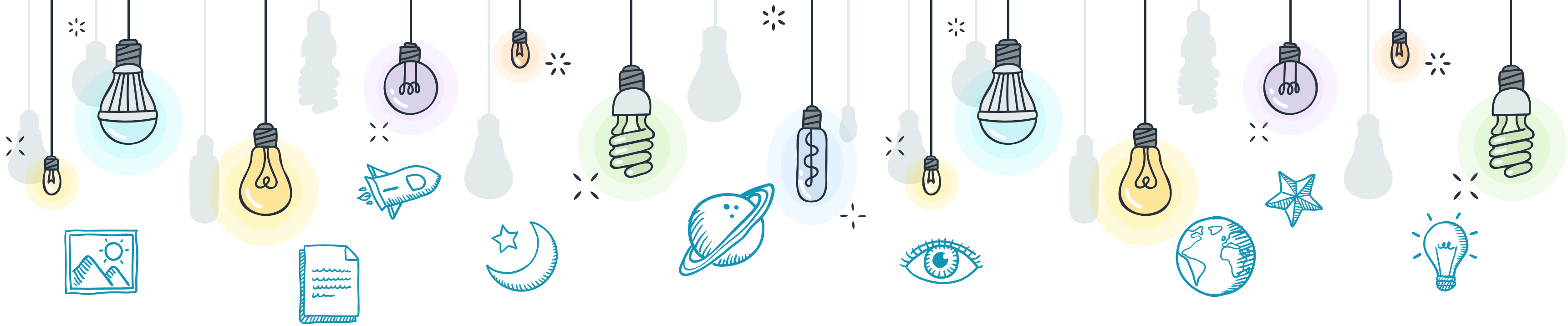


From Multimodal LLM to Human-level AI

Modality, *Instruction*, *Reasoning*, *Efficiency* **and Beyond**

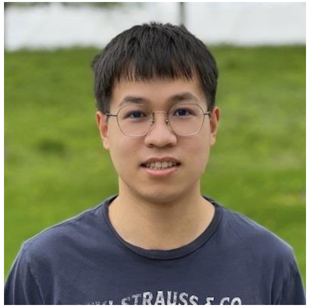
https://mllm2024.github.io/CVPR2024/

CVPR JUNE 17-21, 2024 SEATTLE, WA

Hao Fei
National University of Singapore

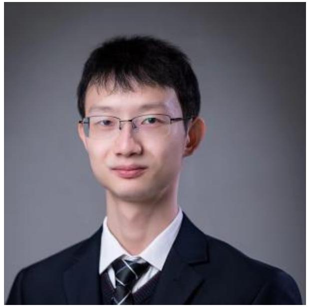
Yuan Yao
National University of Singapore

Ao Zhang
National University of Singapore

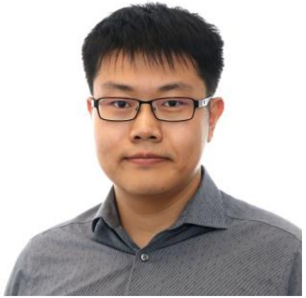


Haotian Liu
University of Wisconsin-Madison

Fuxiao Liu
University of Maryland, College Park



Zhuosheng Zhang
Shanghai Jiao Tong University

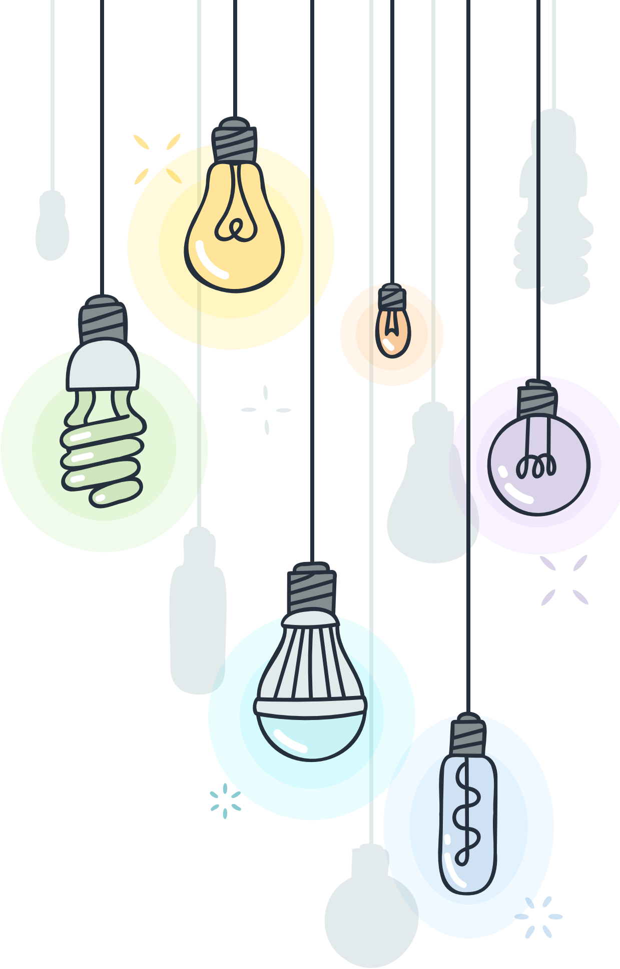


Hanwang Zhang
Nanyang Technological University

Shuicheng Yan
Kunlun 2050 Research, Skywork AI

Part-II

MLLM Design: Architecture



Yuan Yao

Research Fellow

National University of Singapore

https://yaoyuanthu.github.io/

Table of Content

1 Architecture

1

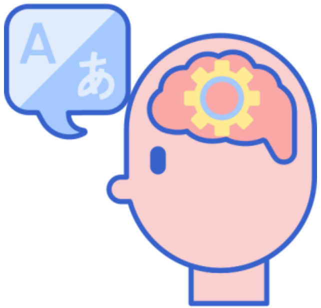
Architecture of MLLM

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.



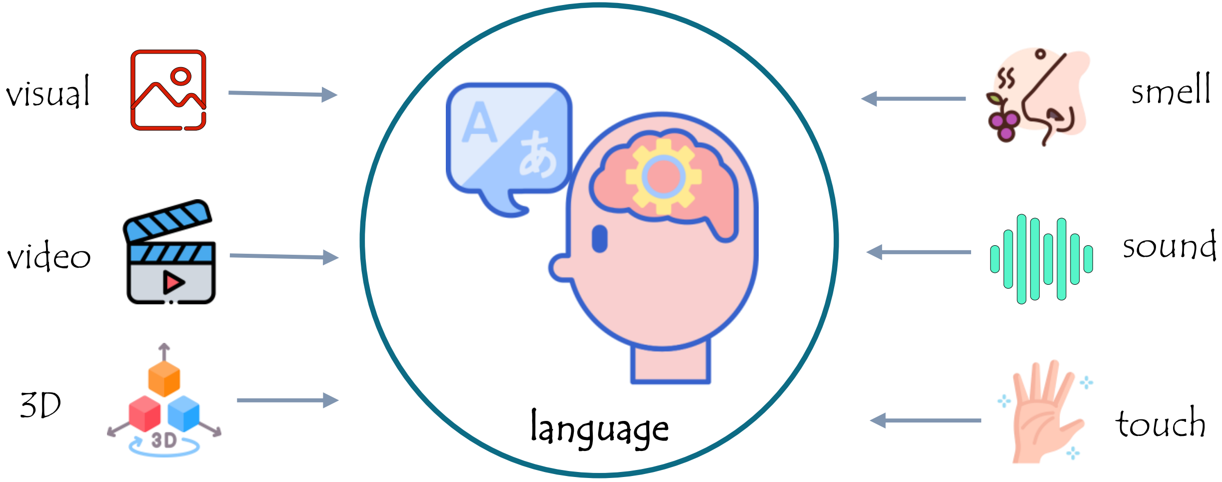
language

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



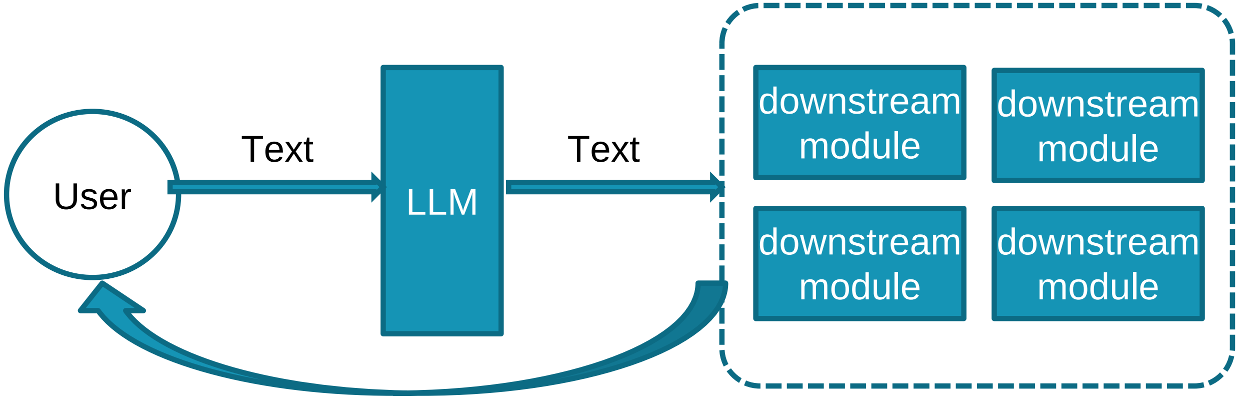
Overview of MLLM Architecture

- ## 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.

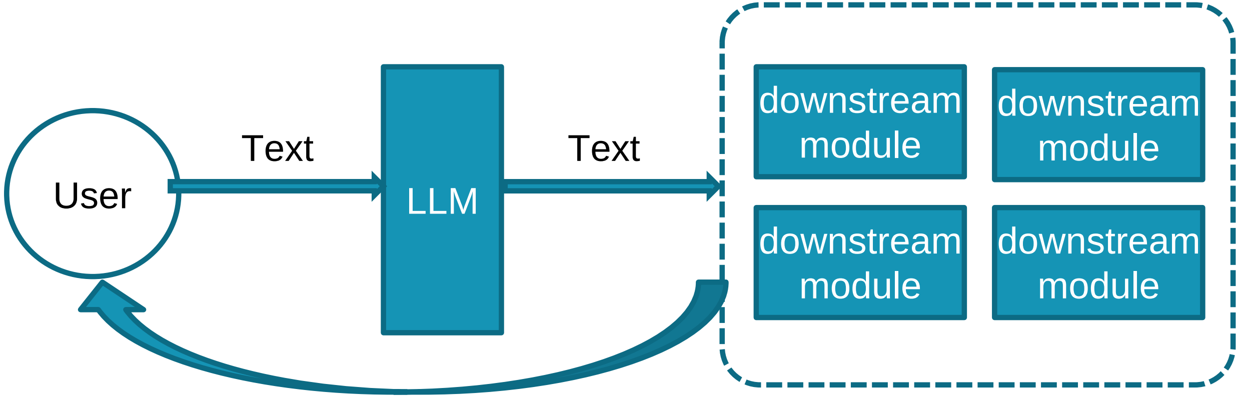
- Key feature:

  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.



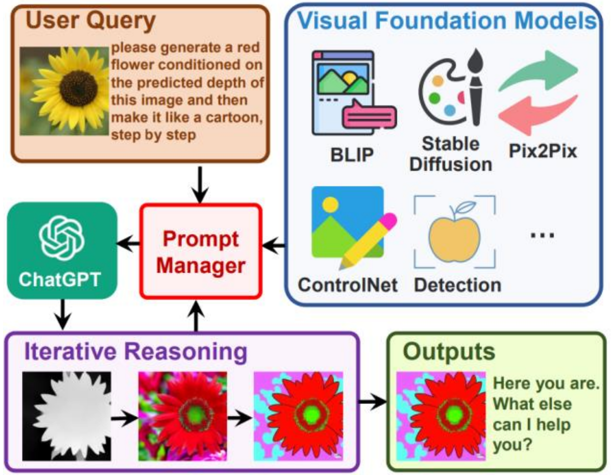
Overview of MLLM Architecture

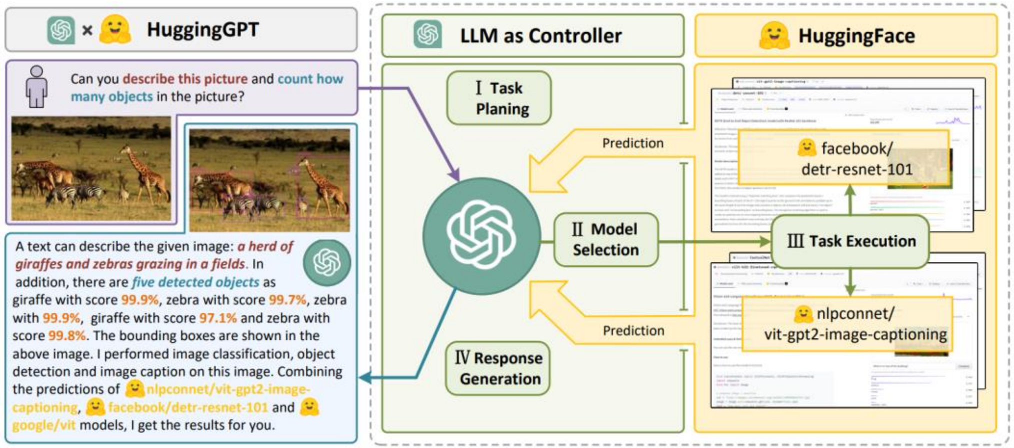
- ## Architecture-I: LLM as Discrete Scheduler/Controller
  - Representative MLLMs:
    - Visual-ChatGPT
    - HuggingGPT
    - MM-REACT
    - ViperGPT
    - AudioGPT
    - LLaVA-Plus
    - …



Overview of MLLM Architecture

- Architecture-I: LLM as **Discrete Scheduler/Controller**
  - Visual-ChatGPT
  - HuggingGPT





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

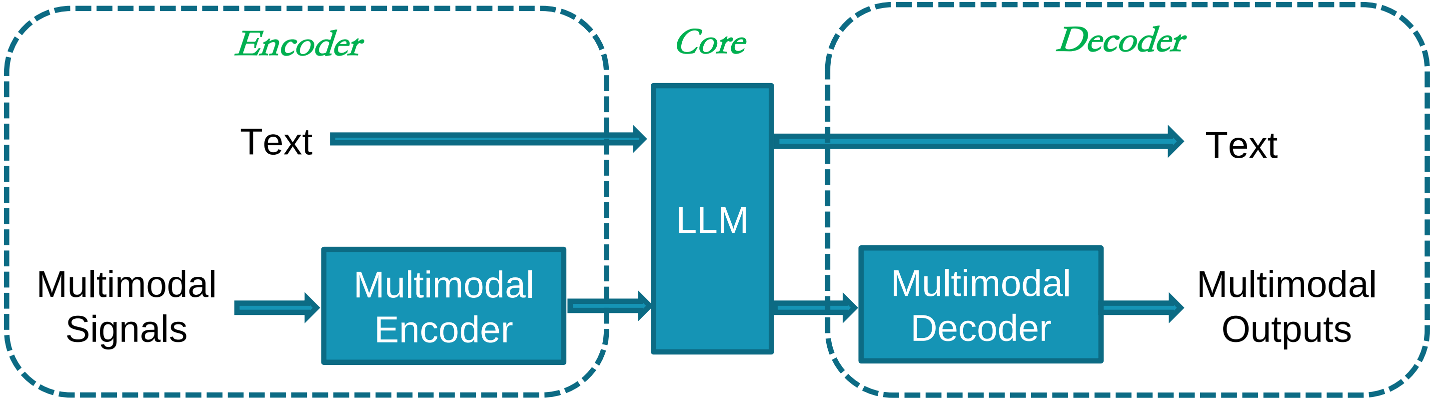
Overview of MLLM Architecture

Architecture-II: LLM as Joint Part of System

☞ The role of the LLM is to perceive multimodal information, and **react by itself**, in an 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.

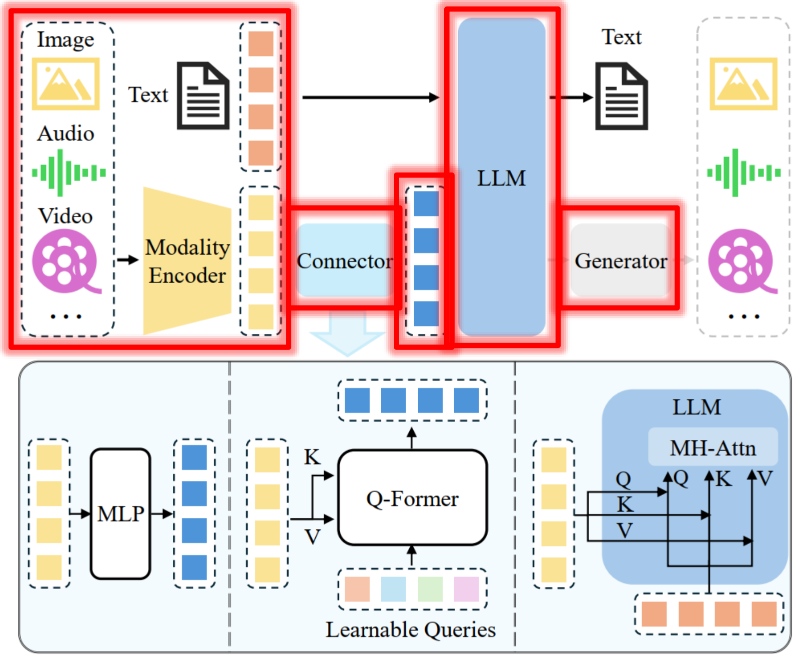


Overview of MLLM Architecture

- **Architecture-II: LLM as Joint Part of System**

More promising

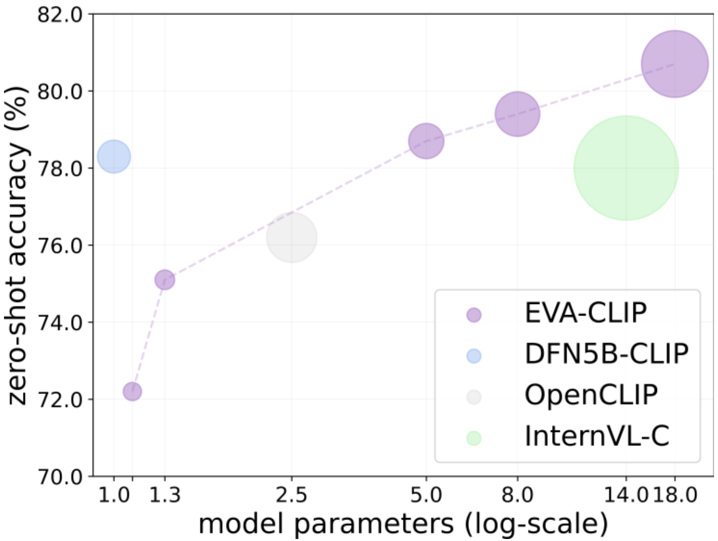
  - > 90% MLLMs belong to this category.
  - Higher upper-bound, better integrated into a unified model



[1] A Survey on Multimodal Large Language Models. https://github.com/BradyFU/Awesome-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

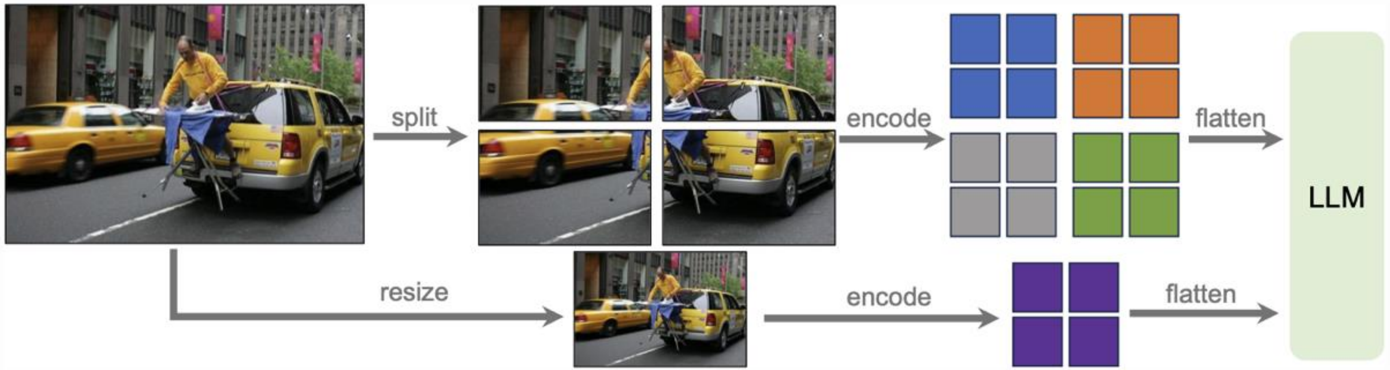


Illustration of dynamic high resolution scheme: a grid configuration of 2×2

Multimodal Encoding

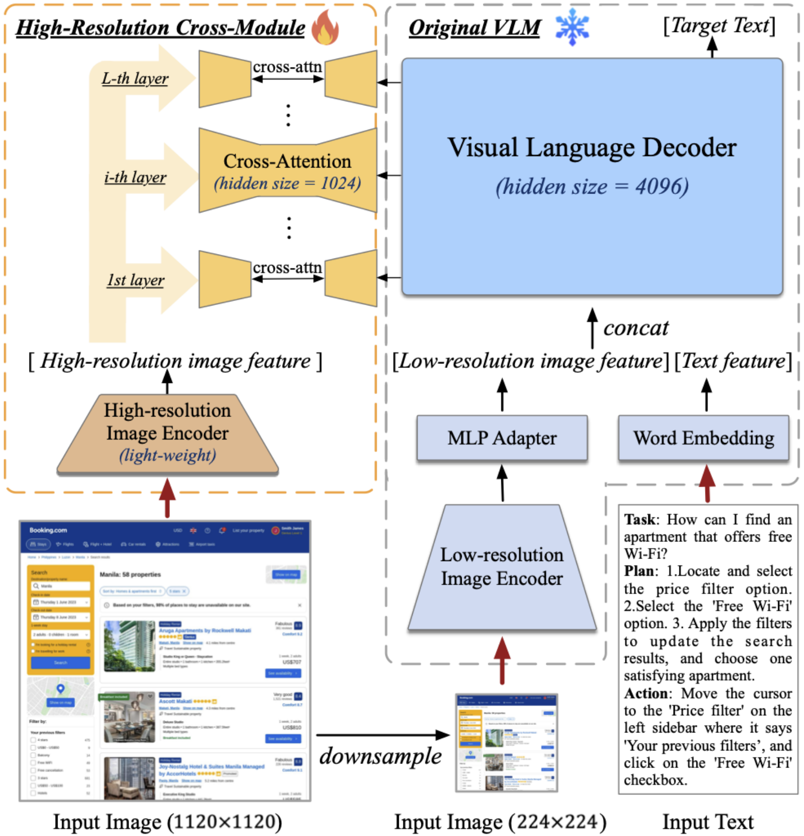
Visual Encoder

- 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	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**	56.1	**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

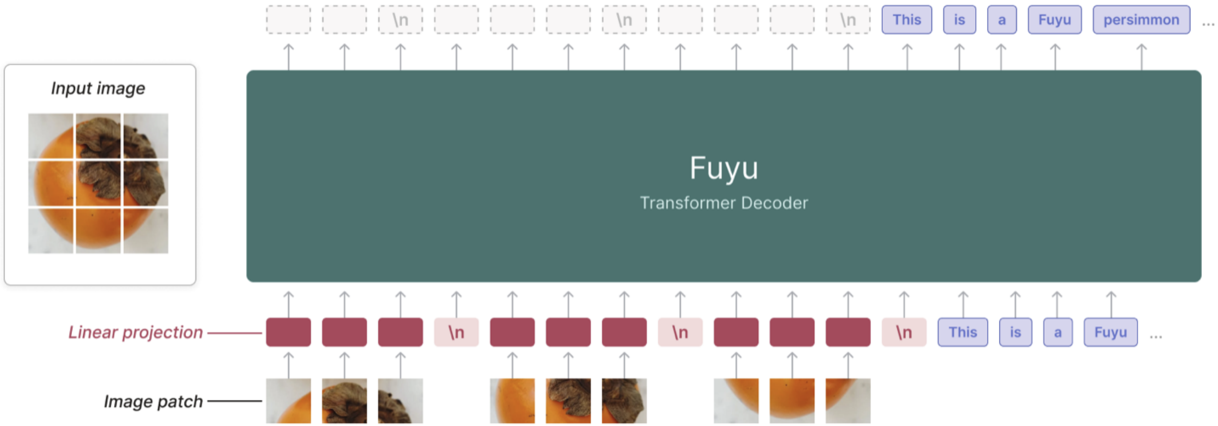
Multimodal Encoding

- **Visual Encoder**
  - High-resolution Multimodal LLMs
    - Dual branch encoders
    - Representatives
      - CogAgent
      - Mini-Gemini
      - DeepSeek-VL
      - LLaVA-HR



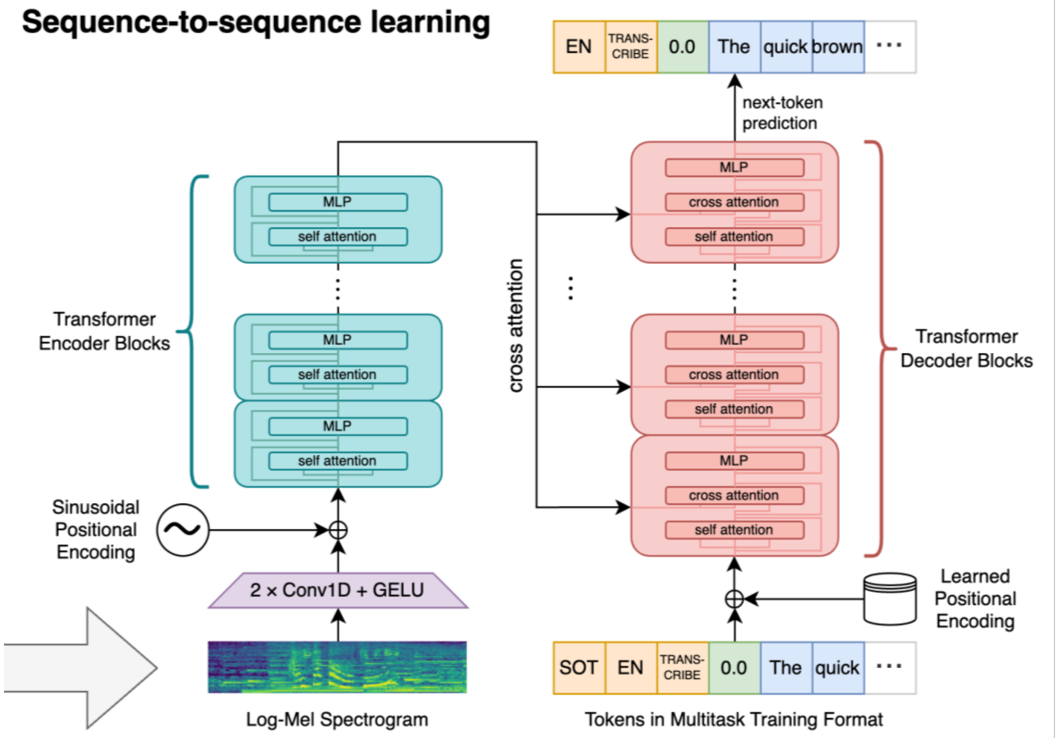
Multimodal Encoding

- **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



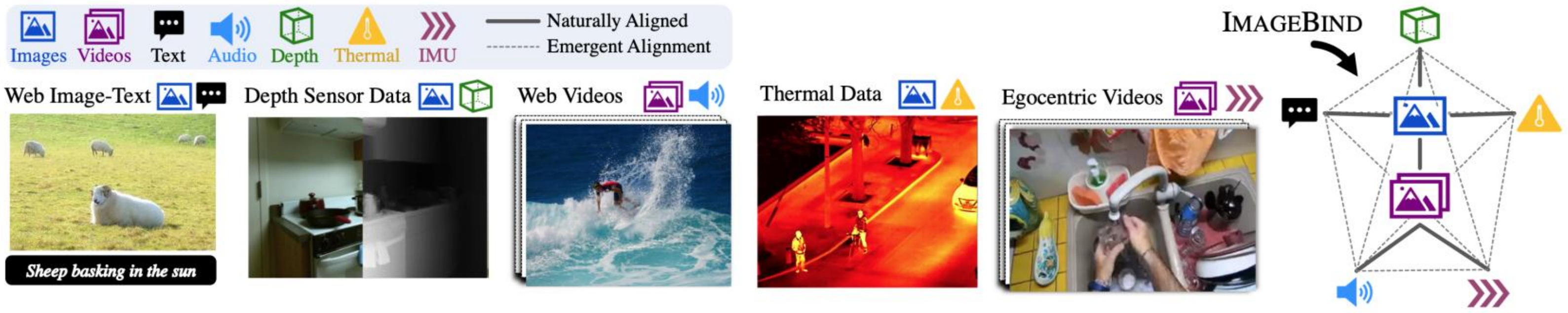
Multimodal Encoding

- ## Non-Visual Encoder
  - **Audio:**
    - Whisper
    - AudioCLIP
    - HuBERT
    - BEATs
  - **3D Point:**
    - Point-BERT



Multimodal Encoding

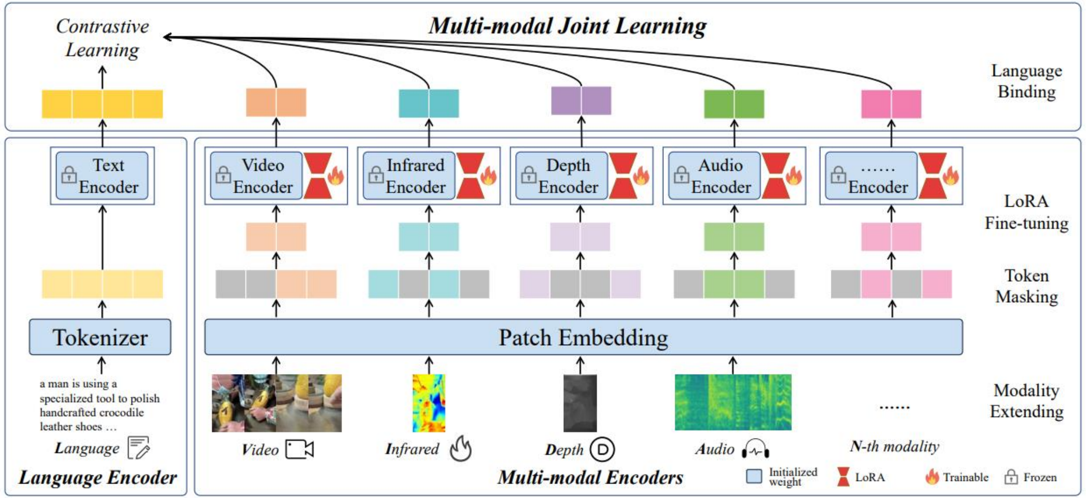
- ## Unified Multimodal Encoder
  - ImageBind:
    - Embedding all modalities into a joint representation space of Image.
    - Well aligned modality representations can benefit LLM understanding



[1] *ImageBind: One Embedding Space To Bind Them All. 2023*

Multimodal Encoding

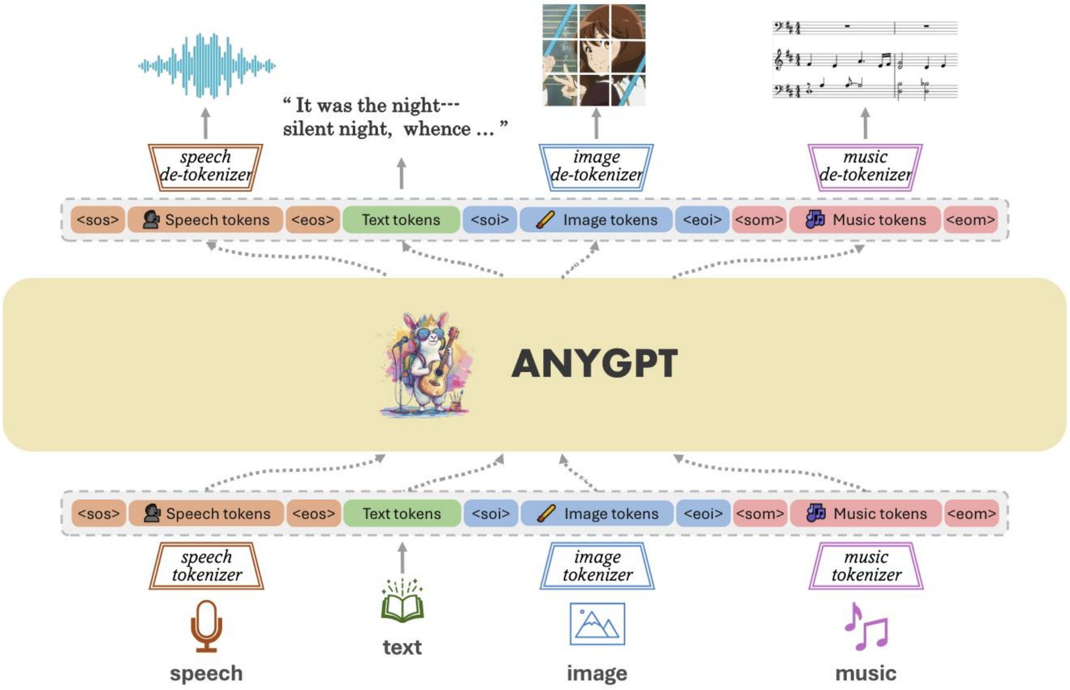
- **Unified Multimodal Encoder**
  - LanguageBind:
    - Embedding all modalities into a joint representation space of Language.
    - Well aligned modality representations can benefit LLM understanding



[1] *LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. 2023*

Multimodal Signal Tokenization

- ## Tokenization
  - AnyGPT



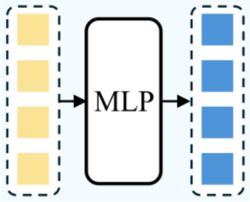
[1] AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. 2023

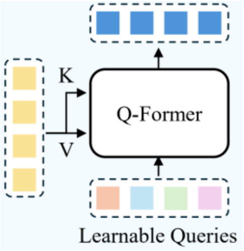
Multimodal Signal Tokenization

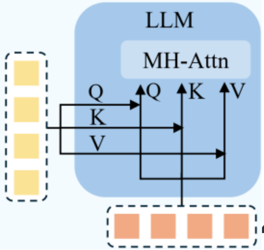
- ## Tokenization in Codebook
  - Represent multimodal signals as discrete tokens in a codebook
    - Advantages: support unified multimodal signal understanding and generation in an auto-regressive next-token prediction framework
    - More commonly used in image synthesize
      - Parti
      - Muse (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
Only using a subset of InstructBLIP training data						
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	$\mathrm{VQA}^{\mathrm{v2}}$	GQA	$\mathrm{VQA}^{\mathrm{T}}$
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

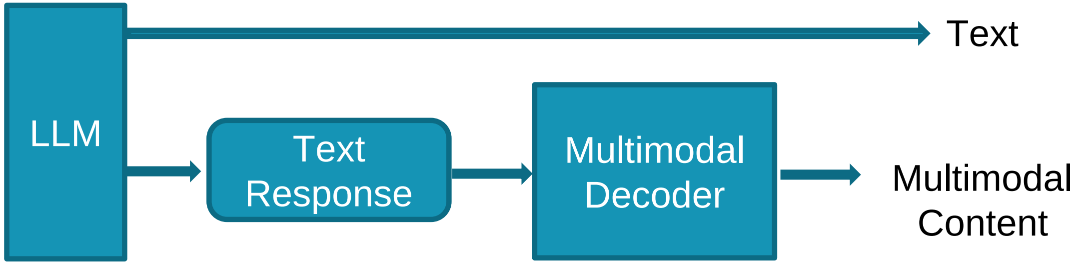
- ## Some Insights
  - Agreement: Number of visual token matters! Especially for efficiency
    - Resampler/Q-Former/C-Abstractor yield less visual tokens than MLP/Linear
    - Favorable in high-resolution image understanding

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**	56.1	**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.

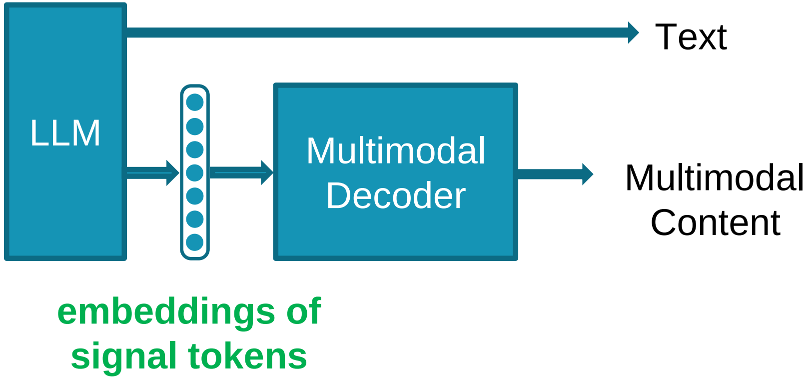
[1] Visual-ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. 2023
[2] HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. 2023
[3] ModaVerse: Efficiently Transforming Modalities with LLMs. 2024
[4] VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, 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*



[1] Generating Images with Multimodal Language Models. 2023
[2] NExT-GPT: Any-to-Any Multimodal LLM. 2023

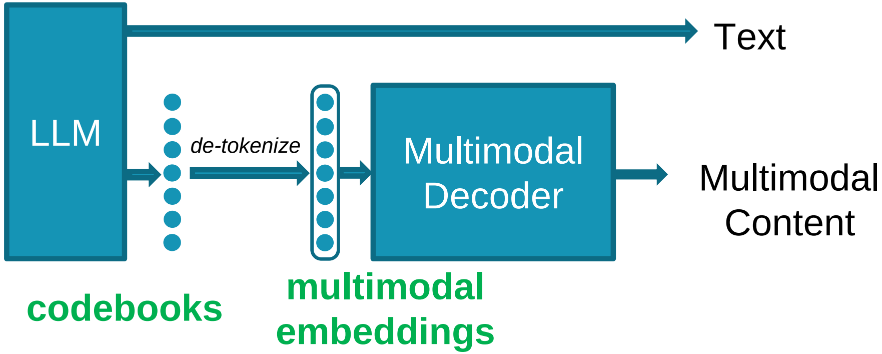
Decoding-side Connection

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

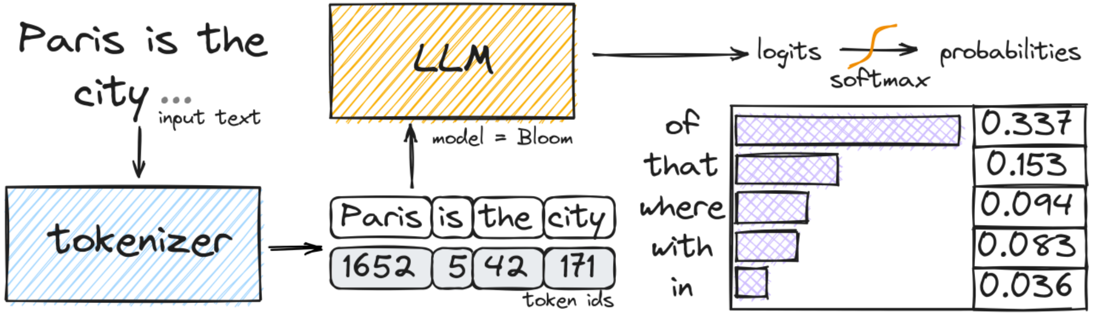


[1] Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action. 2023
[2] LVM: Sequential Modeling Enables Scalable Learning for Large Vision Models. 2023
[3] AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. 2024
[4] VideoPoet: A Large Language Model for Zero-Shot Video Generation. 2024

Multimodal Generation

- **Text Generation**
  - LLMs naturally support direct text generation

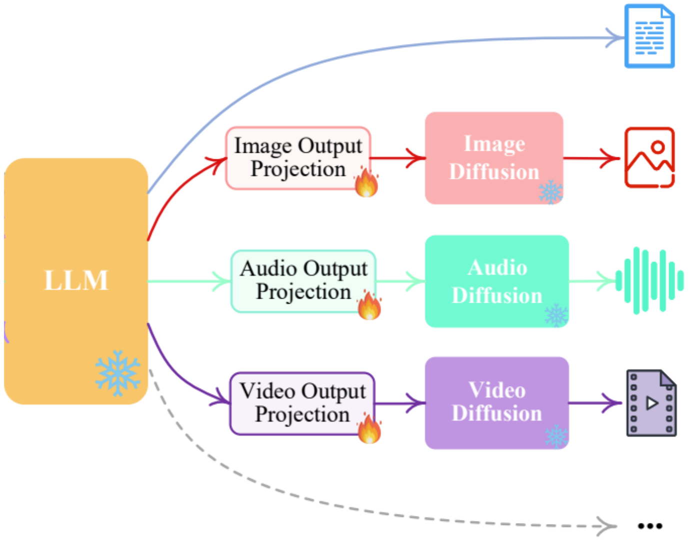
    via e.g., BPE decoding, Beam search, ...



Multimodal Generation

- ## Generation via Diffusion Models

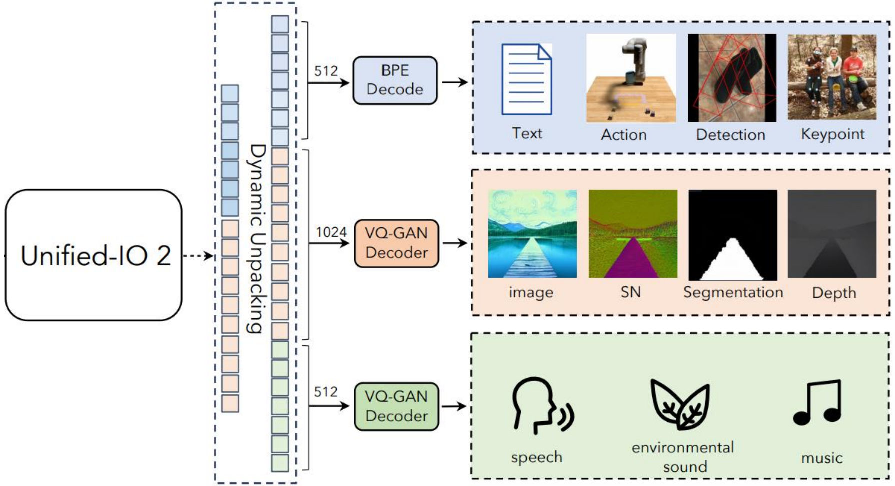
  - Visual (Image/Video) Generator
    - Image Diffusion
    - Video Diffusion
  - Audio Generator
    - Speech Diffusion
    - Audio Diffusion



[1] NExT-GPT: Any-to-Any Multimodal LLM. 2023

Multimodal Generation

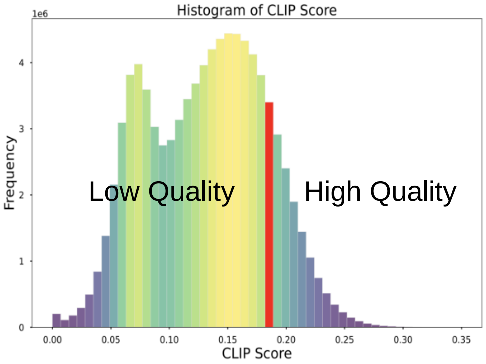
- **Generation via Codebooks**
  - Visual (Image/Video) Generator
    - VQ-VAE + Codebooks
    - VQ-GAN + Codebooks
  - Audio Generator
    - SpeechTokenizer + Residual Vector Quantizer
    - SoundStream + Residual Vector Quantizer



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

Multilingual Multimodal LLMs

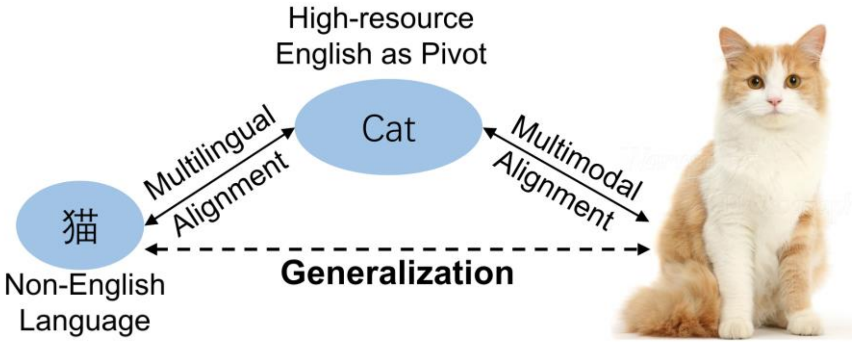
- **MLLMs are mostly strong in English, weak in other languages**
  - Limited scale and quality of multimodal data in non-English languages
  - Huge computation cost for each language even if sufficient data available
  - Why not machine translation pipeline ?
    - Another LLM for translation: double 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



[1] Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages. 2024.

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!

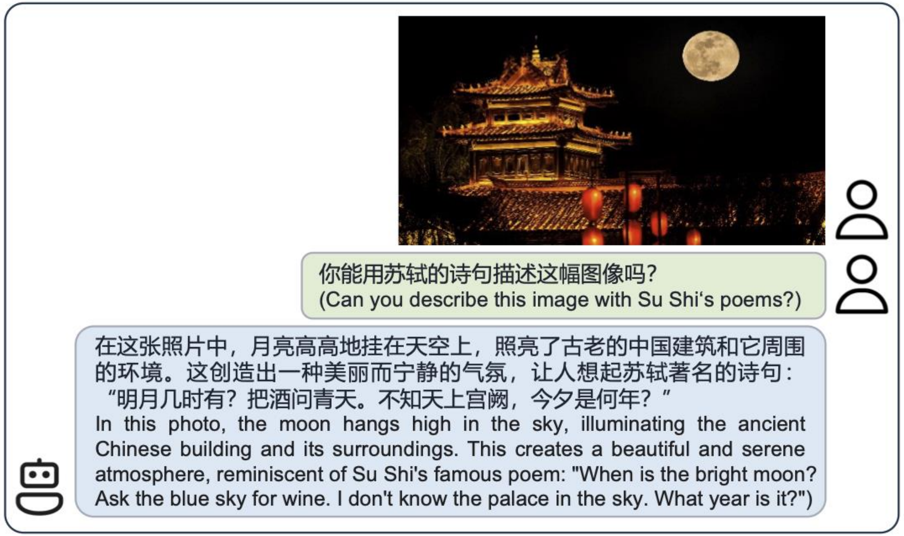
Was ist an diesem Bild ungewöhnlich?

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.

[1] Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages. 2024.

Multilingual Multimodal LLMs

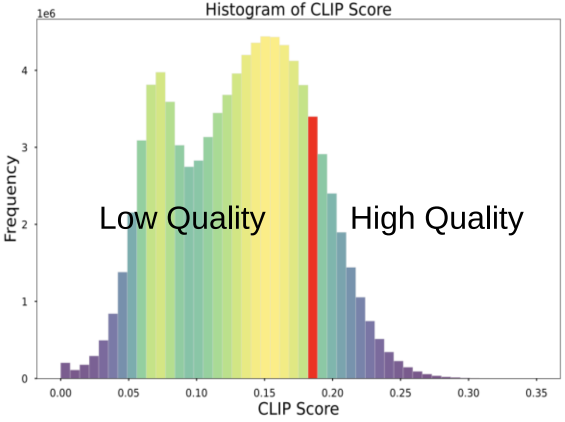
- MLLMs are mostly strong in English, weak in other languages
  - Generalizing multimodal capabilities to non-English languages via multilingual LLM as pivots
  - Lightweight SFT is sufficient to switch to the target language
  - Generalize to cultures, etc.



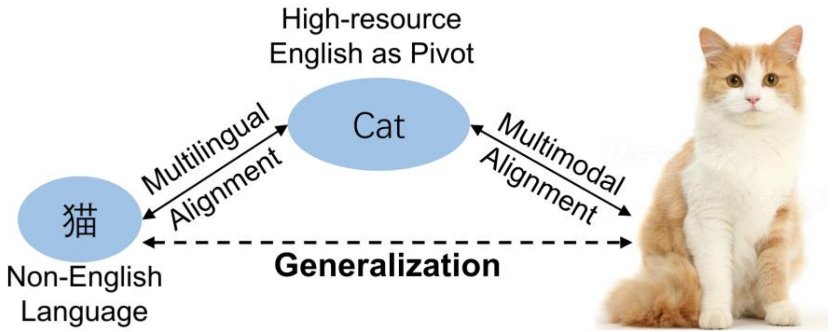
[1] Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages. 2024.

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:

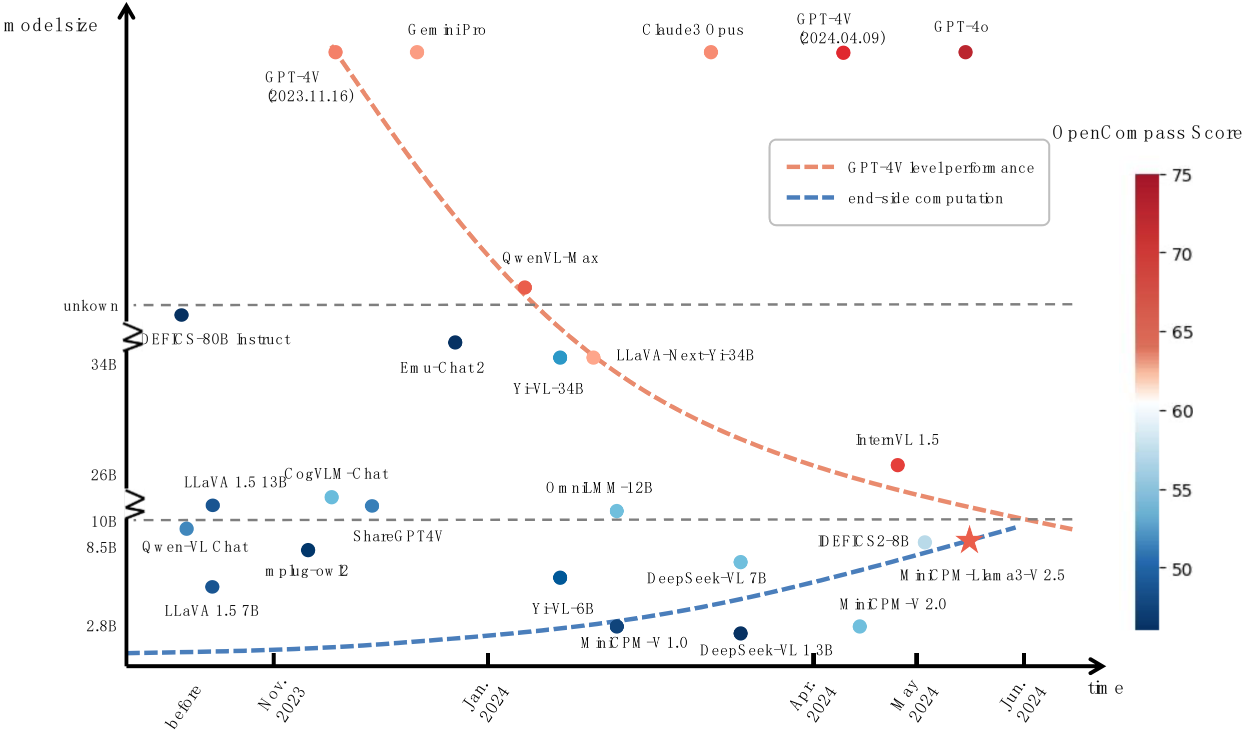


Chinese Image-Text Data Quality Distribution



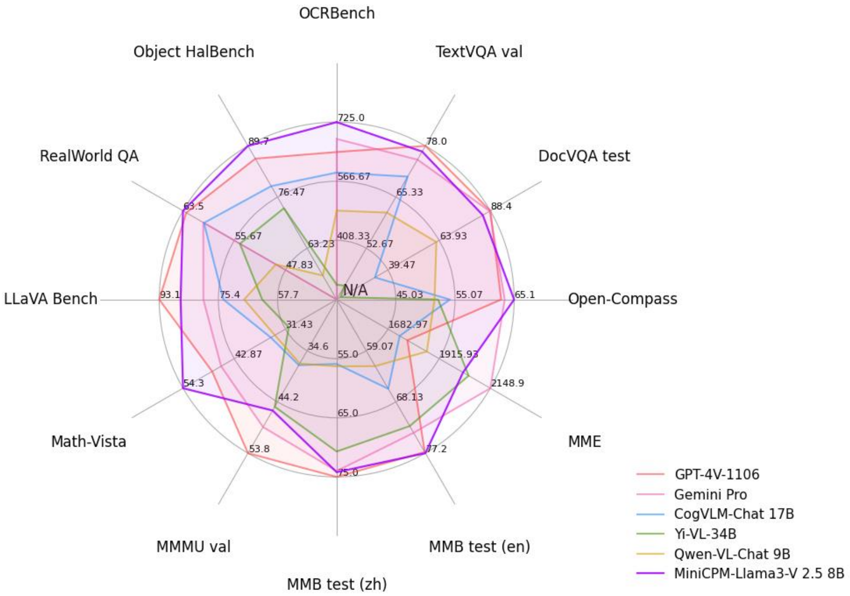
End-side MLLM

- End-side MLLMs show promising potentials
  - Promising performance: Matching larger LLMs



End-side MLLM

- ## End-side MLLMs show promising potentials
  - Promising performance: Matching larger MLLMs
  - Representatives
    - MiniCPM-V 1.0/2.0/2.5
    - DeepSeek-VL 1.3B
    - Mini-Gemini 2B
    - Bunny
    - MobileVLM V1/V2
    - Imp V1/V2
    - GLM-4V
    - Phi-3-vision
    - PaliGemma



End-side MLLM

- **End-side MLLMs show promising potentials**
  - End-side deployment, e.g., mobile phones
    - High efficiency
    - Better accessibility
    - Better privacy protection
    - Run without network connection

Thanks!

Any questions?