From Multimodal LLM to Human-level AI

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

https://mllm2024.github.io/COLING2024

LREC-COLING 2024

NUS National University of Singapore

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

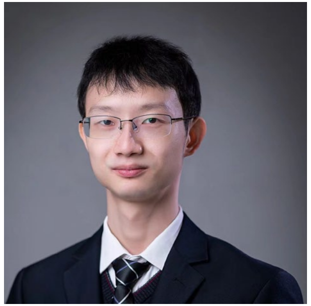
UNIVERSITY OF MARYLAND

Hao Fei

National University of Singapore

Yuan Yao

National University of Singapore



Zhuosheng Zhang

Shanghai Jiao Tong University

Fuxiao Liu

University of Maryland, College Park

Ao Zhang

National University of Singapore

Tat-Seng Chua

National University of Singapore

Part-II

MLLM Design: Architecture and Modality

Hao Fei
Research Fellow
National University of Singapore
http://haofei.vip/

Yuan Yao
Research Fellow
National University of Singapore
https://yaoyuanthu.github.io/

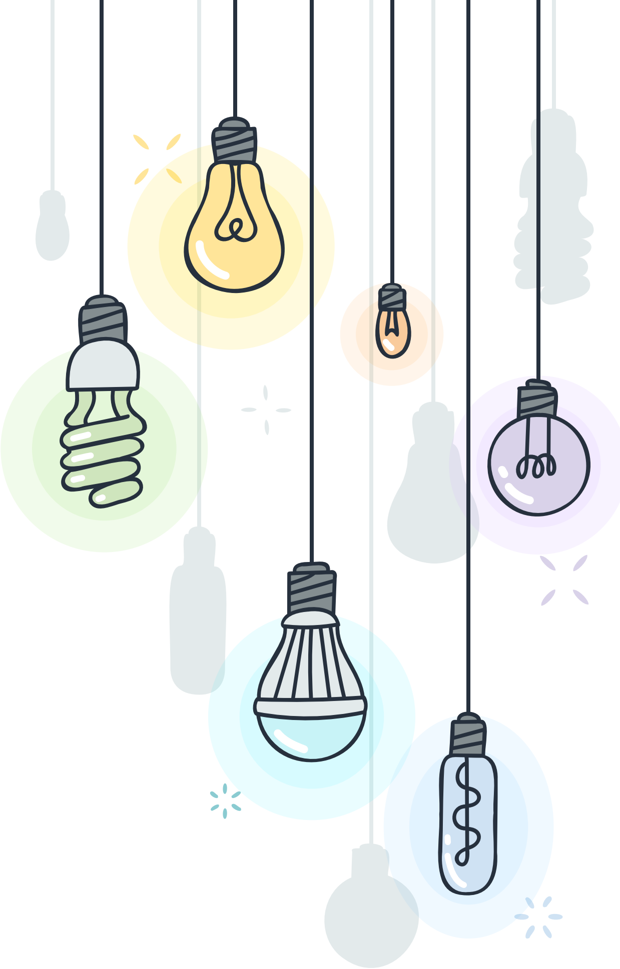


Table of Content

1

Architecture of MLLM

How to design an MLLM?

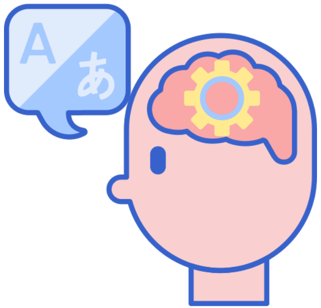
Overview of MLLM Architecture

- ## Preliminary Idea: Intelligence over Language

☞ *Due to the scaling law, 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.*



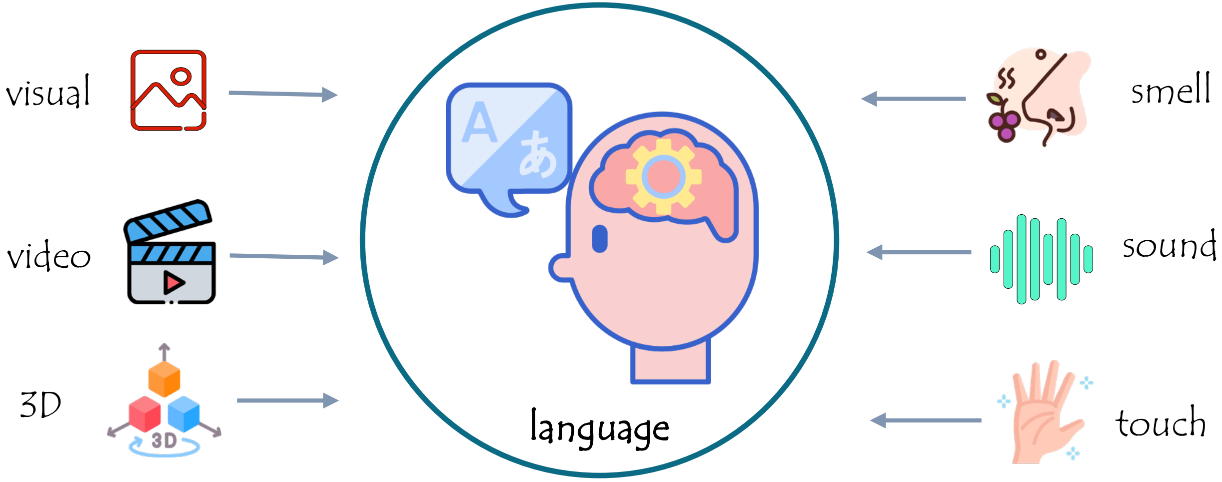
language

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 or encoders, LLMs are enabled with multimodal perceptual/operation abilities.*



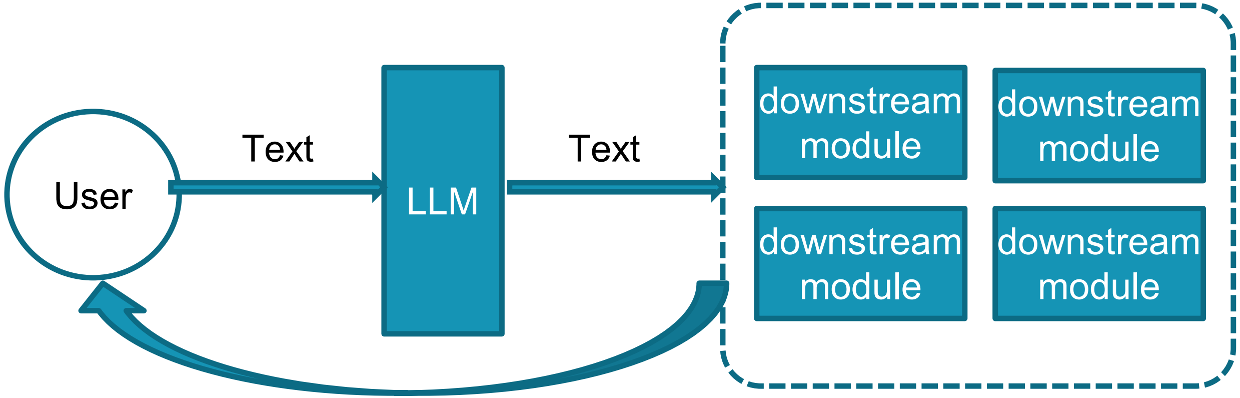
Overview of MLLM Architecture

• 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.*

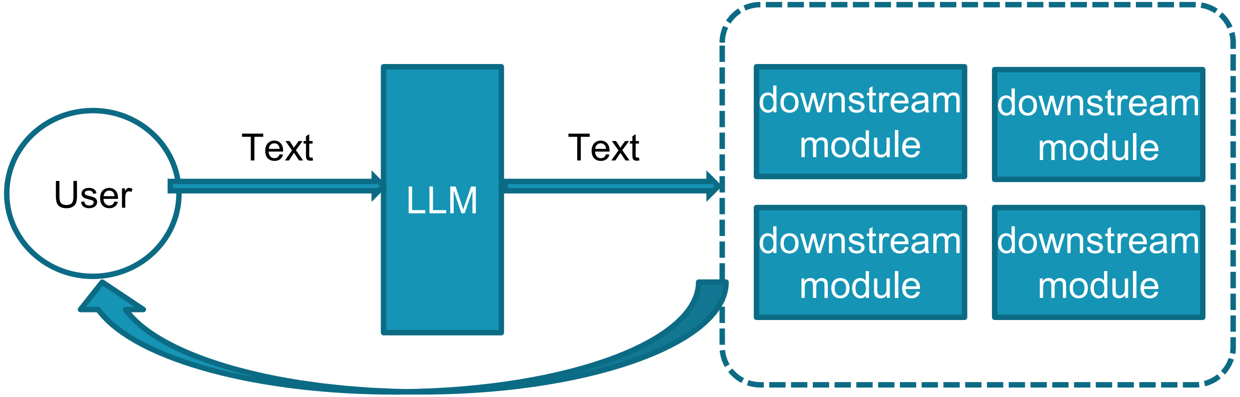
- Key feature:

  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.



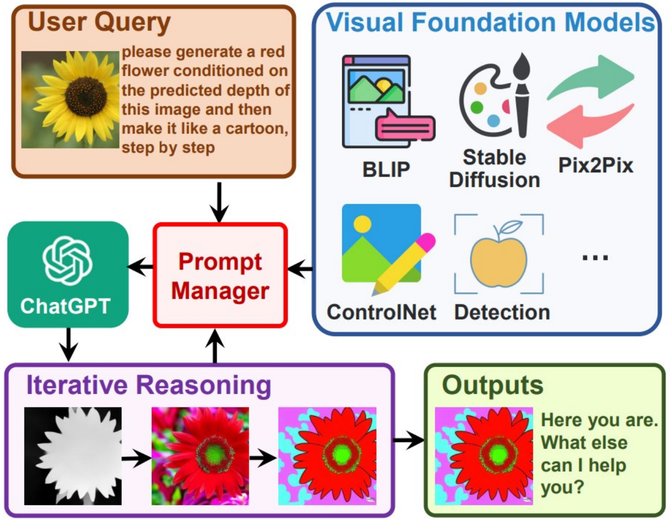
Overview of MLLM Architecture

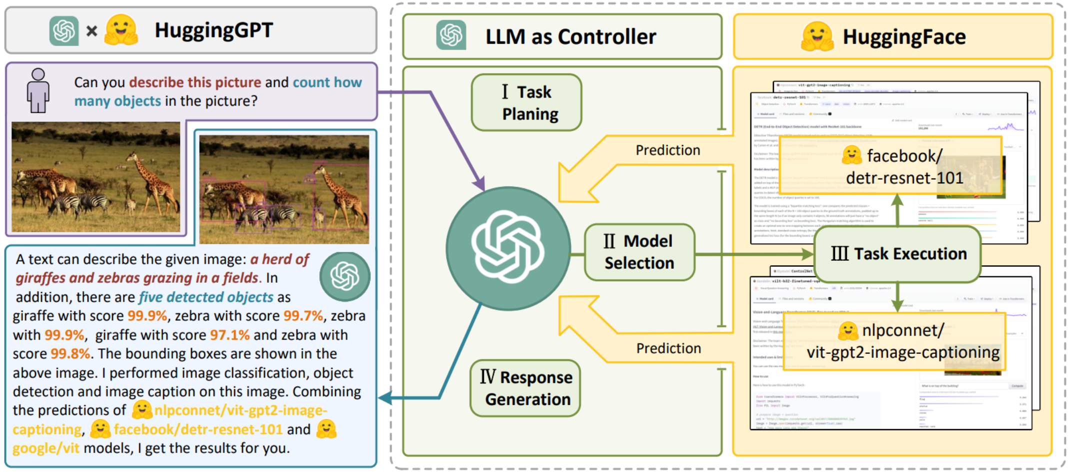
- Architecture-I: LLM as Discrete Scheduler/Controller
  - Representative MLLMs:
    - Visual-ChatGPT
    - HuggingGPT
    - MM-REACT
    - ViperGPT
    - AudioGPT
    - LLaVA-Plus
    - …



Overview of MLLM Architecture

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





[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*

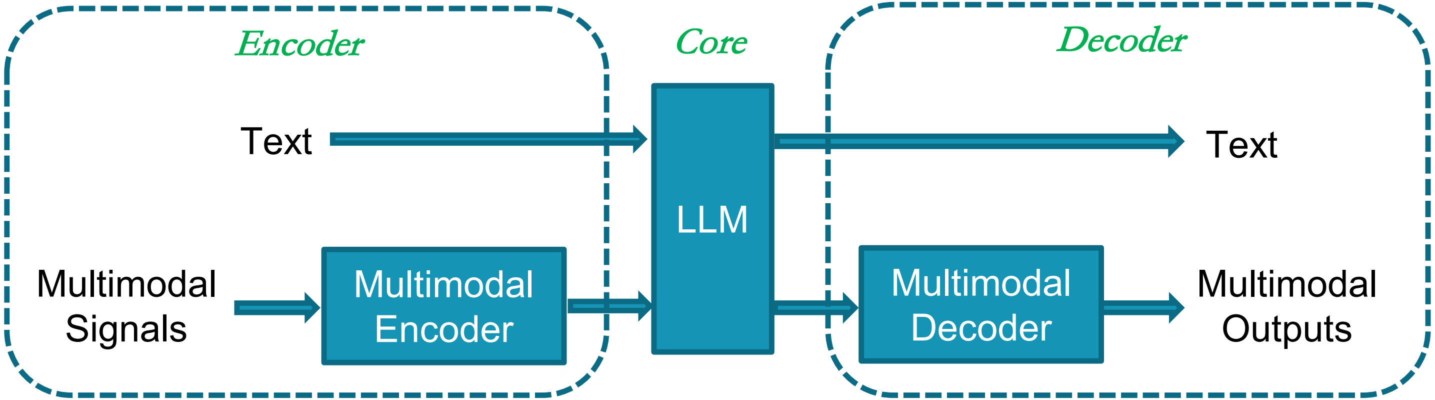
Overview of MLLM Architecture

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.

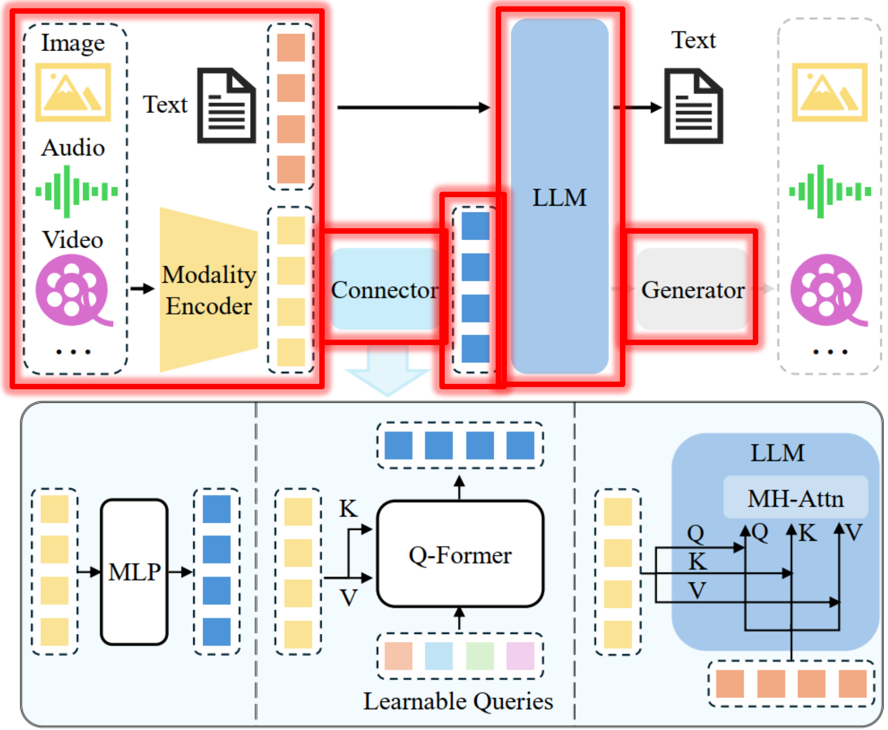


Overview of MLLM Architecture

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

More promising

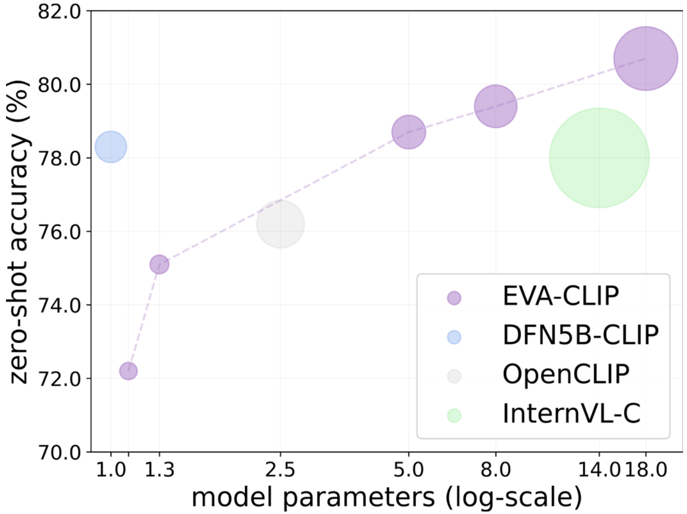
  - ≈ 96% MLLMs belong to this category.



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

Multimodal Encoding

- ## Visual (Image&Video) Encoder
  - **CLIP-ViT** is the most popular choice for vision-language models.
  - Advantages:
    - Providing image representations well aligned with text space.
    - Scale well with respect to parameters and data.



Multimodal Encoding

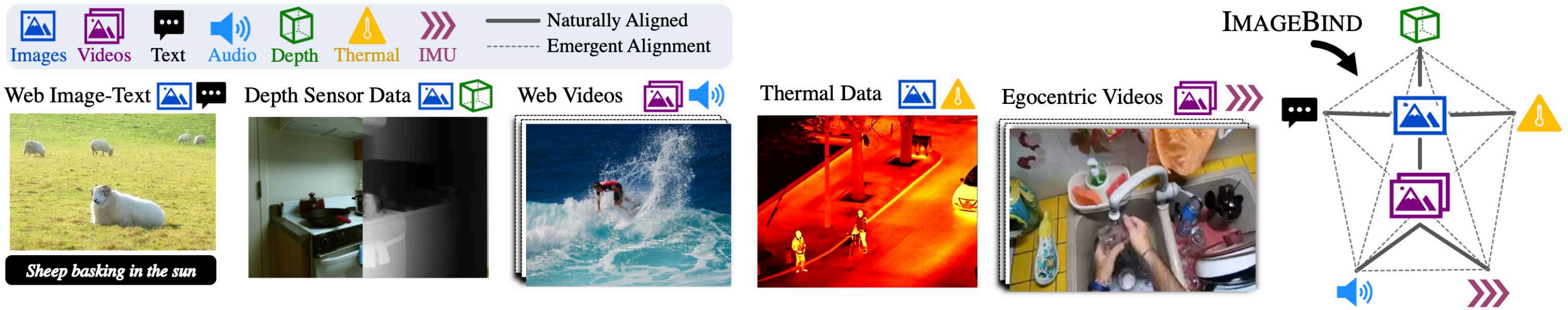
- ## Visual (Image&Video) Encoder
  - **CLIP-ViT** is the most popular choice for vision-language models.
  - Limitations:
    - Fixed low-resolution (224x224 or 336x336) in square shape
  - High-resolution perception is essential, especially for OCR capability!
  - High-resolution Multimodal LLMs
    - Image slice-based: GPT-4V, LLaVA-NeXT, MiniCPM-V 2.0/2.5, LLaVA-UHD, mPLUG-DocOwl 1.5, SPHINX, InternLM-XComposer2-4KHD, Monkey
    - Dual branch encoders: CogAgent, Mini-Gemini, DeepSeek-VL, LLaVA-HR
    - ViT-free: Fuyu, OtterHD

Multimodal Encoding

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

Multimodal Encoding

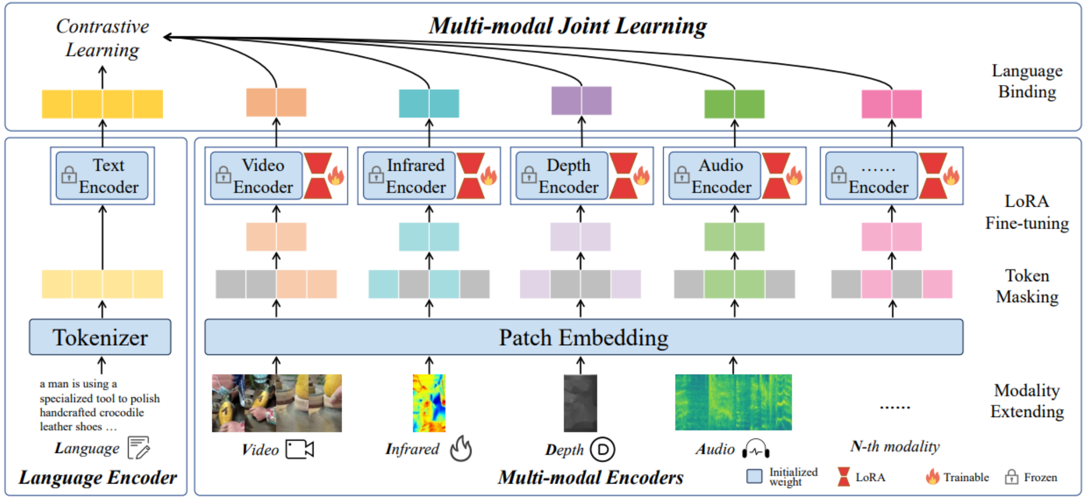
- ## Unified Multimodal Encoder
  - ImageBind:
    - Embedding all modalities into a joint representation space of Image.
    - Well aligned modality representations can benefit LLM understanding



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

Multimodal Encoding

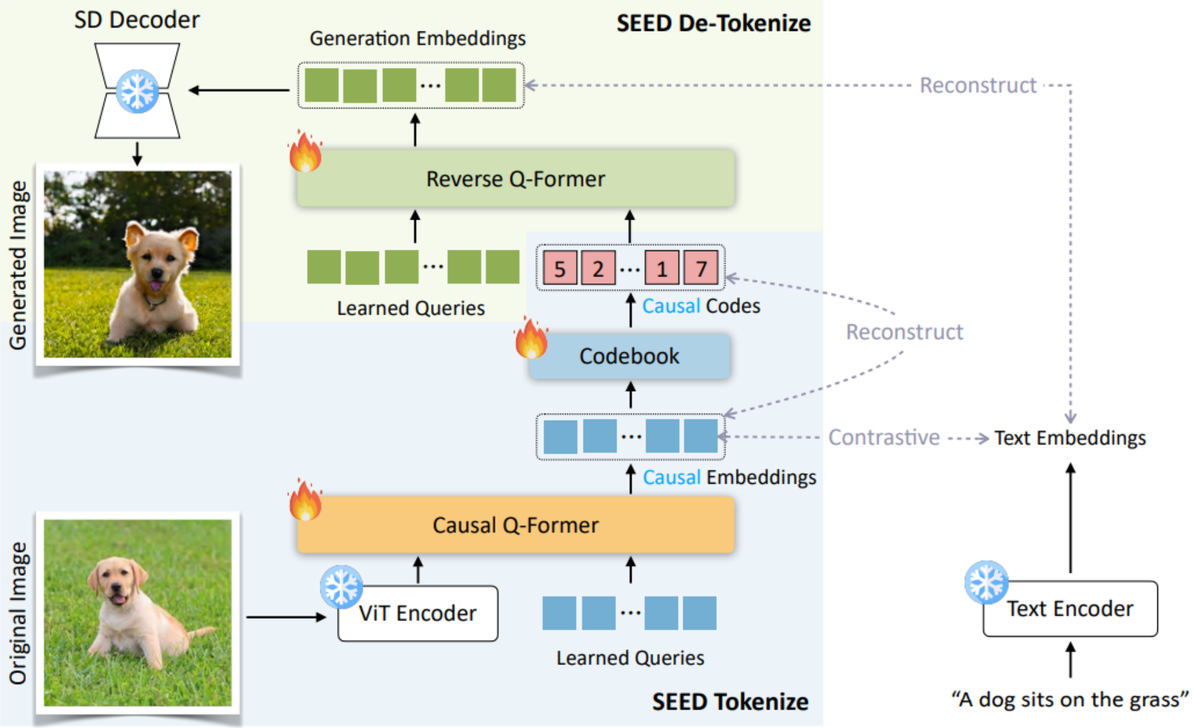
- ## Unified Multimodal Encoder
  - LanguageBind:
    - Embedding all modalities into a joint representation space of Language.
    - Well aligned modality representations can benefit LLM understanding



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

Multimodal Signal Tokenization

- Tokenization
  - SEED

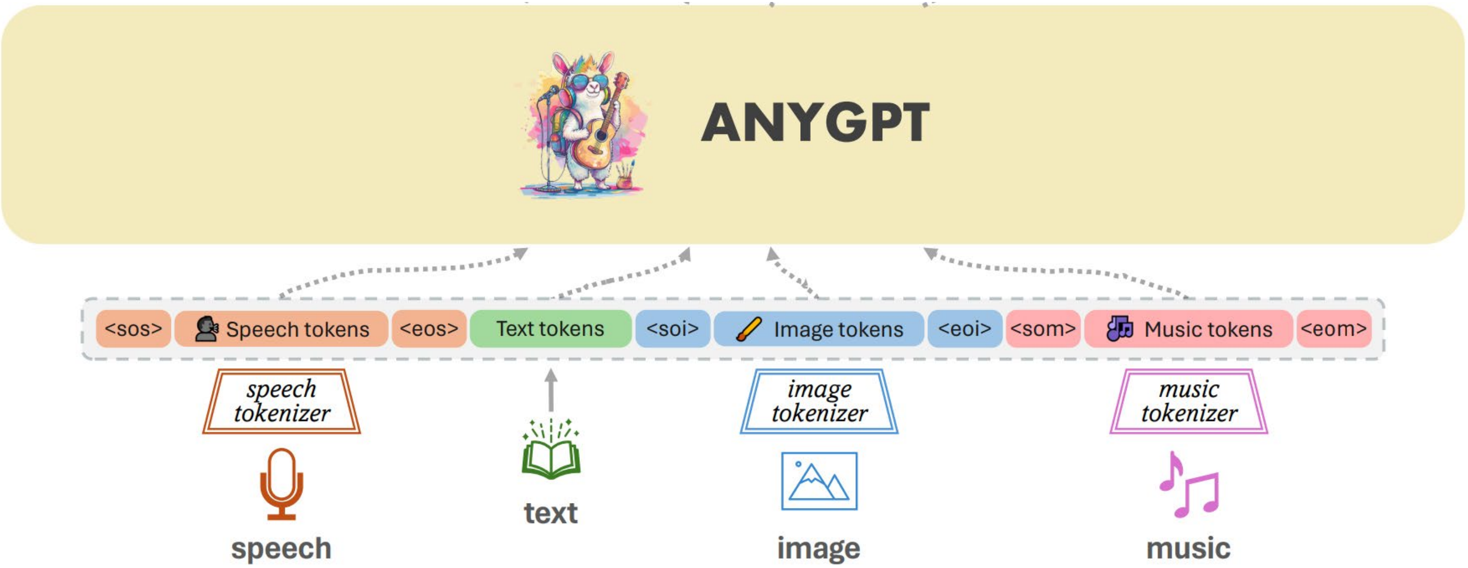


[1] *Planting a SEED of Vision in Large Language Model. 2023*

Multimodal Signal Tokenization

- ## Tokenization

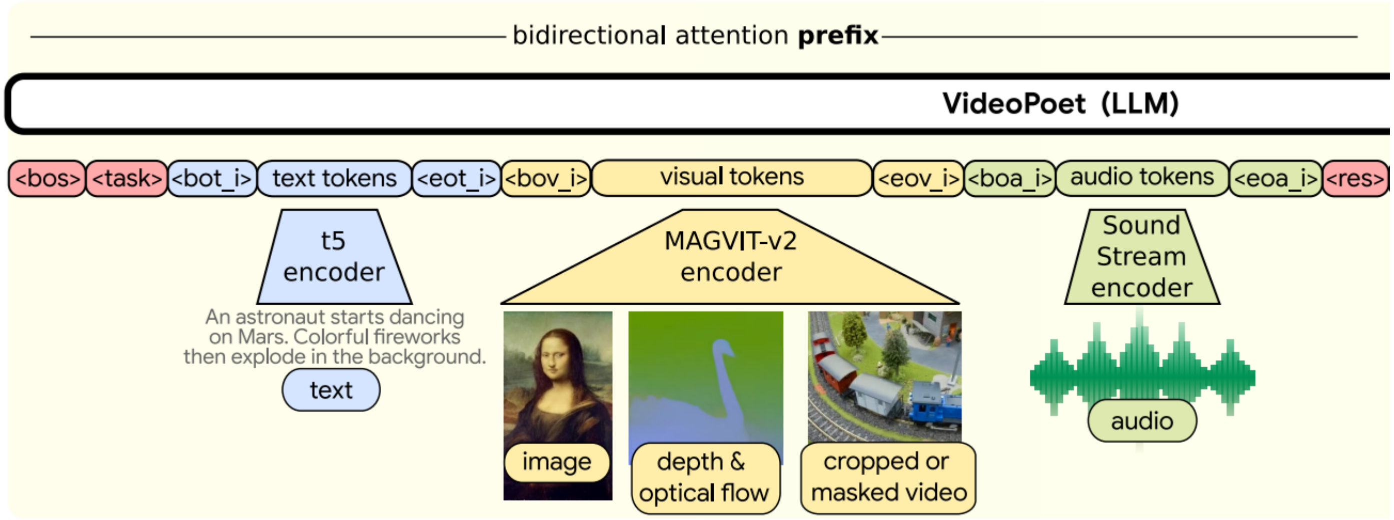
  - AnyGPT



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

Multimodal Signal Tokenization

- ## Tokenization
  - VideoPoet



[1] *VideoPoet: A Large Language Model for Zero-Shot Video Generation. 2023*

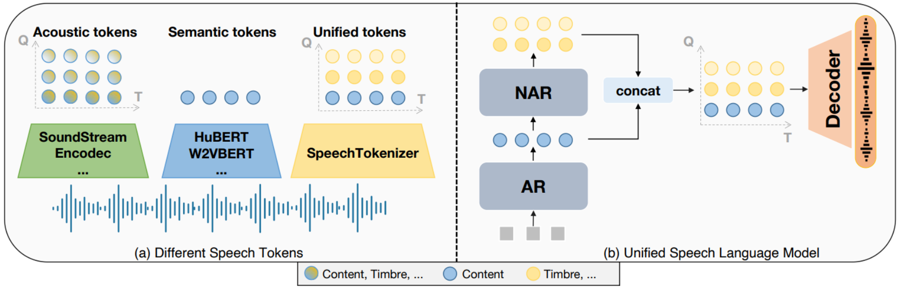
Multimodal Signal Tokenization

- ## Visual (Image&Video) 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

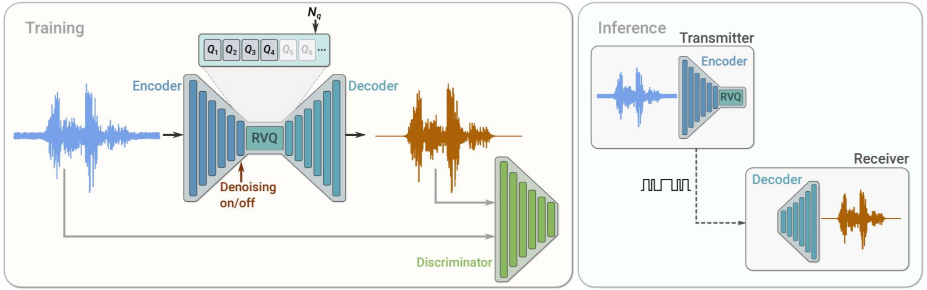
Multimodal Signal Tokenization

Audio Tokenization

- SpeechTokenizer +RVQ-VAE



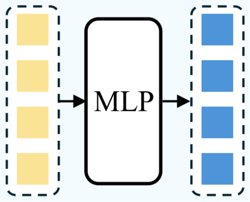
- SoundStream +RVQ-VAE

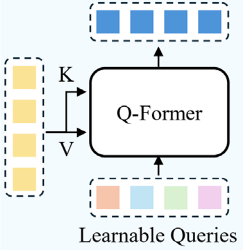


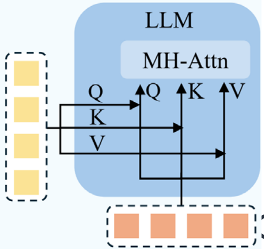
[1] *SpeechTokenizer: Unified Speech Tokenizer for Speech Large Language Models. 2023*
[2] *SoundStream: An End-to-End Neural Audio Codec. 2021*

Input-side Projection

- **Methods to Connect Multimodal Representation with LLM**
  - Projecting multimodal (e.g., image) representations into LLM semantic space
    - 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
    - Q-Former: BLIP-2, InstructBLIP, VisCPM, VisualGLM
    - C-Abstractor: HoneyBee, MM1



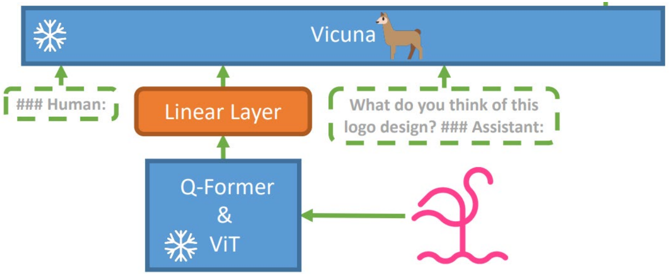




Input-side Projection

- ## Some Insights
  - Different papers have different conclusions about other projection methods.
  - Two-layer MLP is better than linear projection. (LLaVA)
  - Linear projection is more useful than Q-former layers. (MiniGPT-4)

	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	AOK-VQA	GQA
MiniGPT-4	58.2	32.2
(a) MiniGPT-4 w/o Q-Former	56.9	33.4
(b) MiniGPT-4 + 3 Layers	49.7	31.0
(c) MiniGPT-4 + Finetune Q-Former	52.1	28.0

[1] Improved Baselines with Visual Instruction Tuning. 2023
[2] MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. 2021

Backbone LLMs

- ## Open-source Language-based LLMs

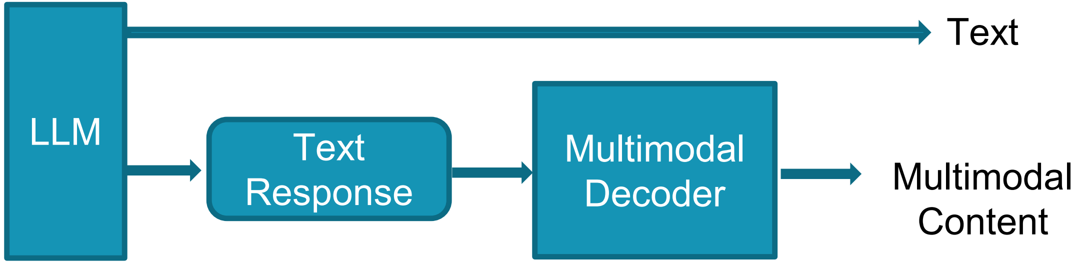
LLM	Size (B)	Data Scale (T)	Date	Language	Architecture
Flan-T5	3/11	-	Oct-2022	en, fr, de	Encoder-Decoder
LLaMA	7/13	1.4	Feb-2023	en	Decoder
Alpaca	7	-	Mar-2023	en	Decoder
Vicuna	7/13	1.4	Mar-2023	en	Decoder
LLaMA-2	7/13	2	Jul-2023	en	Decoder
GLM	2/10	0.4	Oct-2022	en	Decoder
Qwen	1.8/7/14	3	Sep-2023	en, zh	Decoder
Skywork	13	3.2	Oct-2023	en	Decoder

[1] A Survey of Large Language Models. https://github.com/RUCAIBox/LLMSurvey, 2023

Decoding-side Connection

Message passing via 1) discrete token of language

- 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

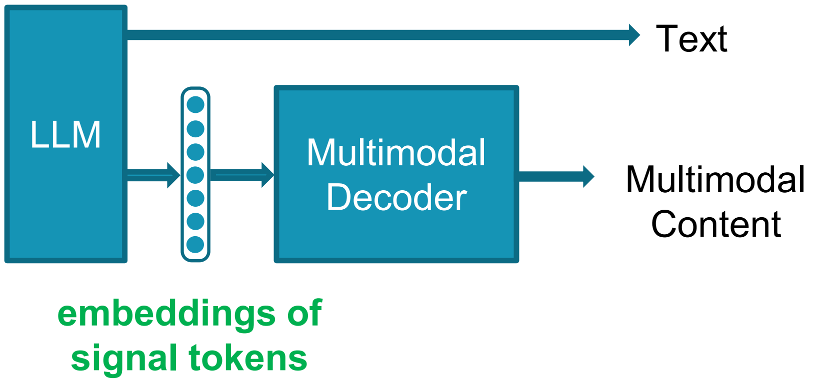
Decoding-side Connection

• 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 visualspatial 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 1) discrete token of language
  - *Instruction: There were originally 7 apples on the table, but I ate one. Then, how many apples are lef now? Please generate a picture to describe the result.*

Response: Sure, I'd be happy to help you understand the process and results.
Based on your statement, "There were originally 7 apples on the table, but I ate one", I inferred that there are now 6 apples left on the table.
To visualize this, I used a text-to-image model to generate an image based on your description. The model I used is called 'runwayml/stable-diffusion-v1-5', which is currently the only available model for this task. The image generated can be found at the following path: '/images/ec4d.jpg'.
Please note that the image is a representation of the text and may not be 100% accurate. If there's anything else you need, feel free to ask!

HuggingGPT

{*Intermedia Caption:* 6 apples left on the table}

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

Decoding-side Connection

- ## Message passing via 2) continuous embedding
  - *Instruction: There were originally 7 apples on the table, but I ate one. Then, how many apples are lef now? Please generate a picture to describe the result.*

Response: Certainly! In the end, you have the following outcomes: 6 apples on the table.

NExT-GPT

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

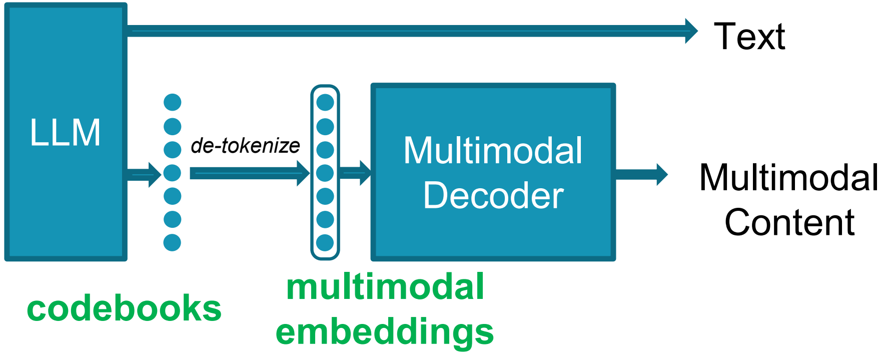
Decoding-side Connection

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



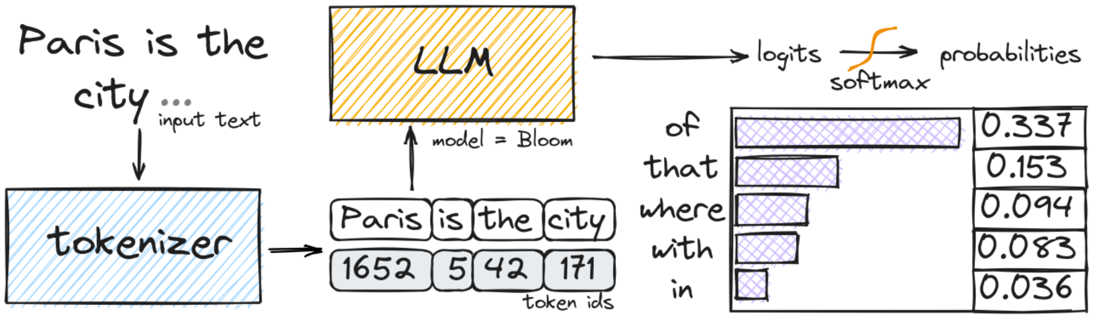
[1] Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action. 2023
[2] LVM: Sequential Modeling Enables Scalable Learning for Large Vision Models. 2023
[3] AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. 2024
[4] VideoPoet: A Large Language Model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

Multimodal Generation

- ## Text Generation

  - LLMs naturally support direct text generation

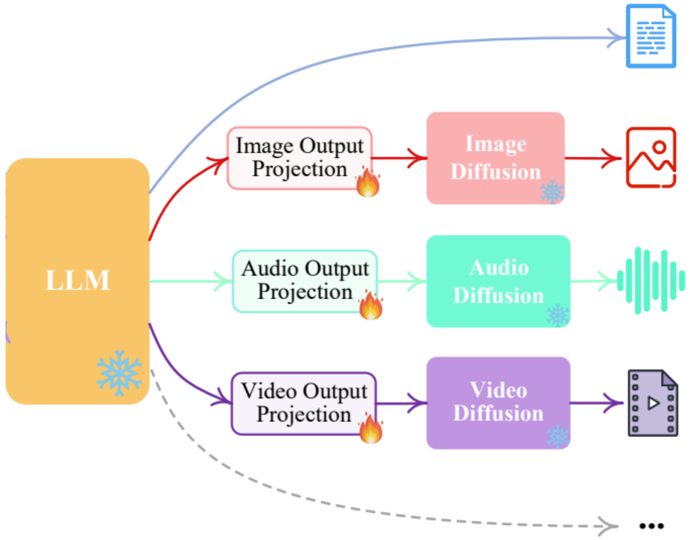
    via e.g., BPE decoding, Beam search, …



Multimodal Generation

- ## Generation via Diffusion Models

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

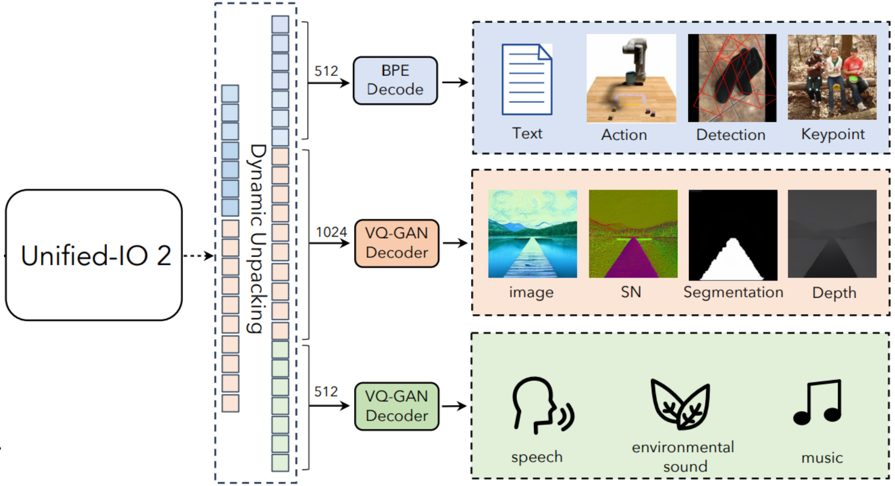


[1] NExT-GPT: Any-to-Any Multimodal LLM. 2023

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*

Multimodal Generation

Generation via Codebooks

- VQ-GAN in Stable-diffusion
  - 64 × 64 × 3 or 32 × 32 × 4

Encoder	Decoder
$x \in \mathbb{R}^{H \times W \times C}$	$z_{\mathbf{q}} \in \mathbb{R}^{h \times w \times n_z}$
Conv2D $\to \mathbb{R}^{H \times W \times C'}$	Conv2D $\to \mathbb{R}^{h \times w \times C''}$
$m\times$ { Residual Block, Downsample Block} $\to \mathbb{R}^{h \times w \times C''}$	Residual Block $\to \mathbb{R}^{h \times w \times C''}$
Residual Block $\to \mathbb{R}^{h \times w \times C''}$	Non-Local Block $\to \mathbb{R}^{h \times w \times C''}$
Non-Local Block $\to \mathbb{R}^{h \times w \times C''}$	Residual Block $\to \mathbb{R}^{h \times w \times C''}$
Residual Block $\to \mathbb{R}^{h \times w \times C''}$	$m\times$ { Residual Block, Upsample Block} $\to \mathbb{R}^{H \times W \times C'}$
GroupNorm, Swish, Conv2D $\to \mathbb{R}^{h \times w \times n_z}$	GroupNorm, Swish, Conv2D $\to \mathbb{R}^{H \times W \times C}$

Table 7. High-level architecture of the encoder and decoder of our *VQGAN*. The design of the networks follows the architecture presented in [25] with no skip-connections. For the discriminator, we use a patch-based model as in [28]. Note that $h = \frac{H}{2^m}$, $w = \frac{W}{2^m}$ and $f = 2^m$.

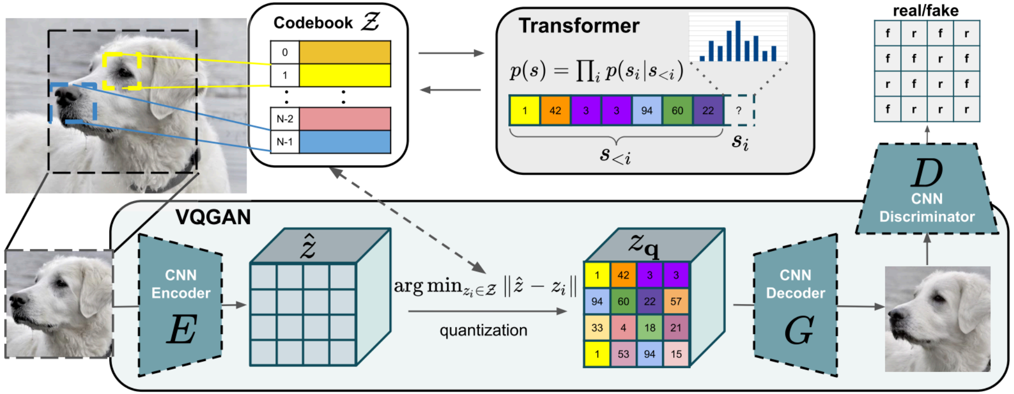


Figure 2. Our approach uses a convolutional *VQGAN* to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer based high resolution image synthesis.

Model	Stage-1 (latent space learning)	Latent Space	Stage-2 (prior learning)
VQ-VAE	VQ-VAE	Discrete (after quantization)	Autoregressive PixelCNN
VQGAN	VQGAN (VQ-VAE + GAN + Perceptual Loss)	Discrete (after quantization)	Autoregressive GPT-2 (Transformer)
VQ-Diffusion	VQ-VAE	Discrete (after quantization)	Discrete Diffusion
Latent Diffusion (VQ-reg)	VAE or VQGAN	Continuous (before quantization)	Continuous Diffusion

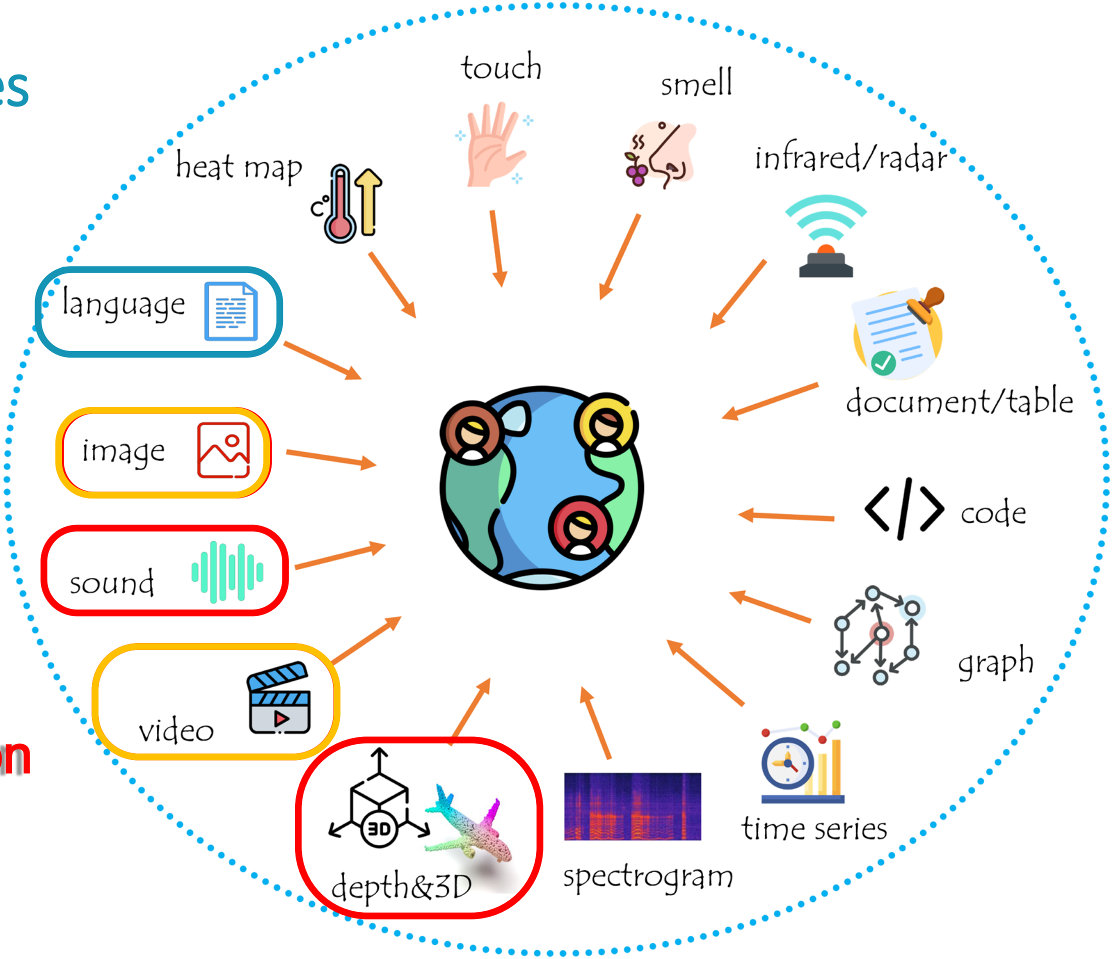
2

Modality and Functionality

What are MLLMs capable of?

Overview of Modality and Functionality

- **Modalities**



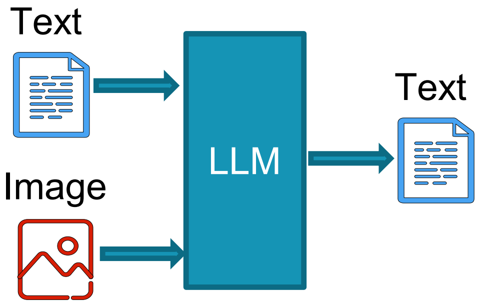
Language+ Vision

Overview of Modality and Functionality

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

Multimodal Perceiving

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



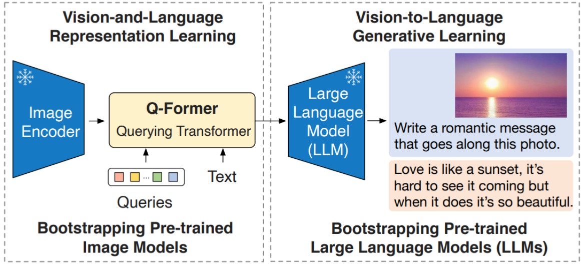
☞ *Encode input images with external image encoders, generating LLM-understandable visual feature, which is then fed into the LLM. LLM then interprets the input images based on the input text instructions and produces a textual response.*

[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
…

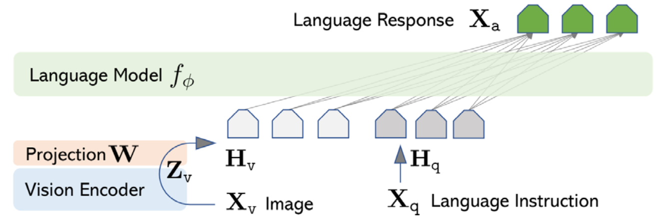
Multimodal Perceiving

Image-perceiving MLLM

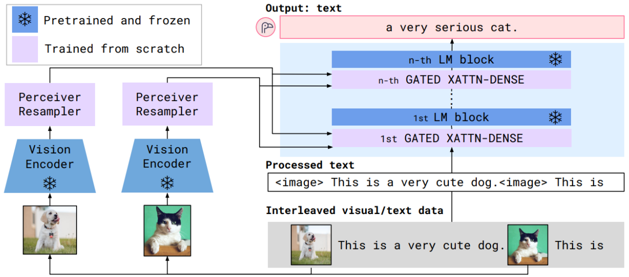
- Blip2



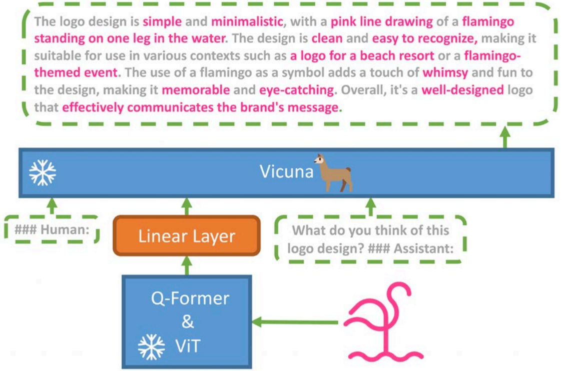
- LLaVA



- Flamingo



- Mini-GPT4

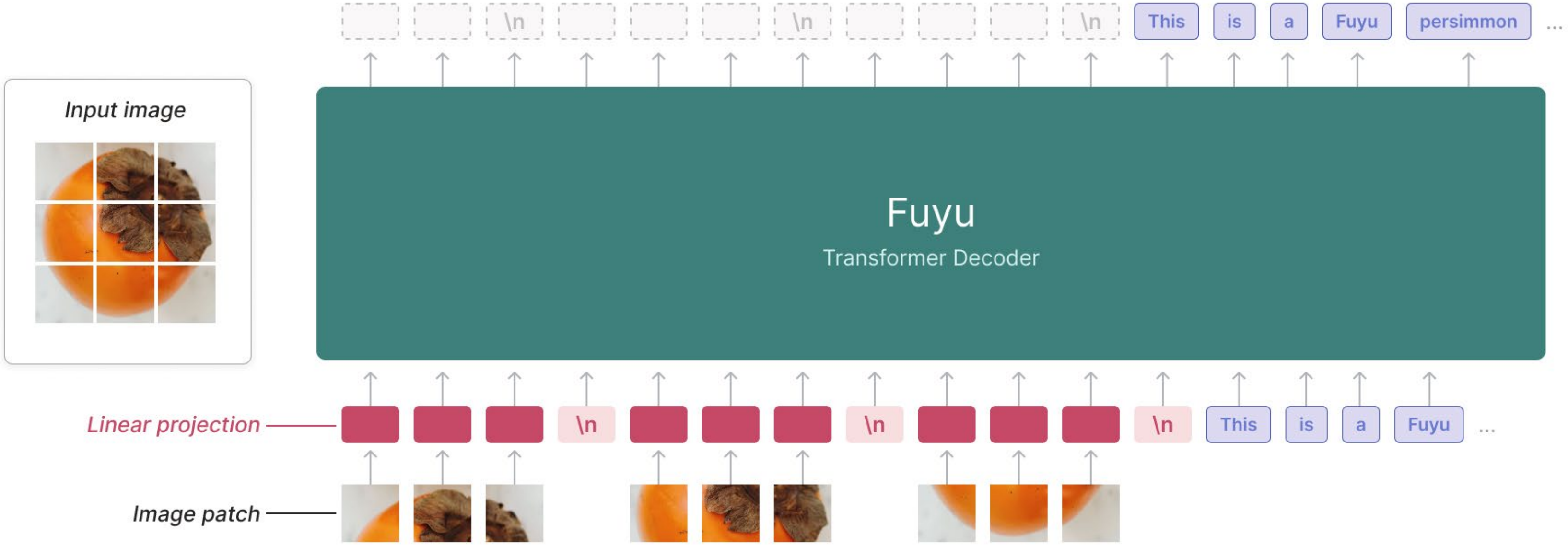


[1] *Flamingo: a Visual Language Model for Few-Shot Learning. 2022*
[2] *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. 2023*
[3] *Visual Instruction Tuning. 2023*
[4] *A Survey on Multimodal Large Language Models. https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models, 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.*

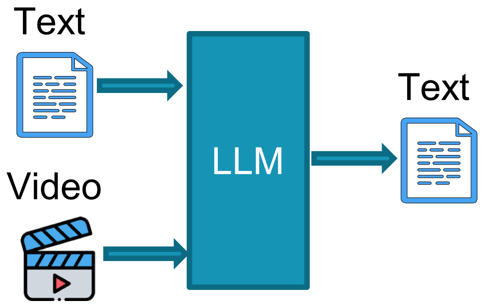


[1] Fuyu-8B. https://www.adept.ai/blog/fuyu-8b, 2023.

Multimodal Perceiving

- ## Video-perceiving MLLM

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



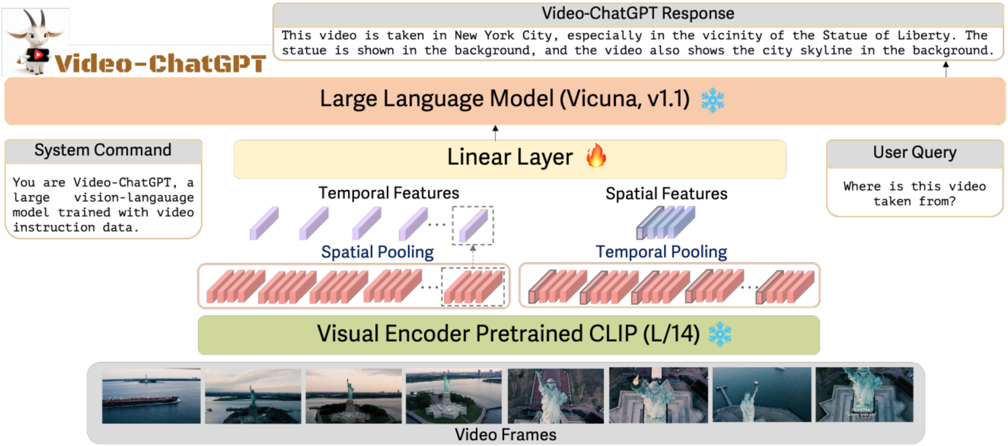
☞ *Encode input videos with external video encoders, generating LLM-understandable visual feature, feeding into LLM, which then interprets the input videos based on the input text instructions and produces a textual response.*

[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
…

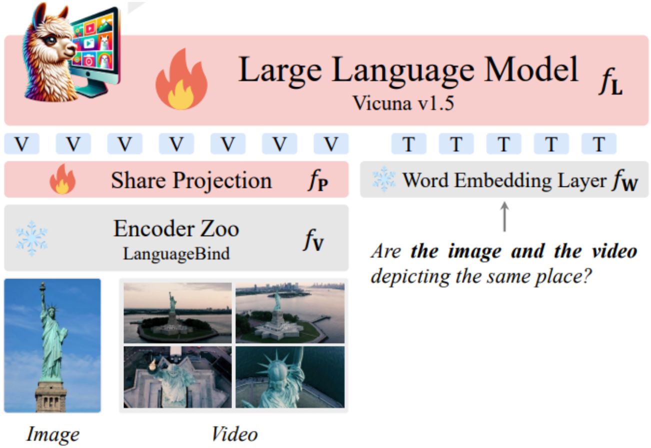
Multimodal Perceiving

Video-perceiving MLLM

Video-ChatGPT



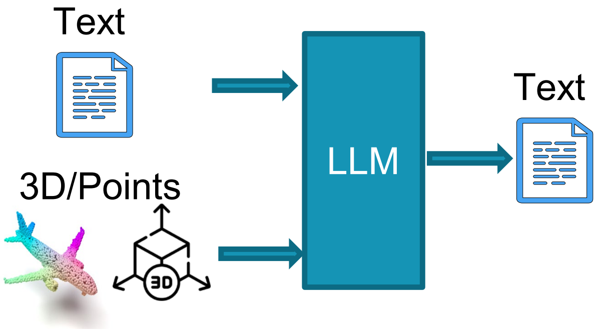
Video-LLaVA



[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
  - …



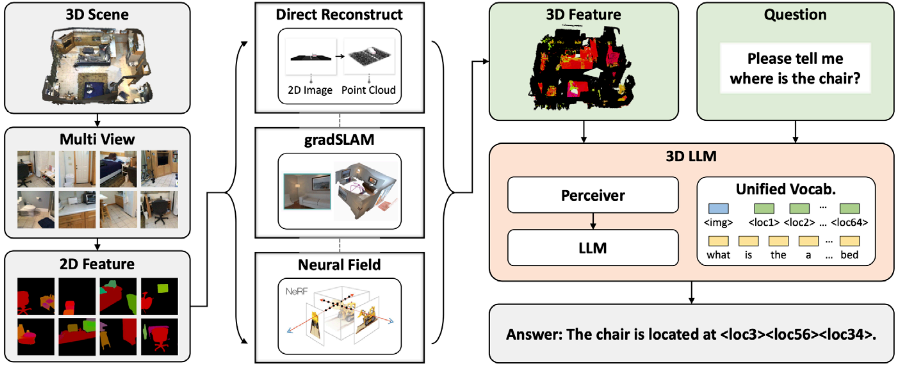
Encode input 3D information with external encoders, generating LLM-understandable 3D feature, feeding into LLM, which then interprets the input 3D/points based on the input text instructions and produces a textual response.

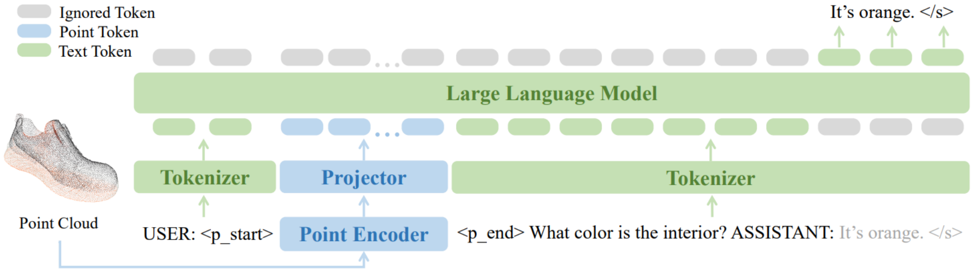
[1] 3D-LLM: Injecting the 3D World into Large Language Models. 2023
[2] 3D-GPT: Procedural 3D Modeling with Large Language Models. 2023
[3] LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning. 2023
[4] PointLLM: Empowering Large Language Models to Understand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

Multimodal Perceiving

- ## 3D-perceiving MLLM

  - ### 3D-LLM
  - ### 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

  - AudioGPT,
  - SpeechGPT,
  - VIOLA,
  - AudioPaLM
  - SALMONN
  - MU-LLaMA
  - …

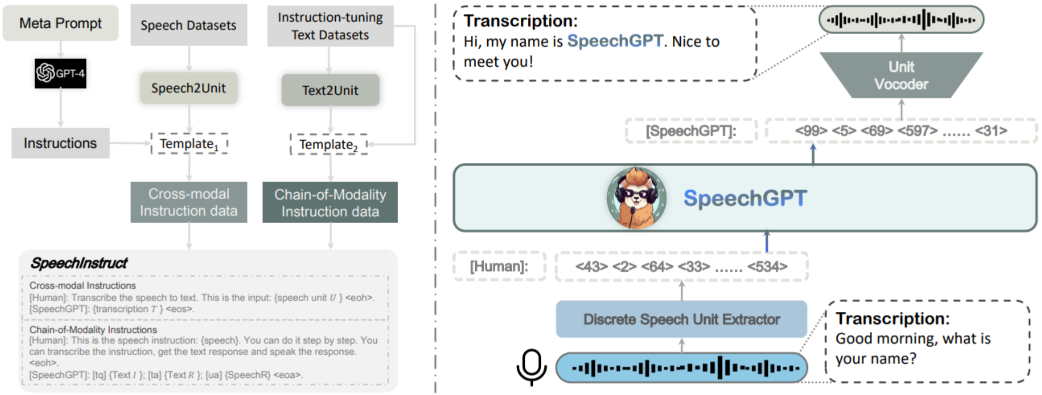
Encode input audio signals with external encoders, generating LLM-understandable signal features, feeding into LLM, which then interprets the audio based on the input text instructions and produces a textual response.

[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
…

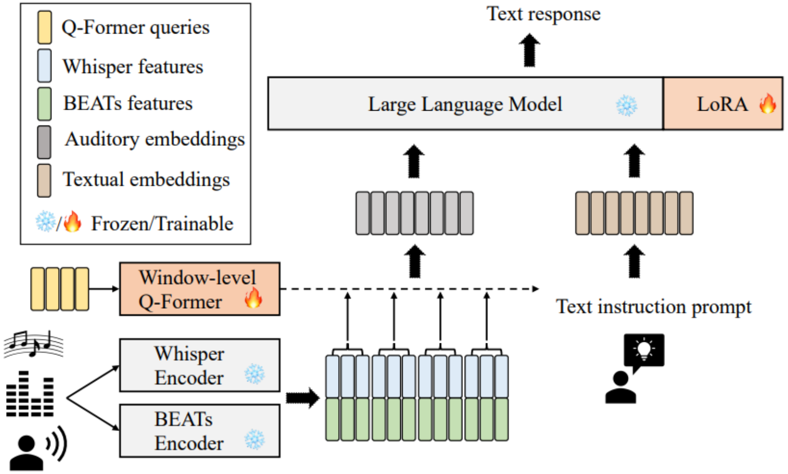
Multimodal Perceiving

Audio-perceiving MLLM

SpeechGPT



SALMONN



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

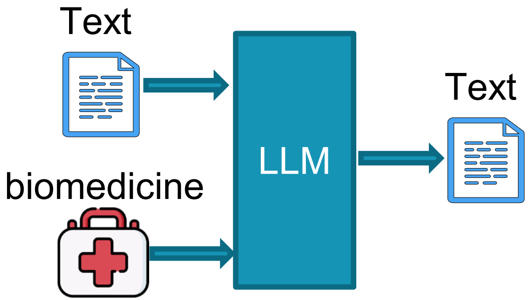
[2] *SALMONN: Towards Generic Hearing Abilities for Large Language Models. 2023*

[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

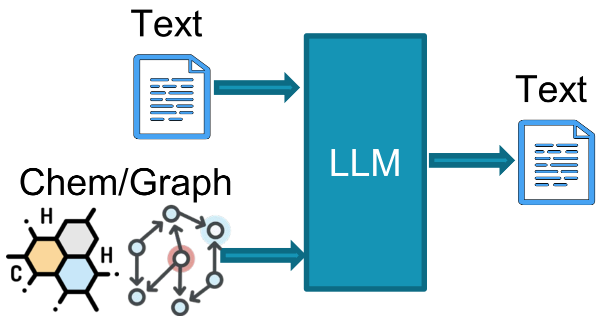


- BioGPT
- DrugGPT
- BioMedLM
- OphGLM
- GatorTron
- GatorTronGPT
- MEDITRON
- DoctorGLM
- BianQue
- ClinicalGPT
- Qilin-Med
- ChatDoctor
- BenTsao
- HuatuoGPT
- MedAlpaca
- AlpaCare
- Zhongjing
- PMC-LLaMA
- CPLLM
- MedPaLM 2
- BioMedGPT

[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
    - MolXPT
    - MolSTM
    - 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

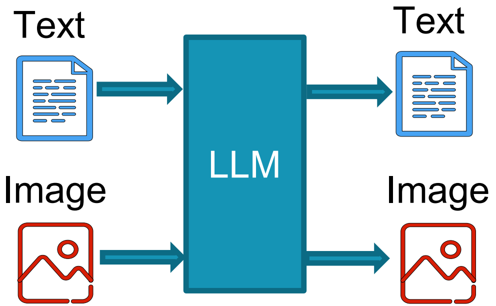
- ## Scenarios

☞ *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
- …

Unified MLLM: Perceiving + Generation

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



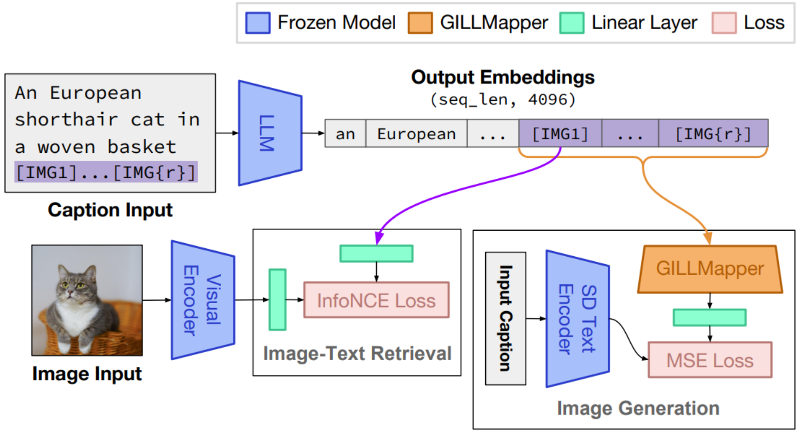
☞ *Central LLMs take as input both texts and images, after semantics comprehension, and generate both texts and images.*

[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
...

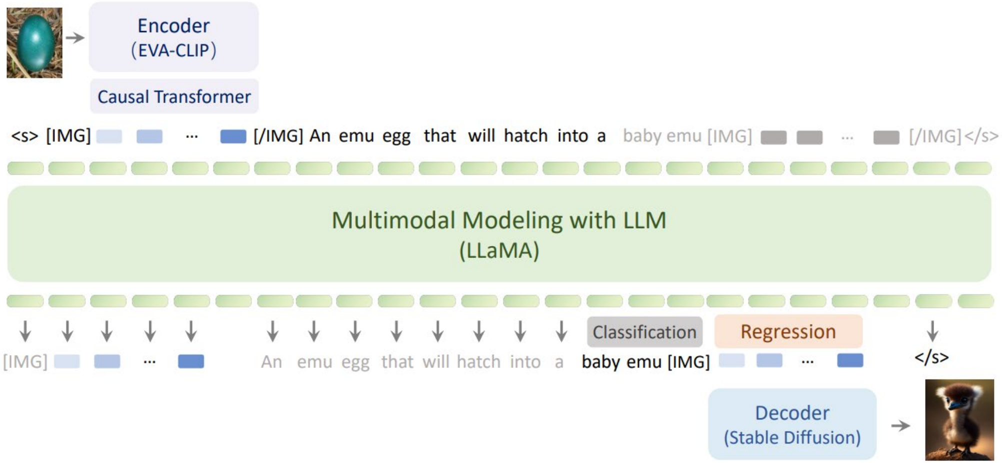
Unified MLLM: Perceiving + Generation

- Image

  - GILL



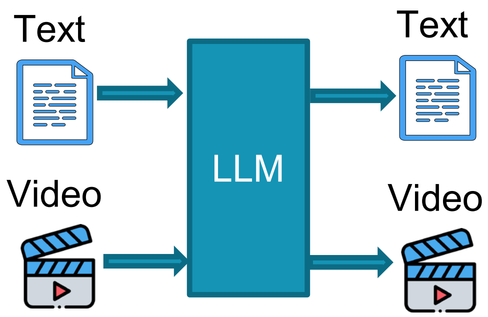
  - EMU



[1] Generating Images with Multimodal Language Models. 2023
[2] Generative Pretraining in Multimodality. 2023

Unified MLLM: Perceiving + Generation

- Video
  - GPT4Video
  - VideoPoet
  - Video-LaVIT
  - …



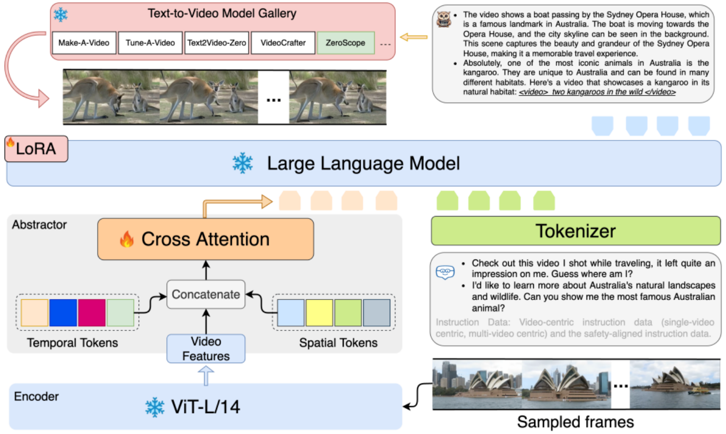
Central LLMs take as input both texts and videos, after semantics comprehension, and generate both texts and videos.

[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
…

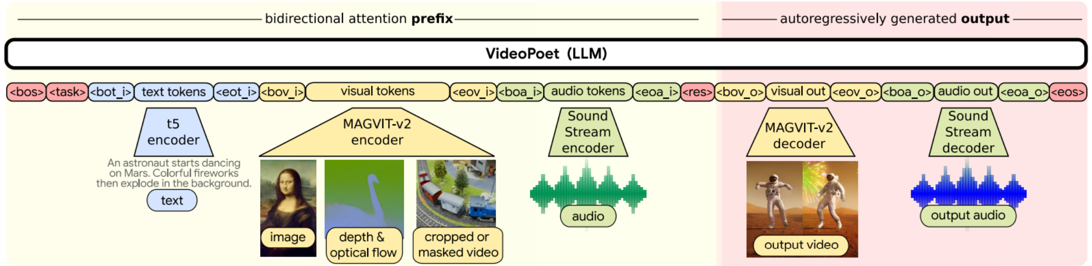
Unified MLLM: Perceiving + Generation

- Video

  - GPT4Video



  - VideoPoet

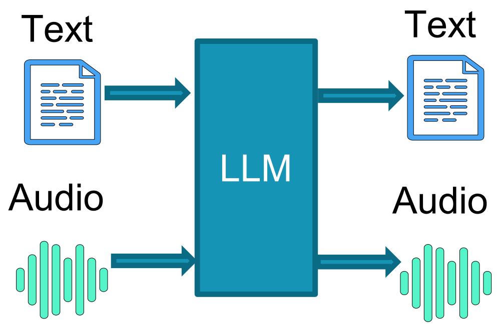


[1] *GPT4Video: A Unified Multimodal Large Language Model for Instruction-Followed Understanding and Safety-Aware Generation. 2023*

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

Unified MLLM: Perceiving + Generation

- Audio
  - AudioGPT,
  - SpeechGPT,
  - VIOLA,
  - AudioPaLM,
  - …



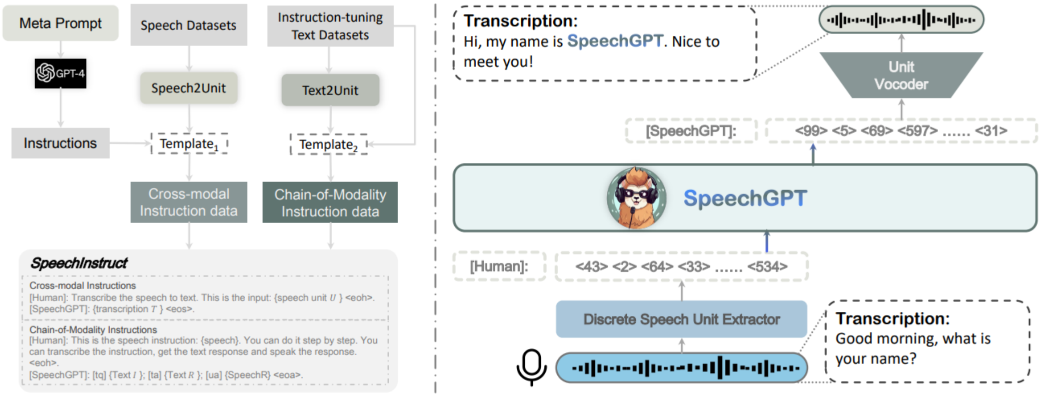
☞ *Central LLMs take as input both texts and audio, after semantics comprehension, and generate both texts and audio.*

[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
…

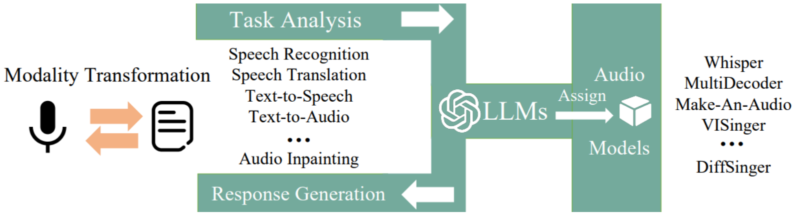
Unified MLLM: Perceiving + Generation

- ## 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

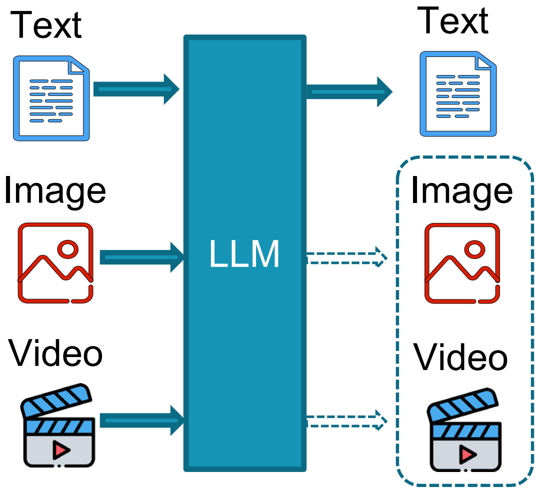
- ## Scenarios:

☞ *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
  - …

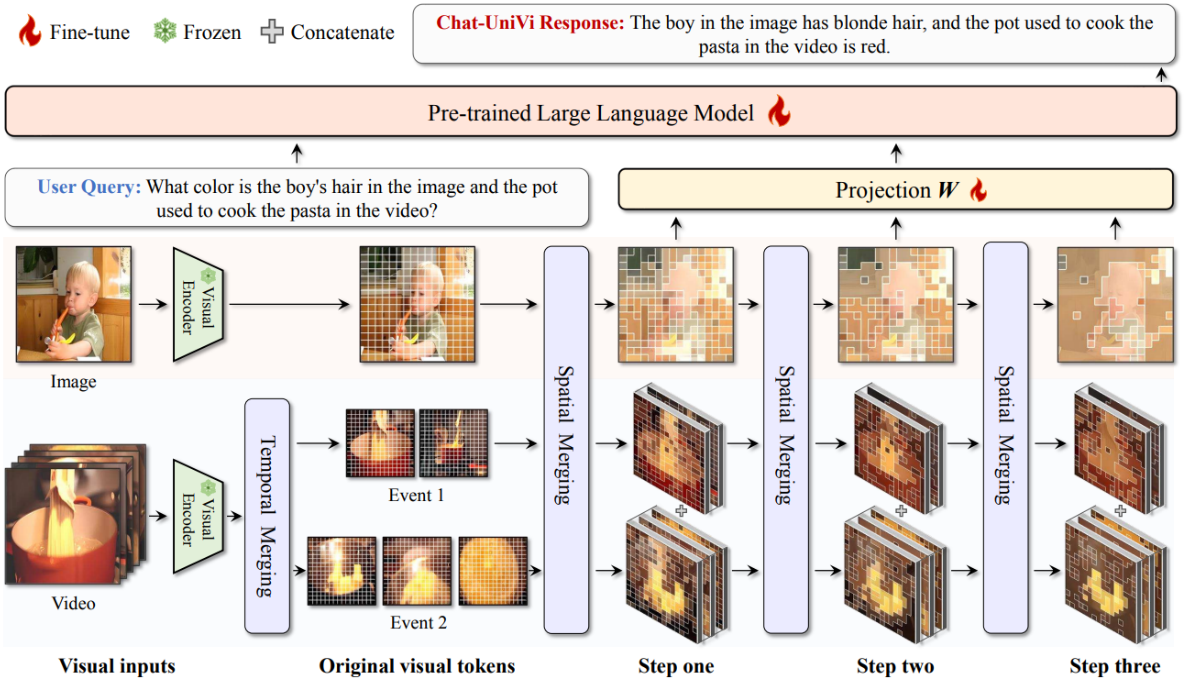


Central LLMs take as input texts, image and video, after semantics comprehension, and generate texts (maybe also image and video, or combination).

[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

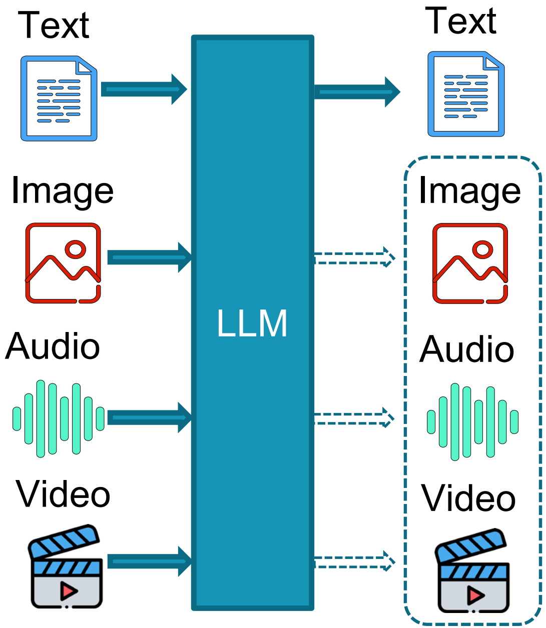
- ## Text+Image+Video
  - Chat-UniVi



[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
  - ...

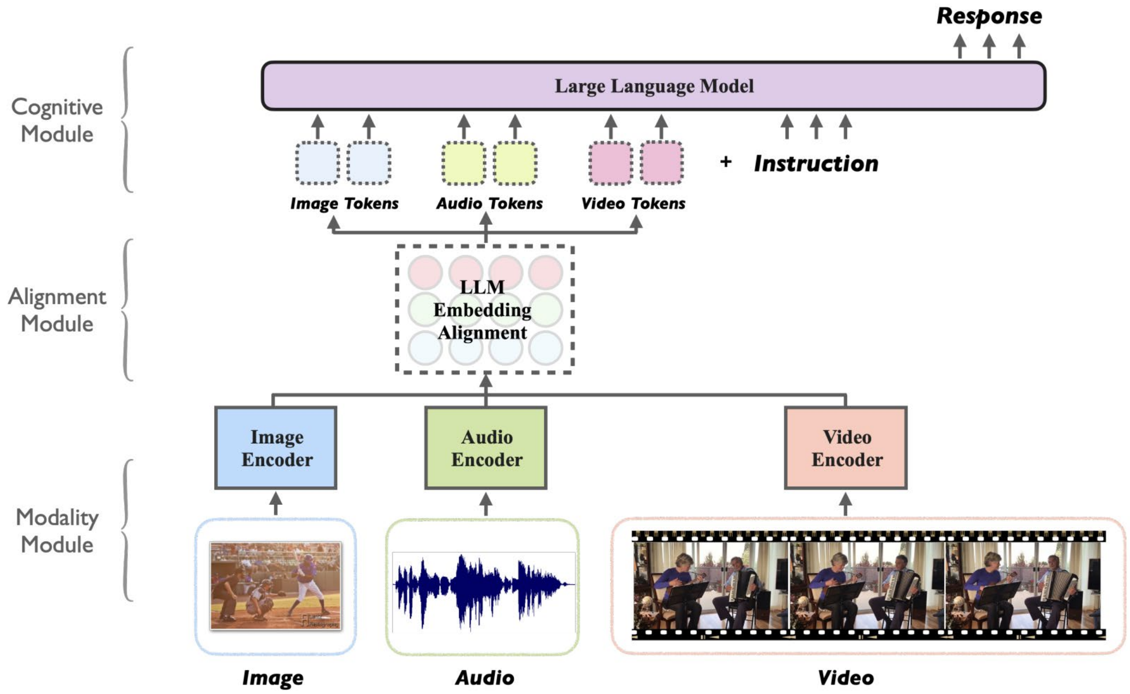


Central LLMs take as input texts, audio, image and video, and generate texts (maybe also audio, image and video, or combination).

[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

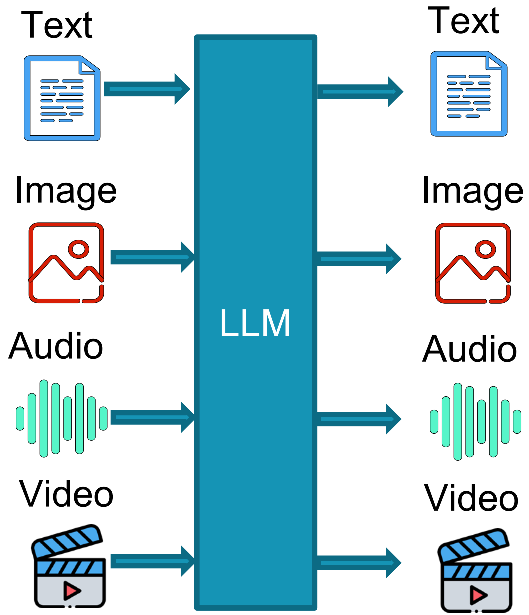
- ## Text+Image+Video+Audio
  - Macaw-LLM



[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
  - …

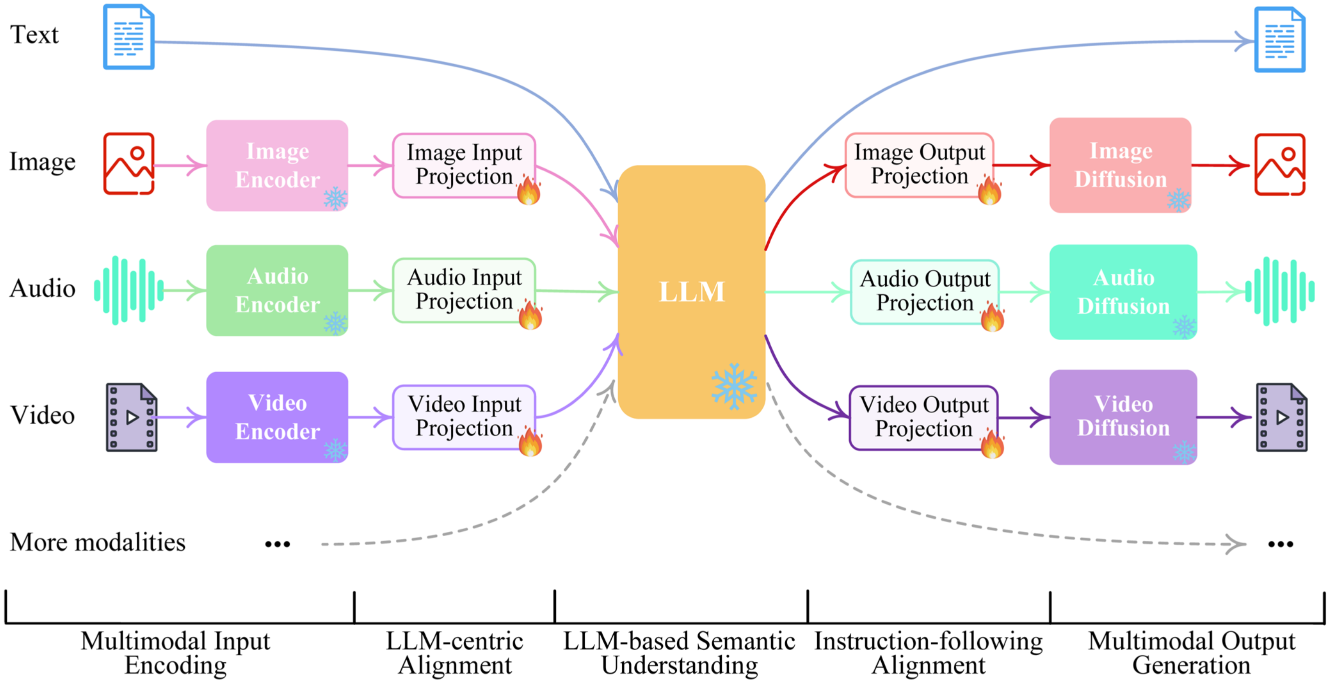


☞ *Central LLMs take as input texts, audio, image and video, and freely generate texts, audio, image and video, or combination.*

[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
  - NExT-GPT



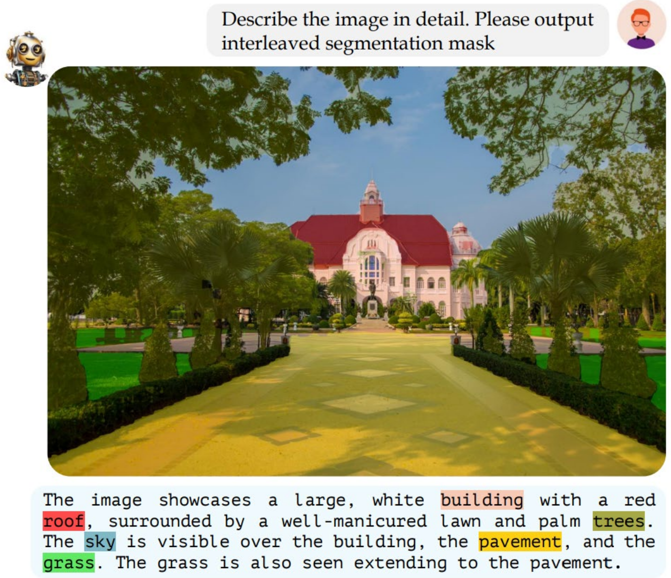
[1] NExT-GPT: Any-to-Any Multimodal LLM. 2023

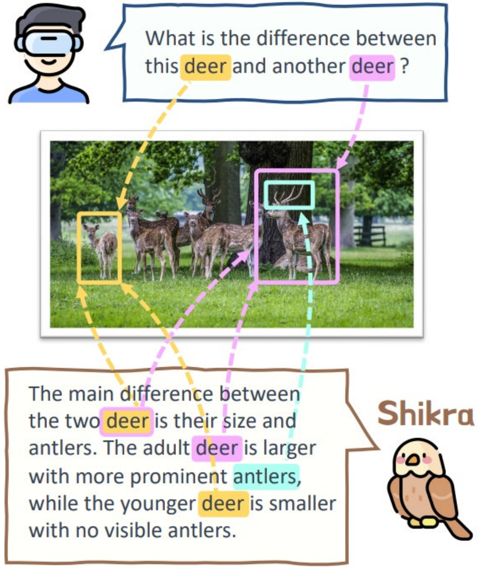
Fine-grained Capability of MLLM

• 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**.*

- Visual Grounding
- Visual Segmentation
- Visual Editing
- Visual Inpainting
- …



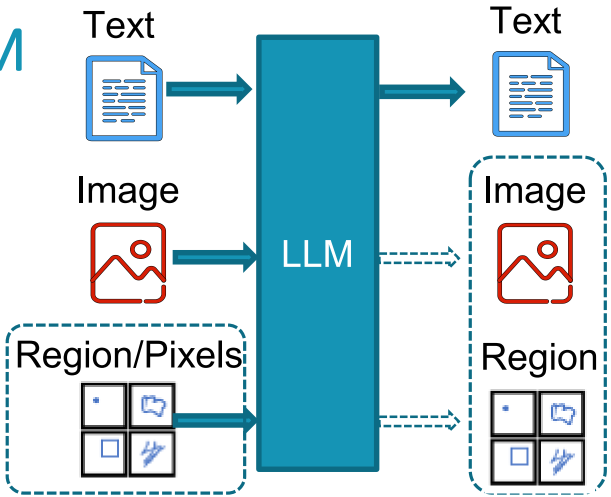


Fine-grained Capability of MLLM

Image-oriented Pixel-wise Regional MLLM

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

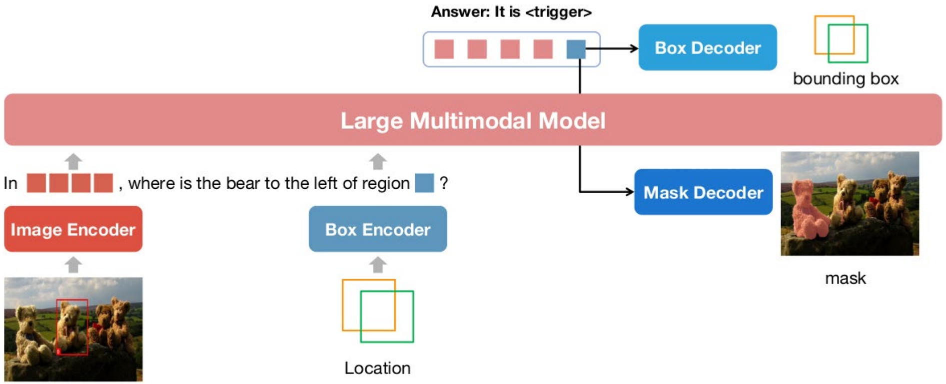
Users input an image (potentially specifying a region), and the LLM outputs content based on its understanding, grounding the visual content to specific pixel-level regions of the image.



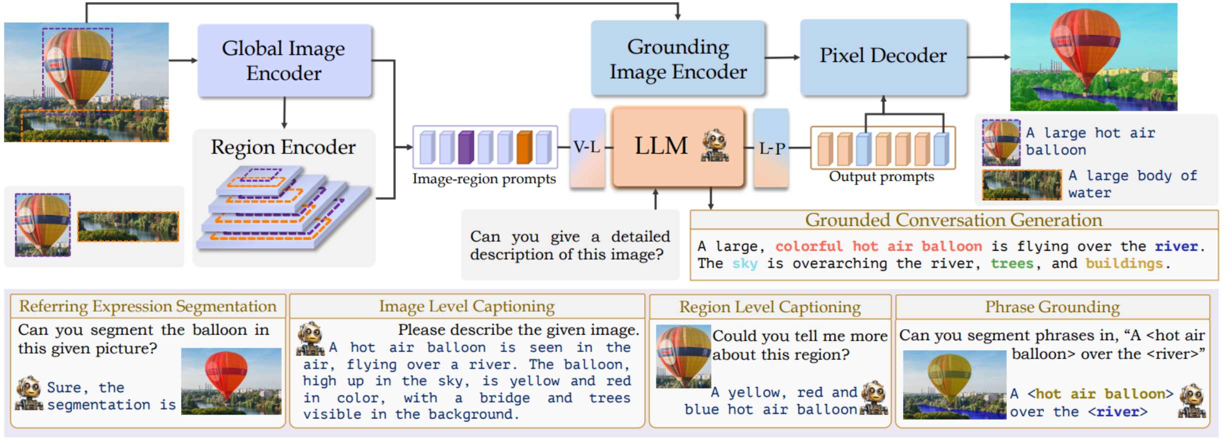
[1] GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. 2023
[2] NExT-Chat: An LMM for Chat, Detection and Segmentation. 2023
[3] MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. 2023
[4] Osprey: Pixel Understanding with Visual Instruction 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
...

Fine-grained Capability of MLLM

- Image-oriented Pixel-wise
  - NExT-Chat



  - GLaMM



Fine-grained Capability of MLLM

- Image-oriented Pixel-wise Regional MLLM

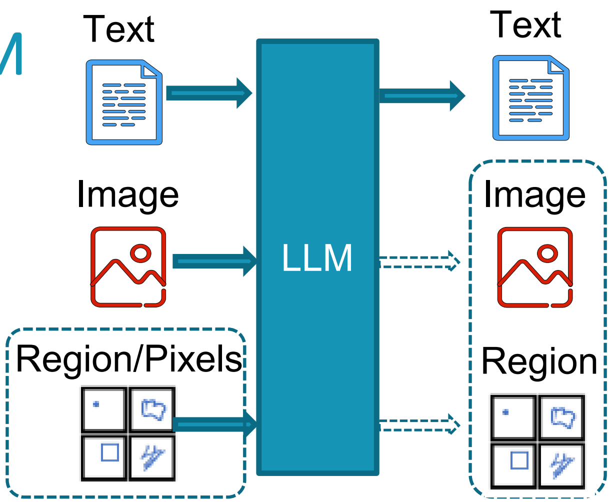
☞ Pixel-level Awareness at Input/Output

- Output-side Only Pixel-wise Awareness

  LISA, PixelLM, DetGPT, MiniGPT-v2, LION

- Input-&Ouput-side Pixel-wise Awareness

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



Fine-grained Capability of MLLM

- Image-oriented Pixel-wise Regional MLLM

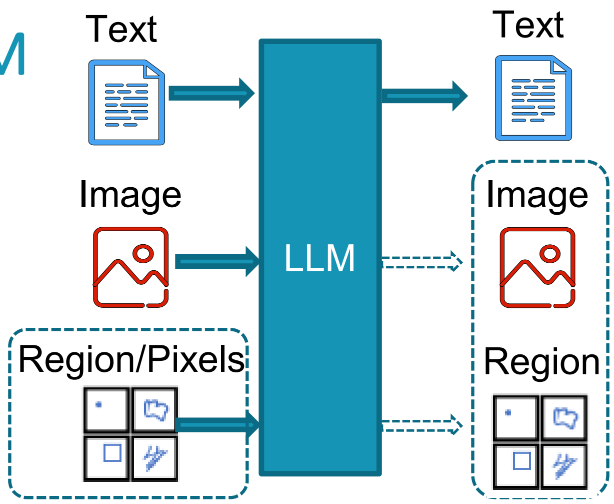
☞ Pixel Granularity

- 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

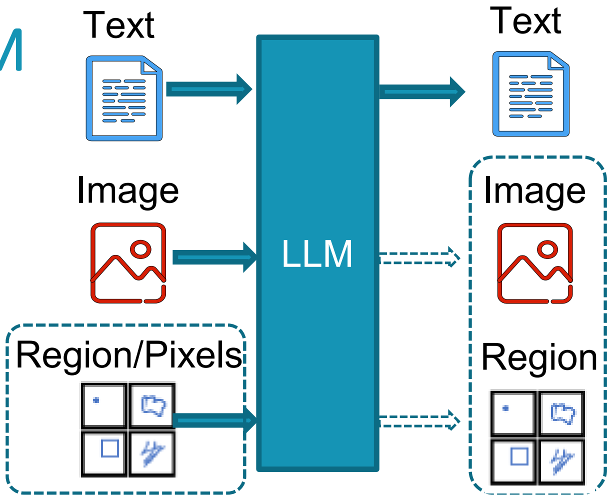
  LISA, PixelLM, DetGPT, MiniGPT-v2, LION

- Bounding-box Coordinates

  GPT4RoI, Shikra, KOSMOS-2, GLaMM

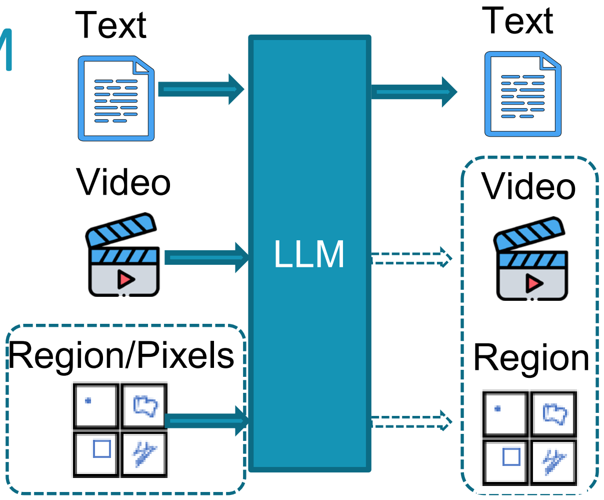
- User Sketches

  NExT-Chat, Osprey,



Fine-grained Capability of MLLM

- **Video-oriented Pixel-wise Regional MLLM**
  - PG-Video-LLaVA
  - Merlin
  - MotionEpic
  - …



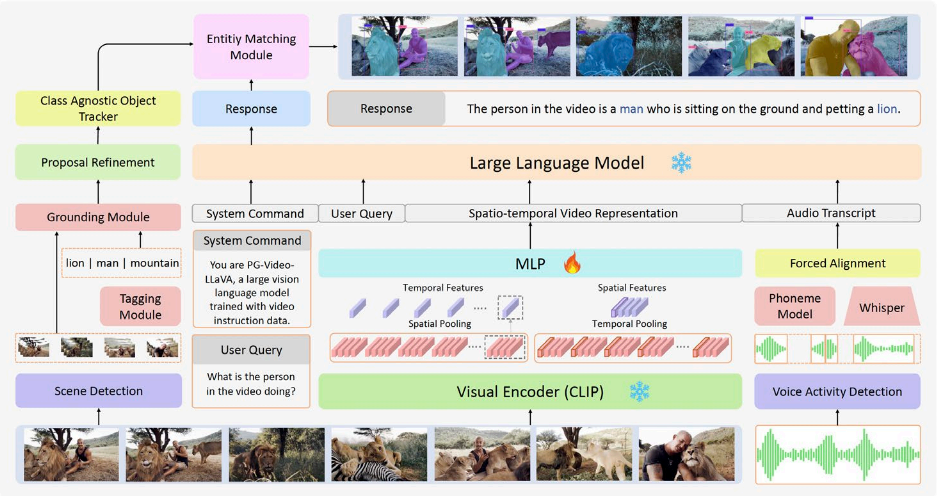
☞ *Users input an video (potentially specifying a region), and the LLM outputs content based on its understanding, grounding or tracking the content to specific pixel-level regions of the video.*

[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
…

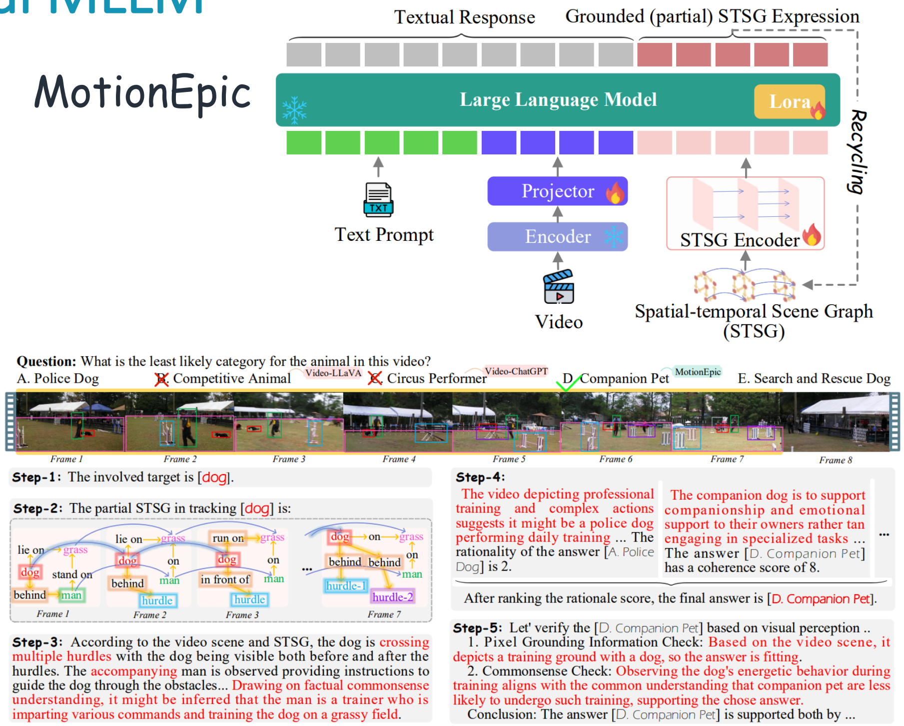
Fine-grained Capability of MLLM

Video-oriented Pixel-wise Regional MLLM

PG-Video-LLaVA



MotionEpic



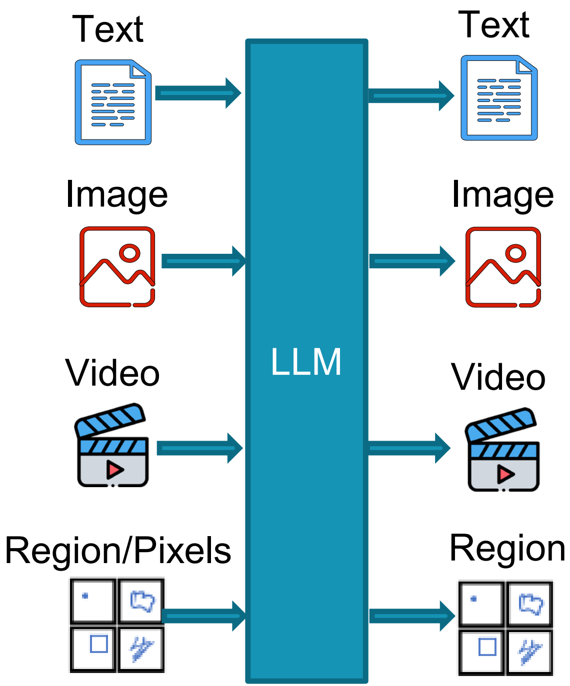
[1] *PG-Video-LLaVA: Pixel Grounding in Large Multimodal Video Models. 2023*

[2] *Video-of-Thought: Step-by-Step Video Reasoning from Perception to Cognition. 2024*

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.*



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

Fine-grained Capability of MLLM

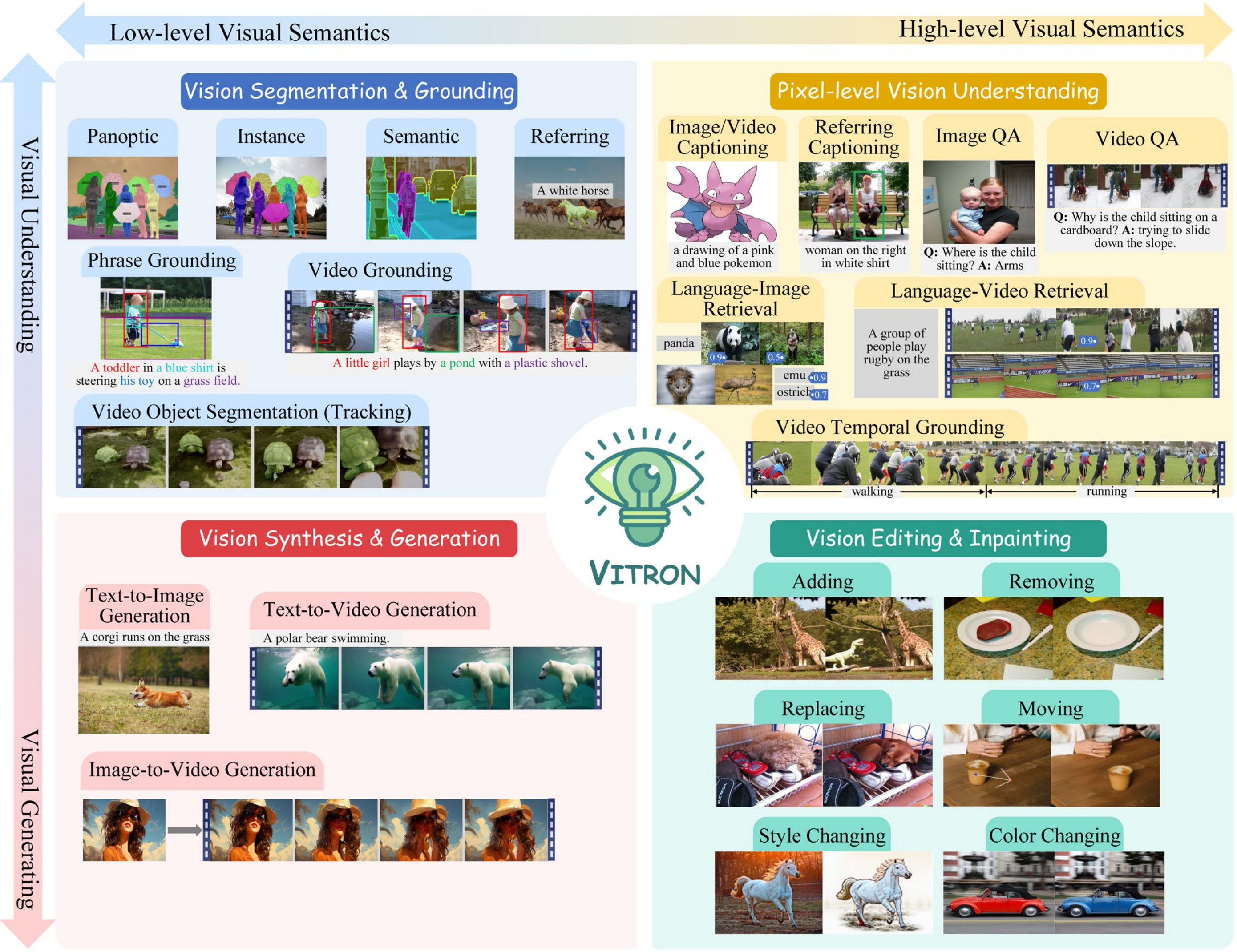
- **Unified**
  - Vitron

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)	✓	✓	✓	✓	✓	✓

[1] VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing. 2024

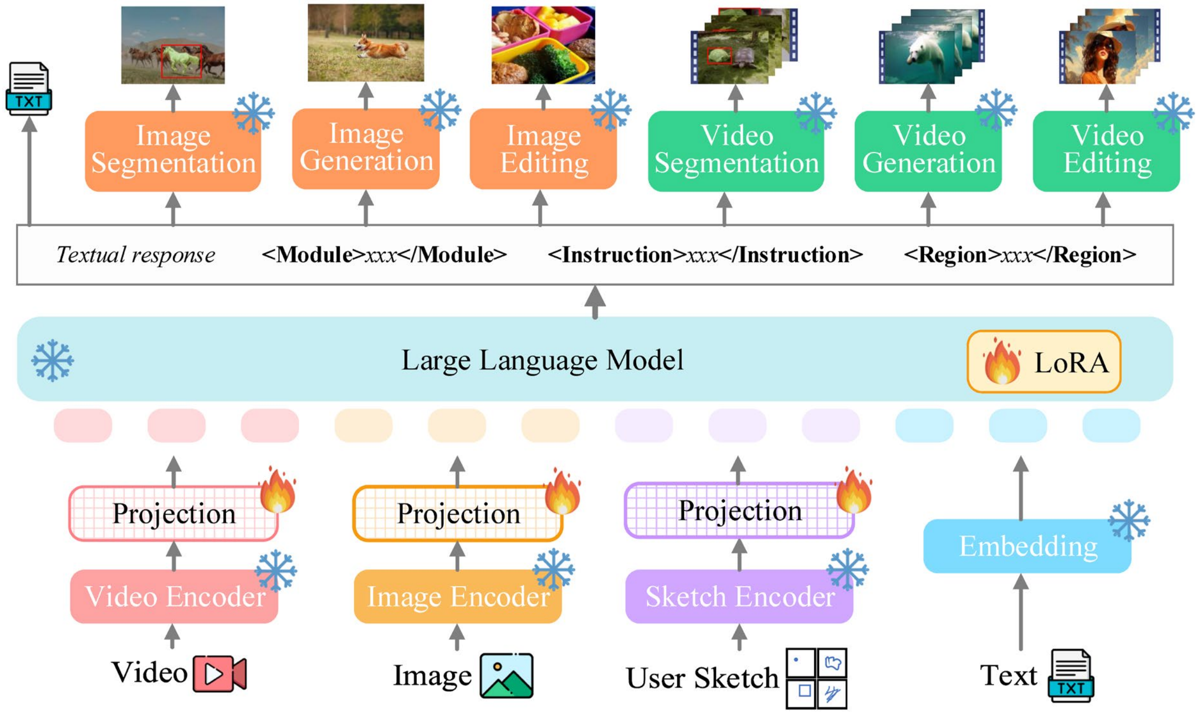
Fine-

- Unified
  - Vitron



Fine-grained Capability of MLLM

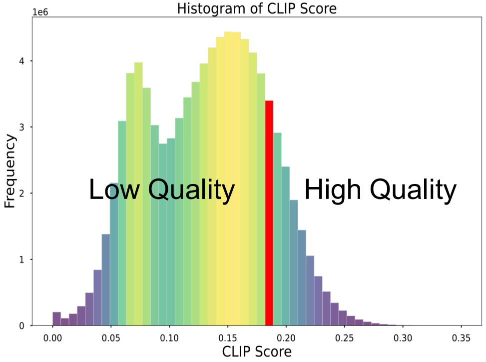
- ## Unified Pixel-wise MLLM
  - Vitron



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

Multilingual Multimodal LLMs

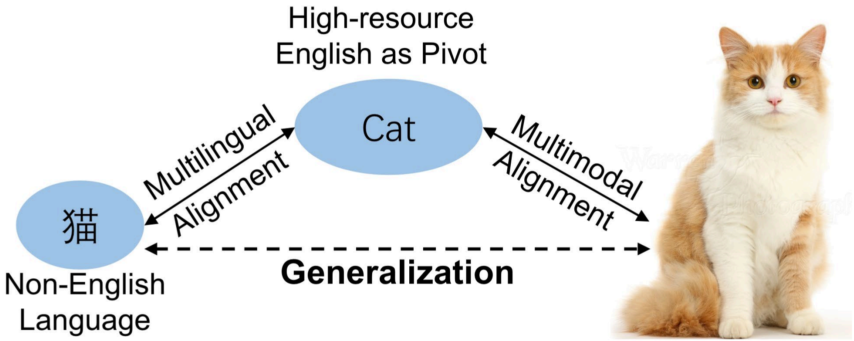
- MLLMs are mostly strong in English, weak in other languages
  - Limited scale and quality of multimodal data in non-English languages
  - Huge computation cost for each language even if sufficient data available
  - Why not machine translation pipeline ?
    - × Another LLM for translation: double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

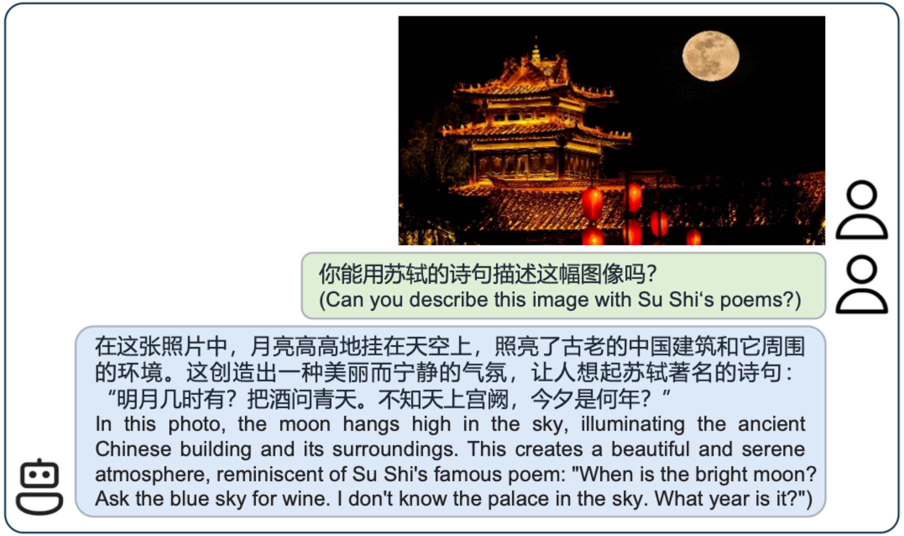
Was ist an diesem Bild ungewöhnlich?

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. ICLR 2024.

Multilingual Multimodal LLMs

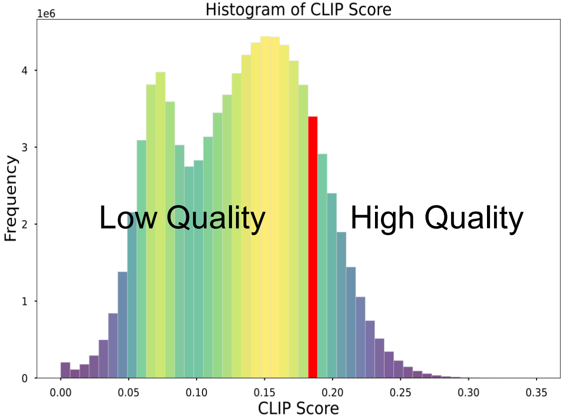
- ## MLLMs are mostly strong in English, weak in other languages
  - Generalizing multimodal capabilities to non-English languages via multilingual LLM as pivots
  - Lightweight SFT is sufficient to switch to the target language
  - Generalize to cultures, etc.



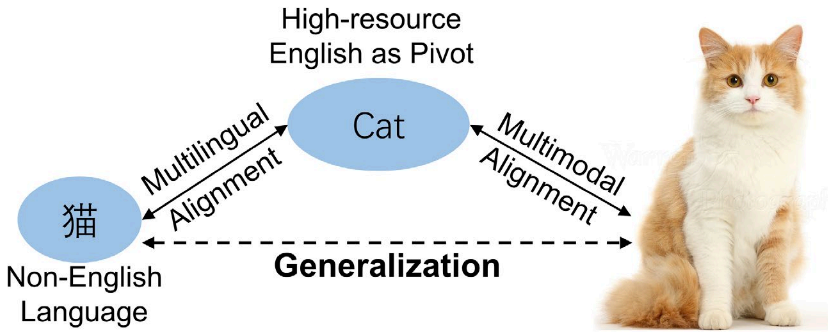
[1] Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages. ICLR 2024.

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, MiniCPM-V, etc.

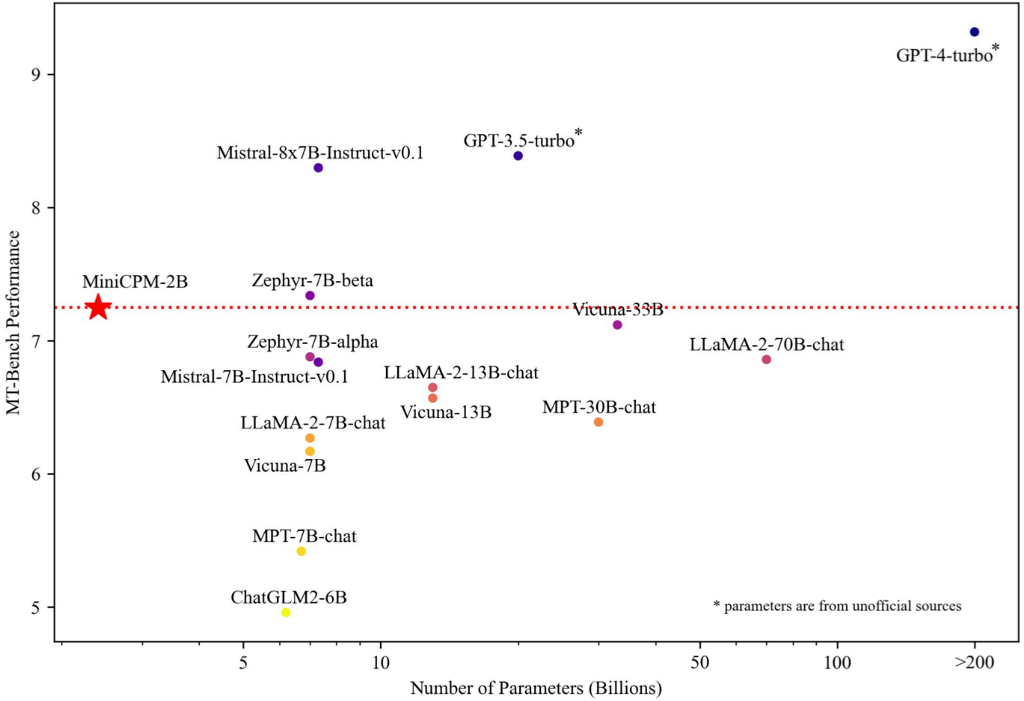


Chinese Image-Text Data Quality Distribution



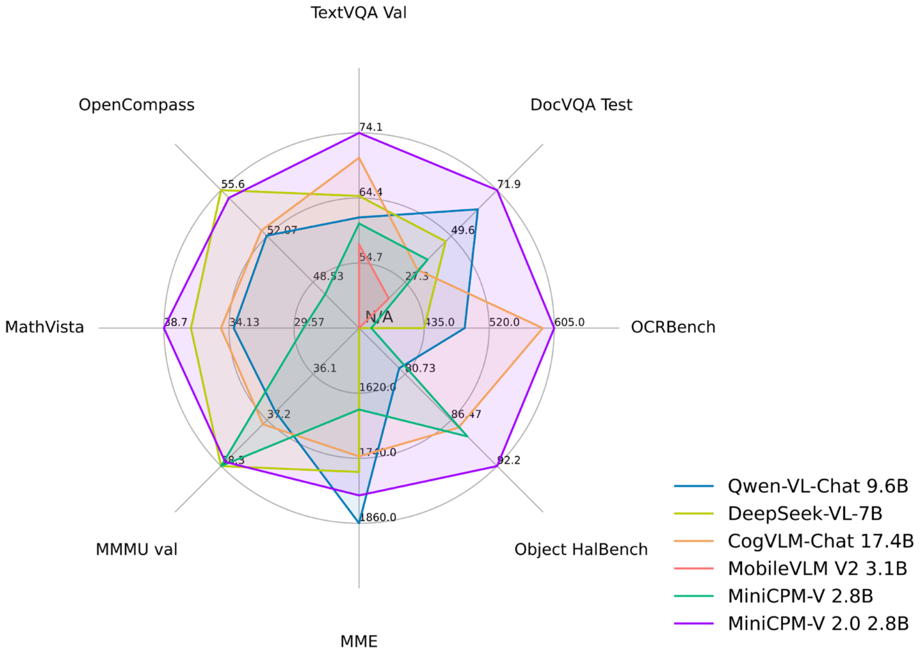
End-side MLLM

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



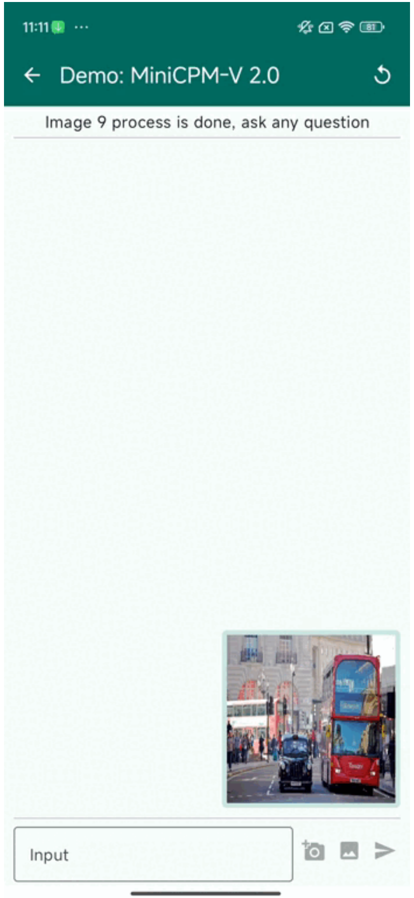
End-side MLLM

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



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



3

MLLM Generalist

Going beyond MLLM, what to do next?

How far from MLLM to AGI?

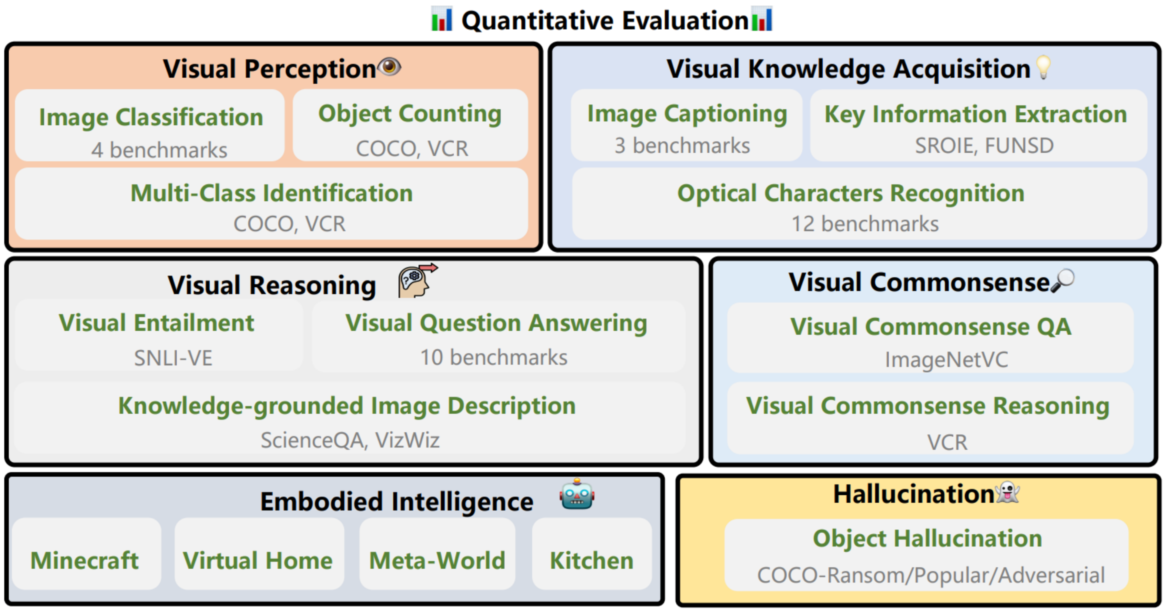
MLLM Evaluation Benchmark

- **MME**
  - Modalities: **Text + Image**
  - 14 choice tasks (Yes or No).
  - Measuring perception and cognition abilities.
  - Manual annotations.

[1] MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. 2023

MLLM Evaluation Benchmark

- ## LVLM-eHub
  - Modalities: **Text + Image**
  - 6 task group of multimodal capability.
  - 47 standard text-related visual benchmarks.
  - Collected from existing datasets.

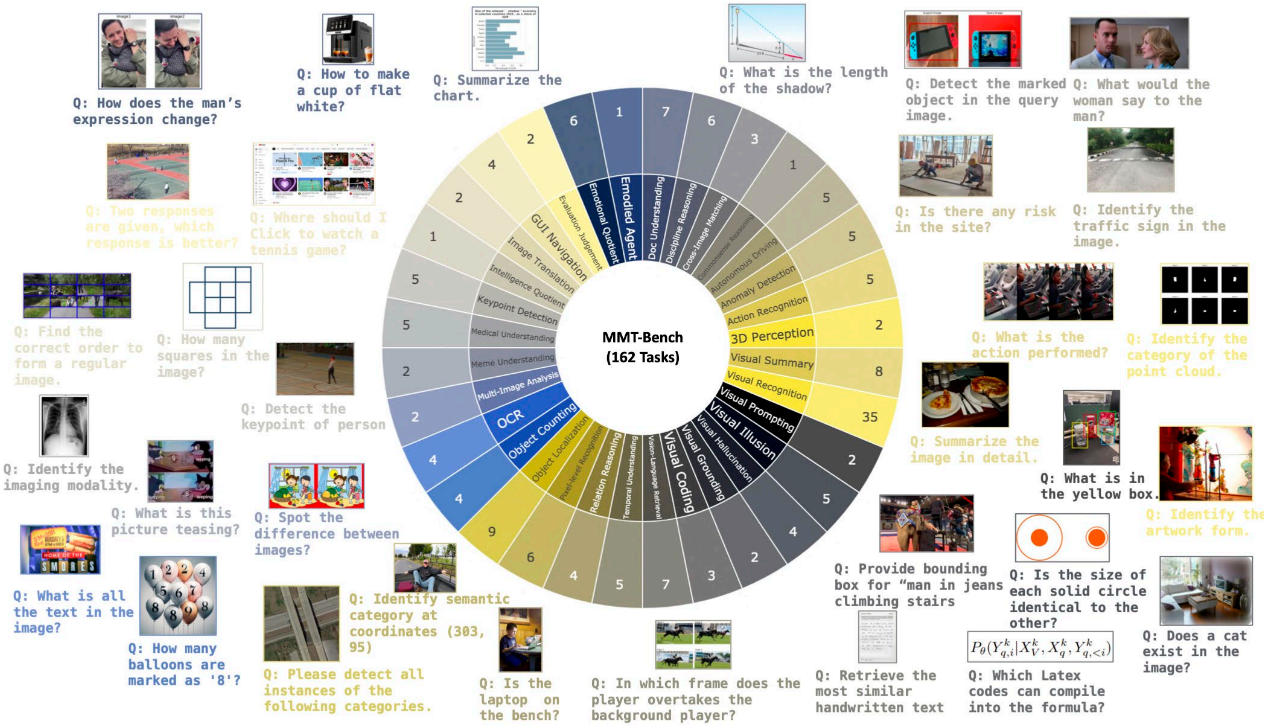


[1] *LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models. 2023*

MLLM Evaluation Benchmark

- ## MMT-Bench

  - Modalities: Image + Text + Video + Point cloud
  - Multi-choice visual questions.
  - 32 task group of multimodal capability.
  - 162 subtasks.
  - Collected from existing datasets.



[1] *MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI. 2023*

MLLM Evaluation Benchmark

Soul Questions

- Q1. Can we simply assume that **the stronger the performance of MLLMs** on these benchmarks, the closer they are to **achieving AGI**?

- Q2. Can it be said that the **more modalities and task types supported by MLLMs**, the closer they are to **AGI**?

- Q3. How should MLLMs evolve to ultimately reach AGI?

Levels of Multimodal Generalist

- ## Key Points from MLLM to Human-level AI

  - Human-level AI will come with the form of a multimodal generalist.
  - Human-level AI will support as many modalities as possible, as well as a broader range of functionalities.
  - Human-level AI will have strong generalizabilities and abductive reasoning, enabling it to make decisions and reasoning from limited information.

Levels of Multimodal Generalist

Checkpoints	Description	Capability (Multimodality, Multi-task)
Level-0: Separate Specialists	Various current models, fine-tuned on different modalities and specific task datasets, produce specialists capable of handling particular tasks within those modalities.	Tasks across various modalities such as language, image, video, etc., including classification, text generation, image generation, video segmentation, speech recognition, and more.
Upgrading Conditions: ***Combining with LLM as the central decision maker***		
Level-1: Generalist of Comprehension	General multimodal comprehension ability, being able to effectively filter out irrelevant information and features from the input multimodal data to solve questions. Key characteristic: the generalist is weak than individual specialists.	The comprehension process is primarily a process of converting multimodality to text. Classification tasks across different modalities; text generation tasks for various modalities (such as image/video/audio captioning, QA, etc.).
Upgrading Conditions: ***Under condition of preserving the core capabilities of LLM, sharing cross-modal and cross-task invariant features***		
Level-2: Generalist of Comprehension with Multimodality & Multi-task Synergy	Building on level 1, different modalities and tasks exhibit a **synergistic** effect, achieving a result where the whole is greater than the sum of its parts (1+1>2). By mastering a few select modalities and tasks, this capability can be transferred to understanding other unseen modalities and tasks, resulting in enhanced abilities. Key characteristic: the generalist is stronger than individual specialists.	Same to Level-1's capability in task and modality supporting, but with stronger performance.
Upgrading Conditions: ***During generation, reconstructing the necessary multimodal detail clues from the LLM output (i.e., multimodal tokens)***		
Level-3: Generalist of both Comprehension and Generation	Simultaneously possesses general multimodal comprehension and generation ability, allowing for filtering out irrelevant information during the understanding process while providing the necessary information required for generation. Key characteristic: the generalist is weak than individual specialists.	The process mainly involves converting multimodality to multimodality. Supporting all existing tasks, e.g., classification and QA tasks across various modalities; generation, segmentation, and editing tasks within visual modalities; cross-modal reasoning tasks, and more.
Upgrading Conditions: ***Acquiring the capable of abductive reasoning***		
Level-4: Generalist of both Comprehension and Generation with Multimodality & Multi-task Synergy	Different modalities and tasks, as well as the comprehension and generation processes, can mutually assist each other with **synergy**. For instance, learning from process A (a modality or task) can facilitate mastering capabilities B, C, D, etc., through analogical reasoning. Key characteristic: the generalist is stronger than individual specialists.	Same to Level-3's capability in task and modality supporting, but with stronger performance.

Levels of Multimodal Generalist

	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, … **Level 1**	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, Ospr **Maybe Few in Level 2**	[Pixel-wise] PG-Video-LLaVA, Merlin, MotionEpic, …	–	–
	Video-LLaVA, Chat-		–	–
	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 **Level 3**		–	–
	NExT-GPT, Unified-IO 2, AnyGPT, CoDi-2, Modaverse, ViT-Lens, …			–

Levels of Multimodal Generalist

Checkpoints	Description	Capability (Multimodality, Multi-task)
Level-0: Separate Specialists	Various current models, fine-tuned on different modalities and specific task datasets, produce specialists capable of handling particular tasks within those modalities.	Tasks across various modalities such as language, image, video, etc., including classification, text generation, image generation, video segmentation, speech recognition, and more.
Upgrading Conditions: ***Combining with LLM as the central decision maker***		
Level-1: Generalist of Comprehension	General multimodal comprehension ability, being able to effectively filter out irrelevant information and features from the input multimodal data to solve questions. Key characteristic: the generalist is <u>weak</u> than individual specialists.	The comprehension process is primarily a process of converting multimodality to text. Classification tasks across different modalities; text generation tasks for various modalities (such as image/video/audio captioning, QA, etc.).
Upgrading Conditions: ***Under condition of preserving the core capabilities of LLM, sharing cross-modal and cross-task invariant features***		
Level-2: Generalist of Comprehension with Multimodality & Multi-task Synergy	Building on level 1, different modalities and tasks exhibit a **synergistic** effect, achieving a result where the whole is greater than the sum of its parts (1+1>2). By mastering a few select modalities and tasks, this capability can be transferred to understanding other unseen modalities and tasks, resulting in enhanced abilities. Key characteristic: the generalist is <u>stronger</u> than individual specialists.	Same to Level-1' s capability in task and modality supporting, but with stronger performance.
Upgrading Conditions: ***During generation, reconstructing the necessary multimodal detail clues from the LLM output (i.e., multimodal tokens)***		
Level-3: Generalist of both Comprehension and Generation	Simultaneously possesses general multimodal comprehension and generation ability, allowing for filtering out irrelevant information during the understanding process while providing the necessary information required for generation. Key characteristic: the generalist is <u>weak</u> than individual specialists.	The process mainly involves converting multimodality to multimodality. Supporting all existing tasks, e.g., classification and QA tasks across various modalities; generation, segmentation, and editing tasks within visual modalities; cross-modal reasoning tasks, and more.
Upgrading Conditions: ***Acquiring the capable of abductive reasoning***		
Level-4: Generalist of both Comprehension and Generation with Multimodality & Multi-task Synergy	Different modalities and tasks, as well as the comprehension and generation processes, can mutually assist each other with **synergy**. For instance, learning from process A (a modality or task) can facilitate mastering capabilities B, C, D, etc., through analogical reasoning. Key characteristic: the generalist is <u>stronger</u> than individual specialists.	Same to Level-3' s capability in task and modality supporting, but with stronger performance.

What's Next from Multimodal LLM to AGI

- ## 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:

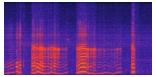
heat map 

infrared/radar 

code 

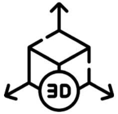
document/table 

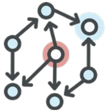
time series 

spectrogram 

touch 

smell 

depth&3D  

graph 

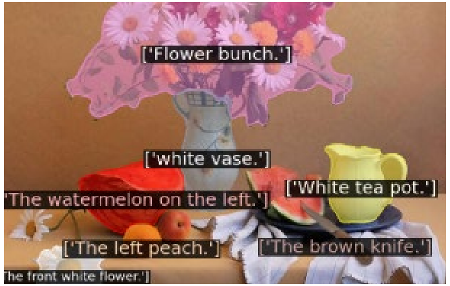
What's Next from Multimodal LLM to AGI

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

Task Perspective: Going Deeper

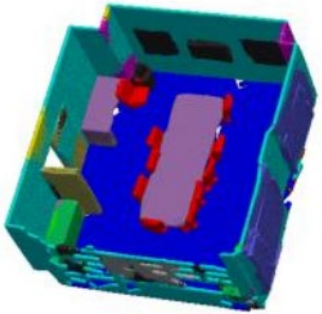
☞ *Vision-based MLLM, **Vitron**, has focused on unifying image and video processing under the scope of pixel-wise tasks, ranging from low-level to high-level.*

☞ *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**.*

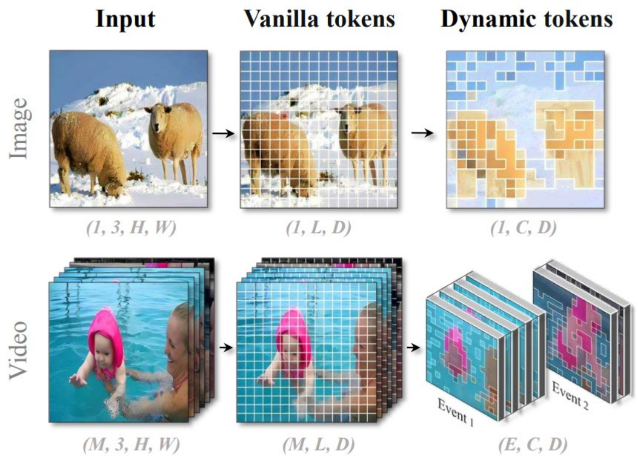
      - Remove the equivalence constraint between pre-LLM and post-LLM, as the roles of input and output multimodal tokens differ.
      - Increase the information content of multimodal tokens to include more high-frequency details.

[1] Auto-Encoding Morph-Tokens for Multimodal LLM. 2024

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

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

What's Next from Multimodal LLM to AGI

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

  - Core Idea

☞ *Master abductive reasoning for analogical thinking.*

Summary

MLLM Architecture

- Overview of MLLM Architecture
- Multimodal Encoding
- Multimodal Signal Tokenization
- Input-side Projection
- Backbone LLMs
- Decoding-side Connection
- Multimodal Generation

MLLM Modality and Functionality

- Overview of Modality and Functionality
- Multimodal Perceiving
- Unified MLLM: Perceiving + Generation
- Unified MLLM: Harnessing Multi-Modalities
- Fine-grained Capability of MLLM
- Multilingual Multimodal LLMs
- End-side MLLM

MLLM Generalist

- MLLM Evaluation Benchmark
- Levels of Multimodal Generalist
- What’s Next from Multimodal LLM to AGI

Thanks!

Any questions?