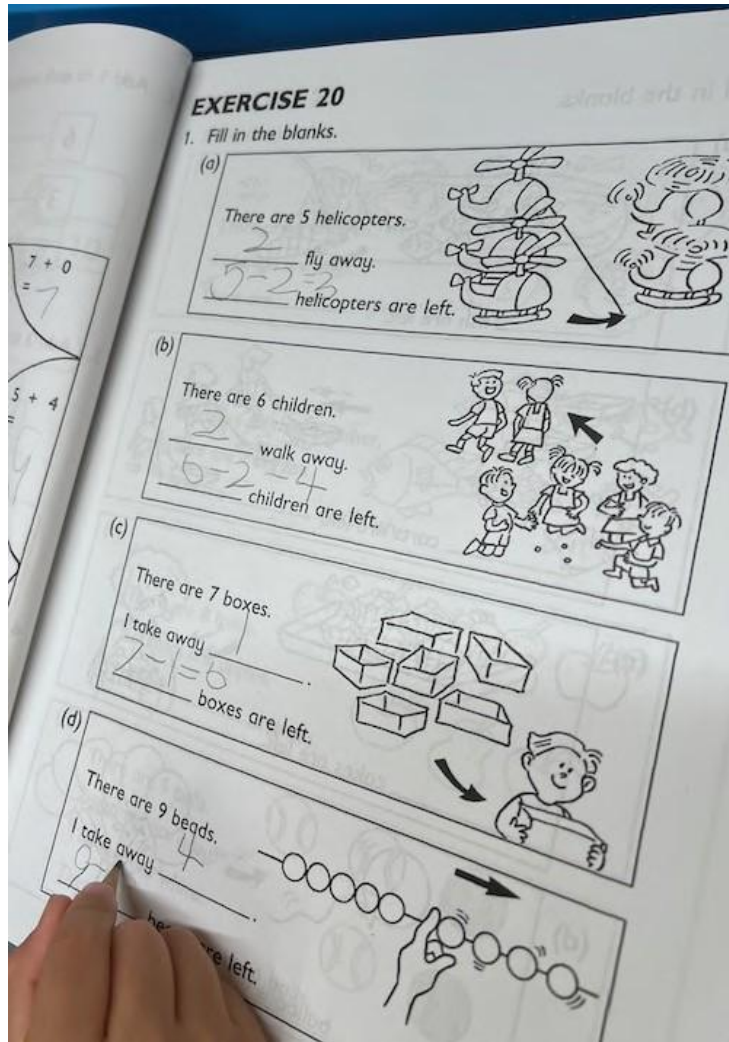


1

Basics of Multimodal Reasoning



Multimodal Reasoning



Prompt

Please create a blog post about a trip to New York, where a dog and his owner had lots of fun. Include a few pictures of the dog posing happily at different landmarks.

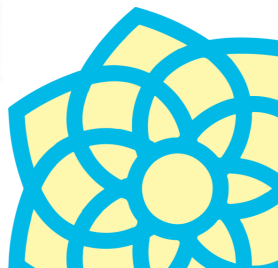
Model Response

Hi, I'm Cooper. I'm a 2 year old Goldendoodle who loves to travel. Last month, my family and I went to New York City. It was my first time in the Big Apple, and I had a blast!

We stayed in a hotel near Times Square. It was so much fun to see all the bright lights and people. We went to see a Broadway show, and I even got to meet some of the actors.

One of my favorite parts of New York City was Central Park. It was so big and beautiful. I loved running around and playing fetch. We also went to the Statue of Liberty and the Empire State Building. I was so excited to see all the famous landmarks.

I had so much fun in New York City. It's definitely one of my favorite places to travel. I can't wait to go back someday!



Multimodal Reasoning

□ **Concept:** a process of deriving high-level conclusions from multiple modalities, possibly via multiple logical steps based on atomic evidences



Accounting : Investment

Question: Consider the three stocks in the following table. Pt represents price at time t, and Qt represents shares outstanding at time t. Stock C splits two for one in the last period. Calculate the rate of return on a price-weighted index of the three stocks for the first period (t0 to t1).

- Option:**
 (A) 4.17%
 (B) 4.03%
 (C) 3.89%

	P ₀	Q ₀	P ₁	Q ₁	P ₂	Q ₂
A	90	100	95	100	95	100
B	50	200	45	200	45	200
C	100	200	110	200	55	400

<image 1>

GPT-4V: To calculate the rate of return on a price-weighted index of the three stocks for the first period (t0 to t1), follow these steps: Calculate the initial index value at t0. Calculate the index value at t1. Compute the rate of return using the two index values.

Initial Index Value at t0:

$$\text{Price-weighted index value at } t_0 = (P_0(A) + P_0(B) + P_0(C)) / 3$$

$$= (90 + 50 + 100) / 3$$

$$= 80$$

Index Value at t1:

$$\text{Price-weighted index value at } t_1 = (P_1(A) + P_1(B) + P_1(C)) / 3$$

$$= (95 + 45 + 110) / 3$$

$$= 83.33$$

Rate of Return:

$$\text{Rate of Return} = (\text{Index Value at } t_1 - \text{Index Value at } t_0) / \text{Index Value at } t_0$$

$$= (83.33 - 80) / 80$$

$$= 0.041625$$

$$= 4.1625\%$$

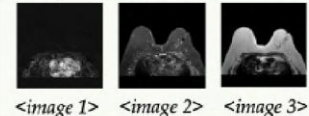


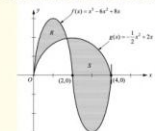
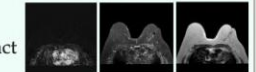

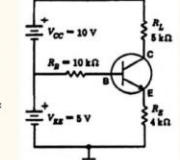
Therefore, the rate of return on a price-weighted index of the three stocks for the first period is approximately 4.17%.

The correct answer is:

(A) 4.17%

Ground Truth: (A) 4.17%

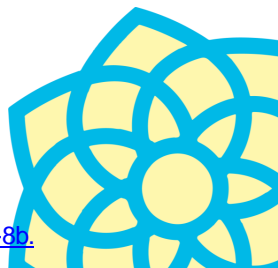
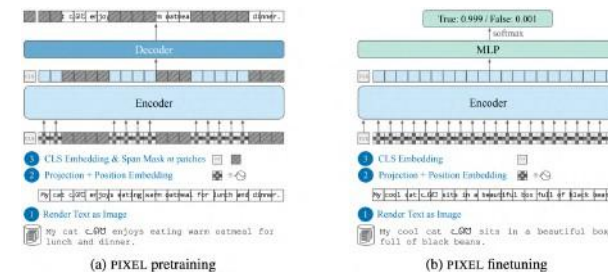
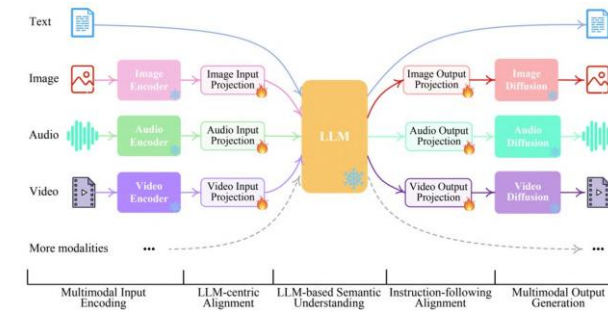
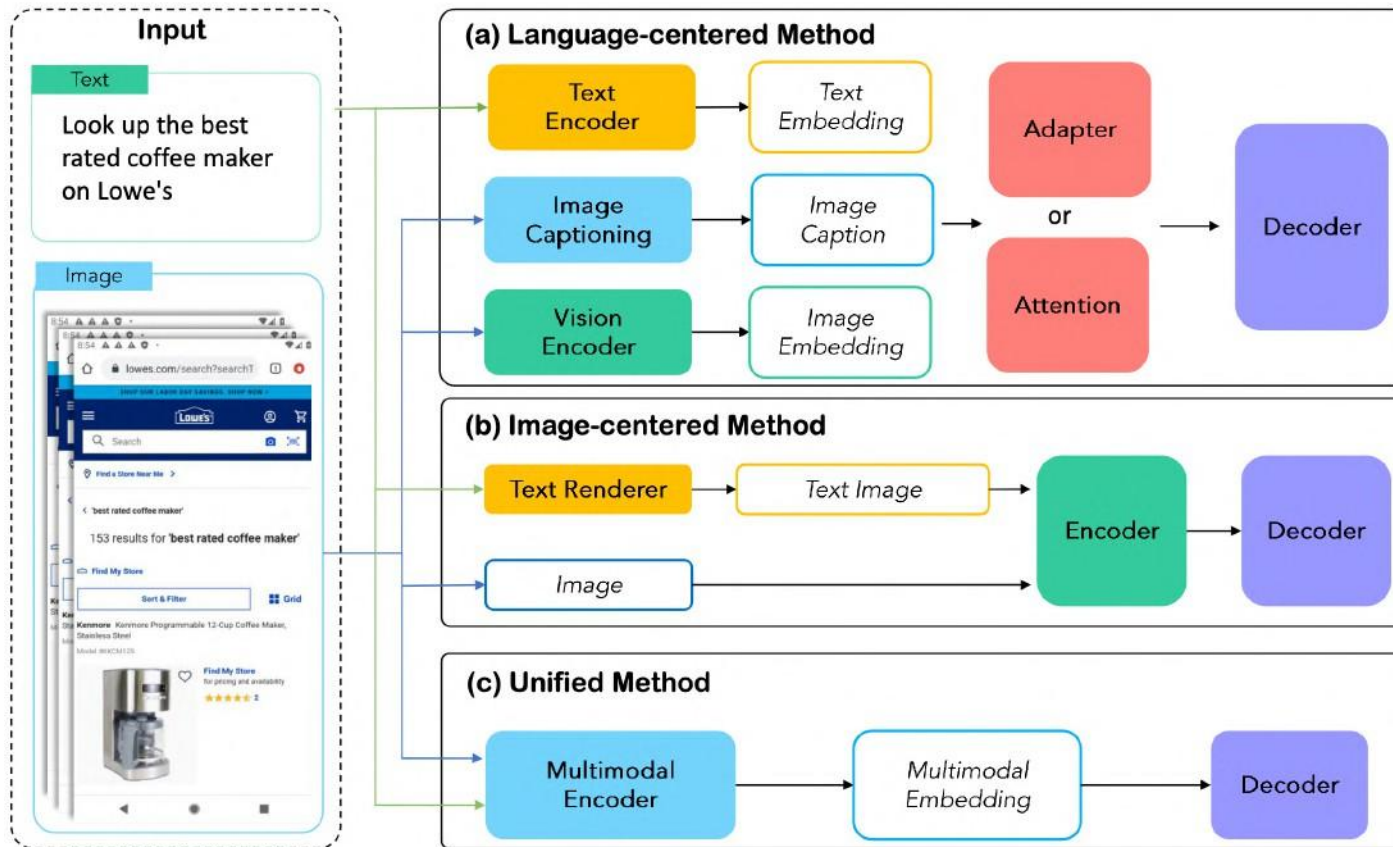
Explanation: At t = 0, the value of the index is: (90 + 50 + 100)/3 = 80. At t = 1, the value of the index is: (95 + 45 + 110)/3 = 83.333. The rate of return is: (83.333/80) - 1 = 4.17%

Comprehensive Disciplines	Heterogeneous Image Types	Interleaved Text and Images	Expert-level Skills Test
<p>Engineering (26%) Art & Design (11%) Business (14%) Science (23%) Humanities & Social Sci. (9%) Medicine (17%)</p>	<p>Diagrams, Tables, Plots and Charts, Photographs, Chemical Structures, Paintings, Medical Images, Sheet Music, Geometric, Pathology images, Microscopic Images, Comics, ...</p>	<p>Question: You are shown subtraction <image 1>, T2 weighted <image 2> and T1 weighted axial <image 3> from a screening breast MRI. What is the etiology of the finding in the left breast?</p> 	<p>Expert-level Visual Perception</p> <p>Perception</p> <p>Knowledge → Reasoning</p> <p>Domain Expertise, World, Linguistic, Visual Knowledge, ... Logical, Spatial Commonsense, Mathematical, ...</p>
<p>Art & Design</p> <p>Question: Among the following harmonic intervals, which one is constructed incorrectly?</p> <p>Options:</p> <p>(A) Major third <image 1></p> <p>(B) Diminished fifth <image 2></p> <p>(C) Minor seventh <image 3></p> <p>(D) Diminished sixth <image 4></p> 	<p>Business</p> <p>Question: ...The graph shown is compiled from data collected by Gallup <image 1>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?</p> <p>Options:</p> <p>(A) 0 (B) 0.2142</p> <p>(C) 0.3571 (D) 0.5</p> 	<p>Science</p> <p>Question: <image 1> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.</p> <p>Options:</p> <p>(A) $\int_0^{1.5} [f(x) - g(x)] dx$</p> <p>(B) $\int_0^{1.5} [g(x) - f(x)] dx$</p> <p>(C) $\int_0^2 [f(x) - g(x)] dx$</p> <p>(D) $\int_0^2 [g(x) - x(x)] dx$</p> 	
<p>Health & Medicine</p> <p>Question: You are shown subtraction <image 1>, T2 weighted <image 2> and T1 weighted axial <image 3> from a screening breast MRI. What is the etiology of the finding in the left breast?</p> <p>Options:</p> <p>(A) Susceptibility artifact</p> <p>(B) Hematoma</p> <p>(C) Fat necrosis (D) Silicone granuloma</p> 	<p>Humanities & Social Science</p> <p>Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <image 1></p> <p>Option:</p> <p>(A) Oppressor</p> <p>(B) Imperialist</p> <p>(C) Savior (D) Isolationist</p> 	<p>Tech & Engineering</p> <p>Question: Find the VCE for the circuit shown in <image 1>. Neglect VBE</p> <p>Answer: 3.75</p> <p>Explanation: ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V</p> 	
<p>Subject: Music; Subfield: Music; Image Type: Sheet Music; Difficulty: Medium</p>	<p>Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts; Difficulty: Medium</p>	<p>Subject: Math; Subfield: Calculus; Image Type: Mathematical Notations; Difficulty: Easy</p>	
<p>Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans: MRI, CT; Difficulty: Hard</p>	<p>Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy</p>	<p>Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard</p>	

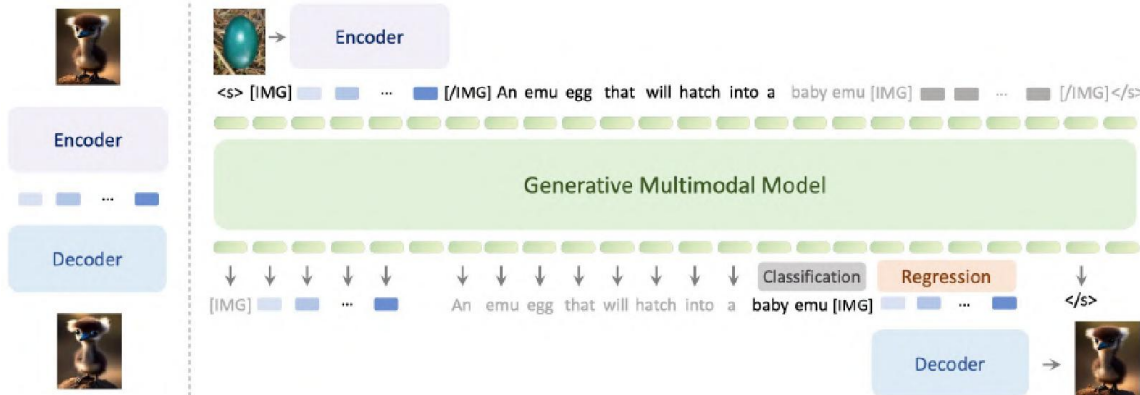
Model Architecture

Three architectures:

(a) language-centered method; (b) image-centered method; (c) unified method



In-Context Learning



Each image in the multimodal sequence is tokenized into embeddings via a visual encoder, and then **interleaved** with text tokens for autoregressive modeling.



Leveraging **few-shot Prompting** for diverse reasoning tasks

	Input Prompt						Completion
In-context Completion	[dog: 1, frisbee: 1].	[burger: 1, glass: 1, bottle: 1].	[cat: 3].	[beer: 3, banana: 2].			[beer: 3, banana: 2].
	The text in the red circle: 'Rights'.	The text in the red circle: 'Ave'.	The text in the red circle: 'Do Not'.	The text in the red circle: 'Lynn'.			The text in the red circle: 'Lynn'.
	motorcycle's wheel.	woman's feet.	car's license plate.	motorcycle's headlight.			motorcycle's headlight.
	a photo of a yellow backpack:	a photo of a blue backpack:	a photo of a red backpack:	a photo of a brown backpack:	a photo of a blue and red backpack:		
The subject A with a city in the background:	The subject A wearing a santa hat:	The subject A in a purple wizard outfit:	The subject A in a rainbow hat:				

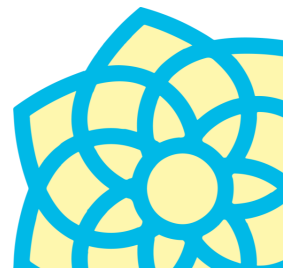
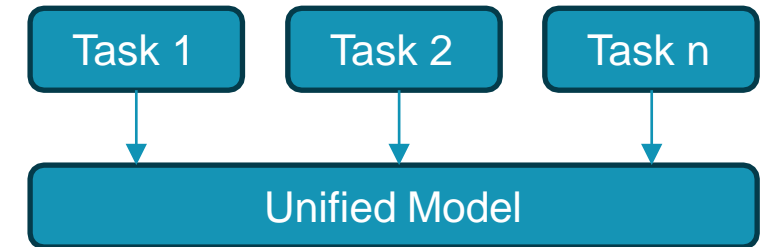
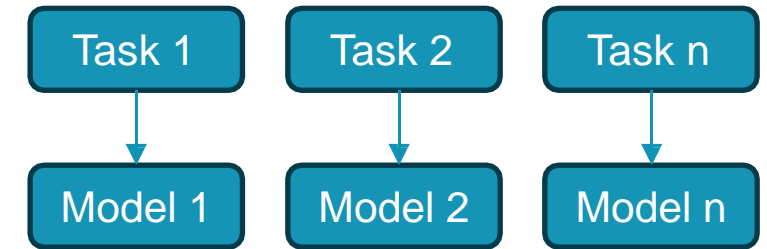
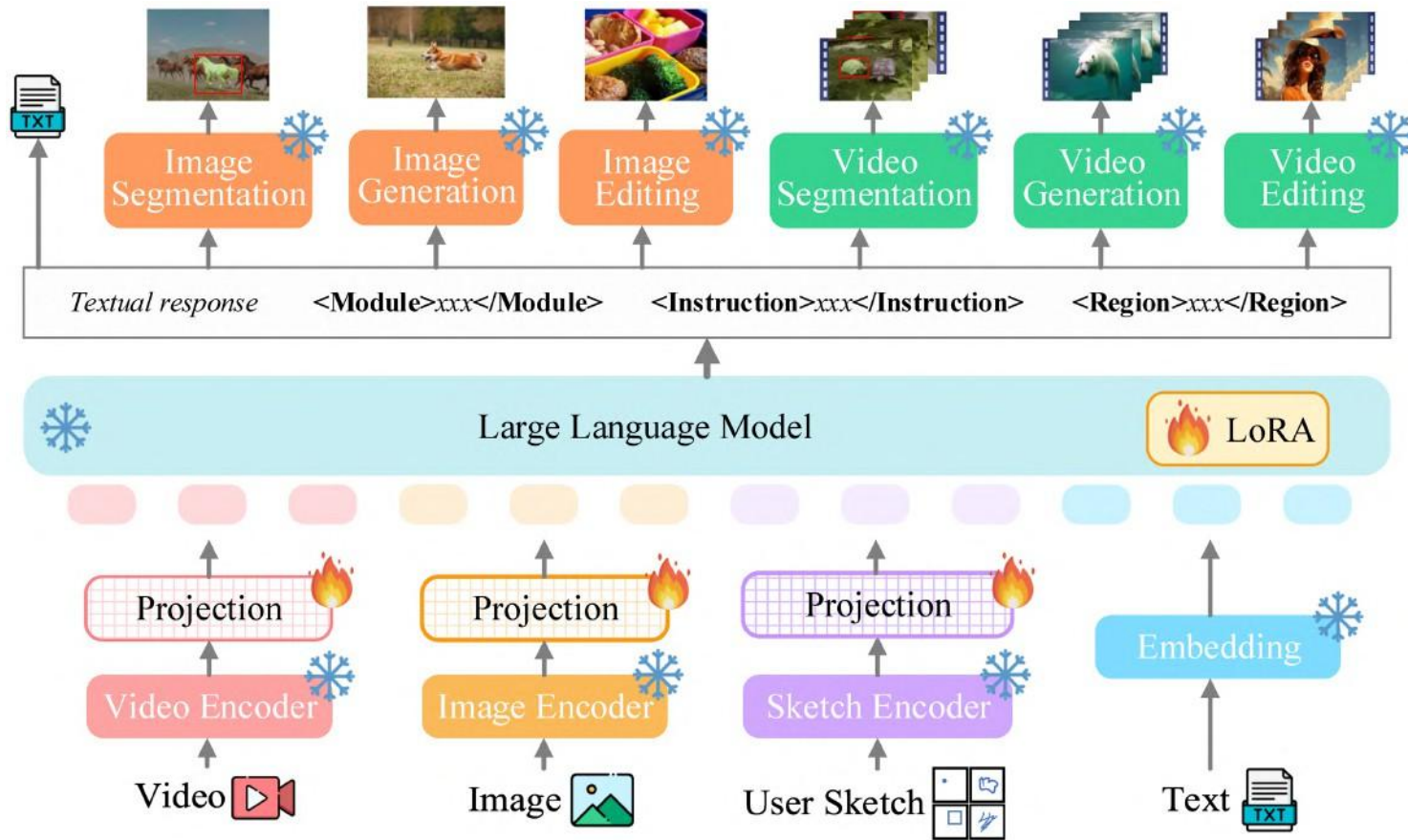
Multimodal understanding	Reasoning	Visual Analysis
Look at this sequence of three shapes. What shape should come as the fourth shape? Explain your reasoning with detailed descriptions of the first shapes.	The first shape is a triangle, the second shape is a square, and the third shape is a pentagon. To find the next shape in the sequence, we need to identify a pattern. In this case, the pattern is that each shape has one more side than the previous shape. Therefore, the next shape in the sequence should have six sides, which is a hexagon .	The funny aspect of this image is that a squirrel is holding a camera and appears to be taking a picture of a bird . This is amusing because animals, such as squirrels and birds, do not typically use cameras or engage in human-like activities like photography. The scene is set up in a way that makes it look like the squirrel is intentionally capturing the bird's image, which adds a humorous twist to the scene. Additionally, the bird is perched on a mushroom , further enhancing the whimsical and playful nature of the image.

Generate from any prompt sequence	Image 1	Prompt	Image 2	Image 3
		impressionist painting of an astronaut in a jungle		An image of <emu> wearing a big sunglasses on the beach
		<a man> is riding <a big horse> on the <water> in front of <mountain>		An anime draw of <a baby emu> and <a dog> , in the style of
		<a dog> and <a dog> running on the city street		



Evolution of Multimodal Reasoning

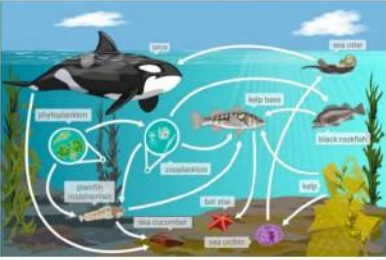
From task-specific to centralized paradigms



Evolution of Multimodal Reasoning

From (implicit) single-step prediction to (explicit) multi-step reasoning

Vision



Text

Question: Which of these organisms contains matter that was once part of the phytoplankton?

Context: Below is a food web from an ocean ecosystem in Monterey Bay, off the coast of California. A food web models how the matter eaten by organisms moves through an ecosystem. The arrows in a food web represent how matter moves between organisms in an ecosystem.

Options: (A) black rockfish (B) sea otter

Rationale


A food web is a model. A food web shows where organisms in an ecosystem get their food. Models can make things in nature easier to understand because models can represent complex things in a simpler way. If a food web showed every organism in an ecosystem, the food web would be hard to understand. So, each food web shows how some organisms in an ecosystem can get their food. Arrows show how matter moves. A food web has arrows that point from one organism to another. Each arrow shows the direction that matter moves when one organism eats another organism...

Answer

The answer is (A).

(a) An example of ScienceQA.

Vision



Text

Question: What should the title of this image be?

Rationale

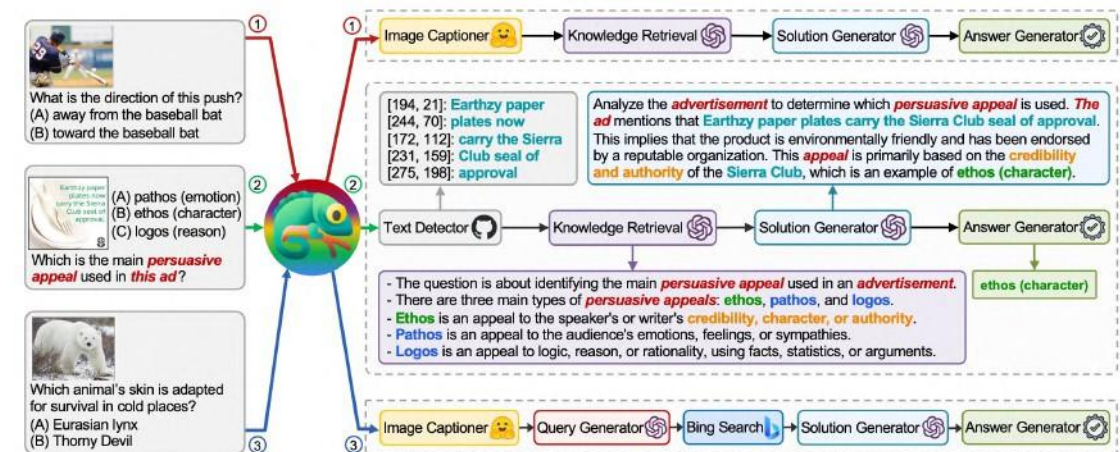
This image shows a woman and a dog sitting on the beach playing a game. The woman is sitting on the ground with her back to the camera, while the dog is sitting next to her, looking up at her with a curious expression. The sun is setting in the background, casting a warm, golden light on the scene. Based on this description, it is likely that the caption for this image is meant to highlight the bond and playfulness between the woman and her dog. The fact that they are playing a game together on the beach suggests that they have a close relationship and enjoy spending time together...

Answer

A woman and a puppy playing a game on the beach.

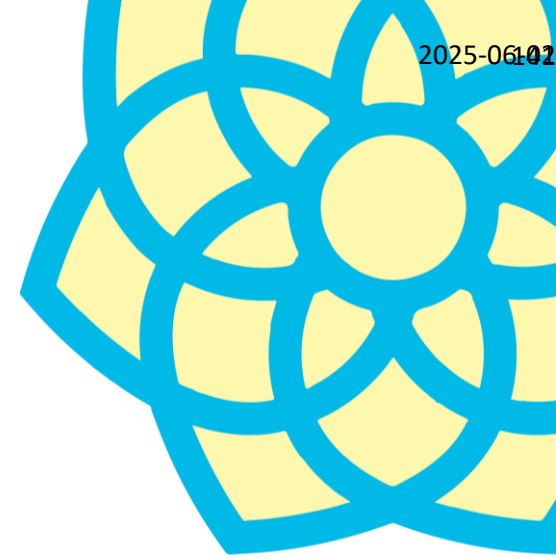
(b) An example of CoCo-MMRD.

- ❑ **Improved Interpretability:** offer an interpretable glimpse into the decision-making process
- ❑ **Improved Controllability:** interfere the reasoning process, e.g., adding complementary information, verifying and correcting mistakes
- ❑ **Improved Flexibility:** allow interactive communications between different models



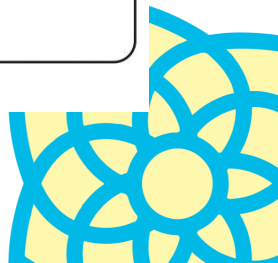
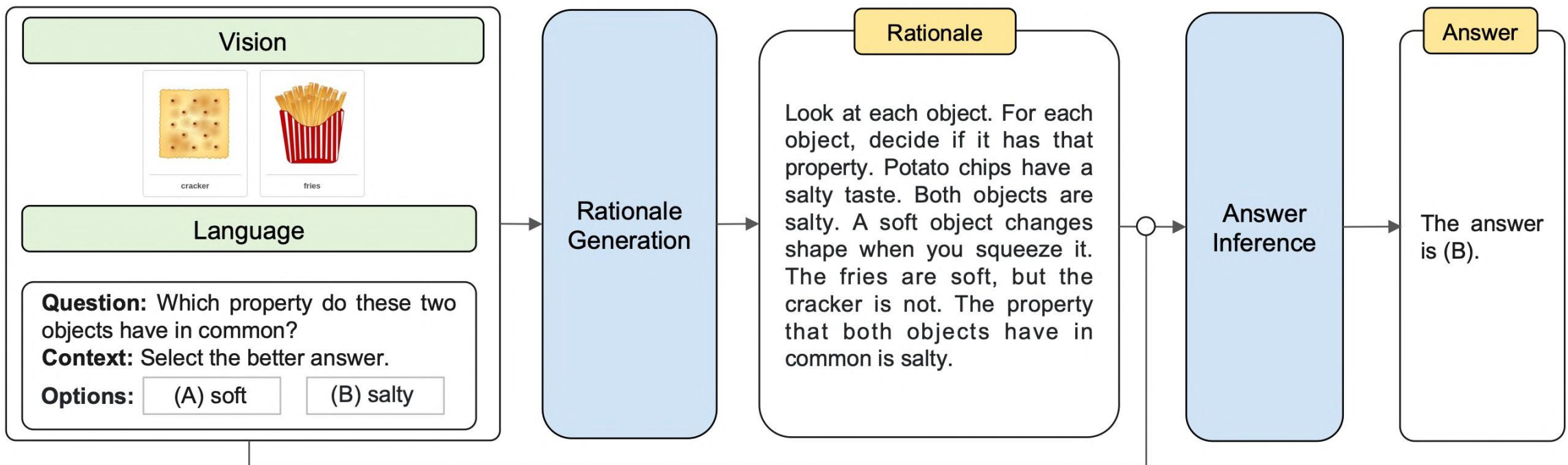
2

Multimodal Chain-of- Thought Reasoning



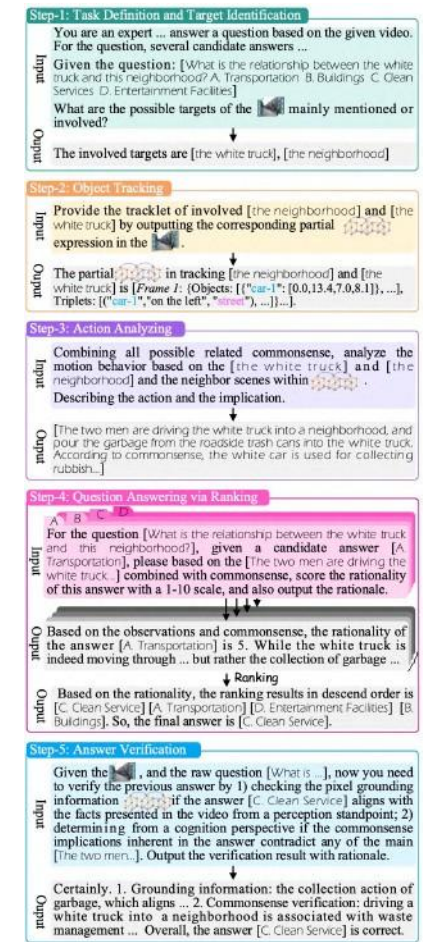
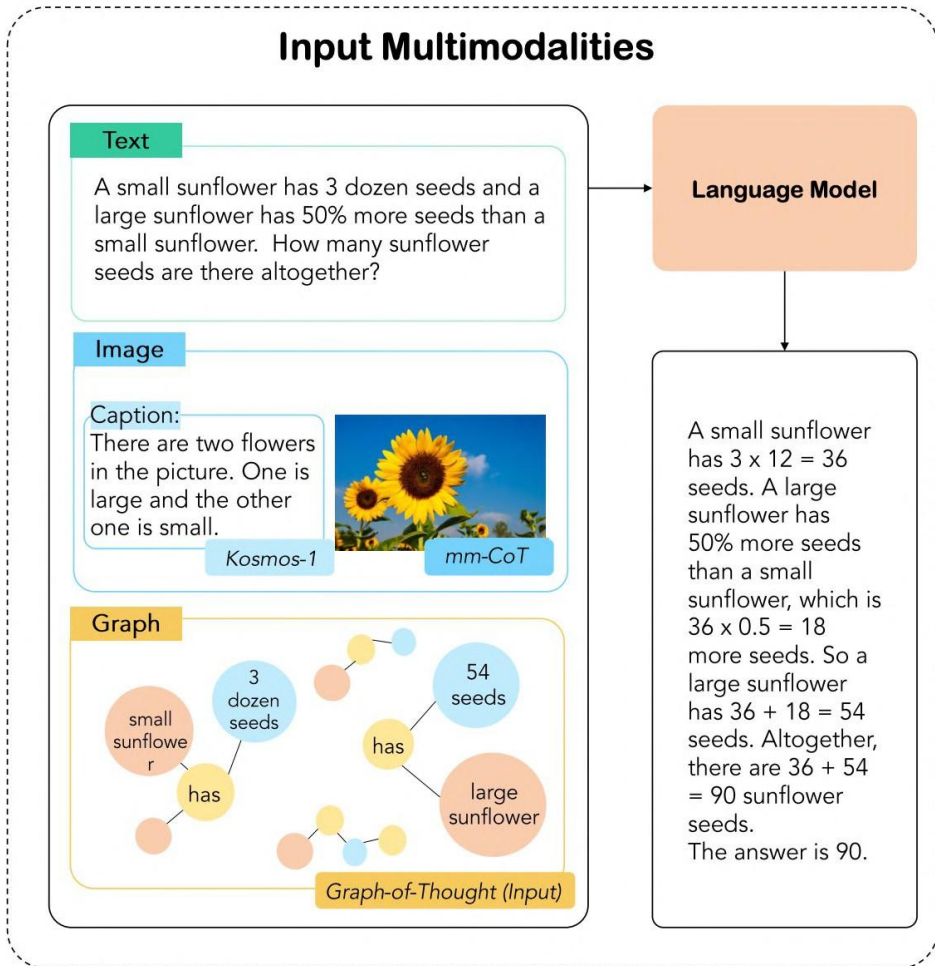
Multimodal Chain-of-Thought Reasoning

- ❑ Think **step by step**, formulate intermediate steps before deriving an answer
- ❑ Paradigm shift of task format
 - Standard Format: <input → output>
 - CoT Format: <input → rationale → output>



Multimodal Chain-of-Thought Reasoning

- Input: Various modalities such as text, image, and graph are incorporated into the model's input
- Output: Multimodalities, including text and image, are generated in the model's output



Video-of-Thought (VoT)



The Role of (Multimodal) Chain-of-Thought

- ❑ Role 1: Introducing more reliable input results in more **convincing reasoning process**
- ❑ **Case studies: 50 error cases**
 - **Imperfect training data: when the vision input is missing**
 - **Generate hallucinated rationales that mislead the answer inference (64%)**

Problem

Question: Will these magnets attract or repel each other?

Context: Two magnets are placed as shown. Hint: Magnets that attract pull together. Magnets that repel push apart.

Options: (A) attract (B) repel

Gold Rationale: Magnets can pull or push on each other without touching. When magnets attract, they pull together. When magnets repel, they push apart. Whether a magnet attracts or repels other magnets depends on the positions of its poles, or ends. Every magnet has two poles, called north and south. Here are some examples of magnets. The north pole of each magnet is marked N, and the south pole is marked S. If different poles are closest to each other, the magnets attract. The magnets in the pair below attract. If the same poles are closest to each other, the magnets repel. The magnets in both pairs below repel. Will these magnets attract or repel? To find out, look at which poles are closest to each other. The north pole of one magnet is closest to the south pole of the other magnet. Poles that are different attract. So, these magnets will attract each other.

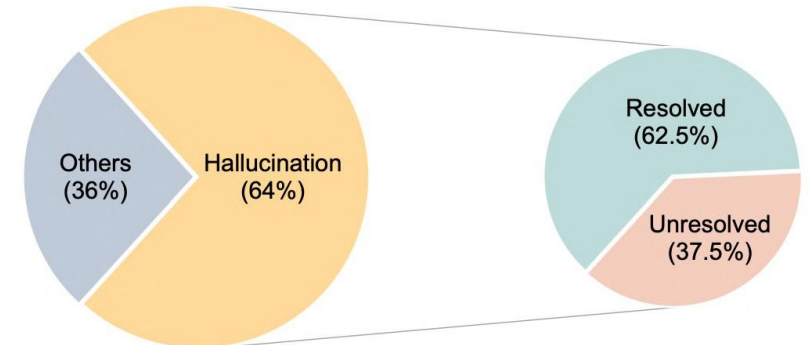
Answer: The answer is (A).



Baseline

Generated Rationale: Magnets can pull or push on each other without touching. When magnets attract, they pull together. When magnets repel, they push apart. Whether a magnet attracts or repels other magnets depends on the positions of its poles, or ends. Every magnet has two poles, called north and south. Here are some examples of magnets. The north pole of each magnet is marked N, and the south pole is marked S. If different poles are closest to each other, the magnets attract. The magnets in the pair below attract. If the same poles are closest to each other, the magnets repel. The magnets in both pairs below repel. Will these magnets attract or repel? To find out, look at which poles are closest to each other. The south pole of one magnet is closest to the south pole of the other magnet. Poles that are the same repel. So, these magnets will repel each other.

Answer: The answer is (B).

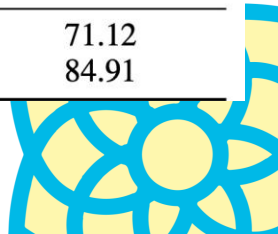


(a) ratio of hallucination mistakes

(b) correction rate w/ vision features

Table 3. Two-stage setting of (i) rationale generation (RougeL) and (ii) answer inference (Accuracy).

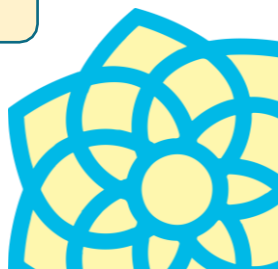
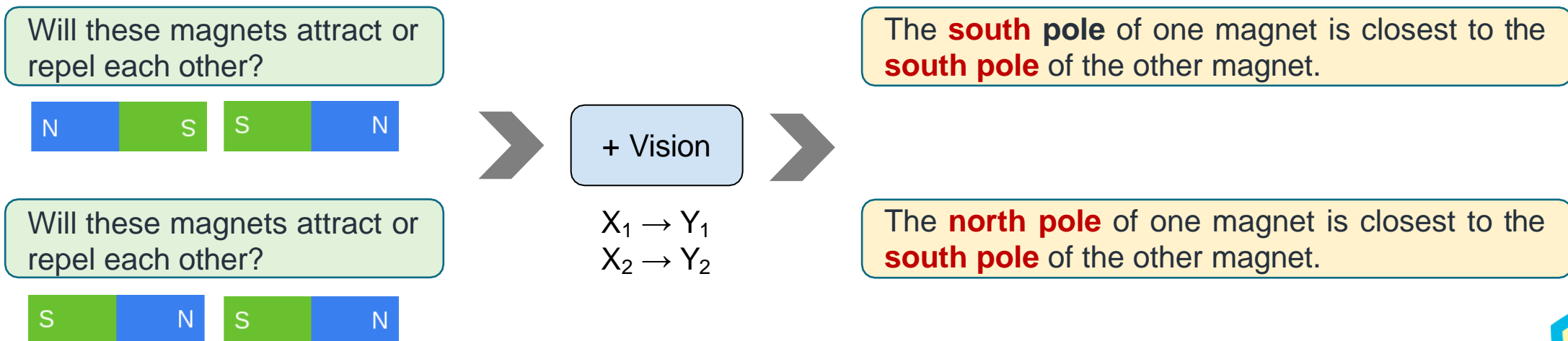
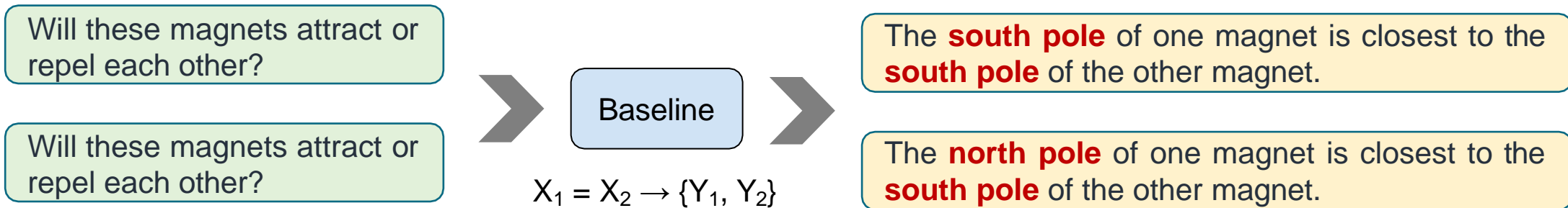
Method	(i) QCM → R	(ii) QCMR → A
Two-Stage Framework	91.76	70.53
w/ Captions	91.85	71.12
w/ Vision Features	96.97	84.91



The Role of (Multimodal) Chain-of-Thought

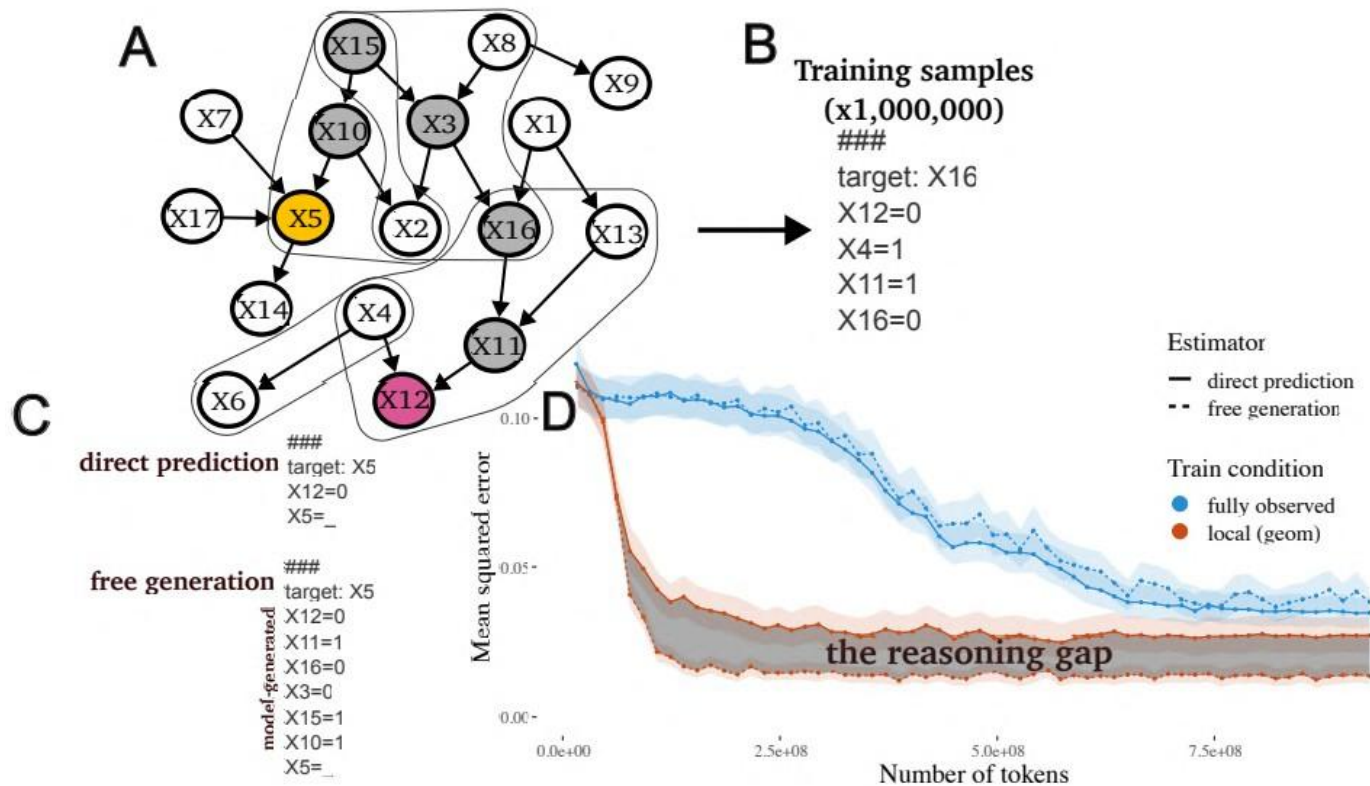
□ Role 1: Introducing more reliable input results in more **convincing reasoning process**

- **More accurate perception, less hallucinations during the reasoning process**



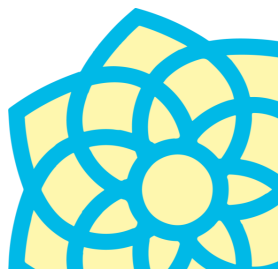
The Role of (Multimodal) Chain-of-Thought

- **Role 2: Breaking complex problems into smaller, manageable sub-problems**
 - identifies atomic knowledge components essential for reasoning processes
 - integrates their relationships, thereby constructing coherent reasoning steps



CoT reasoning is useful because

- 1) direct prediction is inaccurate for some inferences because the relevant variables are rarely seen together in training
- 2) chain-of-thought reasoning improves estimation by incrementally chaining local statistical dependencies that are observed frequently in training



The Role of (Multimodal) Chain-of-Thought

Role 3: Available for stepwise **knowledge update** and **self-correction** (w/ external feedback)

Question Answering

Question: Who won the 2016 Russian national silver medal with another Russian ice dancer born 29 April 1995?
Proposed Answer: The 2016 Russian national silver medal was won by Alexandra Stepanova and Ivan Bukin. Alexandra Stepanova was born 29 April 1995. Ivan Bukin was born 10 April 1993. So the answer is: Ivan Bukin. X

What's the problem with the above answer?
1. Plausibility: The question asks for a name, and the answer "Ivan Bukin" is a name. So it's plausible. Google

2. Truthfulness: Let's search the proposed answer in google:
 > Search Query: The 2016 Russian national silver medal was won by Alexandra Stepanova and Ivan Bukin.
 > Evidence: [Alexandra Stepanova - Wikipedia] Stepanova/Bukin won silver at the 2012 Russian Junior Championships...
 From the evidence, Stepanova/Bukin won silver at the 2012 Russian Junior Championships, **not 2016**. Globe

Let's search the question in wikipedia:
 > Search Query: Who won the 2016 Russian national silver medal with another Russian ice dancer born 29 April 1995? site: wikipedia.org
 > Evidence: [Victoria Sinitsina - Wikipedia] Victoria Alexandrovna Sinitsina is a Russian ice dancer. Globe

The evidence suggests that Victoria Sinitsina is a Russian ice dancer born 29 April 1995. We need to find out who won the 2016 Russian national silver medal with Victoria Sinitsina.
Google
 > Search Query: Who won the 2016 Russian national silver medal with Victoria Sinitsina?
 > Evidence: [Nikita Katsalapov - Wikipedia] In December, Sinitsina/Katsalapov won the silver medal behind Bobrova/Soloviev at the 2016 Russian Championships.

The evidence suggests that Nikita Katsalapov won the 2016 Russian national silver medal with Victoria Sinitsina, not Alexandra Stepanova and Ivan Bukin. The answer is **incorrect**.

Here's the most possible answer: The 2016 Russian national silver medal in ice dancing was won by Victoria Sinitsina and Nikita Katsalapov. Victoria Sinitsina was born on April 29, 1995. So the answer is: Nikita Katsalapov. ✓

Program Synthesis

Question: Ann, Bill, Cate, and Dale each buy personal pan pizzas cut into 4 pieces. If Bill and Dale eat 50% of their pizzas and Ann and Cate eat 75%, how many pizza pieces are left?
Proposed Answer:

```
python
num_pieces_per_pizza, num_pizza = 4, 4
num_pieces = num_pieces_per_pizza * num_pizza
pieces_bill_and_dale = num_pieces * 0.5
pieces_ann_and_cate = num_pieces * 0.75
answer = num_pieces - pieces_bill_and_dale - pieces_ann_cate
```

X
Execution: No runtime error
Output: answer = -4.0 Python

What's the problem with the above code?
1. Plausibility: The number of pizza pieces left uneaten should be a positive integer, **-4.0 < 0**, so the answer is **not reasonable**.

2. Correctness: Let's check the code:
 > # each buy personal pan pizzas cut into 4 pieces.
 > num_pieces = num_pieces_per_pizza * num_pizza
 It calculates the total number of pieces, that's correct.

> pieces_bill_dale = num_pieces * 0.5
 > pieces_ann_cate = num_pieces * 0.75
 Each person only eats their own personal pan pizza, `num_pieces * 0.5` means 50% of the total pieces, this is **wrong**.

Here's a better solution:

```
python
pizza_pieces = 4 # each personal pizza cut into 4 pieces
ann_cate_pieces = 4 * 0.75 * 2 # Ann, Cate eats 75%
bill_dale_pieces = 4 * 0.5 * 2 # Bill, Dale eats 50%
total_eaten = ann_cate_pieces + bill_dale_pieces
answer = pizza_pieces * 4 - total_eaten # pieces left
```

✓

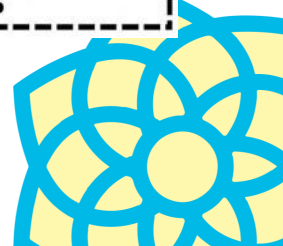
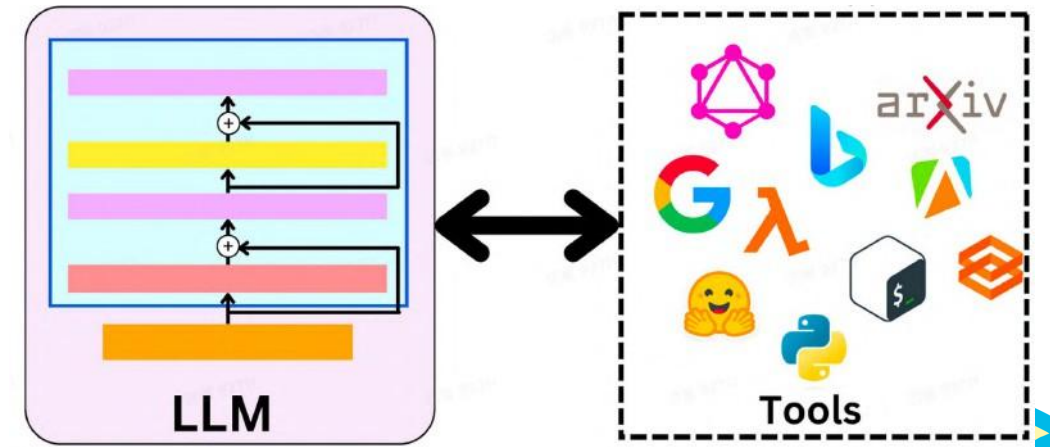
Toxicity Reduction

Previous toxic text: ... last year before hanging himself. X
Perspective

The text has 53% toxicity of threat. To improve content safety while ensuring fluency, here's a **non-toxic continuation** after '\':

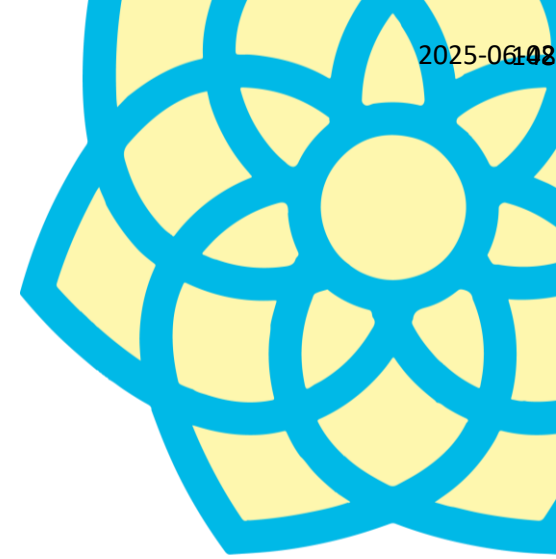
Jason Conroy, 19, strangled fellow resident Melissa Mathieson at Alexandra House in Bristol \ in November 11 2014. It serves as a reminder of the importance of mental health support to... ✓

- Chain-of-Thought w/ Tools
- Plan-Solve, augment and correct intermediate steps with tools
- Overcome the Intrinsic ability deficiency of LLMs such as calculation, searching



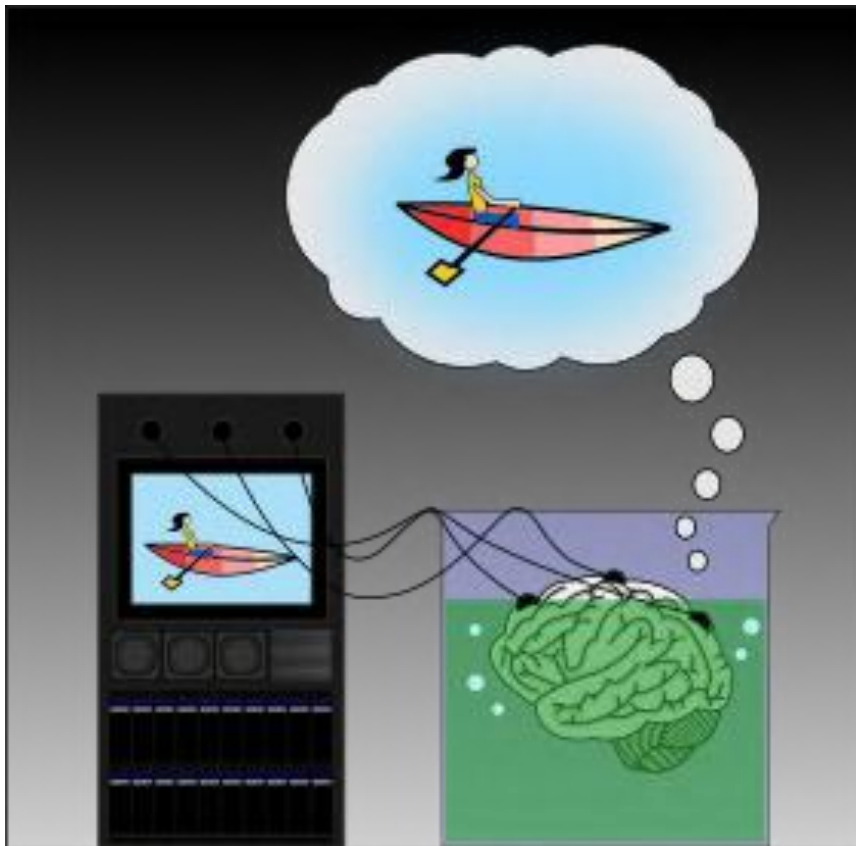
3

Towards Multimodal LLM Agents

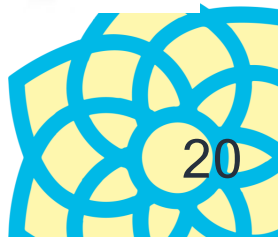
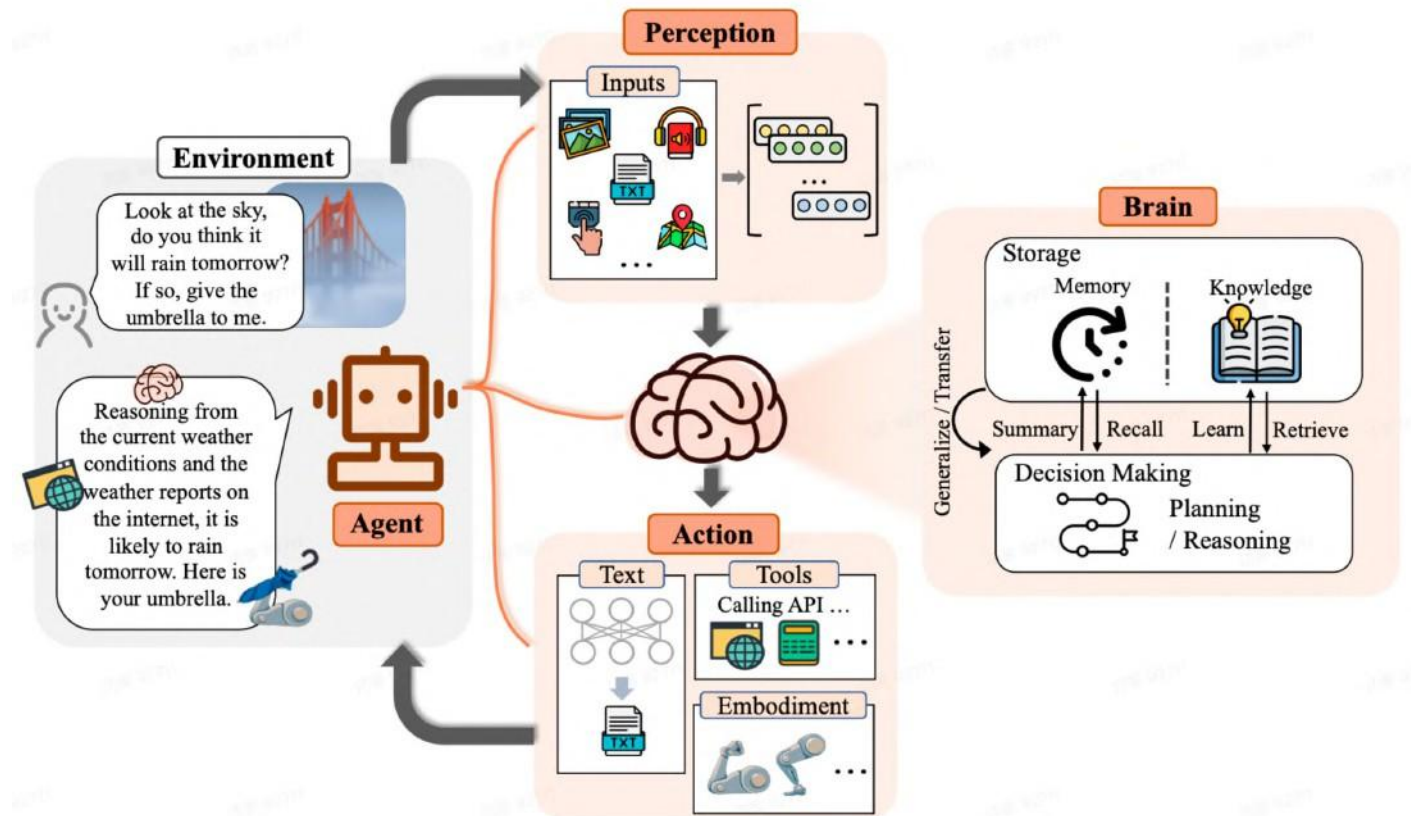


Towards Multimodal LLM Agents

- ❑ From content-based reasoning to behavior control (w/ multimodalities)
- ❑ “Those who know but do not act simply do not yet know”

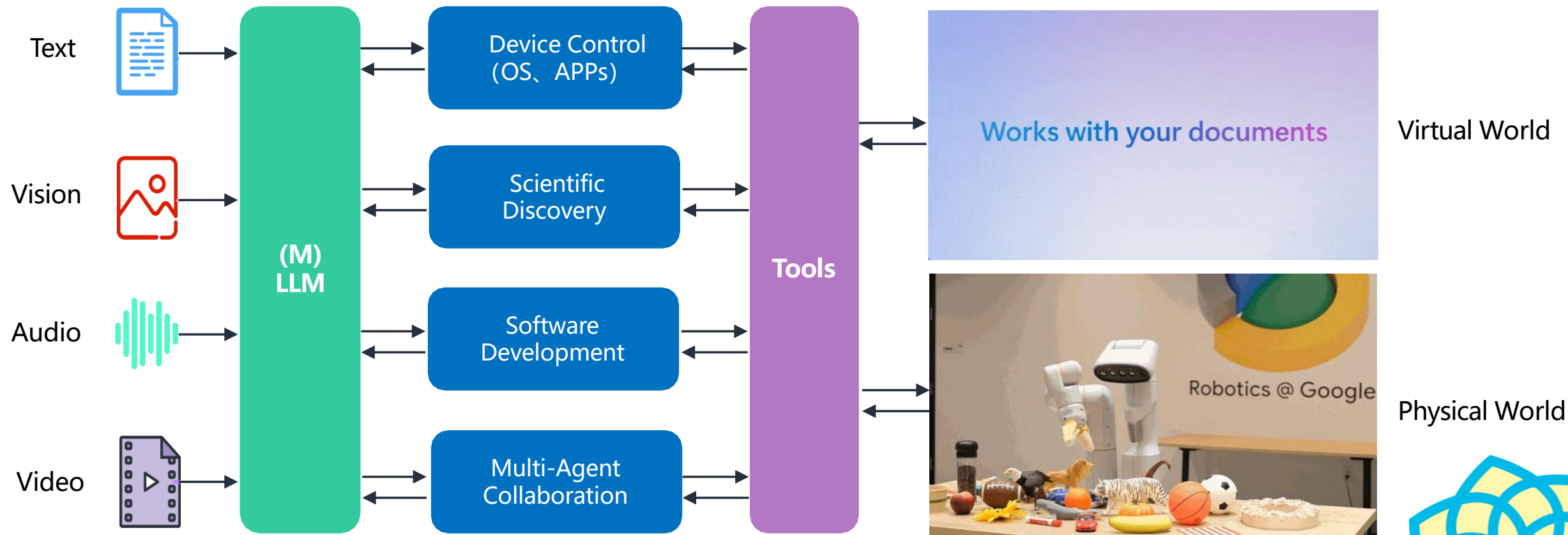


Brain in a Vat

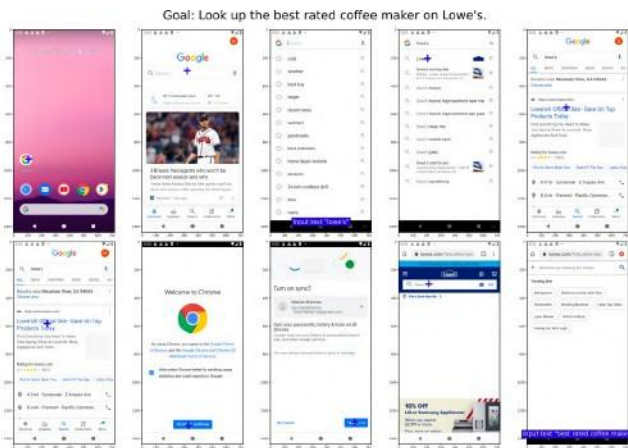


Towards Multimodal LLM Agents

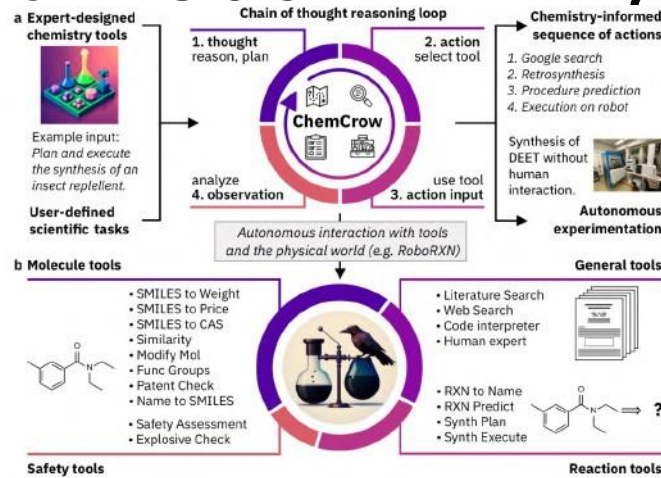
- ❑ **(M)LLM Agents:** follow language instructions and execute actions in environments, possibly use tools
- ❑ **Features:** General, Autonomous, Adaptive, Evolutionary, Socialized



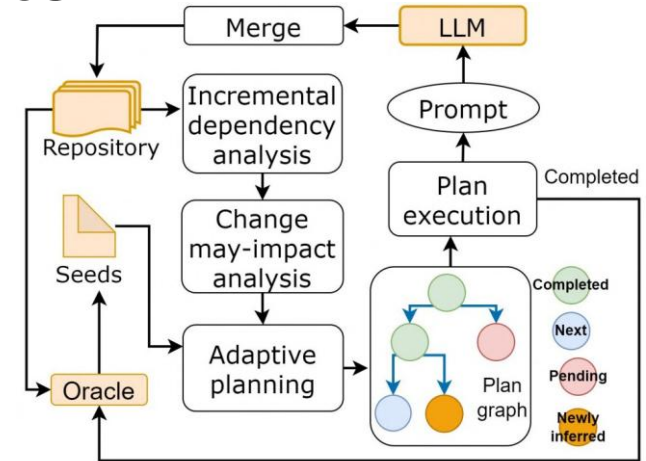
Towards Multimodal LLM Agents



Control: OS and Applications



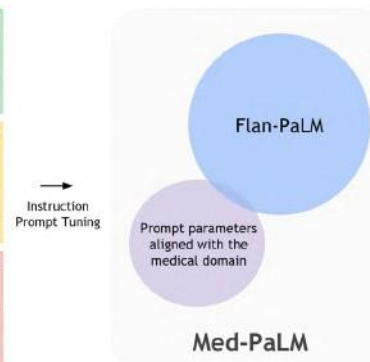
Research: Organic Synthesis



Programming: Code Generation



Control: Embodied Systems



Research: Medical Assistance



Interaction: Multi-Agent Collaboration

Taxonomy of (M)LLM Agents

Autonomous Agents

ADEPT Action Transformer
<https://www.adept.ai/blog/act-1>

Google AITW
https://github.com/google-research/google-research/tree/master/android_in_the_wild



WebArena
<https://webarena.dev>



Auto-UI
<https://github.com/cooelf/Auto-UI>

Communicative Agents



CAMEL
<https://github.com/camel-ai/camel>



Generative Agents
https://github.com/joonspk-research/generative_agents



VOYAGER
<https://voyager.minedojo.org/>



ChatDev
<https://github.com/OpenBMB/ChatDev>

More: AutoGPT, BabyAGI, Meta-GPT, AgentGPT

Taxonomy of (M)LLM Agents

Autonomous Agents: mainly task automation

Mobile Device Automation

User : Hello. Is it cold out today?

Action Executor :



System : The lowest temperature is 10 °C today.

User : What is the chance of rain today?

Action Executor :



System : The chance of rain is 100% today.

Meta-GUI

Webpage Automation

“ Create an efficient itinerary to visit all Pittsburgh's art museums with minimal driving distance starting from CMU. Log the order in my “awesome-northeast-us-travel” repository ”

Search for museums in Pittsburgh

Search for each art museum on the Map

Record the optimized results to the repo

WebArena

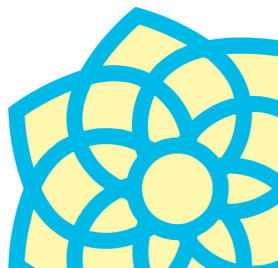
Application Automation

ACT-1

Sun, Liangtai, et al. "META-GUI: Towards Multi-modal Conversational Agents on Mobile GUI." *EMNLP 2022*.

Zhou, Shuyan, et al. "Webarena: A realistic web environment for building autonomous agents." *arXiv preprint arXiv:2307.13854* (2023).

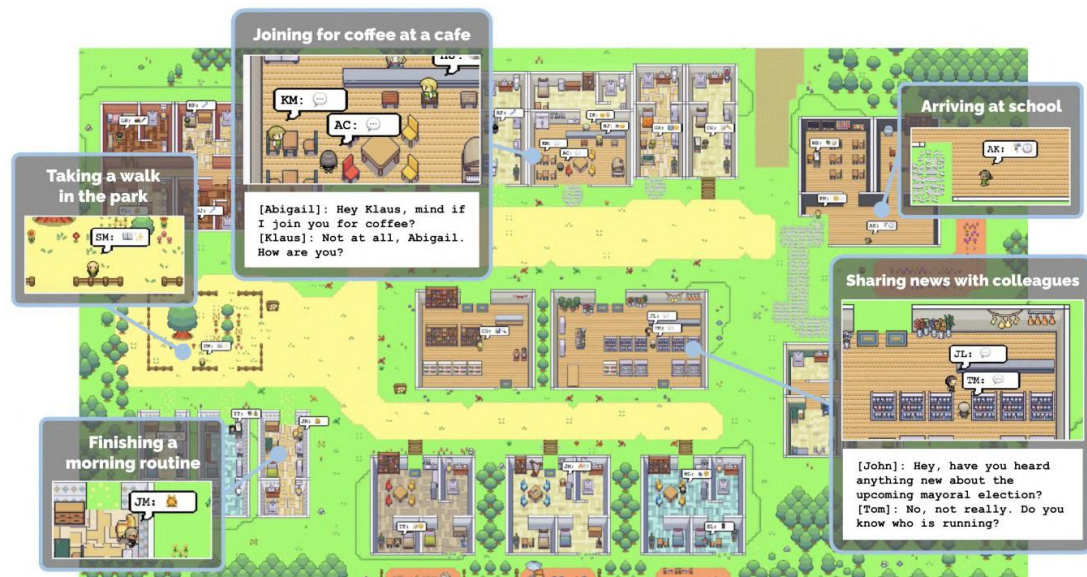
<https://www.adept.ai/blog/act-1>



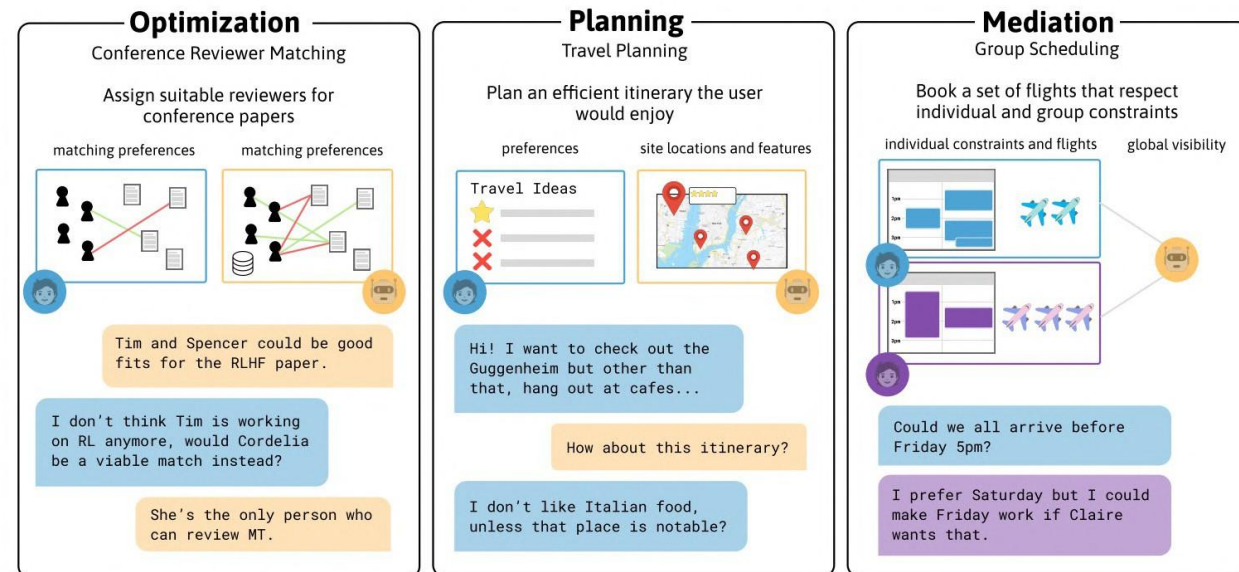
Taxonomy of (M)LLM Agents

Communicative Agents: personalized, socialized, interactive

Agents-Agents



Agents-Human

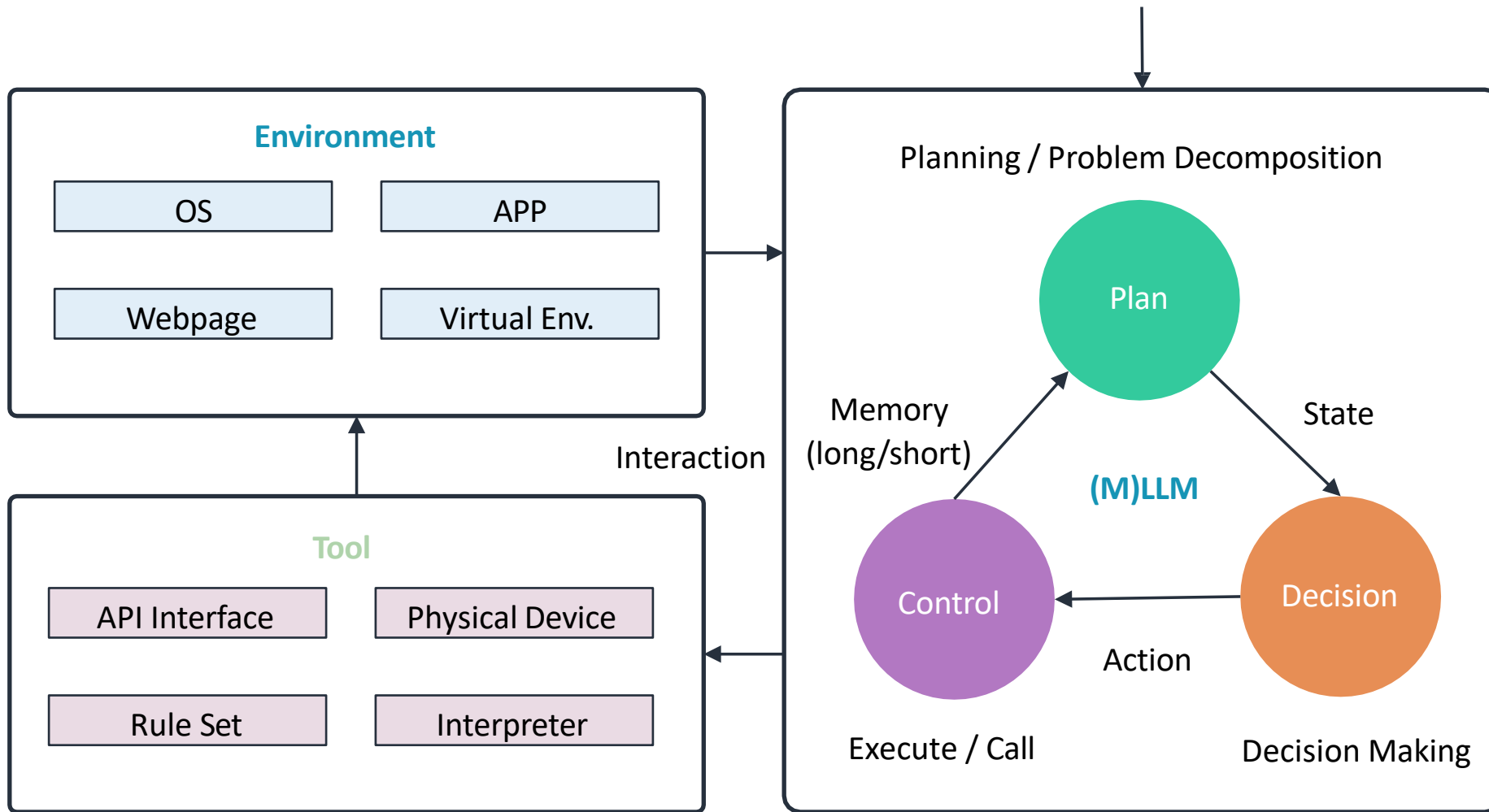


Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." *arXiv preprint arXiv:2304.03442* (2023).

Lin, Jessy, et al. "Decision-Oriented Dialogue for Human-AI Collaboration." *arXiv preprint arXiv:2305.20076* (2023).



Technological Paradigm

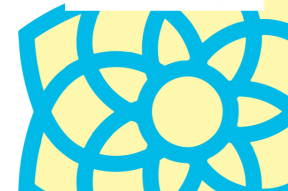
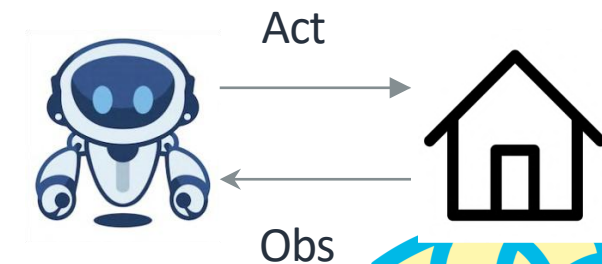


Foundation

- Multimodalities
- Long-context Modeling

Workflow

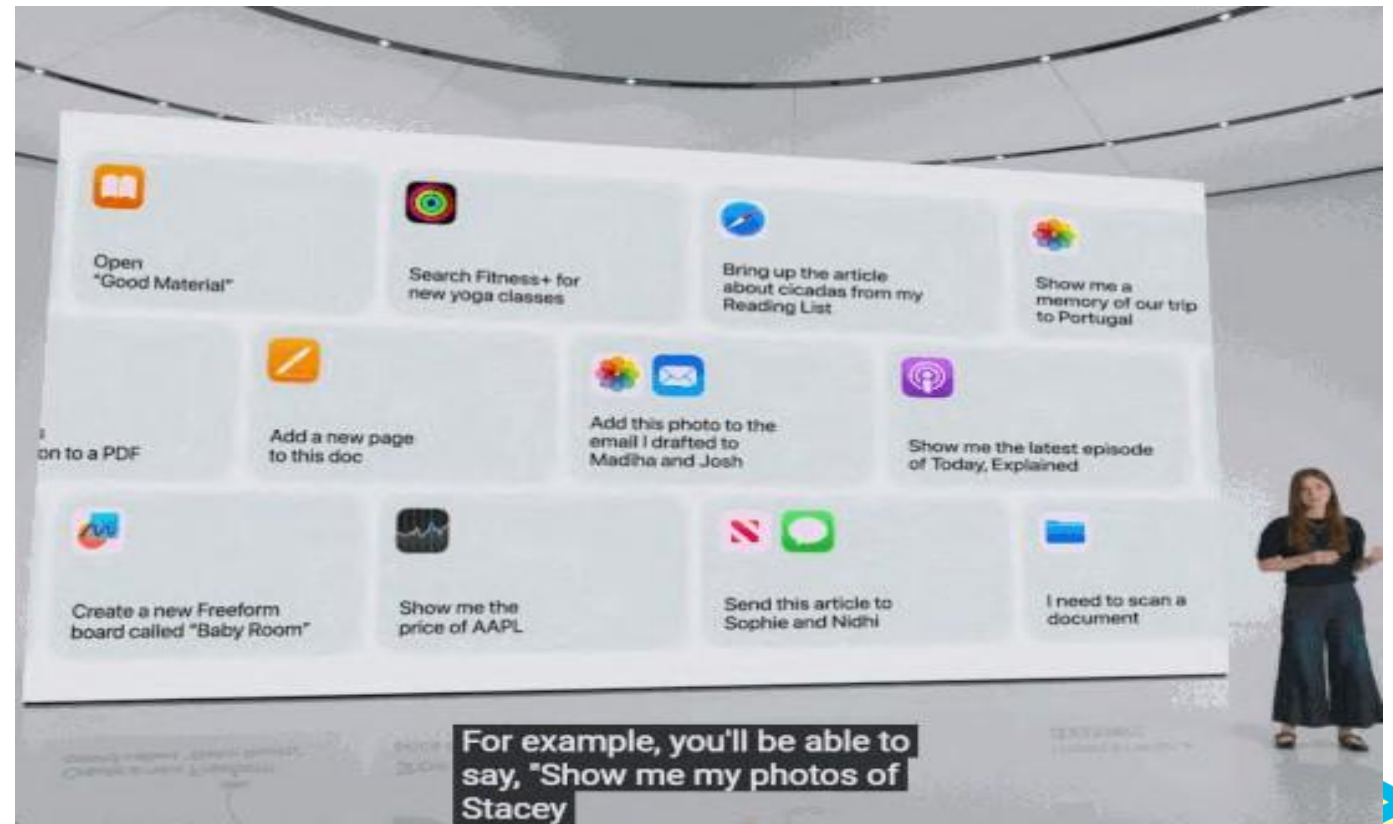
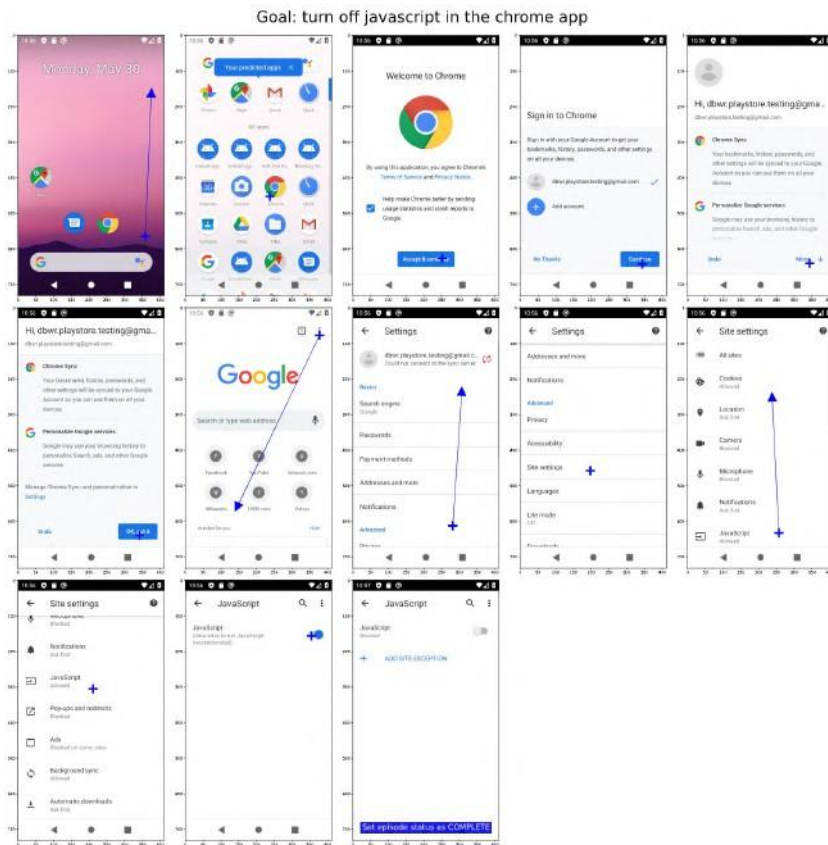
- Perception
- Planning & Decision Making
- Action (w/ Tool Use)
- Interaction
- Memory
- Multi-Agent Collaboration



GUI Agents

Auto-GUI: Multimodal Autonomous Agents for GUI control

- assist users in completing tasks in distinct environments such as operation systems, specific applications, and web browsers
- Imitate human clicking, scrolling, and typing actions, and operate directly with the GUI



Zhuosheng Zhang, Aston Zhang. You Only Look at Screens: Multimodal Chain-of-Action Agents. Findings of ACL 2024.

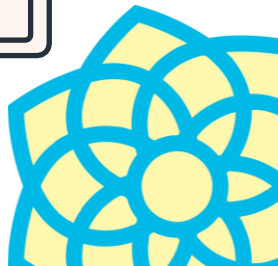
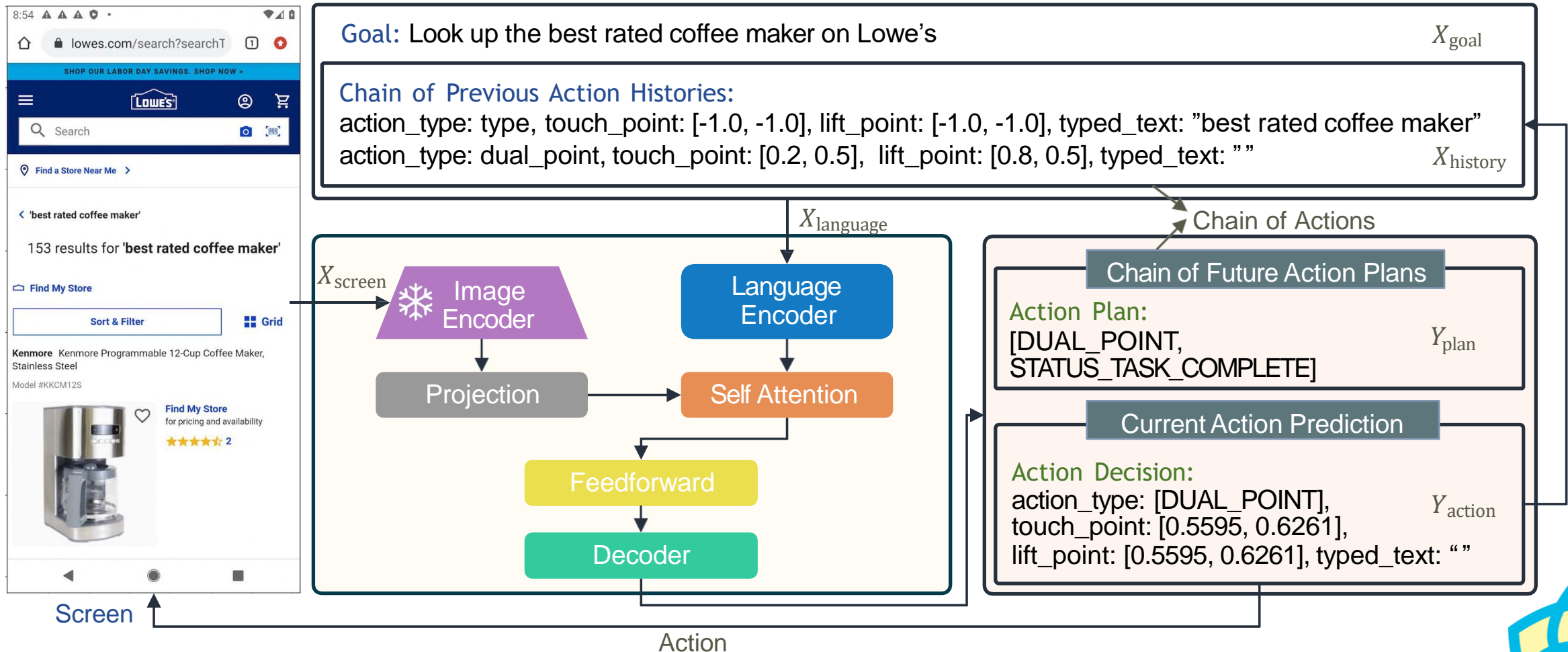
Xinbei Ma, Zhuosheng Zhang, Hai Zhao. Comprehensive Cognitive LLM Agent for Smartphone GUI Automation. Findings of ACL 2024.

<https://maci.apple.com/research/ferret..>

Auto-UI

Multimodal Agent: BLIP2 + FLAN-Alpaca

Chain-of-Action: a series of intermediate previous action histories and future action plans



Results

- ❑ A unified multimodal model out of *first principles thinking* can serve as a strong autonomous agent
 - can be adapted to **different scenarios** without the need to train specific models for each task
 - does not need additional annotations (screen parsing) and is **easy to use**
- ❑ Coverage: 30K unique instructions, 350+ Apps and websites
- ❑ **Action Type Accuracy: 90%+, Action Success Rate: 74%+**

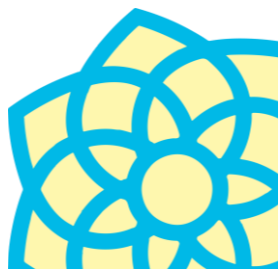
Model	Unified	w/o Anno.	Overall	General	Install	GoogleApps	Single	WebShopping
BC-single	✗	✗	68.7	-	-	-	-	-
BC-history	✗	✗	<u>73.1</u>	<u>63.7</u>	<u>77.5</u>	<u>75.7</u>	<u>80.3</u>	<u>68.5</u>
PaLM 2-CoT	✓	✗	39.6	-	-	-	-	-
ChatGPT-CoT	✓	✗	7.72	5.93	4.38	10.47	9.39	8.42
Fine-tuned Llama 2	✗	✗	28.40	28.56	35.18	30.99	27.35	19.92
Auto-UI _{separate}	✗	✓	74.07	65.94	77.62	76.45	81.39	69.72
Auto-UI _{unified}	✓	✓	74.27	68.24	76.89	71.37	84.58	70.26



Insights

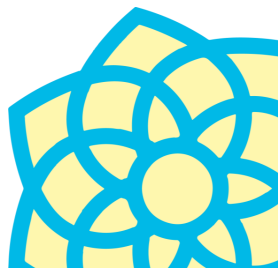
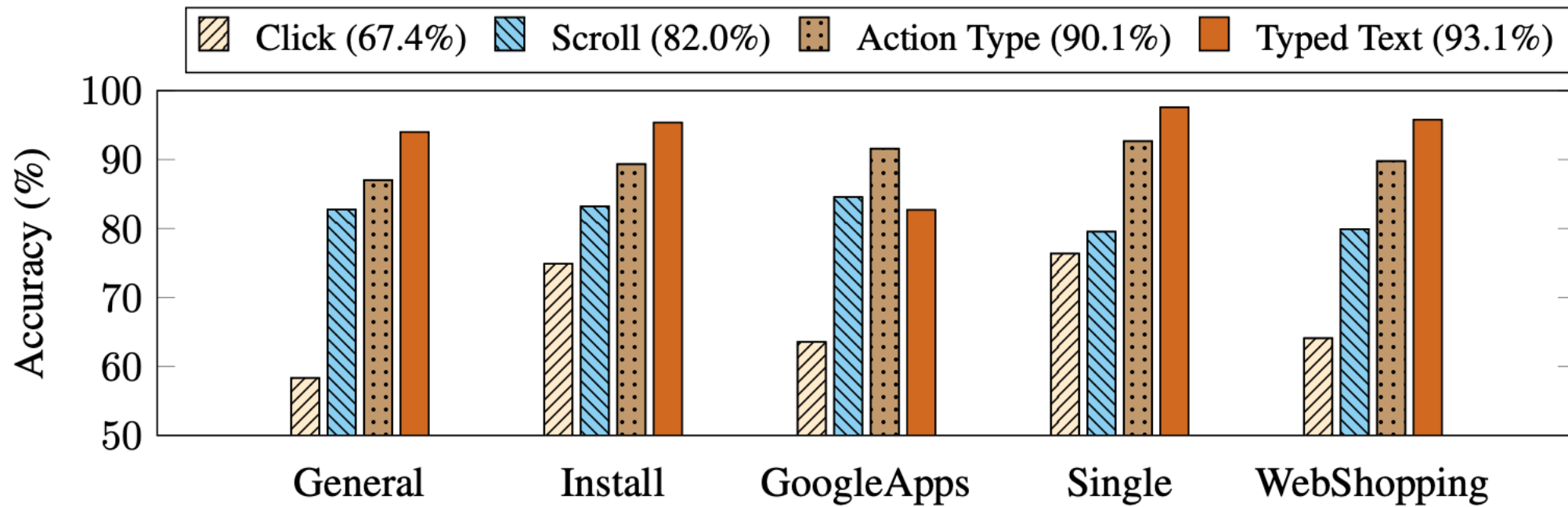
- ❑ The bottleneck seems to be the **multimodal perception**, misleading the reasoning process
 - GUI involves comprehensive elements (interleaved, icons, texts, boxes)
 - Changing vision encoders influences the performance dramatically
- ❑ Scaling does not always improve performance

Model	Overall	General	Install	GoogleApps	Single	WebShopping
Auto-UI on CLIP	71.84	66.28	74.40	69.71	81.60	67.23
Auto-UI on BLIP-2	74.27	68.24	76.89	71.37	84.58	70.26
Auto-UI on Vanilla-T5 _{large}	72.98	66.61	75.40	70.86	83.47	68.54
Auto-UI on FLAN-T5 _{large}	73.36	67.59	76.35	70.71	83.01	69.12
Auto-UI on FLAN-Alpaca _{large}	74.27	68.24	76.89	71.37	84.58	70.26
Auto-UI on FLAN-Alpaca _{small}	71.38	65.26	74.90	68.70	81.20	66.83
Auto-UI on FLAN-Alpaca _{base}	72.84	66.97	75.93	70.29	82.56	68.46
Auto-UI on FLAN-Alpaca _{large}	74.27	68.24	76.89	71.37	84.58	70.26



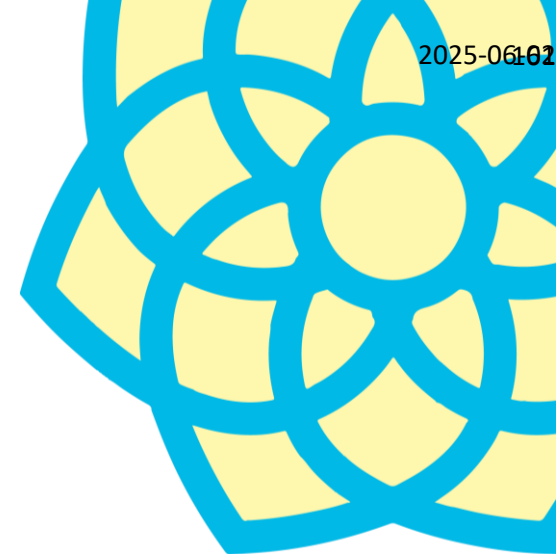
Insights

- Category Accuracy: the major challenges lie within the click region and scroll direction predictions
 - The model tends to click a wrong place or scroll in a wrong direction
- Challenge in “really” understanding the GUI layouts, e.g., relationship between GUI elements



4

Challenges



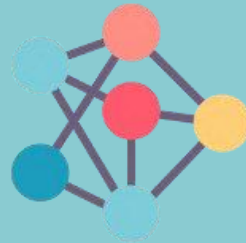
Challenges

- ❑ Multimodal reasoning drives smart MLLMs
 - More broader scenarios (physical and virtual worlds)
 - More comprehensive scenarios (evolutionary, interactive)



Evolutionary Reasoning

- Active explore and evolve from environments
- Learn from (un)successful attempts



Interactive Reasoning

- Human-in-the-loop interference
- Error identification and correction abilities



Reasoning Alignment

- Align both content safety, and behavior safety
- Decide the action trajectory with foresights



Summary

- ❑ **Basics of Multimodal Reasoning**
 - **Concept: derive high-level conclusions from multiple modalities, possibly via multiple logical steps based on atomic evidences**
 - **Developments: (a) From task-specific to centralized paradigms; (b) From single-step prediction to multi-step reasoning**
 - **Popular Approaches: (a) In-Context Learning; (b) Multimodal Chain-of-Thought**
- ❑ **Multimodal Chain-of-Thought Reasoning**
 - **Paradigm Shift: From “<input → output>” to <input → rationale → output>**
 - **Role 1: Introducing more reliable input results in more convincing reasoning process**
 - **Role 2: Breaking complex problems into smaller, manageable sub-problems**
 - **Role 3: Available for stepwise knowledge update and self-correction (w/ external feedback)**
- ❑ **Towards Multimodal LLM Agents**
 - **Taxonomy: Autonomous Agents and Communicative Agents**
 - **Technical Components: Foundation (multimodality & long-context modeling); (b) Workflow (plan, act, memory, feedback)**
- ❑ **Challenges**
 - **Evolutionary Reasoning, Interactive Reasoning, Reasoning Alignment**



Example Planning Labs

Experiment 2:1-2:2

Experiment 1:1

Perplexity

Small model
Large data,
Limited precision

Small model
Small data

Small model
Large data

Experiment 1:2

Experiment 1:3

Experiment 3:1

Experiment 3:2

Experiment 4:1

BPE Tokenizer

Attention Mechanism 1

Large model
Small data

Fine-tuning
with curated data

Instruction
fine-tuning

Fairness Measure

SentencePiece, or
WordPiece

Attention Mechanism 2

Large model
Large data

No fine-tuning

No fine-tuning

Prompt testing on
sensitive and
toxicity

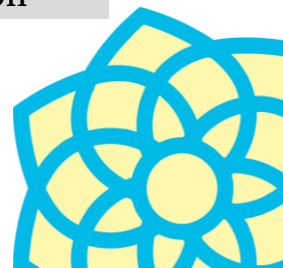
Experiment 4:2

Pre-processing

Training

Fine-tuning

Evaluation



- **LAIM LE10 VT2025: Multi-Modal LLMs**

Architecture

Modality and Functionality

Multi-Modal Instruction Tuning

Multi-modal Reasoning

www.ida.liu.se/~frehe08/llm