From Multimodal LLM to Human-level AI

Architecture, Modality, Function, Instruction, Hallucination, Evaluation, Reasoning and Beyond

https://mllm2024.github.io/ACM-MM2024/

ACM Multimedia 2024

Melbourne, Australia

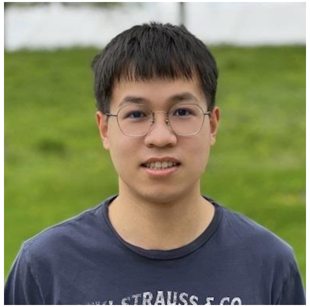
NUS National University of Singapore | ByteDance | XAI | UNIVERSITY OF MARYLAND | 上海交通大學 SHANGHAI JIAO TONG UNIVERSITY | NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | 上海人工智能实验室 Shanghai Artificial Intelligence Laboratory | 2050 RESEARCH

Hao Fei

National University of Singapore

Xiangtai Li

ByteDance/Tiktok

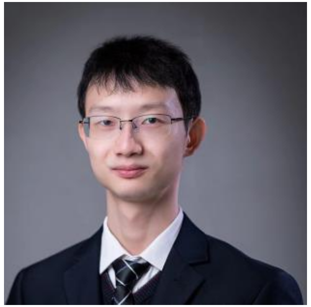


Haotian Liu

xAI

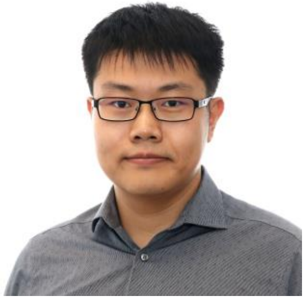
Fuxiao Liu

University of Maryland, College Park



Zhuosheng Zhang

Shanghai Jiao Tong University



Hanwang Zhang

Nanyang Technological University

Kaipeng Zhang

Shanghai AI Lab

Shuicheng Yan

Kunlun 2050 Research, Skywork AI

Part-II

MLLM Architecture&Modality

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

Table of Content

1

Architecture of MLLM

How to design an MLLM?

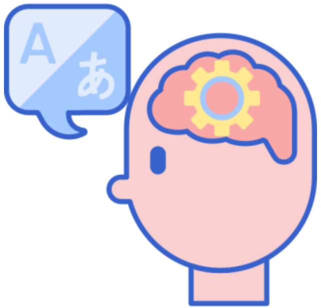
Overview of MLLM Architecture

- ## 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.



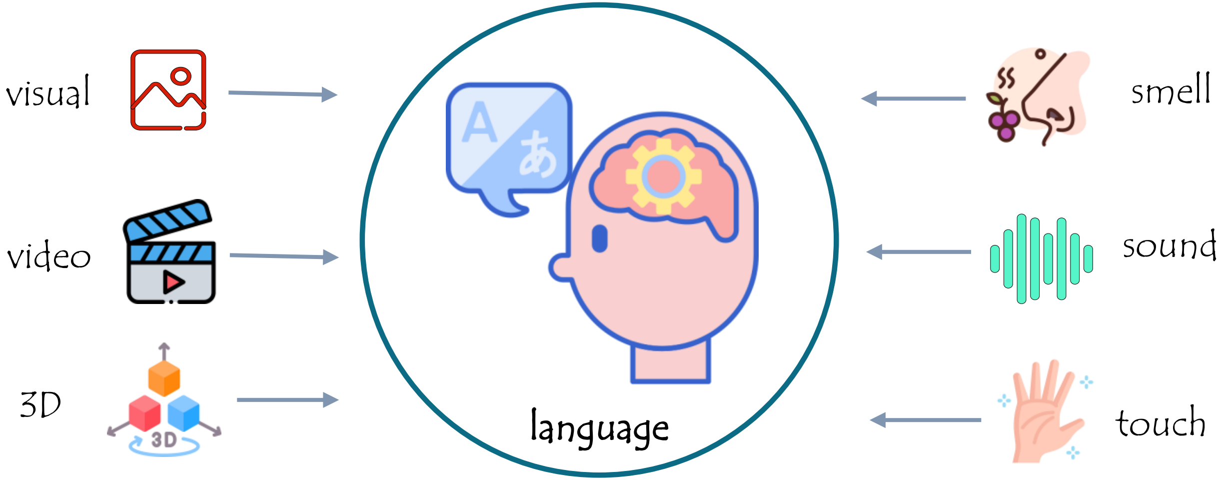
language

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



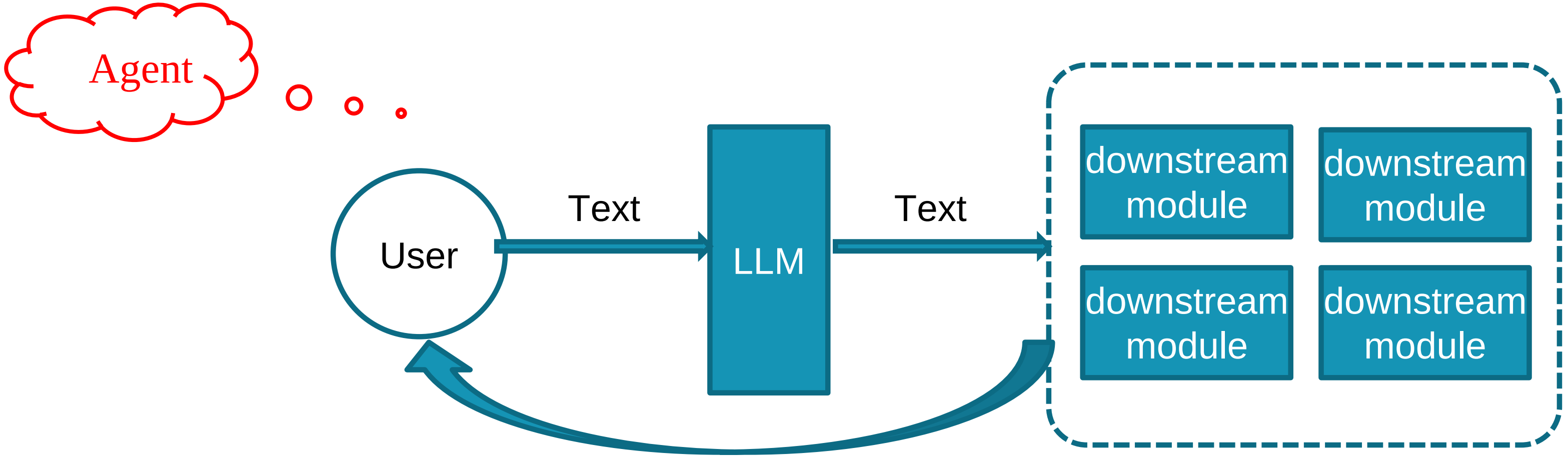
Overview of MLLM Architecture

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.

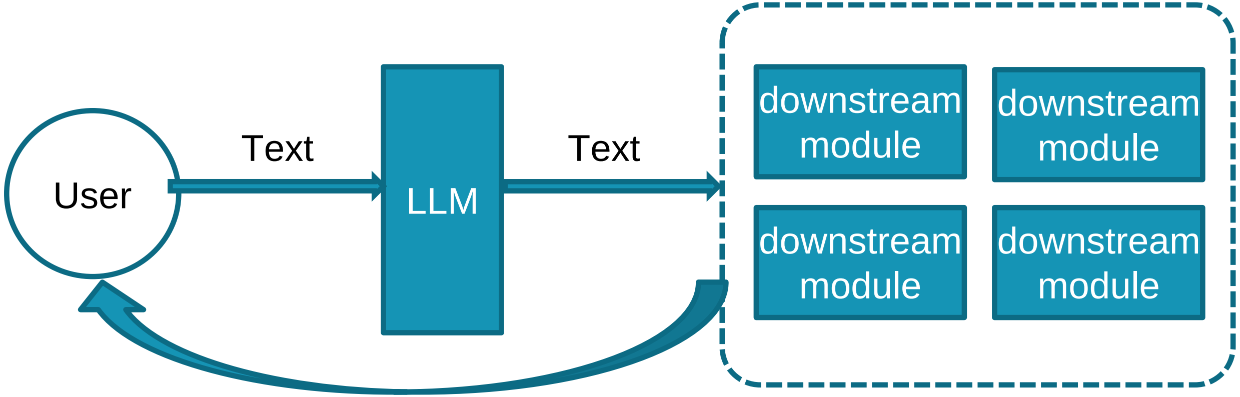
Key feature:

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
    - …

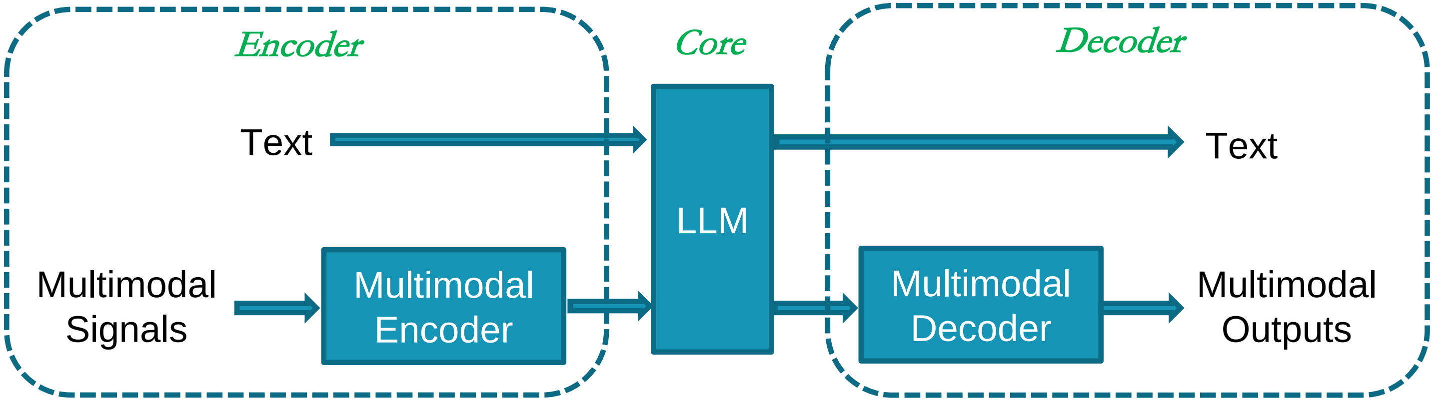


- ## Architecture-II: LLM as Joint Part of System

☞ 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.

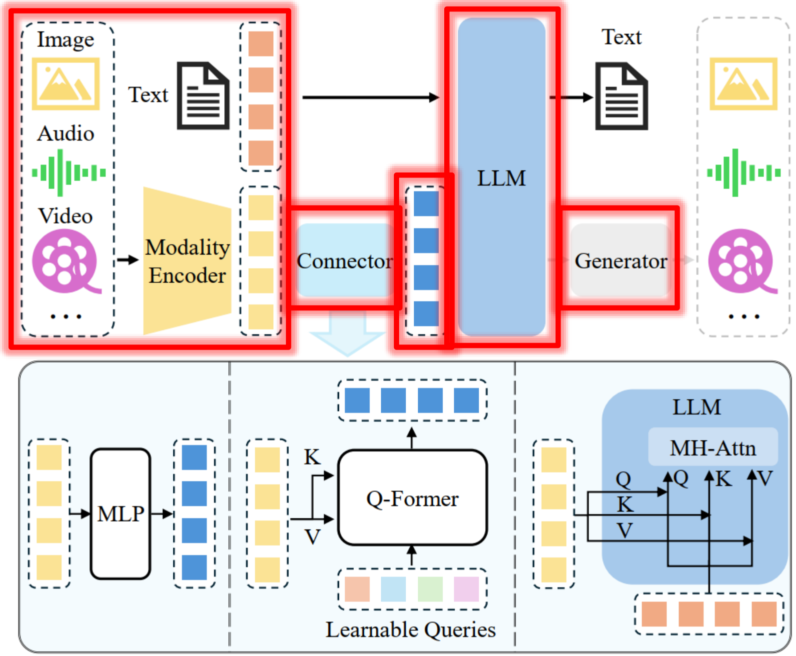


Overview of MLLM Architecture

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

More promising

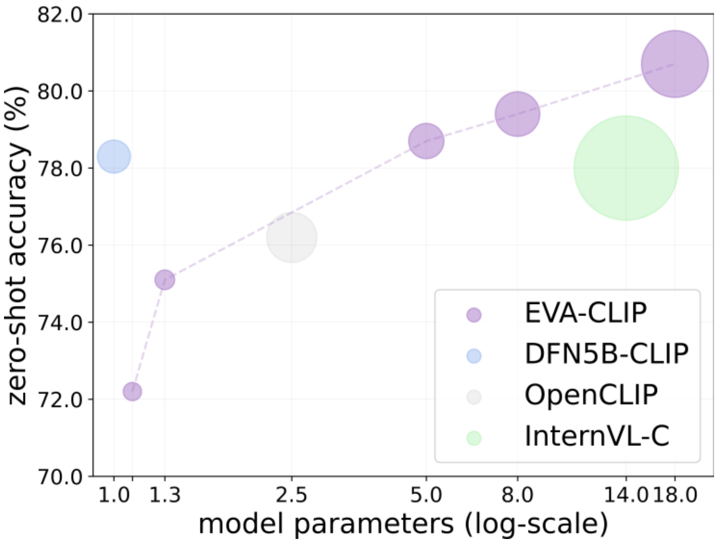
  - > 90% MLLMs belong to this category.
  - Higher upper-bound, better integrated into a unified model



[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

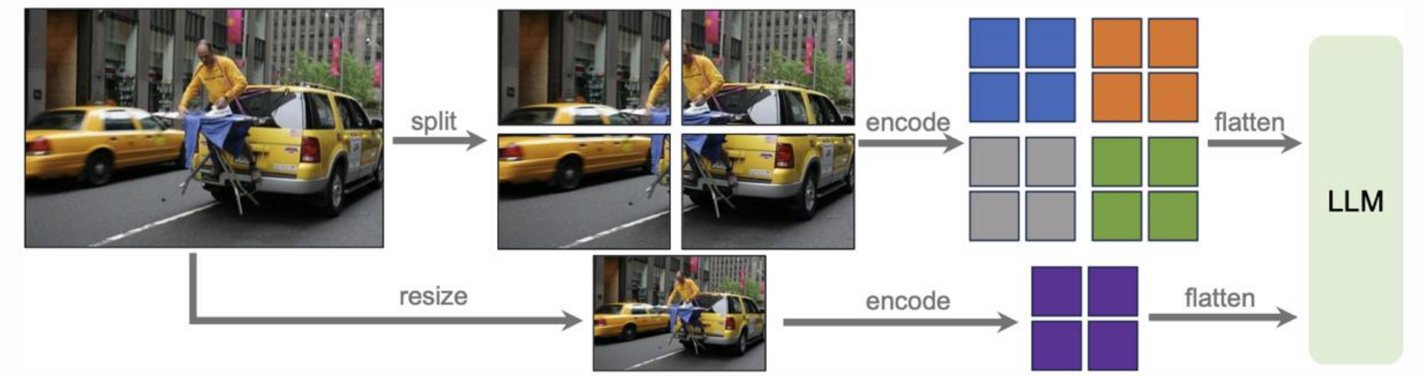


Illustration of dynamic high resolution scheme: a grid configuration of 2×2

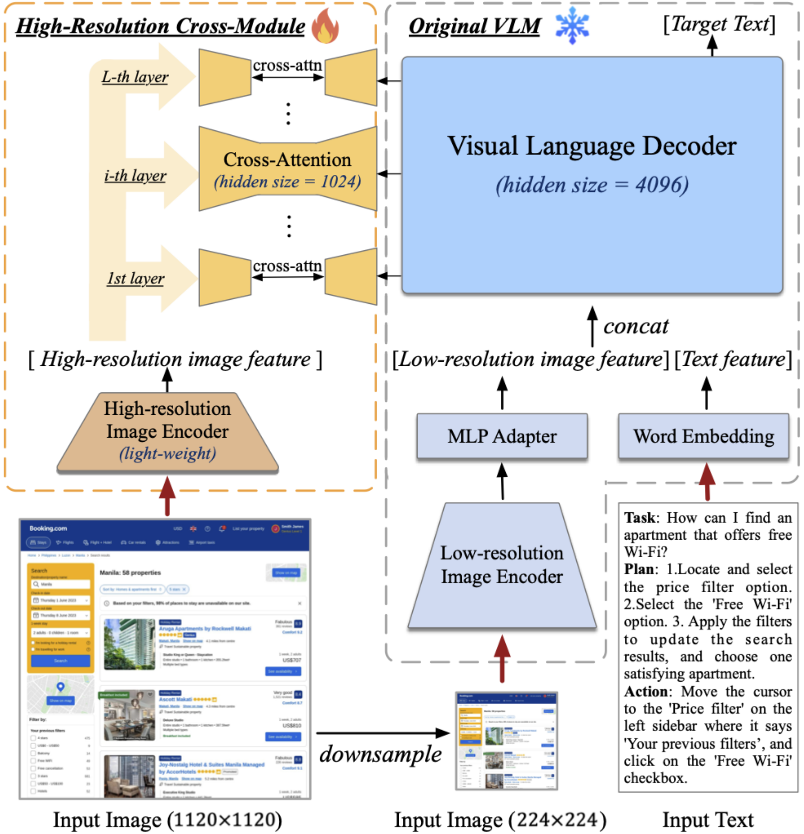
Multimodal Encoding

- ## 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

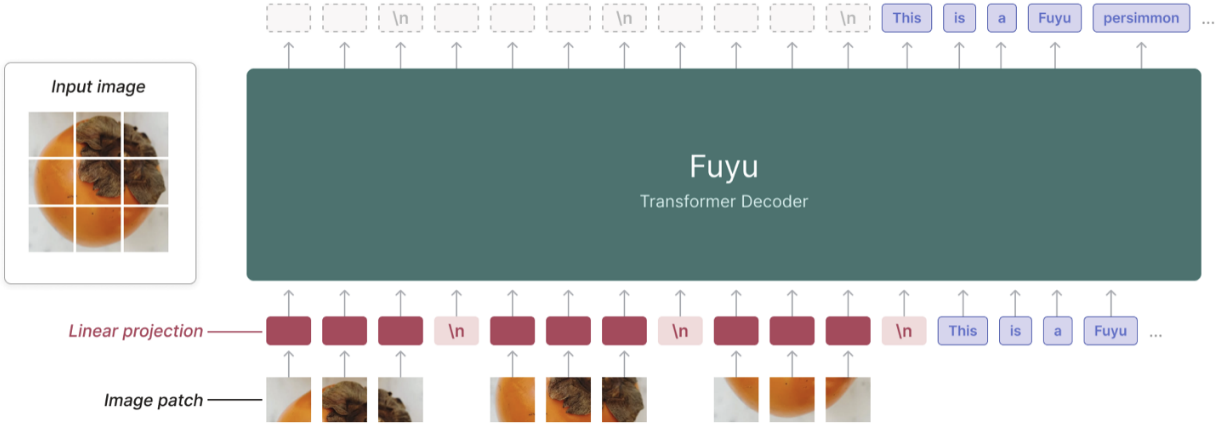
Multimodal Encoding

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



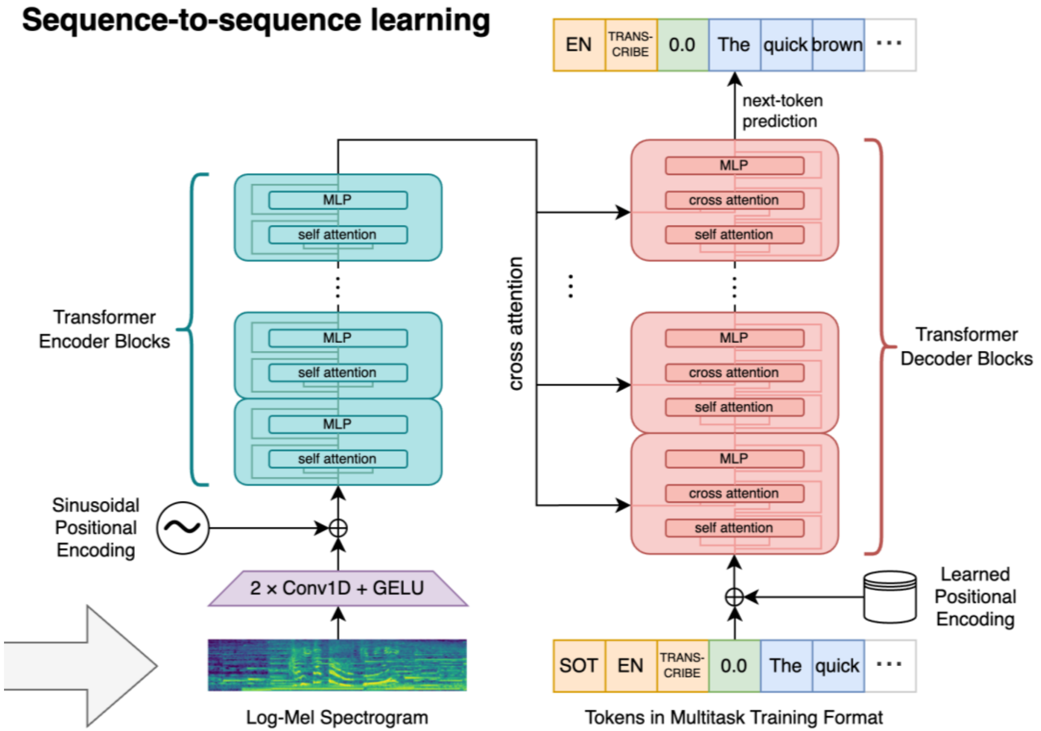
Multimodal Encoding

- ## 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



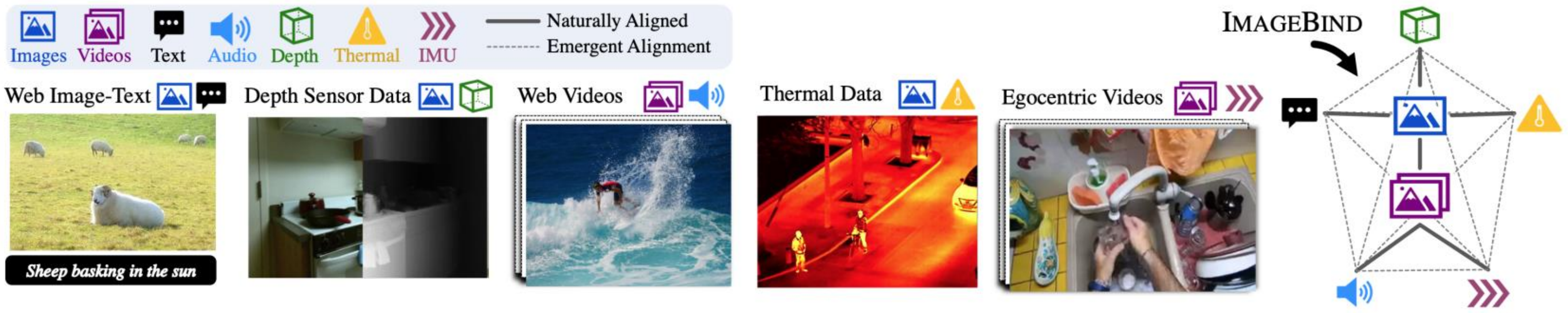
Multimodal Encoding

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



Multimodal Encoding

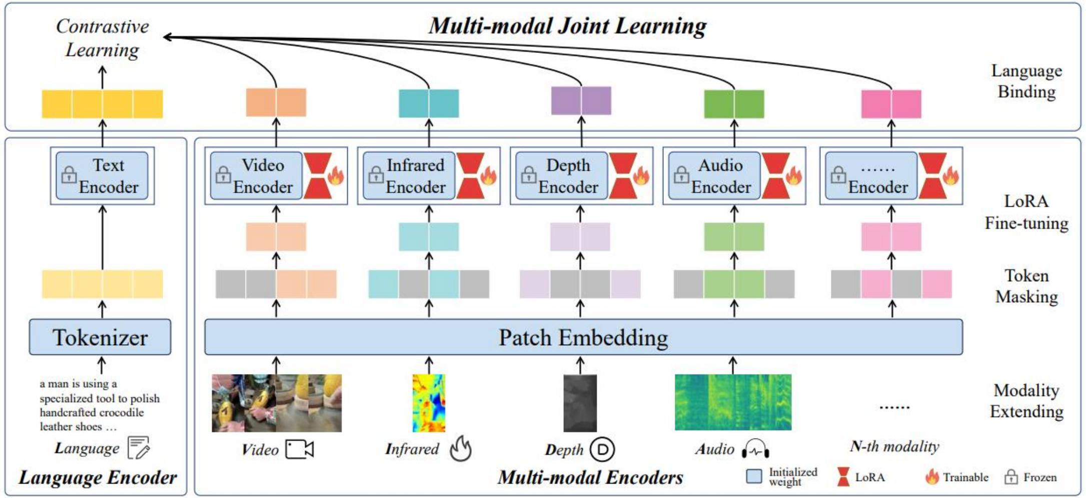
- ## Unified Multimodal Encoder
  - ImageBind:
    - Embedding all modalities into a joint representation space of Image.
    - Well aligned modality representations can benefit LLM understanding



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

Multimodal Encoding

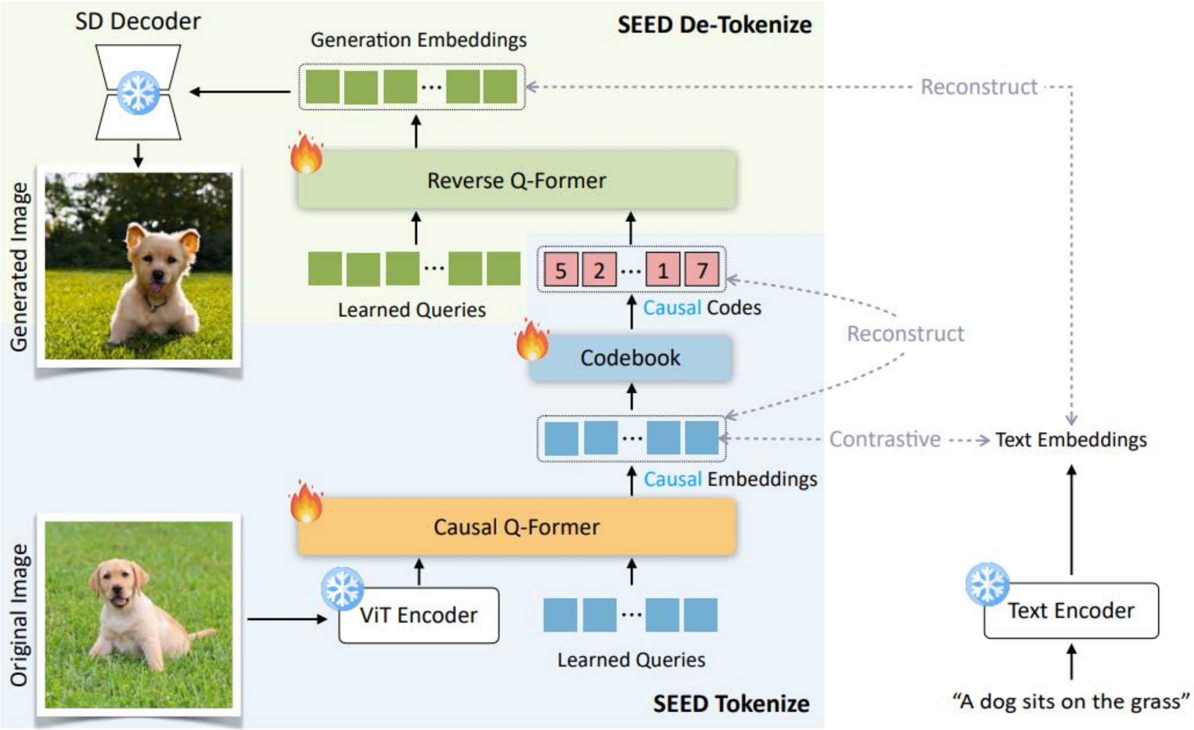
- ## Unified Multimodal Encoder
  - LanguageBind:
    - Embedding all modalities into a joint representation space of Language.
    - Well aligned modality representations can benefit LLM understanding



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

Multimodal Signal Tokenization

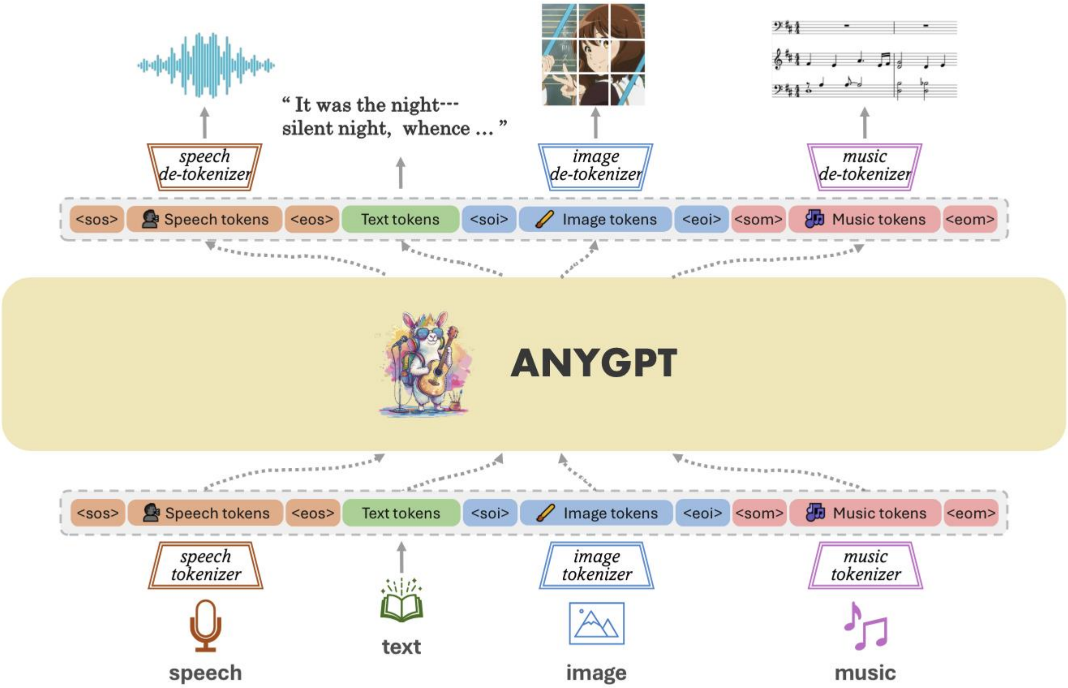
- **Tokenization**
  - SEED



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

Multimodal Signal Tokenization

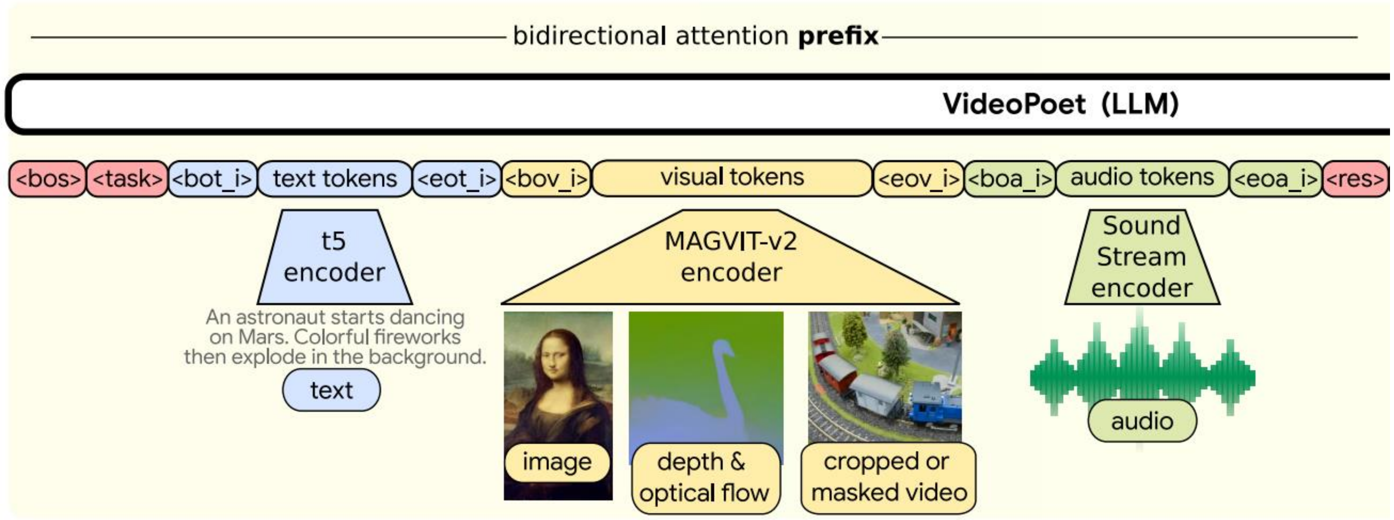
- **Tokenization**
  - AnyGPT



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

Multimodal Signal Tokenization

- **Tokenization**
  - VideoPoet



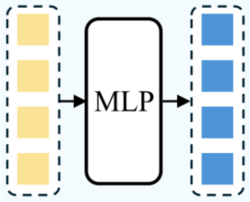
[1] *VideoPoet: A Large Language Model for Zero-Shot Video Generation. 2023*

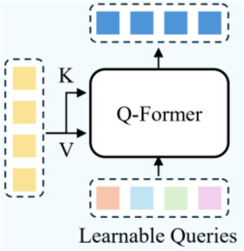
Multimodal Signal Tokenization

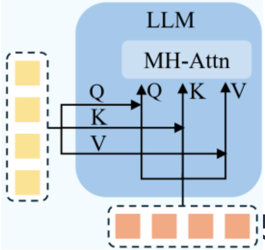
- ## Tokenization in Codebook
  - Represent multimodal signals as discrete tokens in a codebook
    - Advantages: support unified multimodal signal understanding and generation in an auto-regressive next-token prediction framework
    - More commonly used in image synthesize
      - Parti
      - Muse (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	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**	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

Backbone LLMs

- ## Open-source Language-based LLMs

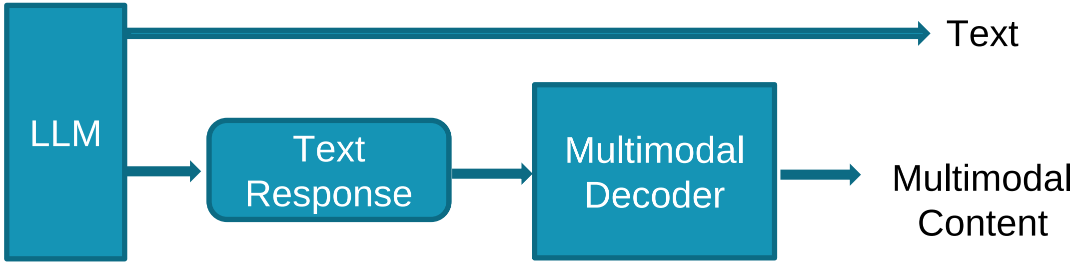
LLM	Size (B)	Data Scale (T)	Date	Language	Architecture
Flan-T5	3/11	-	Oct-2022	en, fr, de	Encoder-Decoder
LLaMA	7/13	1.4	Feb-2023	en	Decoder
Alpaca	7	-	Mar-2023	en	Decoder
Vicuna	7/13	1.4	Mar-2023	en	Decoder
LLaMA-2	7/13	2	Jul-2023	en	Decoder
GLM	2/10	0.4	Oct-2022	en	Decoder
Qwen	1.8/7/14	3	Sep-2023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

- Representative MLLMs:
  - Visual-ChatGPT
  - HuggingGPT
  - GPT4Video
  - MM-REACT
  - ViperGPT
  - ModaVerse
  - Vitron
  - ...



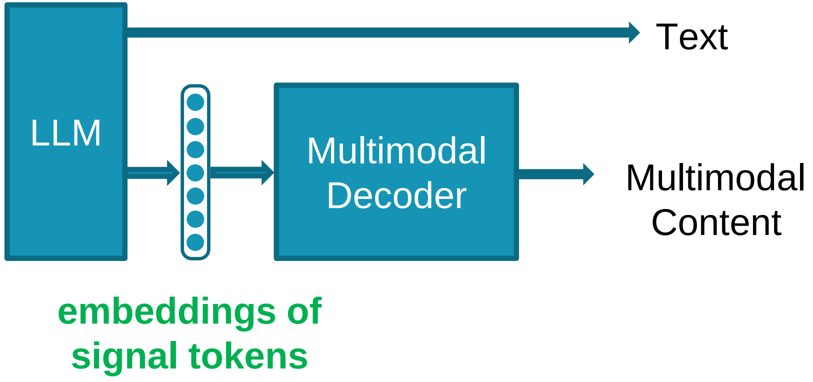
- Pros:
  - High performance lower-bound
  - More efficient, i.e., without tuning
- Cons:
  - Loss of end-to-end tuning capabilities.
  - Performance upper-bound is limited, i.e., some multimodal signals cannot be optimally conveyed through text.

[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**.*



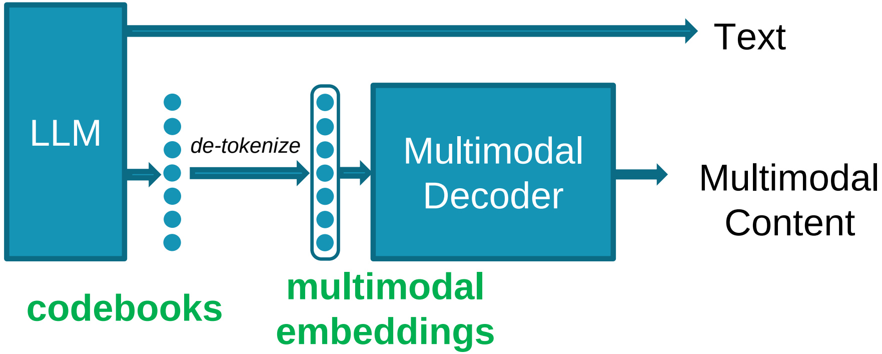
- Merits
  - Capable of end-to-end tuning, resulting in more efficient instruction transmission
  - More able to convey various multimodal signals that text alone cannot express, e.g.,
    - *the numeration of vision*
    - *the visual-spatial relational semantics*

[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

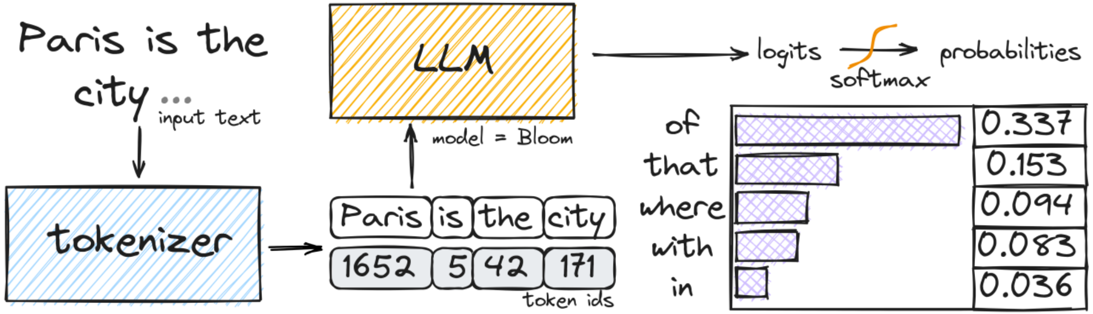
[1] Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action. 2023
[2] LVM: Sequential Modeling Enables Scalable Learning for Large Vision Models. 2023
[3] AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. 2024
[4] VideoPoet: A Large Language Model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

Multimodal Generation

- **Text Generation**

  - LLMs naturally support direct text generation

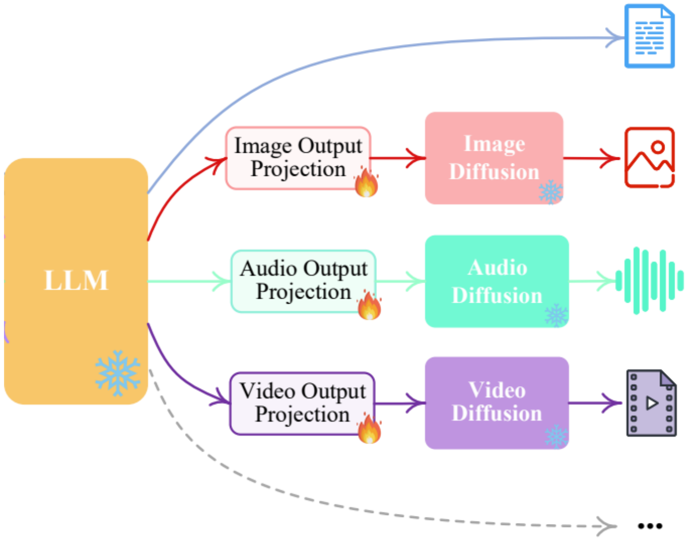
    via e.g., BPE decoding, Beam search, ...



Multimodal Generation

- ## Generation via Diffusion Models

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

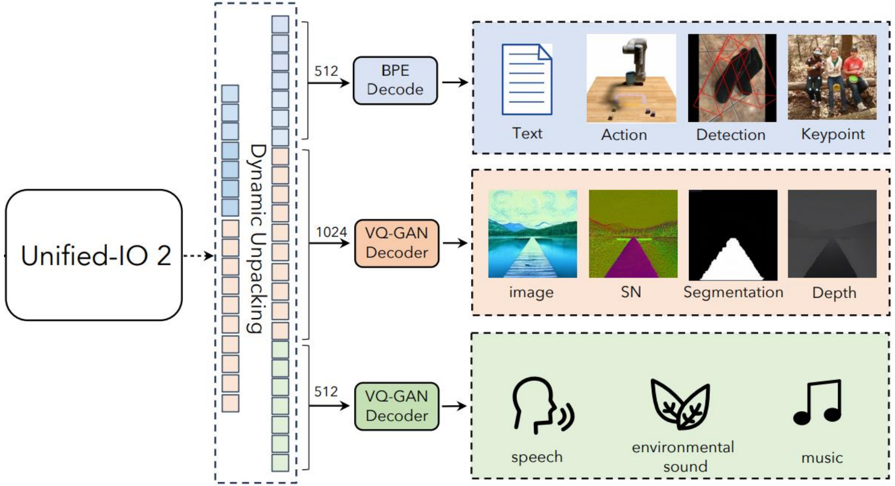


[1] NExT-GPT: Any-to-Any Multimodal LLM. 2023

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*

Multimodal Generation

Generation via Codebooks

- VQ-GAN in Stable-diffusion
  - $64 \times 64 \times 3$ or $32 \times 32 \times 4$

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 $f = 2^m$.

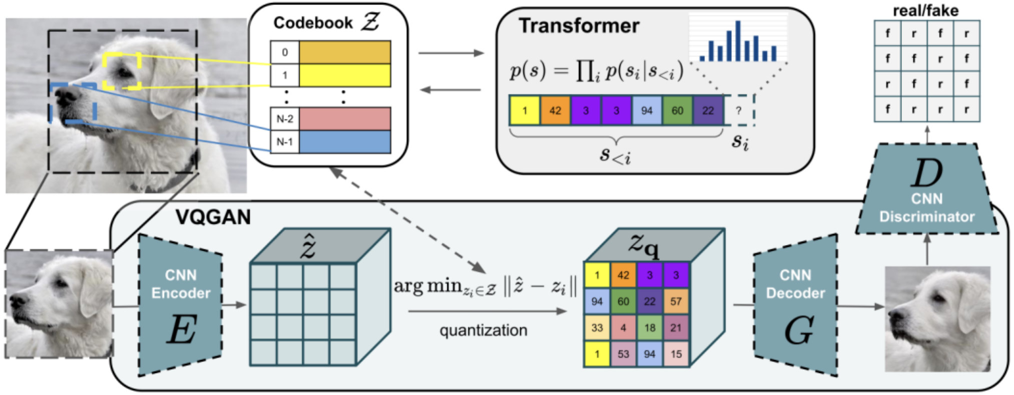


Figure 2. Our approach uses a convolutional *VQGAN* to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables 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

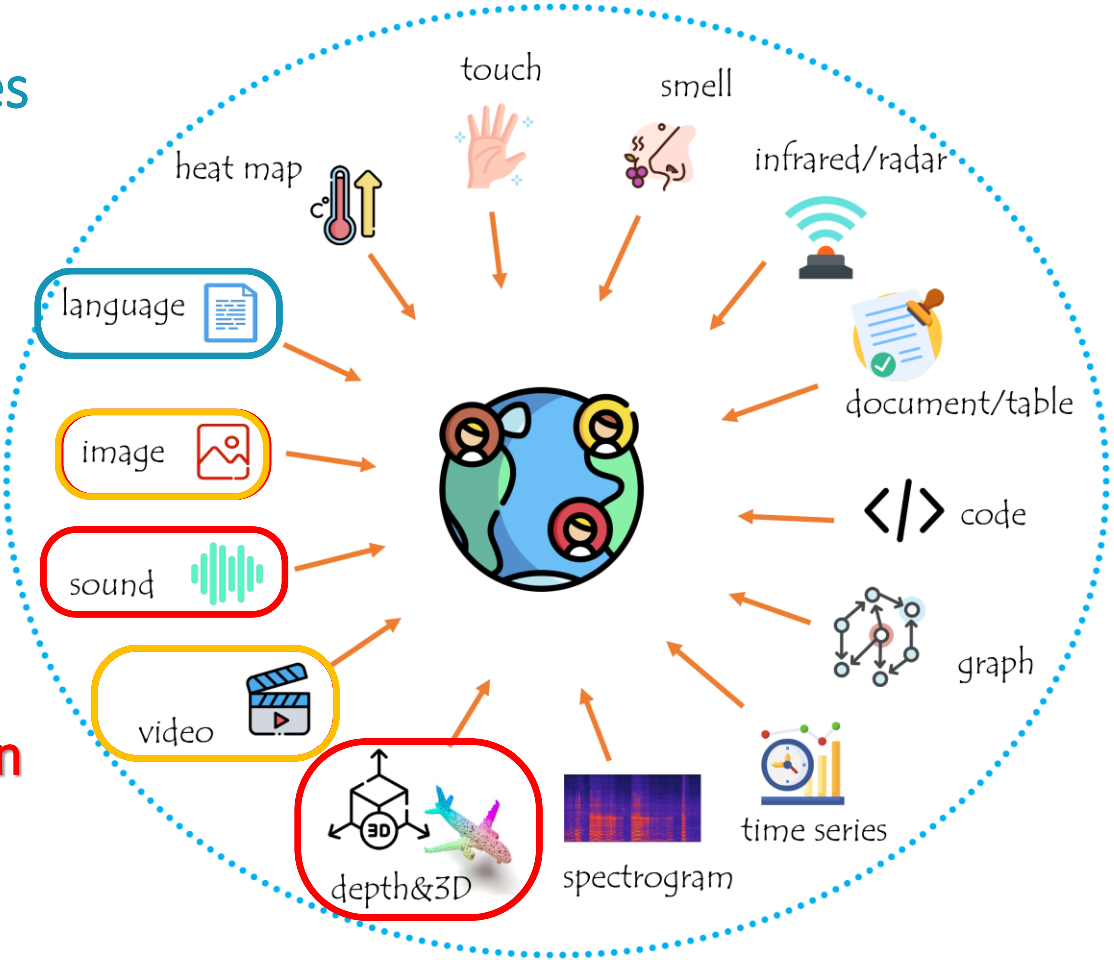
2

Modality of MLLM

What modalities do MLLMs support?

Overview of Modality and Functionality

- **Modalities**



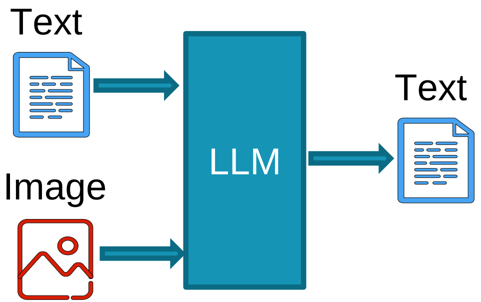
Language + Vision

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,
  - ...



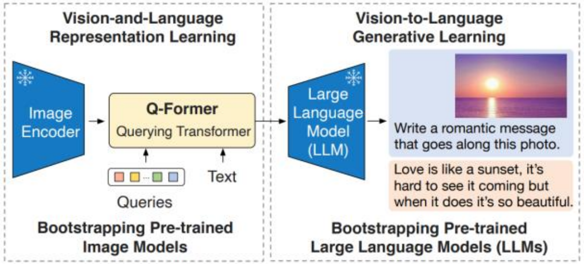
☞ *Encode input images with external image encoders, generating LLM-understandable visual feature, which is then fed into the LLM. LLM then interprets the input images based on the input text instructions and produces a textual response.*

[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
...

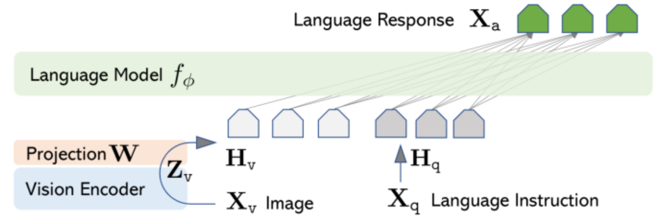
Multimodal Perceiving

- ## Image-perceiving MLLM

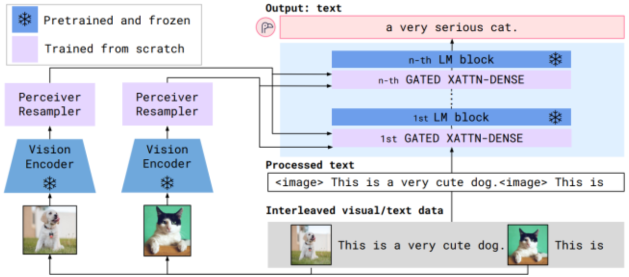
  - Blip2



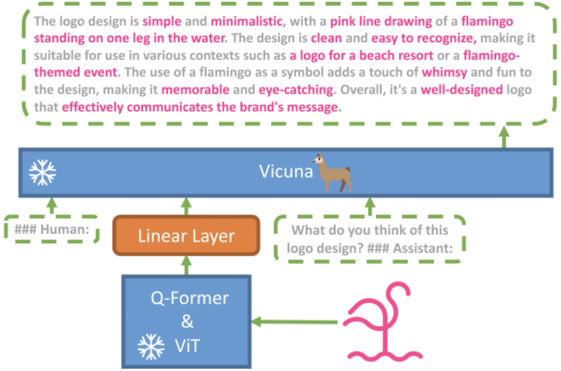
  - LLaVA



  - Flamingo



  - Mini-GPT4

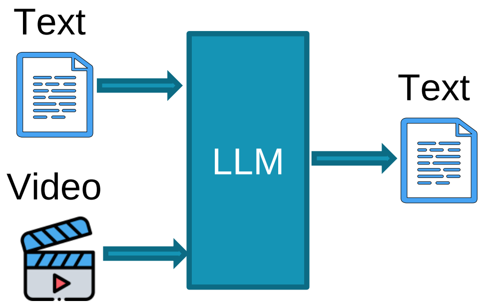


[1] Flamingo: a Visual Language Model for Few-Shot Learning. 2022
[2] BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. 2023
[3] Visual Instruction Tuning. 2023
[4] A Survey on Multimodal Large Language Models. https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models,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

Multimodal Perceiving

- ## Video-perceiving MLLM

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



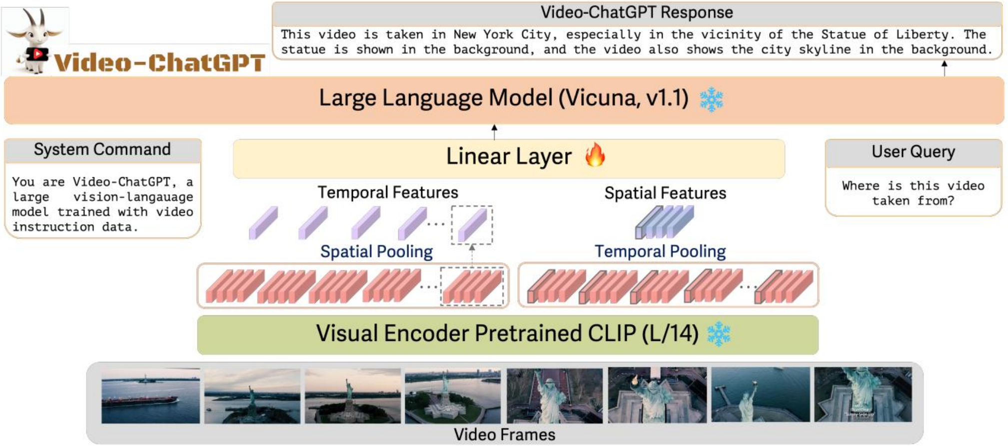
☞ *Encode input videos with external video encoders, generating LLM-understandable visual feature, feeding into LLM, which then interprets the input videos based on the input text instructions and produces a textual response.*

[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
…

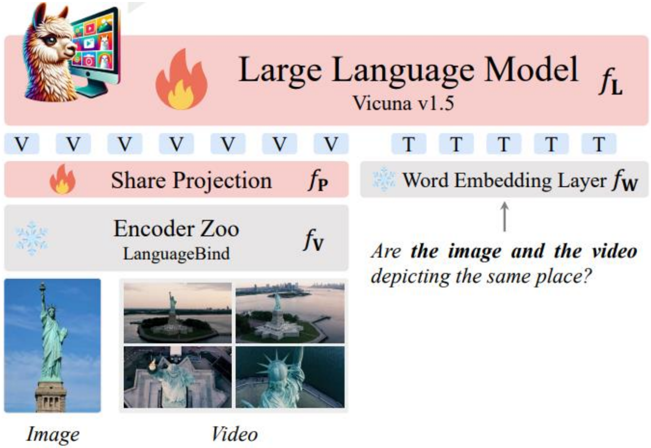
Multimodal Perceiving

- ## Video-perceiving MLLM

  - Video-ChatGPT



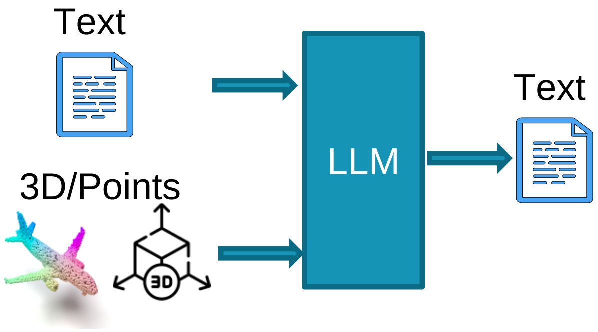
  - Video-LLaVA



[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
  - …



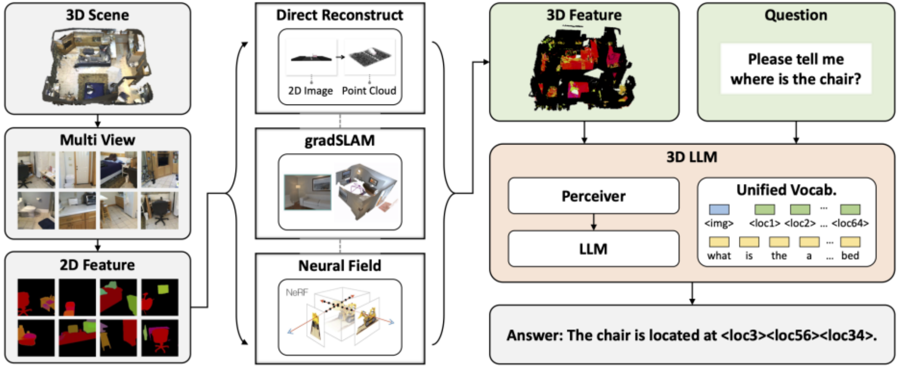
☞ *Encode input 3D information with external encoders, generating LLM-understandable 3D feature, feeding into LLM, which then interprets the input 3D/points based on the input text instructions and produces a textual response.*

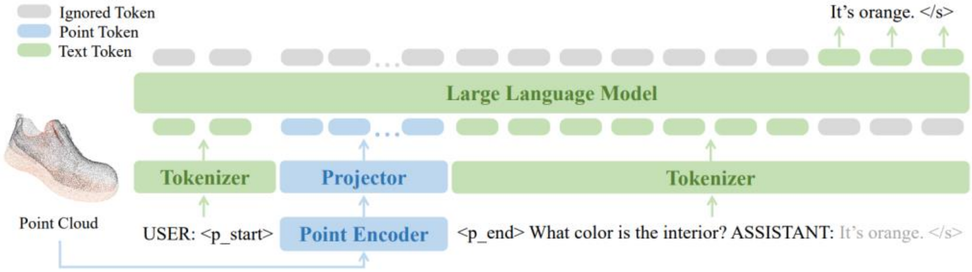
[1] 3D-LLM: Injecting the 3D World into Large Language Models. 2023
[2] 3D-GPT: Procedural 3D Modeling with Large Language Models. 2023
[3] LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning. 2023
[4] PointLLM: Empowering Large Language Models to Understand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

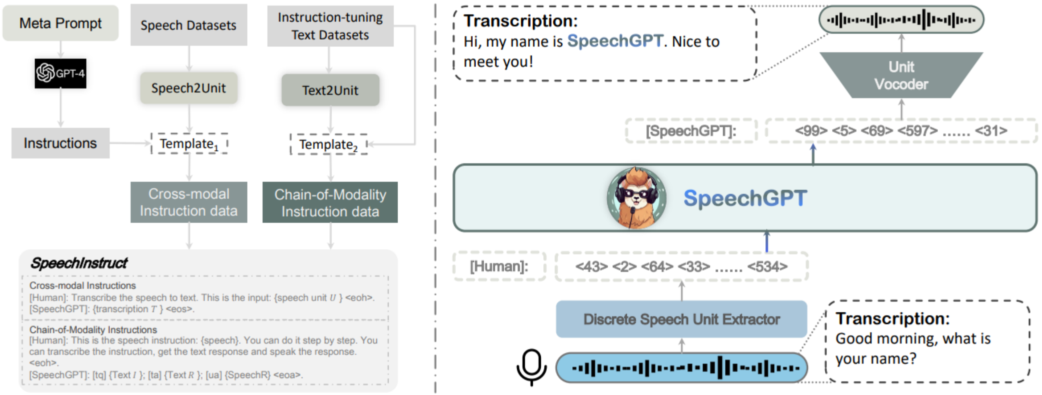
Encode input audio signals with external encoders, generating LLM-understandable signal features, feeding into LLM, which then interprets the audio based on the input text instructions and produces a textual response.

[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
…

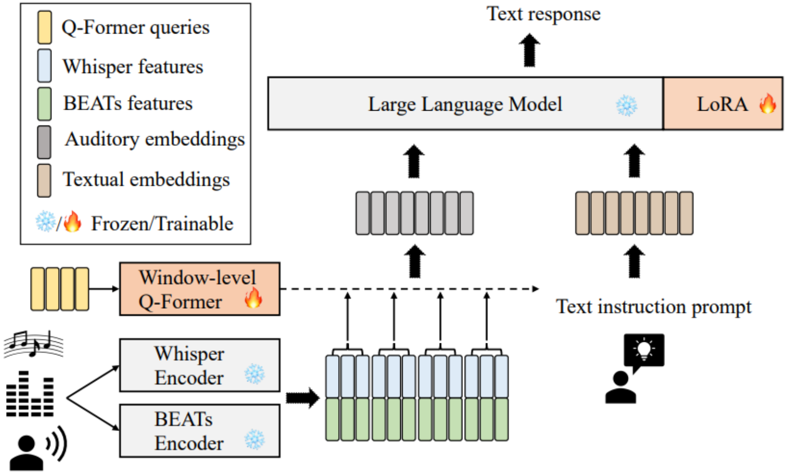
Multimodal Perceiving

Audio-perceiving MLLM

SpeechGPT



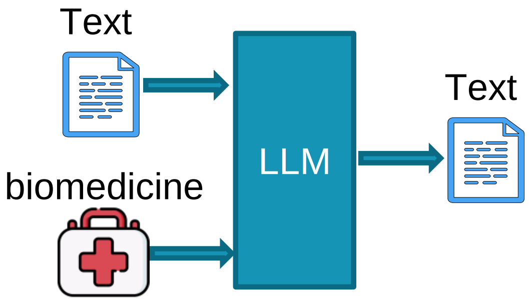
SALMONN



[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

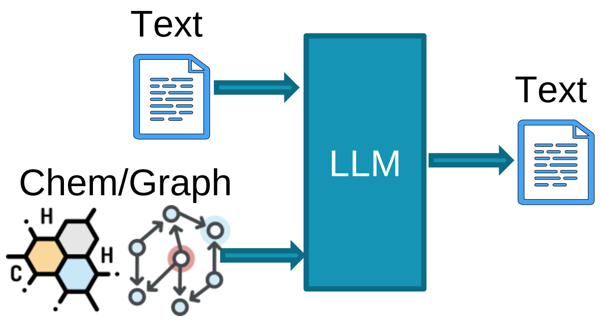
- X-perceiving MLLM
  - Bio-/Medical & Healthcare
    - BioGPT
    - DrugGPT
    - BioMedLM
    - OphGLM
    - GatorTron
    - GatorTronGPT
    - MEDITRON
    - DoctorGLM
    - BianQue
    - ClinicalGPT
    - Qilin-Med
    - ChatDoctor
    - BenTsao
    - HuatuoGPT
    - MedAlpaca
    - AlpaCare
    - 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.*

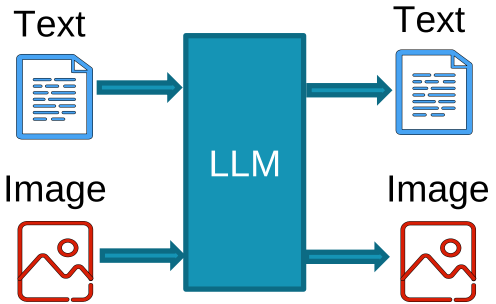
- Image Captioning
- Visual Question Answering
- Text-to-Vision Synthesis
- Vision-to-Vision Translation
- Scene Text Recognition
- Scene Text Inpainting
- …

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
  - …

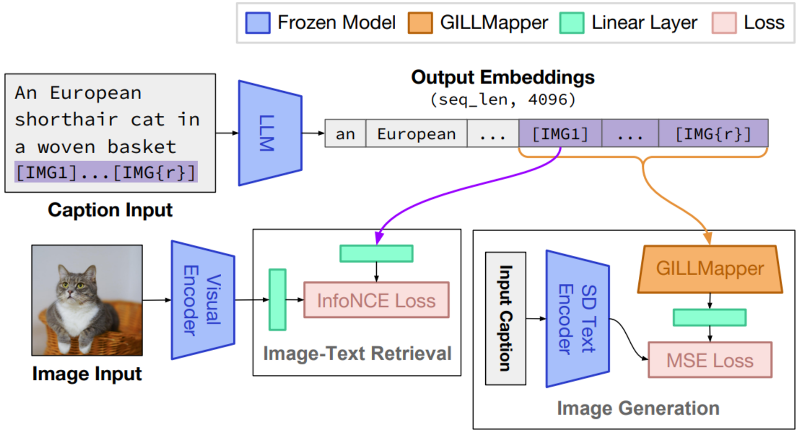


☞ *Central LLMs take as input both texts and images, after semantics comprehension, and generate both texts and images.*

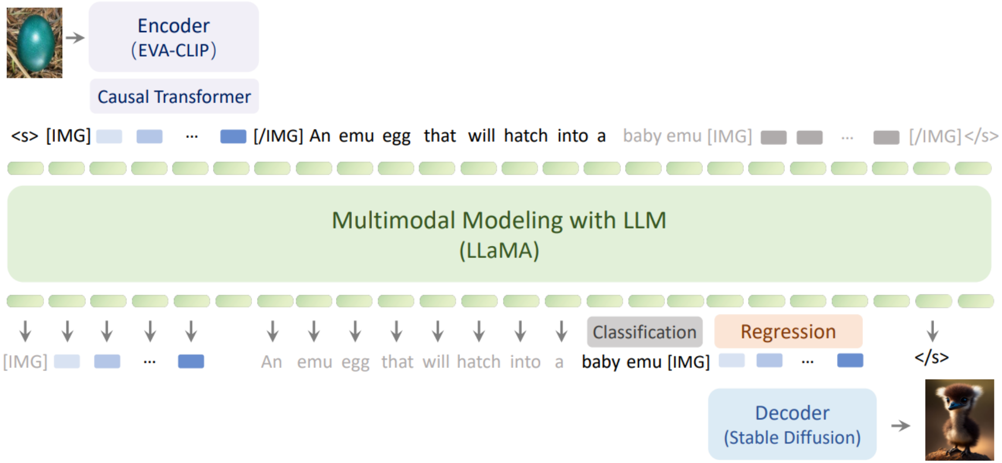
[1] Generating Images with Multimodal Language Models. 2023
[2] Generative Pretraining in Multimodality. 2023
[3] MiniGPT-5: Interleaved Vision-and-Language Generation via Generative Vokens. 2023
[4] DreamLLM: Synergistic Multimodal Comprehension and Creation. 2023
[5] LLaVA-Plus: Learning to Use Tools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

Unified MLLM: Perceiving + Generation

- Image
  - GILL



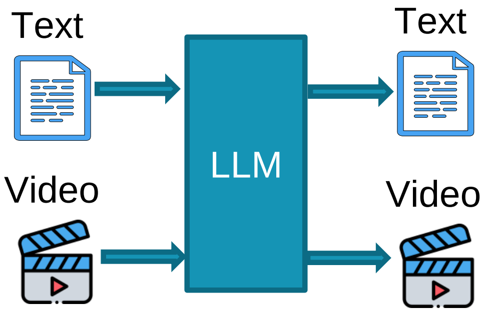
  - EMU



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

Unified MLLM: Perceiving + Generation

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

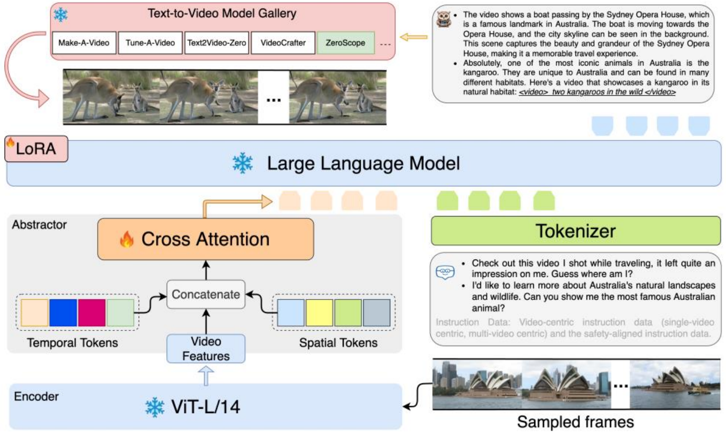


☞ *Central LLMs take as input both texts and videos, after semantics comprehension, and generate both texts and videos.*

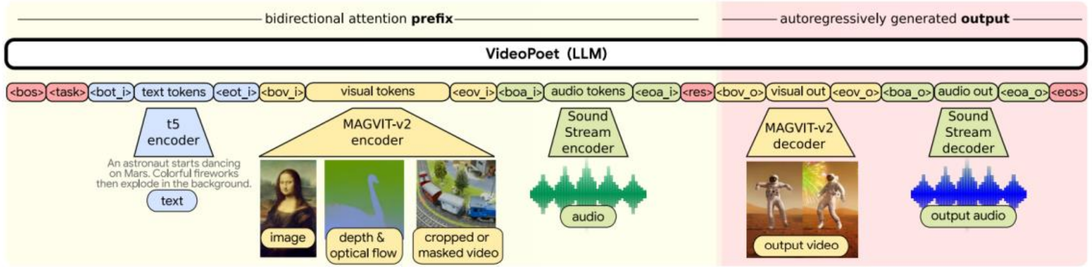
[1] GPT4Video: A Unified Multimodal Large Language Model for Instruction-Followed Understanding and Safety-Aware Generation. 2023
[2] VideoPoet: A Large Language Model for Zero-Shot Video Generation. 2023
[3] Video-LaVIT: Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization. 2024
...

Unified MLLM: Perceiving + Generation

- Video
  - GPT4Video



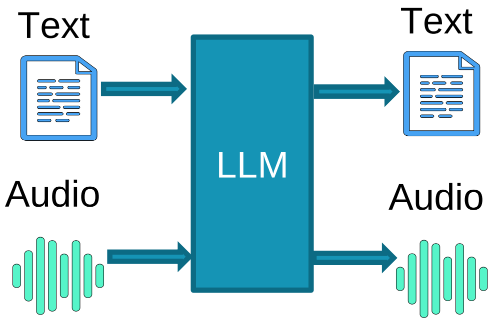
  - VideoPoet



[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,
  - …



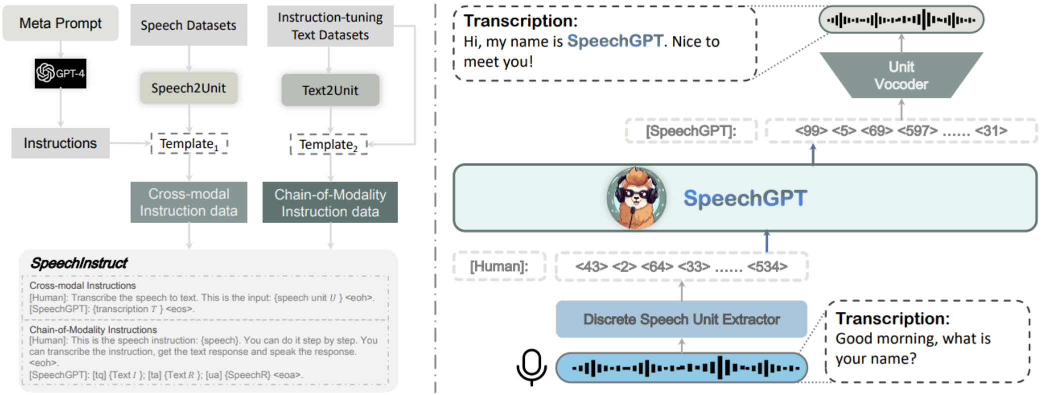
☞ *Central LLMs take as input both texts and audio, after semantics comprehension, and generate both texts and audio.*

[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
…

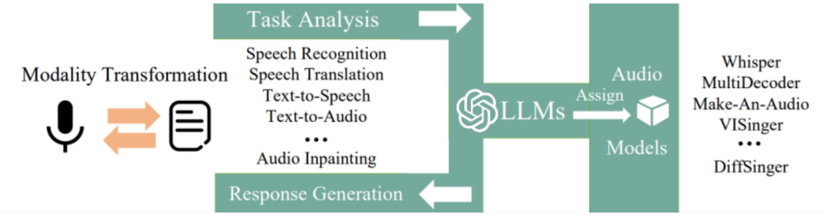
Unified MLLM: Perceiving + Generation

- ## 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 Multimodalities

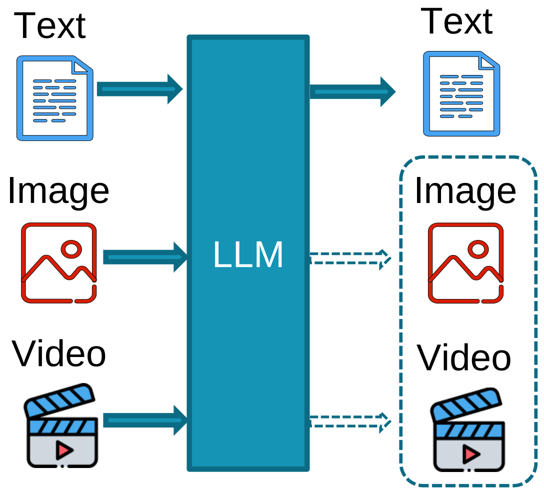
- ## Scenarios:

☞ *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
  - ...

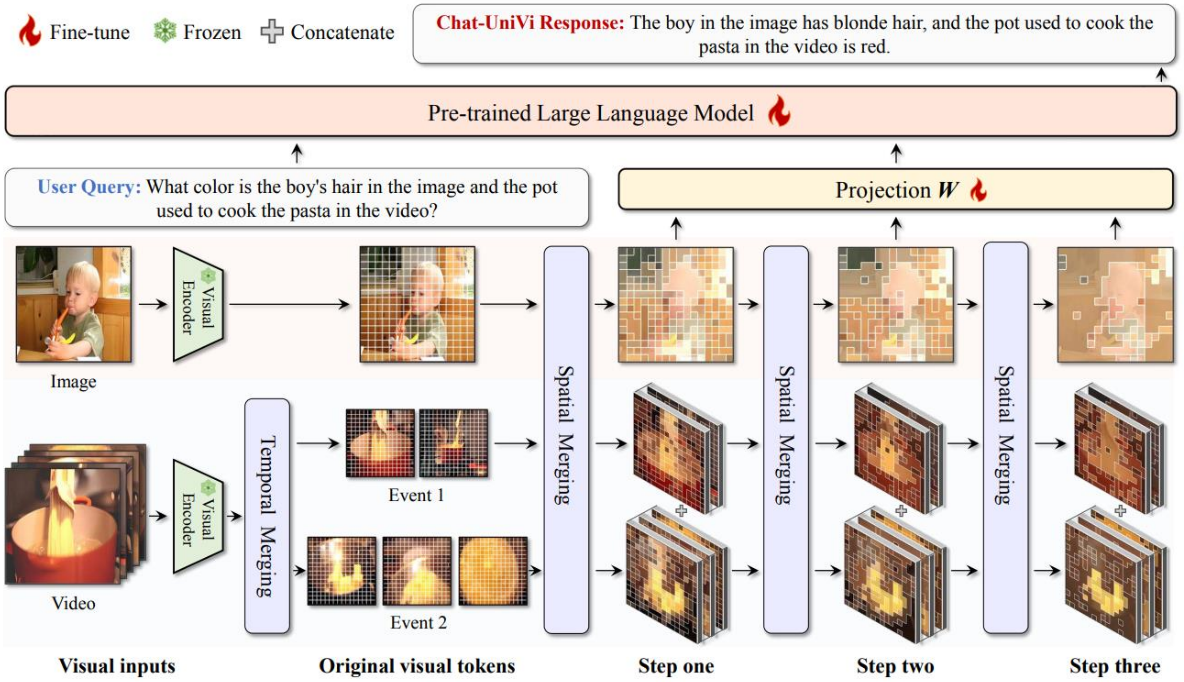


☞ *Central LLMs take as input texts, image and video, after semantics comprehension, and generate texts (maybe also image and video, or combination).*

[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

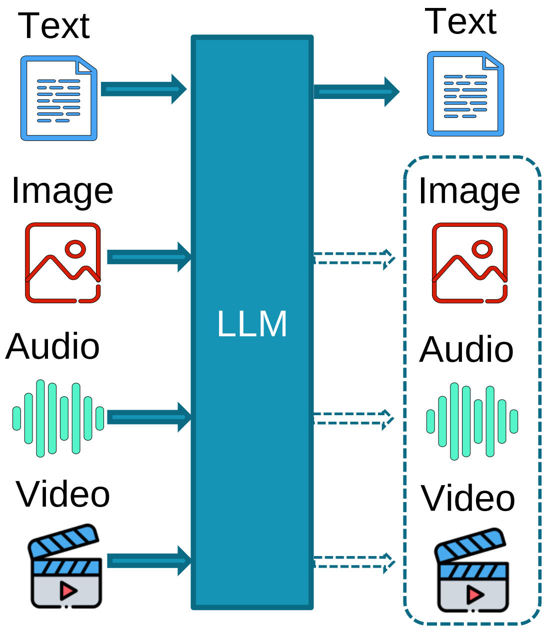
- **Text+Image+Video**
  - Chat-UniVi



[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
  - …

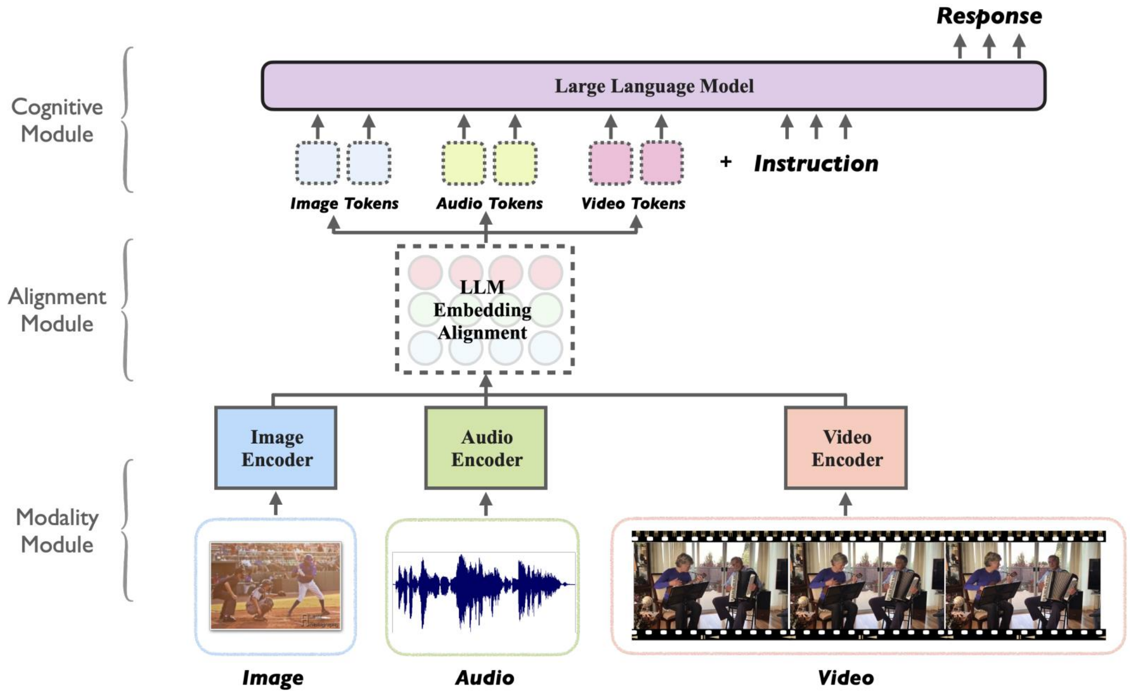


Central LLMs take as input texts, audio, image and video, and generate texts (maybe also audio, image and video, or combination).

[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

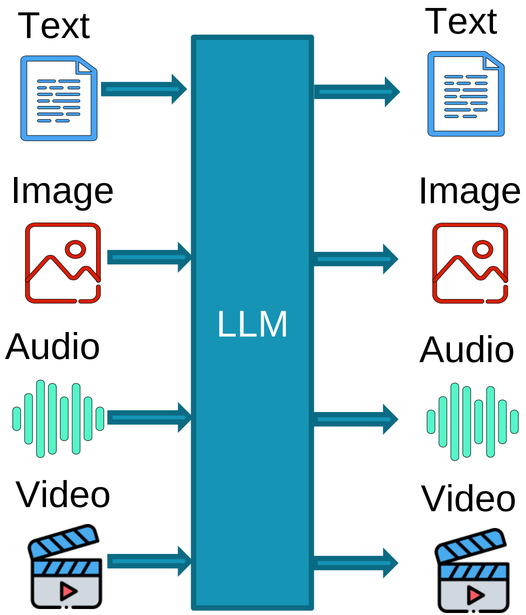
- Text+Image+Video+Audio
  - Macaw-LLM



[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
  - …

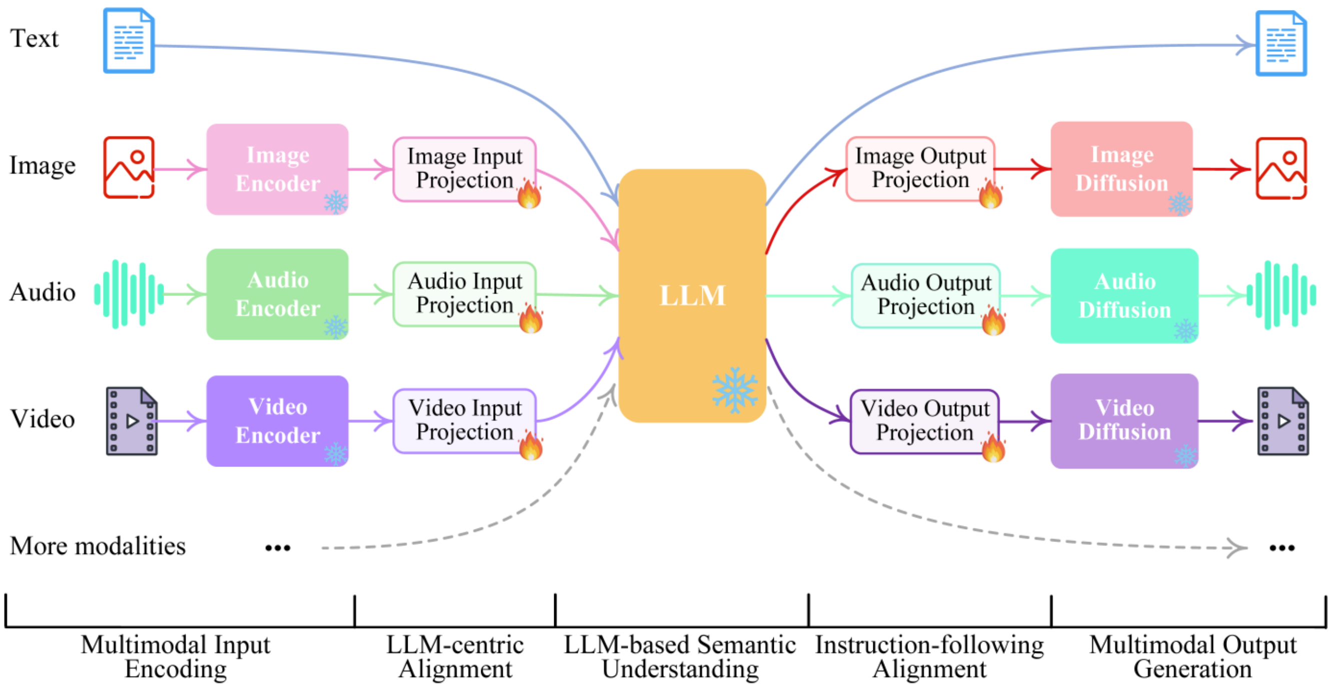


☞ *Central LLMs take as input texts, audio, image and video, and freely generate texts, audio, image and video, or combination.*

[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

- ## 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

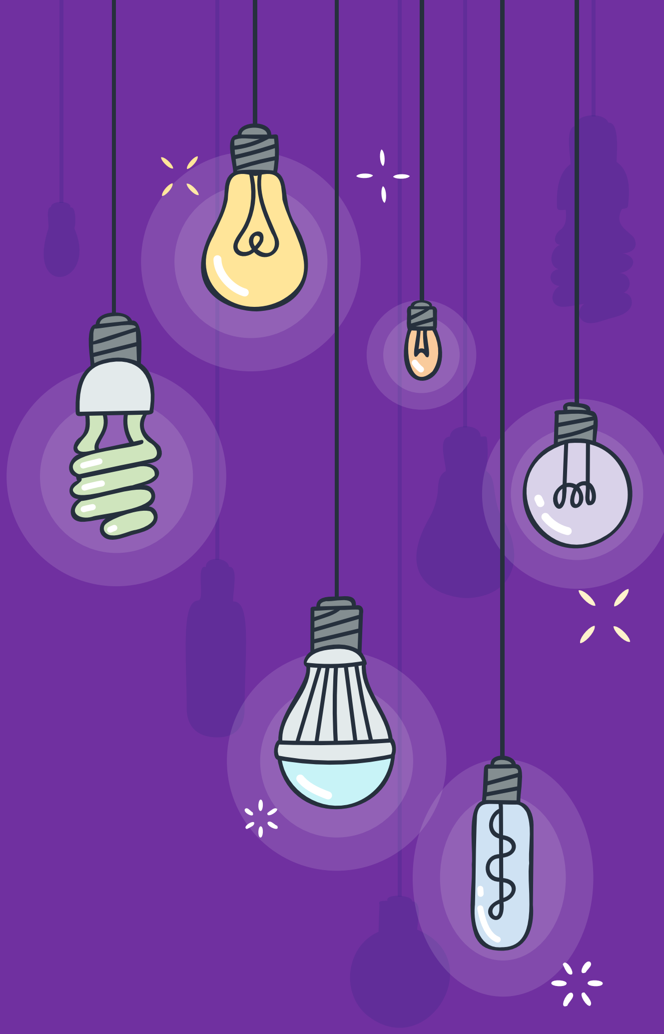
Code: https://github.com/NExT-GPT/NExT-GPT

[1] NExT-GPT: Any-to-Any Multimodal LLM. ICML. 2024

3

Future Direction

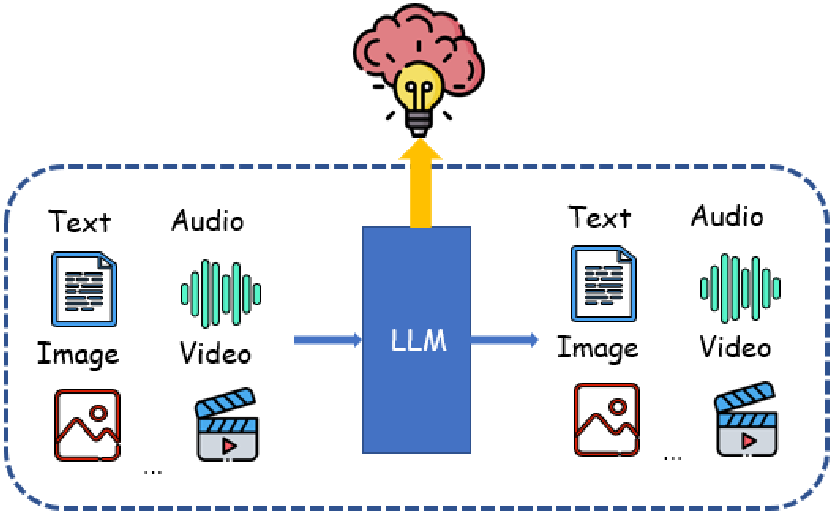
What to do next?



Future Direction

- Multimodal intelligence of MLLM relies on language's intelligence

☞ *The language intelligence of LLMs empowers multimodal intelligence.*

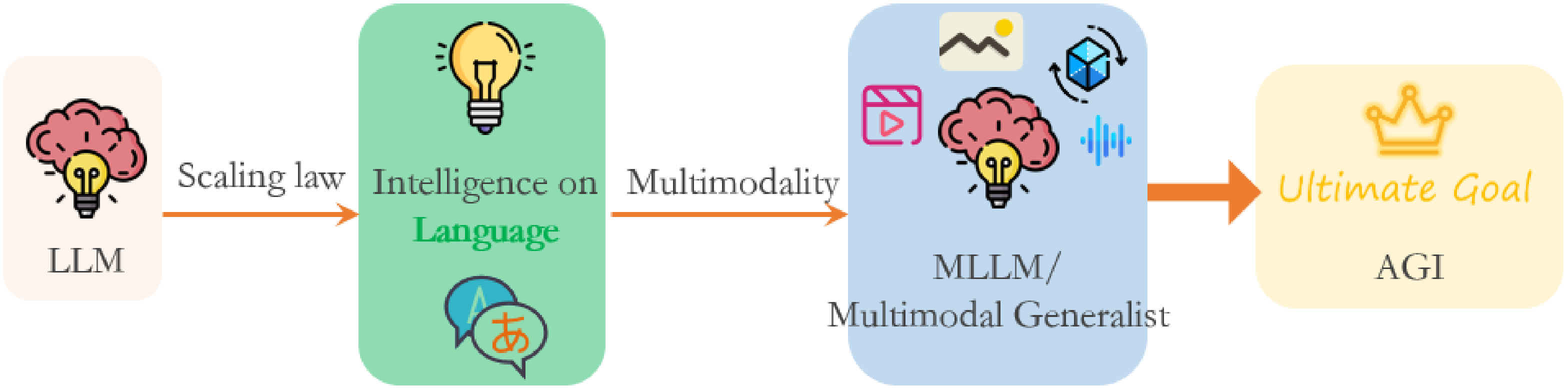


MLLM/Generalist

Future Direction

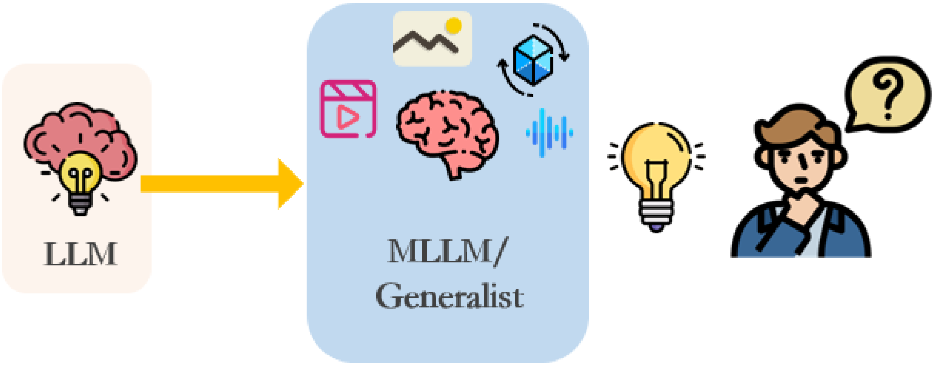
- Multimodal intelligence of MLLM relies on language's intelligence

The language intelligence of LLMs empowers multimodal intelligence.



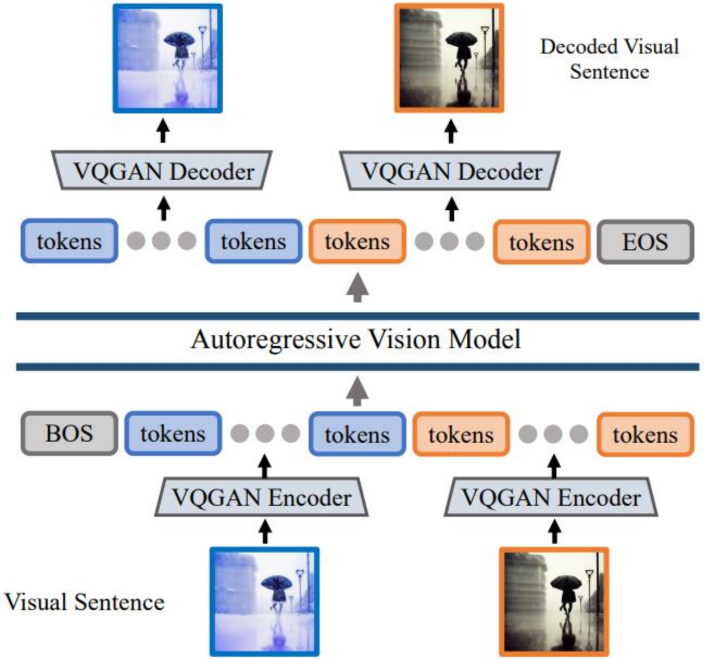
Future Direction

- **Multimodal intelligence of MLLM relies on language's intelligence**
  - *Could the scaling law and emergence success of LLMs be replicated in multimodality to achieve the intelligence of **native MLLMs**?*



Future Direction

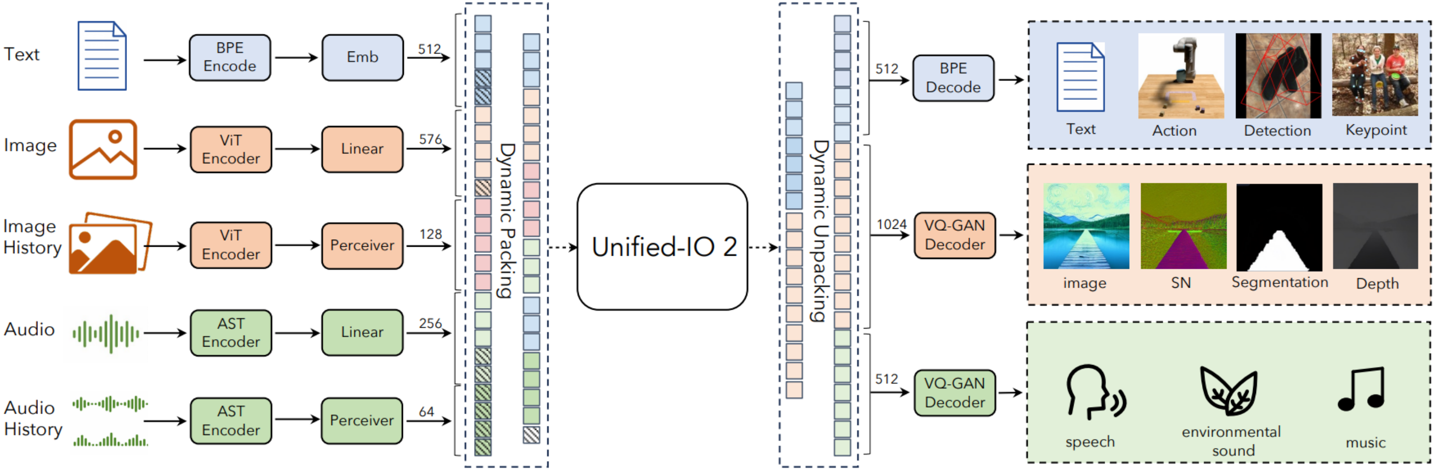
- **Exploration#1**
  - ➢ Large Vision Model (LVM)
    - mimicking LLM pretraining
    - next visual token prediction



[1] Sequential Modeling Enables Scalable Learning for Large Vision Models. CVPR. 2024

Future Direction

- **Exploration#2**
  - ➤ Unified IO-2



- mimicking LLM pretraining
- next visual token prediction

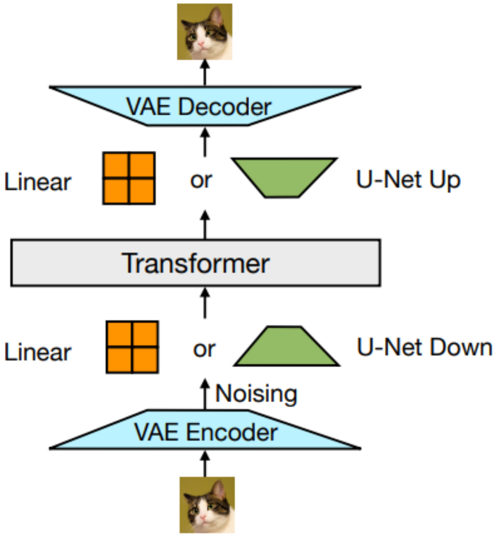
[1] Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision Language Audio and Action. CVPR. 2024

Future Direction

Open Question #1

What is the optimal model architecture under unified MLLM?

- Pipeline Agent
- Joint Encoder+LLM+Diffusion
- Joint LLM^{AR} Tokenization (VQ-VAE)
- Joint LLM^{AR}+Diffusion



1. Autoregressive Image Generation without Vector Quantization. 2024.
2. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. 2024.
3. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. 2024.

Future Direction

- ## Open Question #2

What scale of dataset is required for pre-training from scratch?

Modality	LLM/MLLM	Amount
Language	Chat-GPT4	13 Trillion text tokens
Vision	LVM	420 Billion visual tokens
Multimodalities	Unified-IO 2	1 Trillion text tokens, 1 Billion image-text pairs, 180 Million video clips, 130 Million interleaved image & text, 3 Million 3D assets, 1 Million agent trajectories

Future Direction

- ## Open Question #3

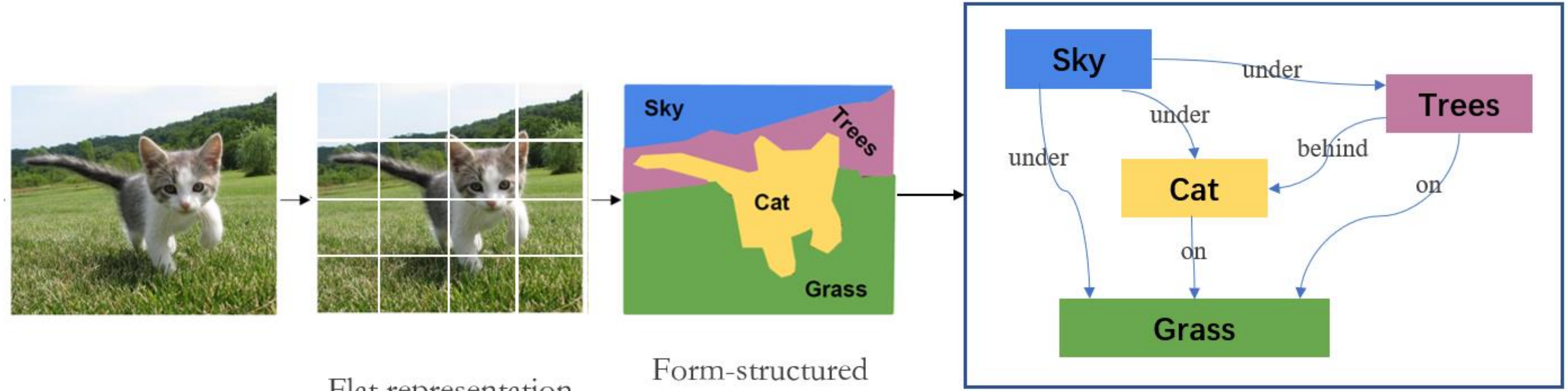
There is a gap of the downstream task performance between native MLLMs and SoTA "LLM+encoder/decoder" architecture MLLMs.

How can this gap be bridged?

Future Direction

- ## Open Question #4

What is the optimal representation method for multimodal data?



Flat representation

Form-structured representation

Semantically-structured representation

[1] Shengqiong Wu, **Hao Fei**, Hanwang Zhang, Tat-Seng Chua, Shuicheng Yan. Towards Semantic Equivalence of Tokenization in Multimodal LLM. 2024

Thank you!

Q&A