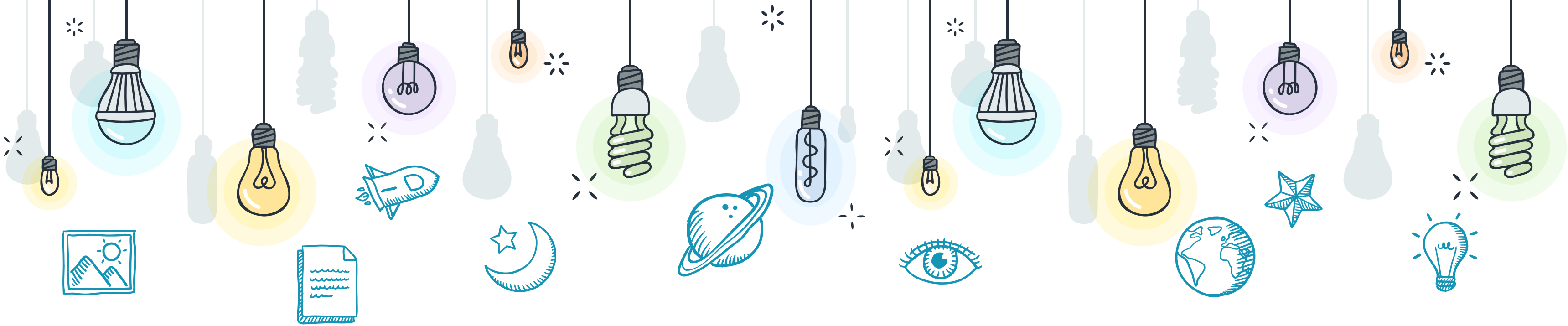


From Multimodal LLM to Human-level AI

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

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

CVPR
JUNE 17-21, 2024
SEATTLE, WA

NUS National University of Singapore · WISCONSIN UNIVERSITY OF WISCONSIN-MADISON · UNIVERSITY OF MARYLAND · 上海交通大學 SHANGHAI JIAO TONG UNIVERSITY · NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE · 2050 RESEARCH

Hao Fei

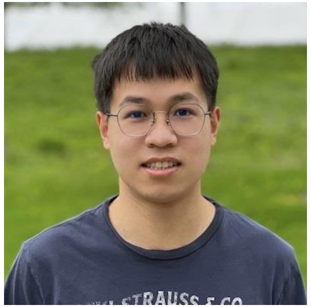
National University of Singapore

Yuan Yao

National University of Singapore

Ao Zhang

National University of Singapore

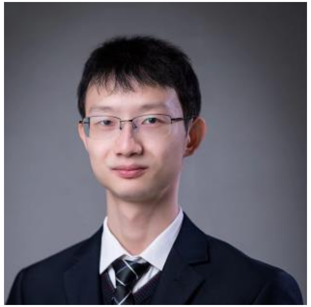


Haotian Liu

University of Wisconsin-Madison

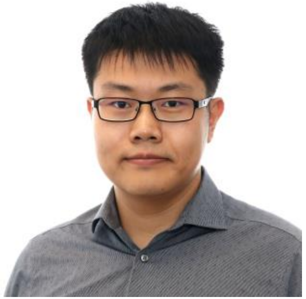
Fuxiao Liu

University of Maryland, College Park



Zhuosheng Zhang

Shanghai Jiao Tong University



Hanwang Zhang

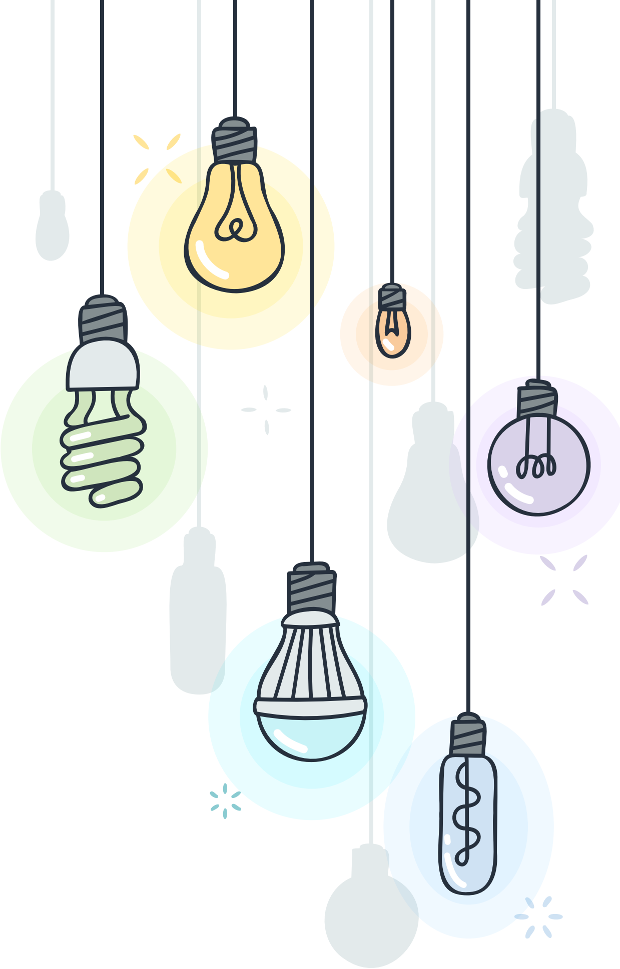
Nanyang Technological University

Shuicheng Yan

Kunlun 2050 Research, Skywork AI

Part-III

Modality and Functionality



Hao Fei

Research Fellow

National University of Singapore

http://haofei.vip/

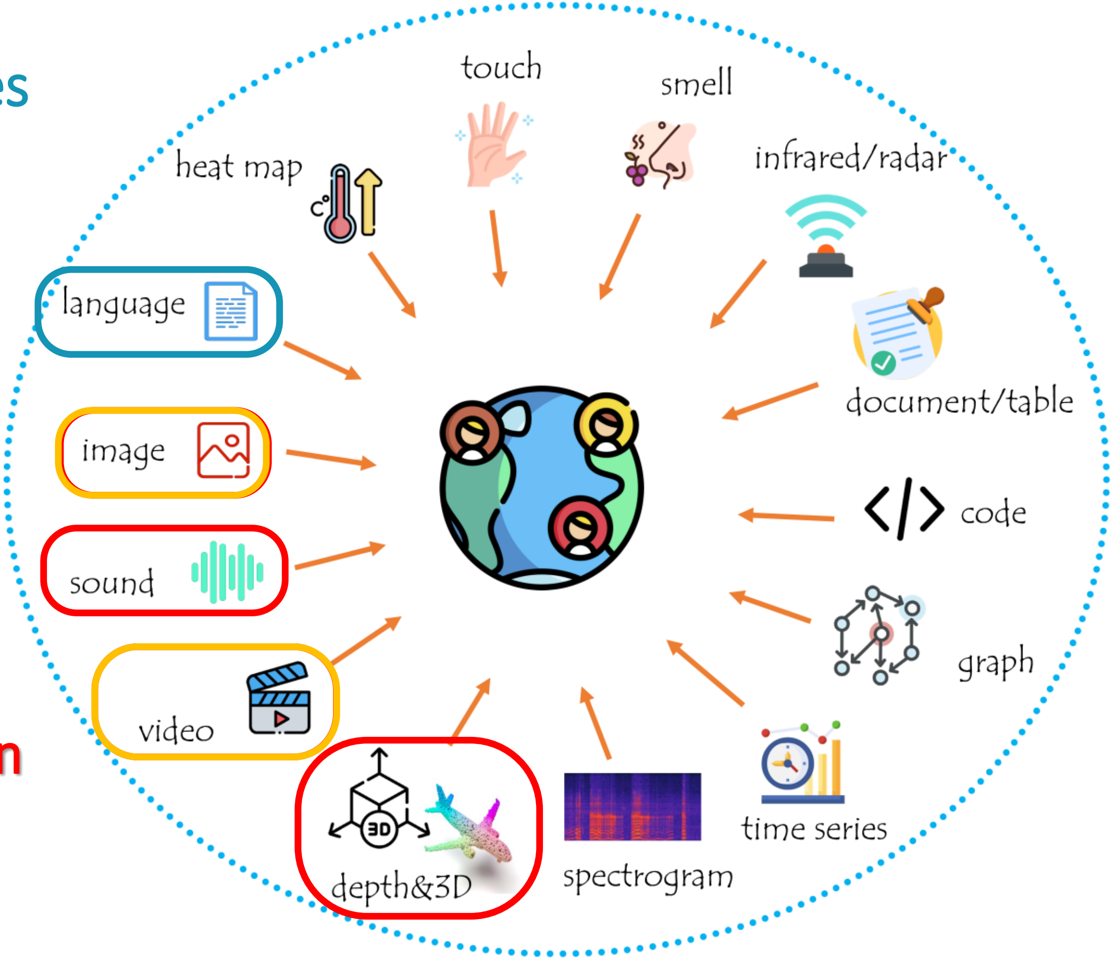
Table of Content

Modality & Functionality

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

Overview of Modality and Functionality

- ## Modalities



Language + Vision

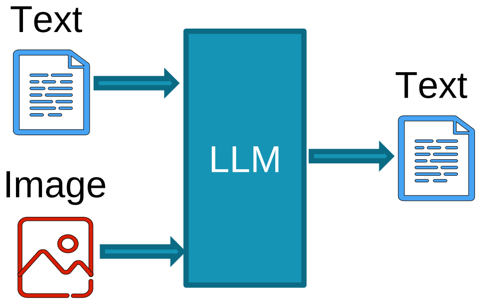
Overview of Modality and Functionality

	Modality (w/ Language)			
	Image	**Video**	**Audio**	**3D**
Input-side Perceiving	Flamingo, Kosmos-1, Blip2, mPLUG-Owl, Mini-GPT4, LLaVA, InstructBLIP, VPGTrans, CogVLM, Monkey, Chameleon, Otter, Qwen-VL, GPT-4v, SPHINX, Yi-VL, Fuyu, …	VideoChat, 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, …			–

Multimodal Perceiving

Image-perceiving MLLM

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



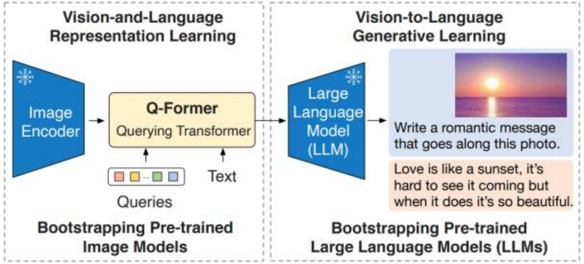
☞ *Encode input images with external image encoders, generating LLM-understandable visual feature, which is then fed into the LLM. LLM then interprets the input images based on the input text instructions and produces a textual response.*

[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*
…

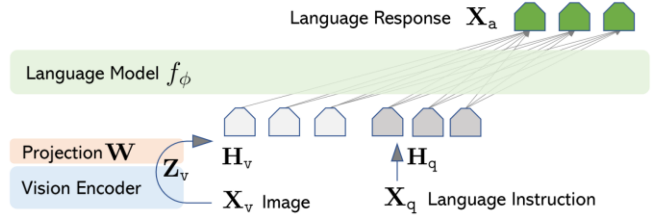
Multimodal Perceiving

- ## Image-perceiving MLLM

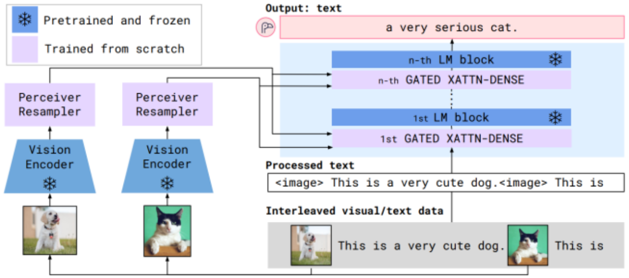
- Blip2


Vision-and-Language Representation Learning
Vision-to-Language Generative Learning
Image Encoder
Q-Former
Querying Transformer
Queries
Text
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained Image Models
Bootstrapping Pre-trained Large Language Models (LLMs)

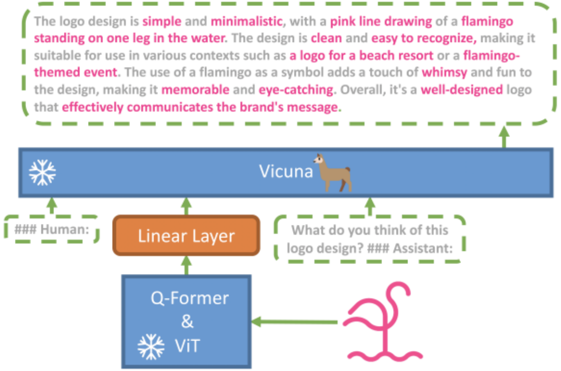
- LLaVA



- Flamingo



- Mini-GPT4


The logo design is simple and minimalistic, with a pink line drawing of a flamingo standing on one leg in the water. The design is clean and easy to recognize, making it suitable for use in various contexts such as a logo for a beach resort or a flamingo-themed event. The use of a flamingo as a symbol adds a touch of whimsy and fun to the design, making it memorable and eye-catching. Overall, it's a well-designed logo that effectively communicates the brand's message.

[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*

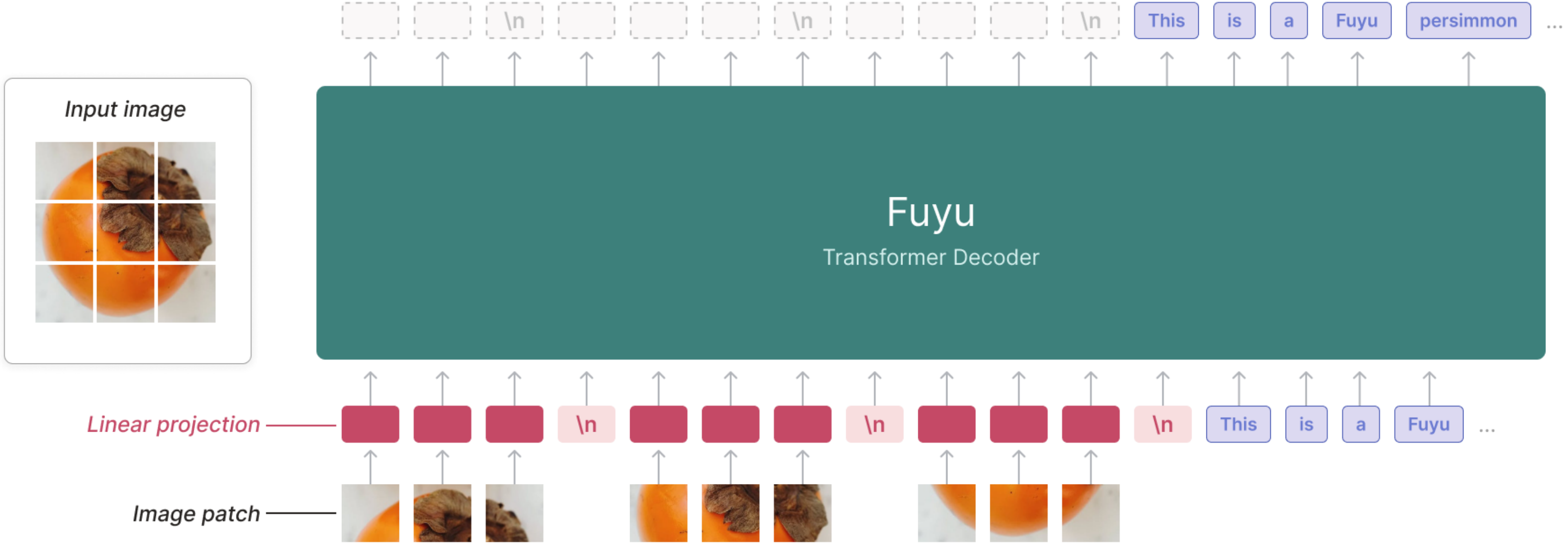
[4] *A Survey on Multimodal Large Language Models. https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models, 2023.*

Multimodal Perceiving

- ## Image-perceiving MLLM

  - Fuyu

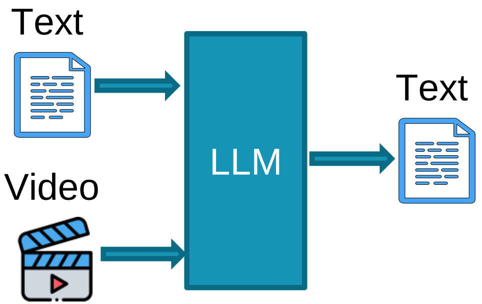
☞ *Unlike all other existing image-oriented MLLMs, Fuyu processes image information without a frontend image encoder, and instead directly inputs image patches into the LLM for interpretation.*



[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
  - …



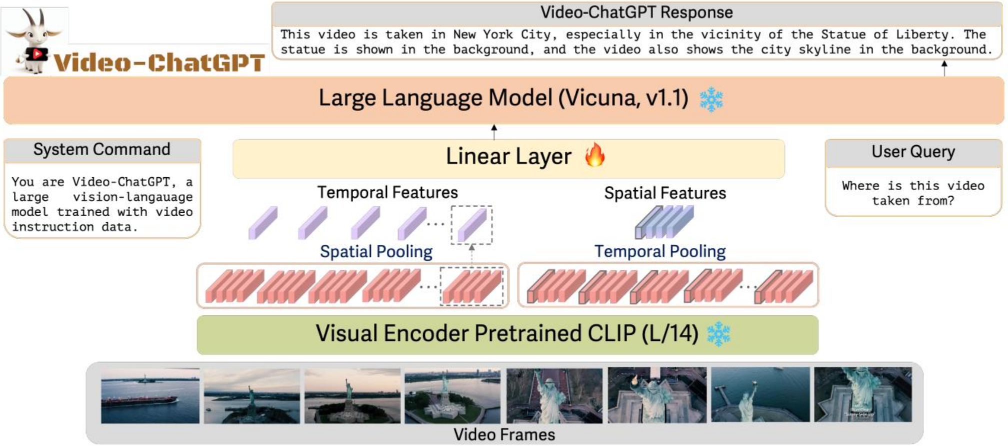
☞ *Encode input videos with external video encoders, generating LLM-understandable visual feature, feeding into LLM, which then interprets the input videos based on the input text instructions and produces a textual response.*

[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

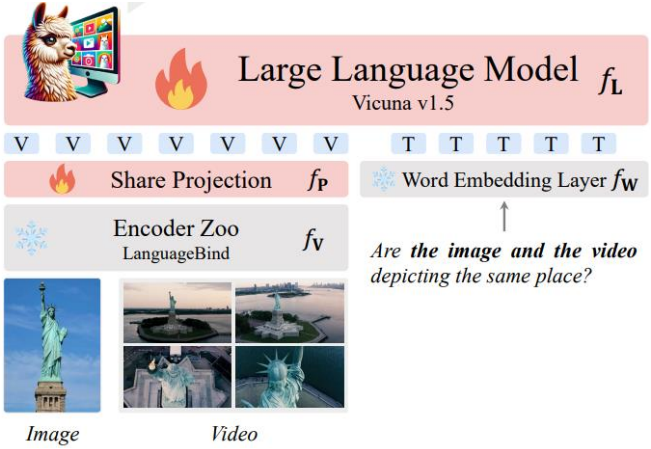


Video-ChatGPT Response: This video is taken in New York City, especially in the vicinity of the Statue of Liberty. The statue is shown in the background, and the video also shows the city skyline in the background.

System Command: You are Video-ChatGPT, a large vision-langauage model trained with video instruction data.

User Query: Where is this video taken from?

Video-LLaVA



Are the image and the video depicting the same place?

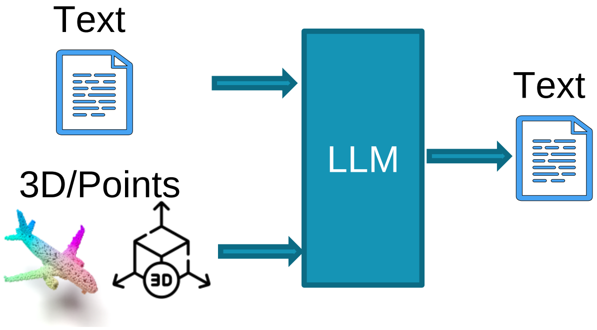
[1] Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. 2023

[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
  - …



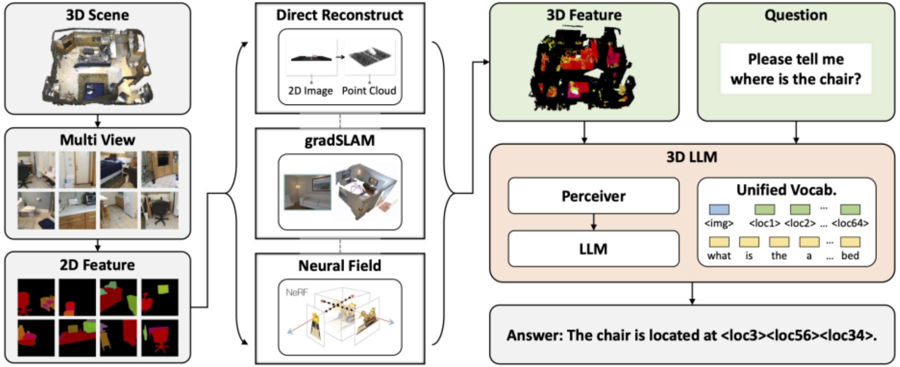
☞ *Encode input 3D information with external encoders, generating LLM-understandable 3D feature, feeding into LLM, which then interprets the input 3D/points based on the input text instructions and produces a textual response.*

[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
…

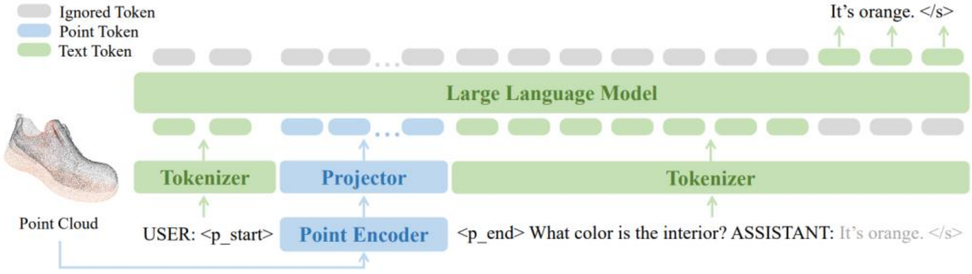
Multimodal Perceiving

- ## 3D-perceiving MLLM

3D-LLM



PointLLM

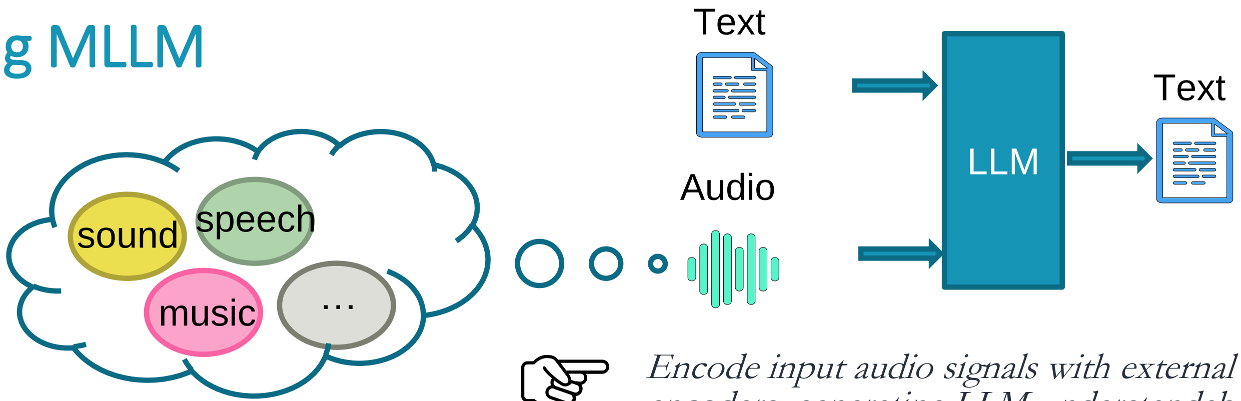


[1] 3D-LLM: Injecting the 3D World into Large Language Models. 2023
[2] PointLLM: Empowering Large Language Models to Understand Point Clouds. 2023

Multimodal Perceiving

- ## Audio-perceiving MLLM

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



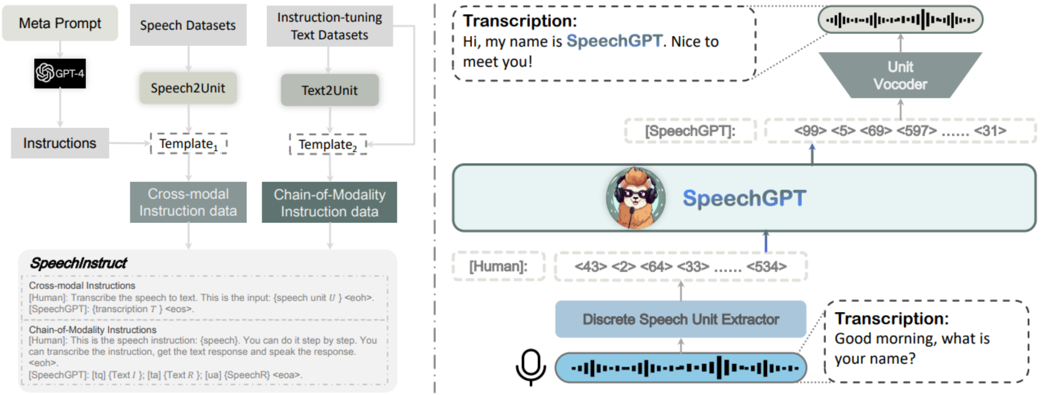
☞ *Encode input audio signals with external encoders, generating LLM-understandable signal features, feeding into LLM, which then interprets the audio based on the input text instructions and produces a textual response.*

[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



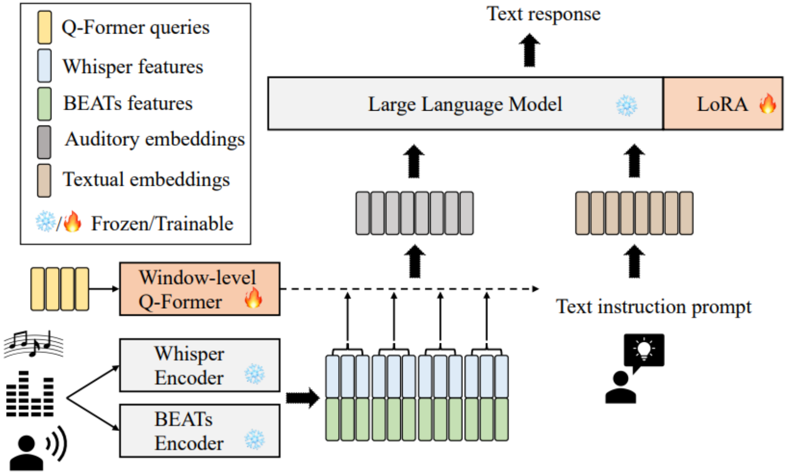
Cross-modal Instructions
[Human]: Transcribe the speech to text. This is the input: {speech unit U } <eoh>.
[SpeechGPT]: {transcription T } <eos>.

Chain-of-Modality Instructions
[Human]: This is the speech instruction: {speech}. You can do it step by step. You can transcribe the instruction, get the text response and speak the response. <eoh>.
[SpeechGPT]: [tq] {Text I }; [ta] {Text R }; [ua] {SpeechR} <eoa>.

Transcription:
Hi, my name is SpeechGPT. Nice to meet you!

Transcription:
Good morning, what is your name?

SALMONN



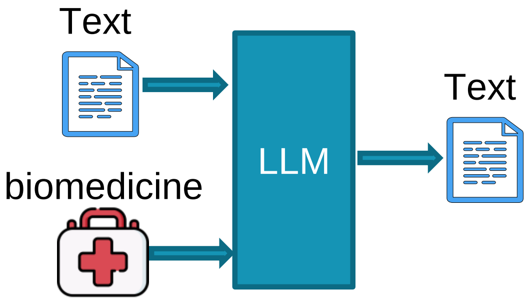
[1] SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. 2023

[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*

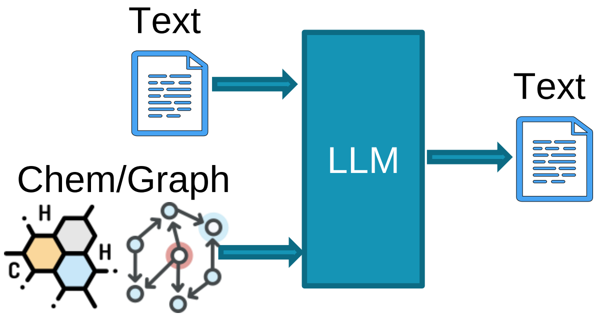
[5] *AlpaCare:Instruction-tuned Large Language Models for Medical Application. 2023*

[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.*

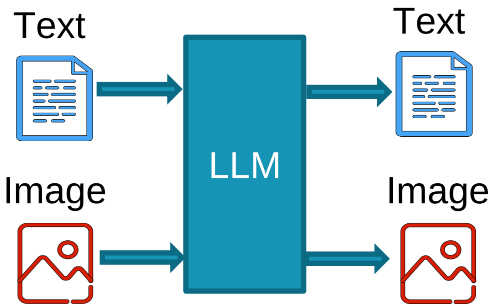
- Image Captioning
- Visual Question Answering
- Text-to-Vision Synthesis
- Vision-to-Vision Translation
- Scene Text Recognition
- Scene Text Inpainting
- …

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
  - …



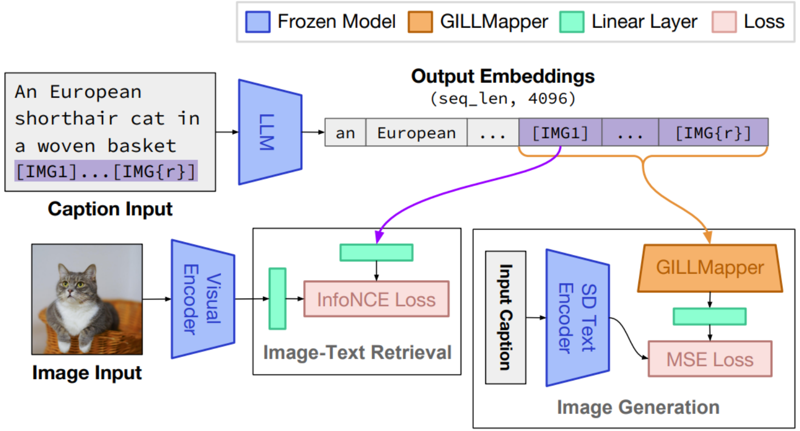
☞ *Central LLMs take as input both texts and images, after semantics comprehension, and generate both texts and images.*

[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
…

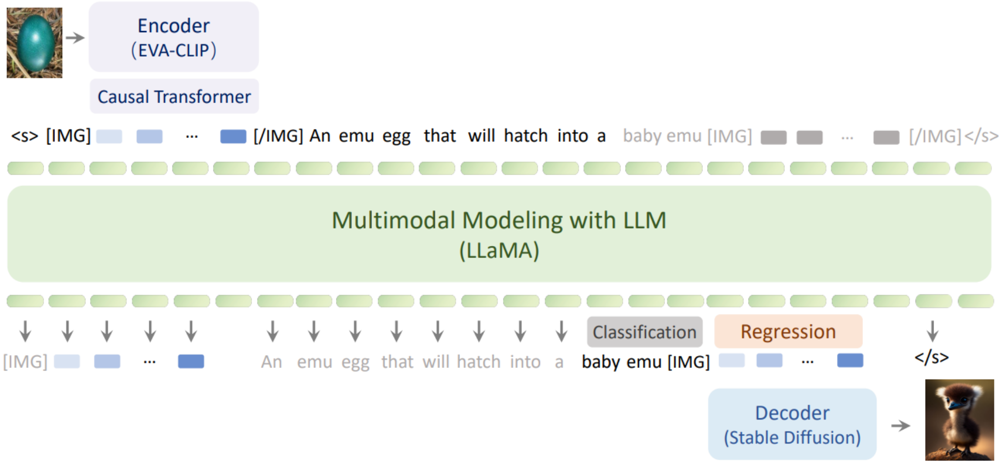
Unified MLLM: Perceiving + Generation

- ## Image

GILL



EMU



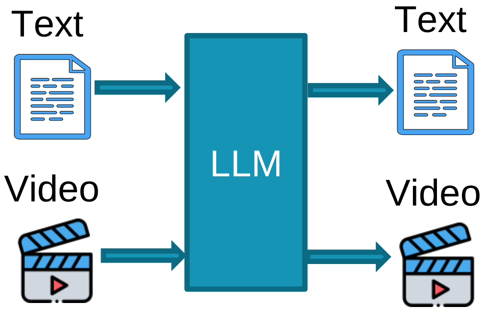
Encoder
(EVA-CLIP)
Causal Transformer
<s> [IMG] ··· [/IMG] An emu egg that will hatch into a baby emu [IMG] ··· [/IMG]</s>
Multimodal Modeling with LLM
(LLaMA)
Classification
Regression
[IMG] ··· An emu egg that will hatch into a baby emu [IMG] ··· </s>
Decoder
(Stable Diffusion)

[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.*

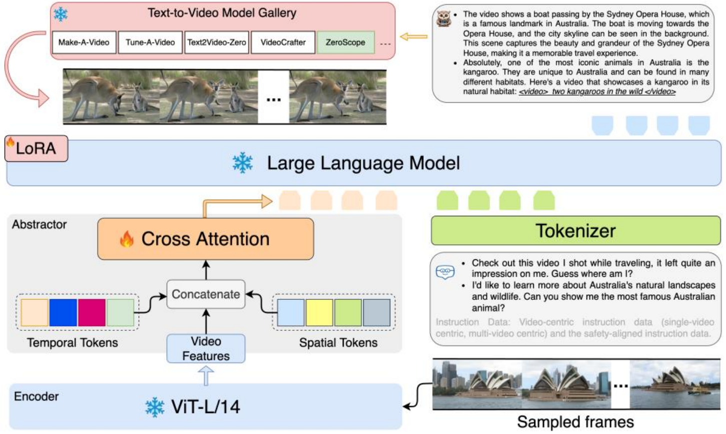
[1] *GPT4Video: A Unified Multimodal Large Language Model for Instruction-Followed Understanding and Safety-Aware Generation. 2023*
[2] *VideoPoet: A Large Language Model for Zero-Shot Video Generation. 2023*
[3] *Video-LaVIT: Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization.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

Unified MLLM: Perceiving + Generation

- ## Video

  - ### GPT4Video

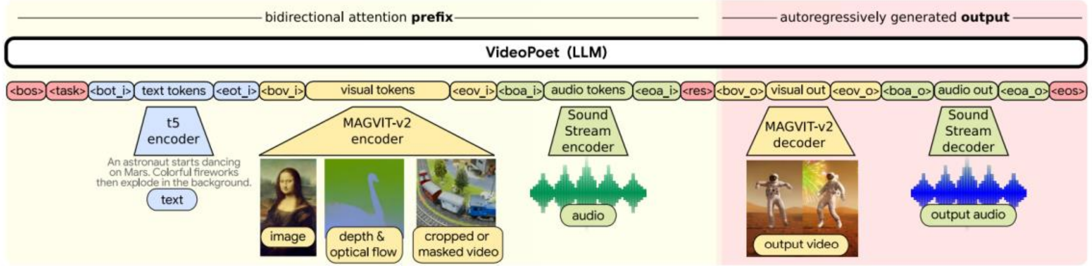
  - ### VideoPoet



- The video shows a boat passing by the Sydney Opera House, which is a famous landmark in Australia. The boat is moving towards the Opera House, and the city skyline can be seen in the background. This scene captures the beauty and grandeur of the Sydney Opera House, making it a memorable travel experience.
- Absolutely, one of the most iconic animals in Australia is the kangaroo. They are unique to Australia and can be found in many different habitats. Here's a video that showcases a kangaroo in its natural habitat: <video> two kangaroos in the wild </video>

- Check out this video I shot while traveling, it left quite an impression on me. Guess where am I?
- I'd like to learn more about Australia's natural landscapes and wildlife. Can you show me the most famous Australian animal?

Instruction Data: Video-centric instruction data (single-video centric, multi-video centric) and the safety-aligned instruction data.

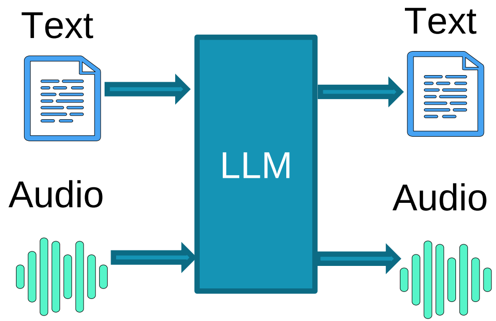


[1] *GPT4Video: A Unified Multimodal Large Language Model for Instruction-Followed Understanding and Safety-Aware Generation. 2023*

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

Unified MLLM: Perceiving + Generation

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



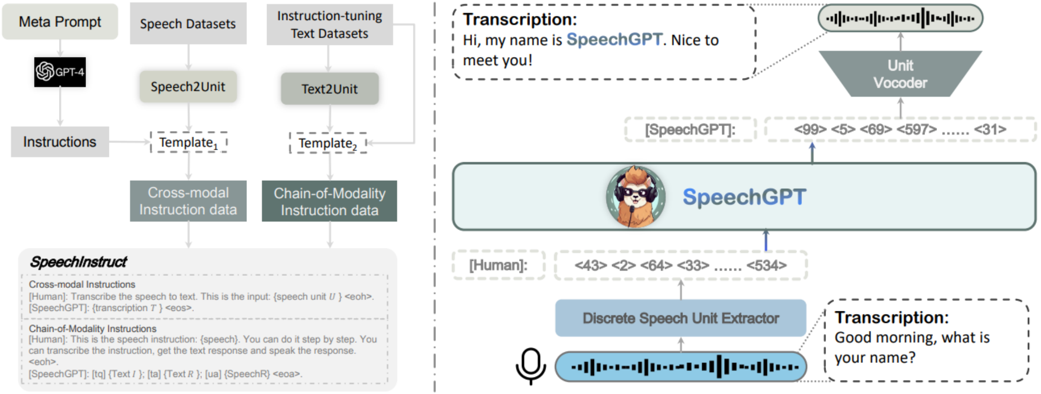
☞ *Central LLMs take as input both texts and audio, after semantics comprehension, and generate both texts and audio.*

[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



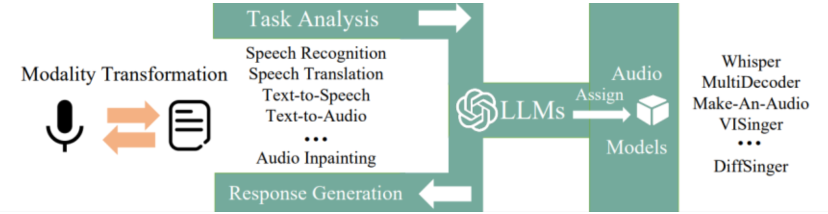
SpeechInstruct

Cross-modal Instructions
[Human]: Transcribe the speech to text. This is the input: {speech unit U } <eoh>.
[SpeechGPT]: {transcription T } <eos>.

Chain-of-Modality Instructions
[Human]: This is the speech instruction: {speech}. You can do it step by step. You can transcribe the instruction, get the text response and speak the response. <eoh>.
[SpeechGPT]: [tq] {Text I }; [ta] {Text R }; [ua] {SpeechR} <eoa>.

Transcription:
Hi, my name is SpeechGPT. Nice to meet you!

Transcription:
Good morning, what is your name?



[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

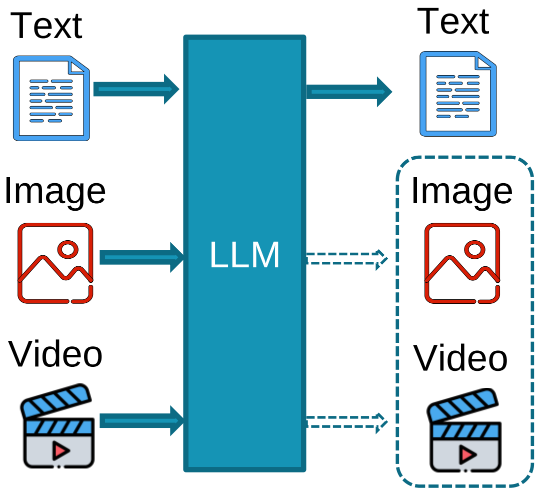
- ## Scenarios:

☞ *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*

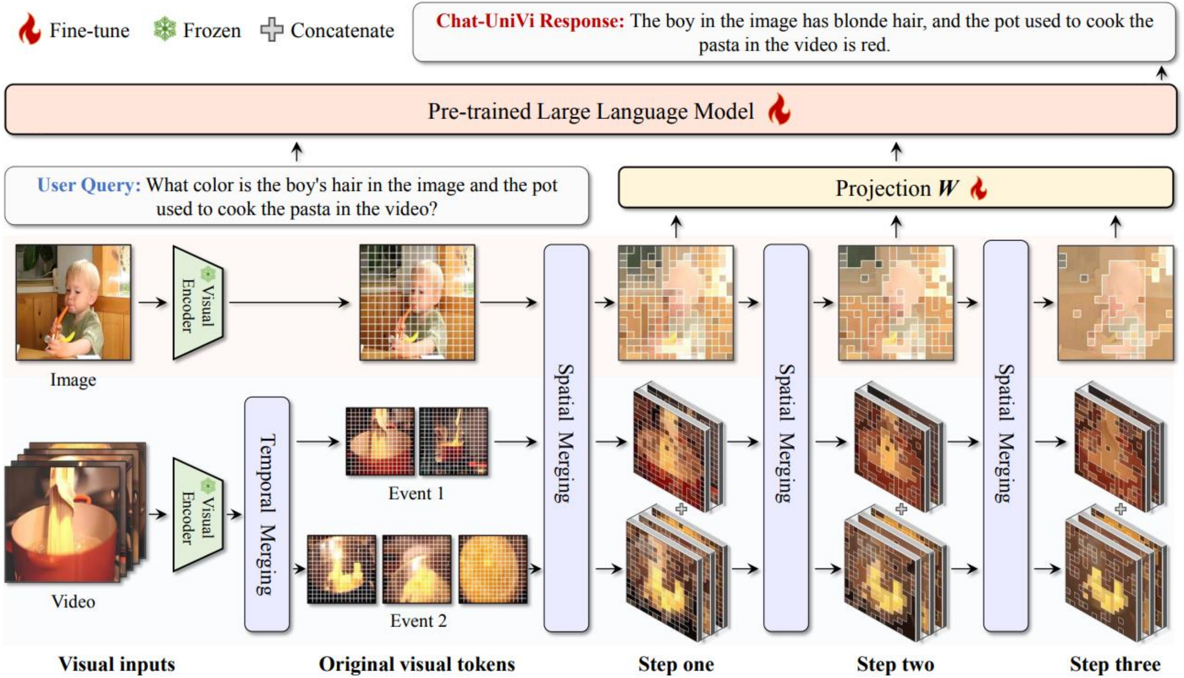
[3] *LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models. 2023*

…

Unified MLLM: Harnessing Multi-Modalities

- ## Text+Image+Video

  - Chat-UniVi

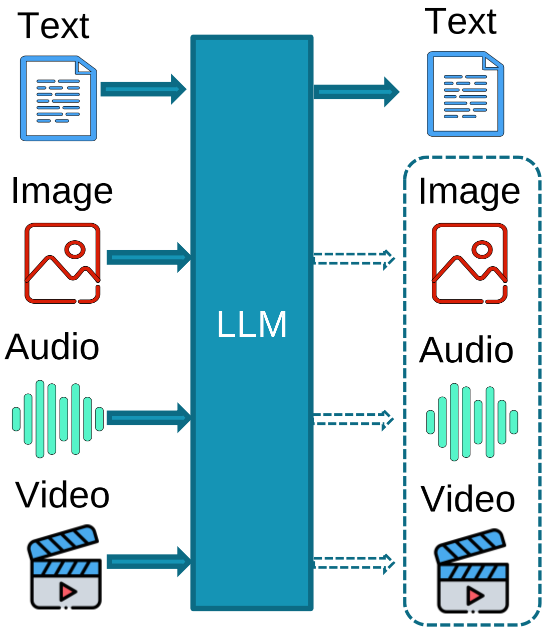

Fine-tune
Frozen
Concatenate
Chat-UniVi Response: The boy in the image has blonde hair, and the pot used to cook the pasta in the video is red.
Pre-trained Large Language Model
User Query: What color is the boy's hair in the image and the pot used to cook the pasta in the video?
Projection W
Image
Visual Encoder
Video
Visual Encoder
Temporal Merging
Event 1
Event 2
Spatial Merging
Spatial Merging
Spatial Merging
Visual inputs
Original visual tokens
Step one
Step two
Step three

[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
  - ...

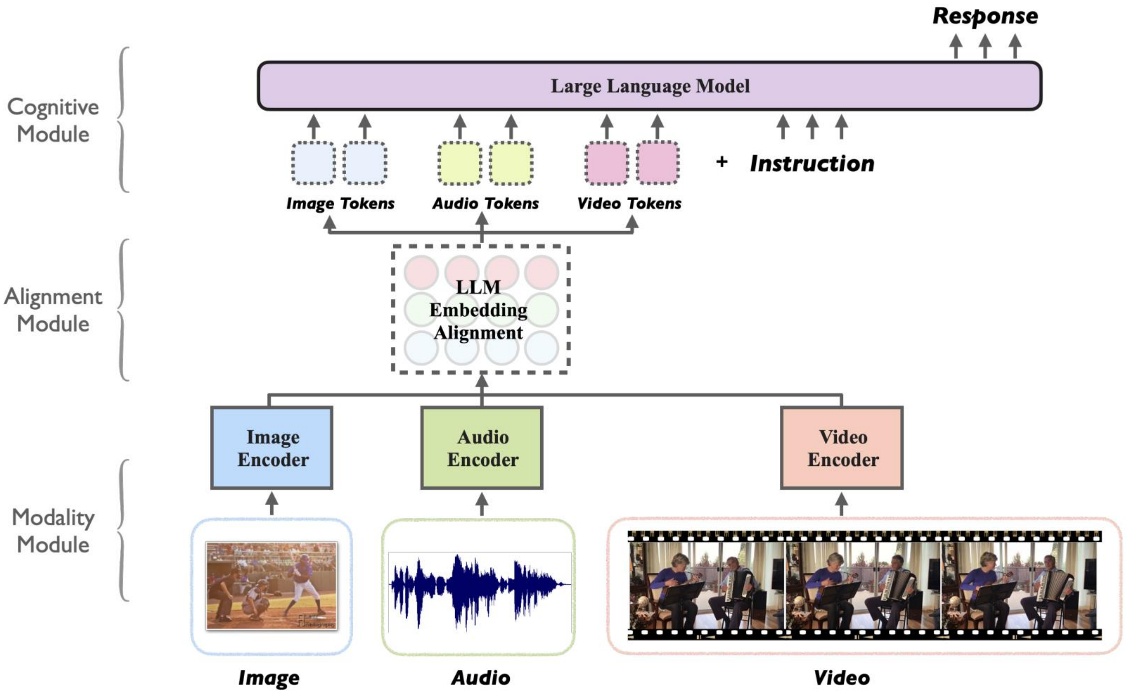


☞ *Central LLMs take as input texts, audio, image and video, and generate texts (maybe also audio, image and video, or combination).*

[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

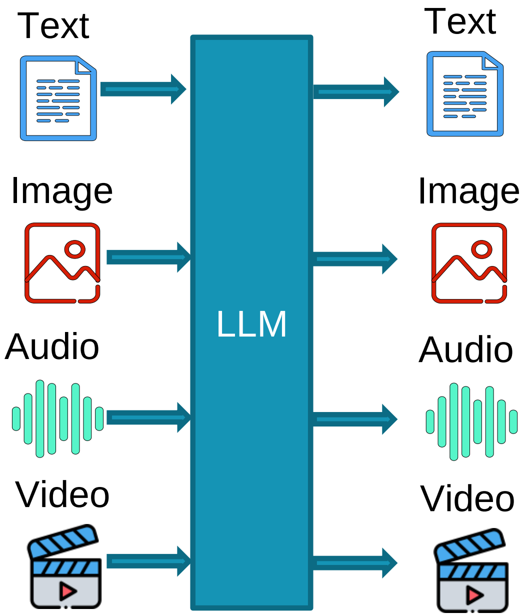
- Text+Image+Video+Audio
  - Macaw-LLM



[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
  - ...

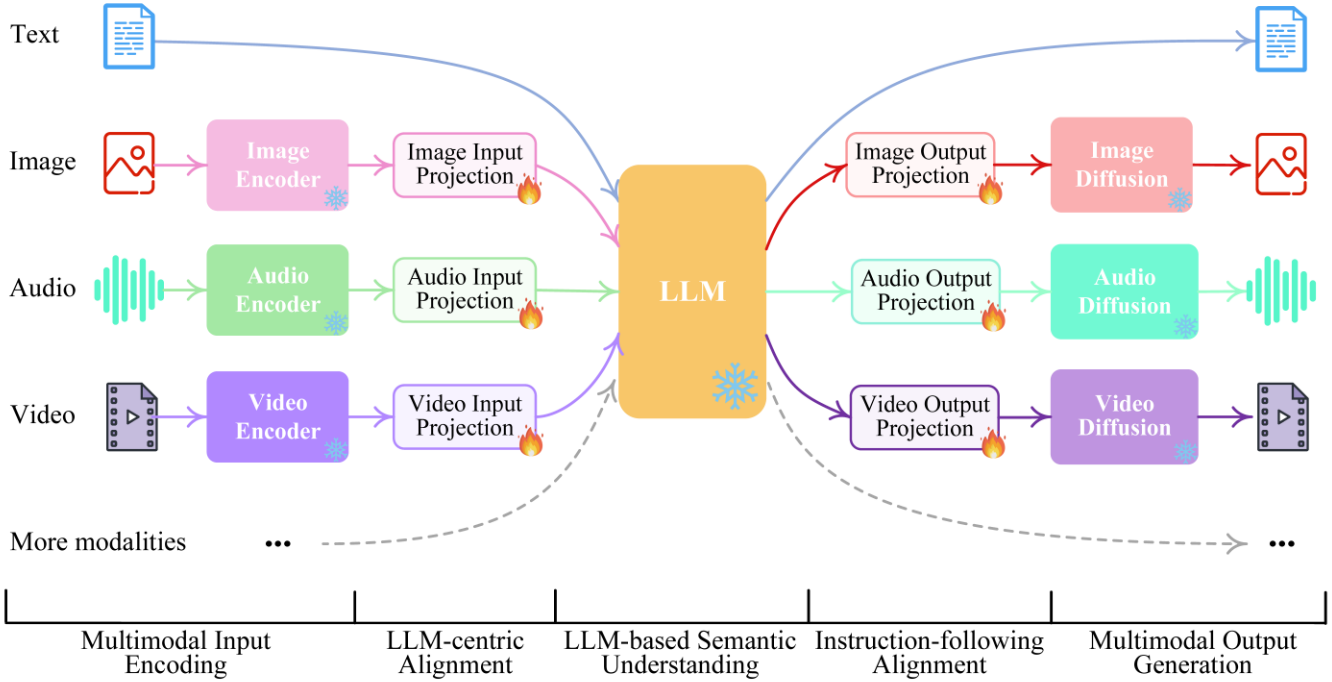


☞ *Central LLMs take as input texts, audio, image and video, and freely generate texts, audio, image and video, or combination.*

[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

Unified MLLM: Harnessing Multi-Modalities

- Any-to-Any MLLM NExT-GPT
  - NExT-GPT

Text + Audio

↓

Text + Image + Video

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

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

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

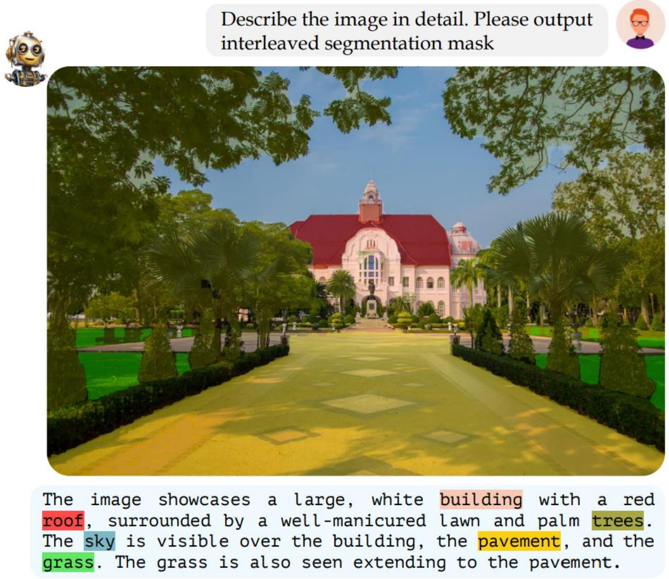
[1] NExT-GPT: Any-to-Any Multimodal LLM. 2023

Fine-grained Capability of MLLM

Pixel-level Vision MLLM

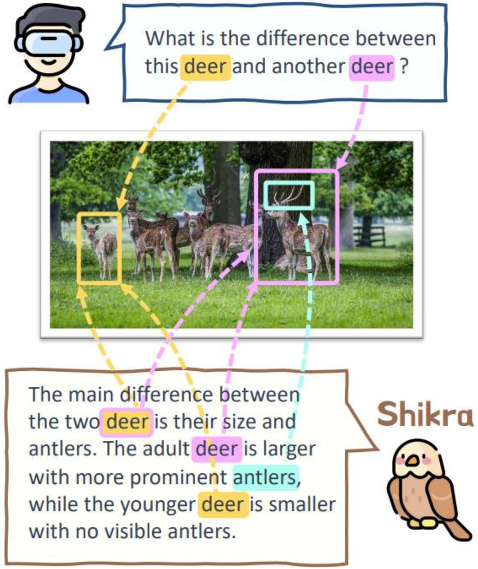
The vision MLLMs described above generally only support coarse-grained, instance-level visual understanding. This can lead to ***imprecise visual interpretations****. Also due to the lack of visual grounding, these MLLMs will potentially* ***produce hallucinations****.*

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



Describe the image in detail. Please output interleaved segmentation mask

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



What is the difference between this deer and another deer ?

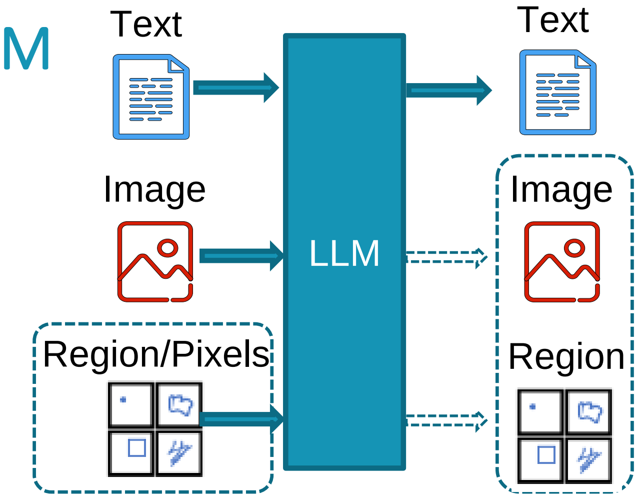
The main difference between the two deer is their size and antlers. The adult deer is larger with more prominent antlers, while the younger deer is smaller with no visible antlers.

Fine-grained Capability of MLLM

Image-oriented Pixel-wise Regional MLLM

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

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



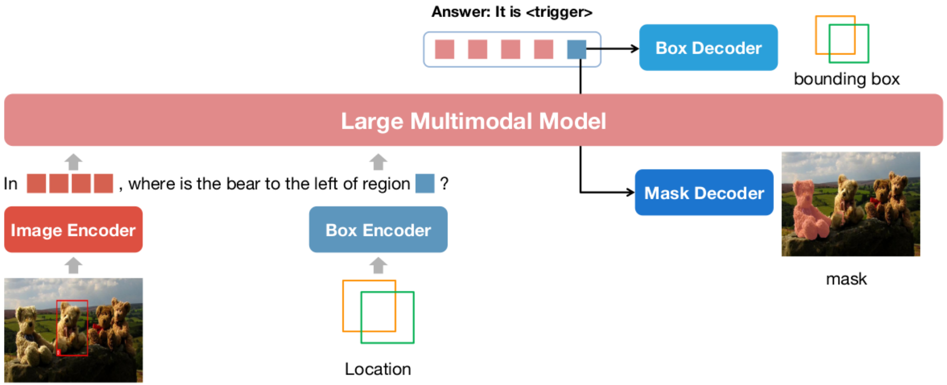
[1] GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. 2023
[2] NExT-Chat: An LMM for Chat, Detection and Segmentation. 2023
[3] MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. 2023
[4] Osprey: Pixel Understanding with Visual Instruction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

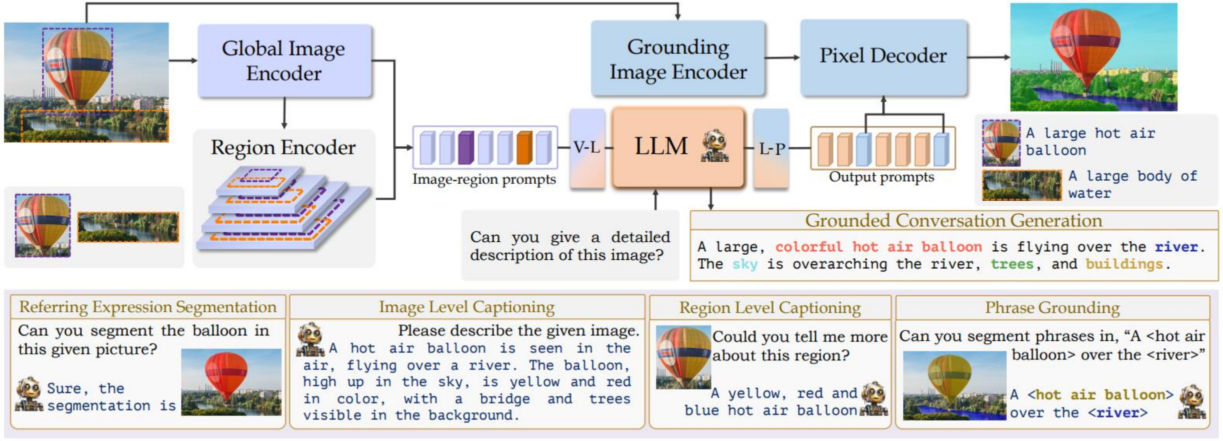
Overview of Modality and Functionality

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

Fine-grained Capability of MLLM

- Image-oriented Pixel-wise
  - NExT-Chat
  - GLaMM





A large hot air balloon

A large body of water

Can you give a detailed description of this image?

A large, colorful hot air balloon is flying over the river. The sky is overarching the river, trees, and buildings.

Can you segment the balloon in this given picture?

Sure, the segmentation is

Please describe the given image.

A hot air balloon is seen in the air, flying over a river. The balloon, high up in the sky, is yellow and red in color, with a bridge and trees visible in the background.

Could you tell me more about this region?

A yellow, red and blue hot air balloon

Can you segment phrases in, "A <hot air balloon> over the <river>"

A <hot air balloon> over the <river>

Fine-grained Capability of MLLM

- Image-oriented Pixel-wise Regional MLLM

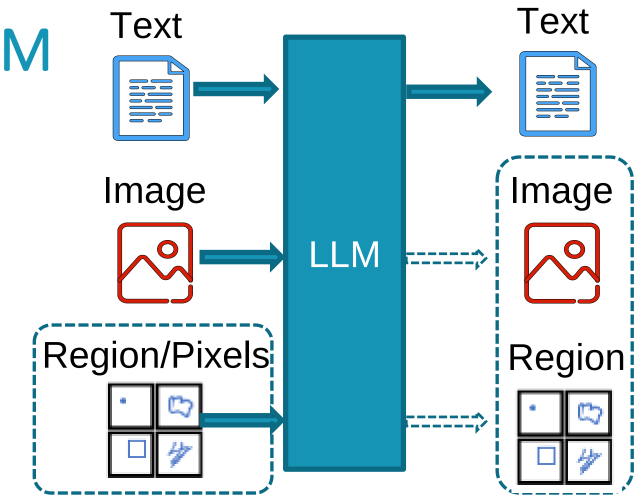
☞ Pixel-level Awareness at Input/Output

- 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

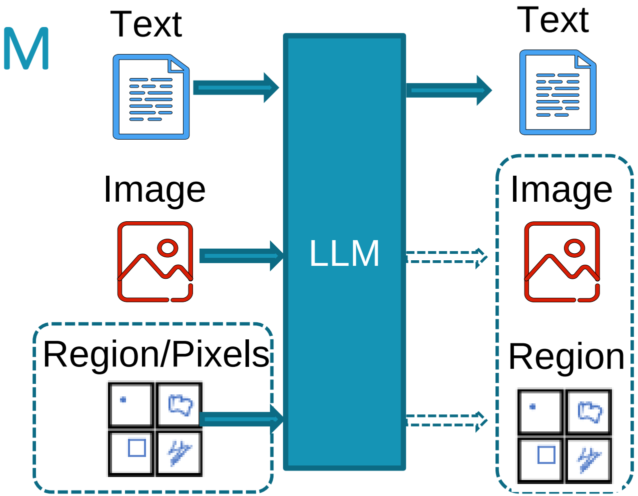
☞ Pixel Granularity

- 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

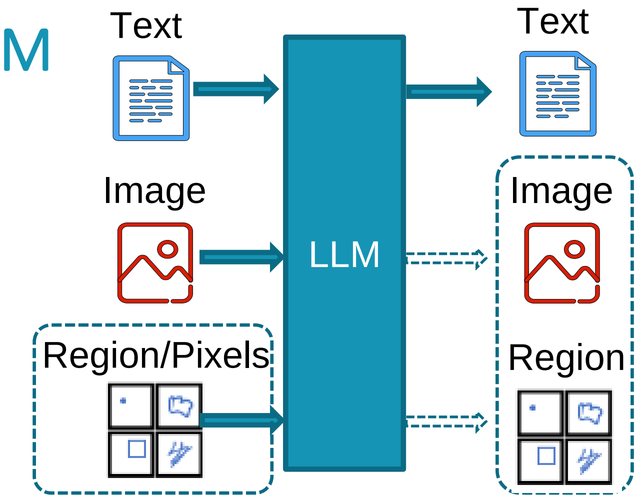
  LISA, PixelLM, DetGPT, MiniGPT-v2, LION

- Bounding-box Coordinates

  GPT4RoI, Shikra, KOSMOS-2, GLaMM

- User Sketches

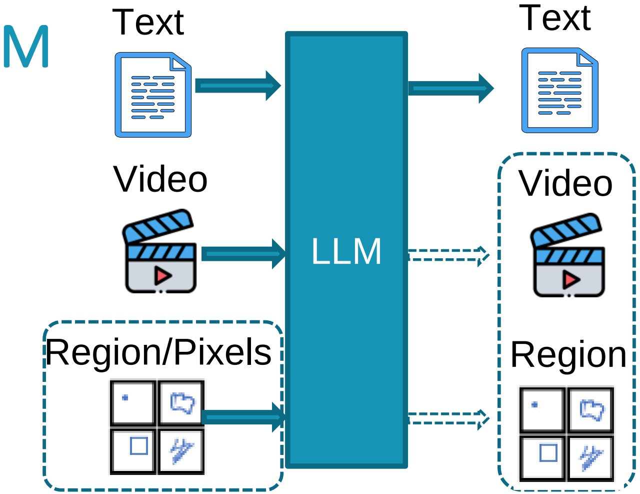
  NExT-Chat, Osprey,



Fine-grained Capability of MLLM

- ## Video-oriented Pixel-wise Regional MLLM

  - PG-Video-LLaVA
  - Merlin
  - MotionEpic
  - …



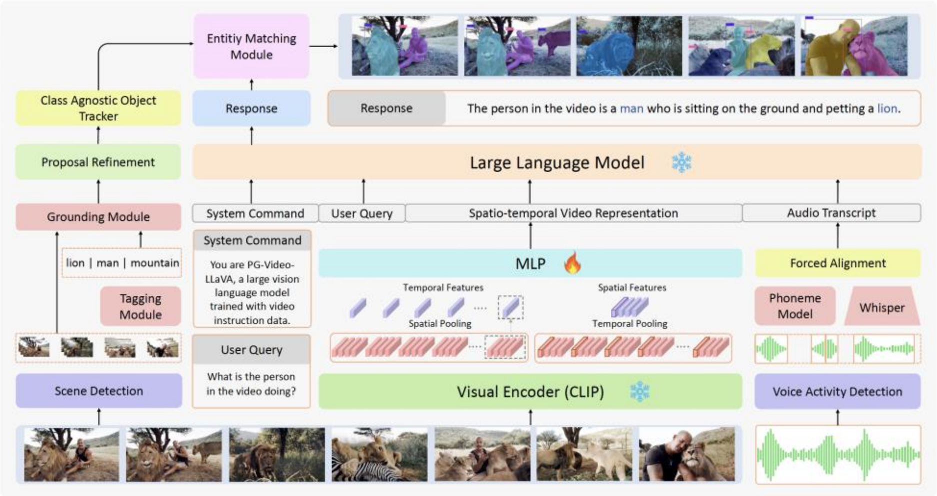
☞ *Users input an video (potentially specifying a region), and the LLM outputs content based on its understanding, grounding or tracking the content to specific pixel-level regions of the video.*

[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

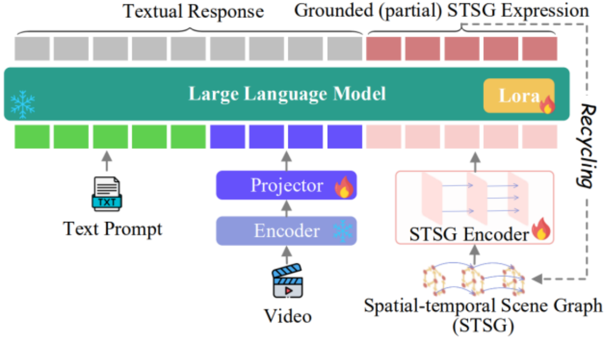


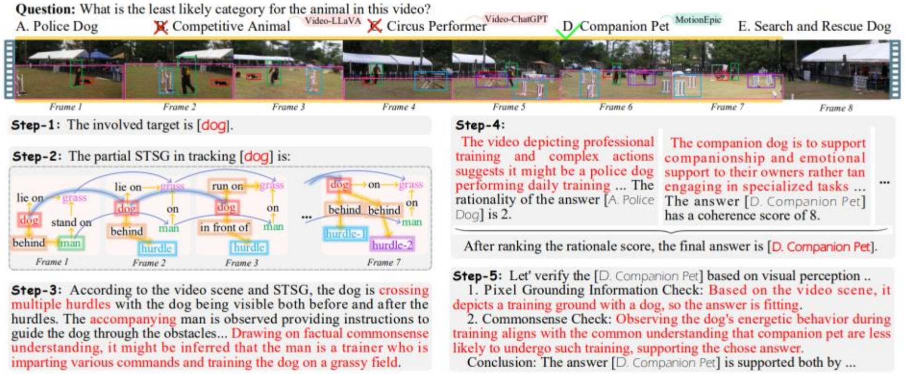
The person in the video is a man who is sitting on the ground and petting a lion.

You are PG-Video-LLaVA, a large vision language model trained with video instruction data.

What is the person in the video doing?

  - ### MotionEpic





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

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

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

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

Step-4:

The video depicting professional training and complex actions suggests it might be a police dog performing daily training ... The rationality of the answer [A. Police Dog] is 2.

The companion dog is to support companionship and emotional support to their owners rather tan engaging in specialized tasks ... The answer [D. Companion Pet] has a coherence score of 8.

...

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

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

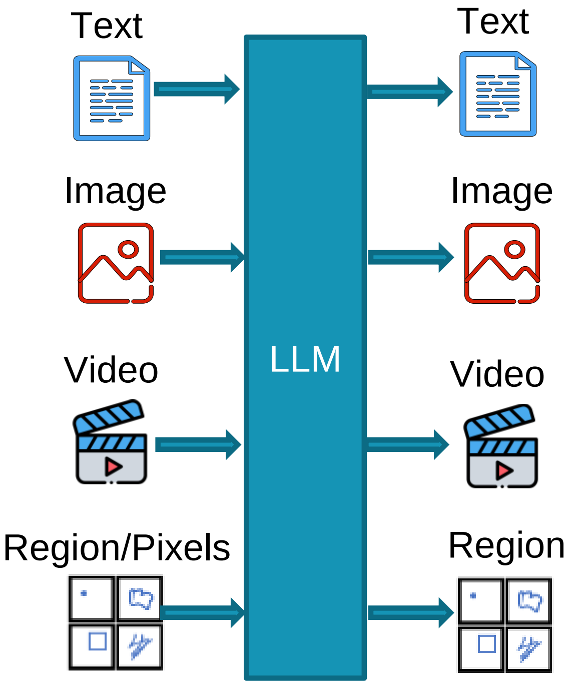
[1] PG-Video-LLaVA: Pixel Grounding in Large Multimodal Video Models. 2023
[2] Video-of-Thought: Step-by-Step Video Reasoning from Perception to Cognition. 2024

Fine-grained Capability of MLLM

- ## Unified Pixel-wise MLLM

  - Vitron

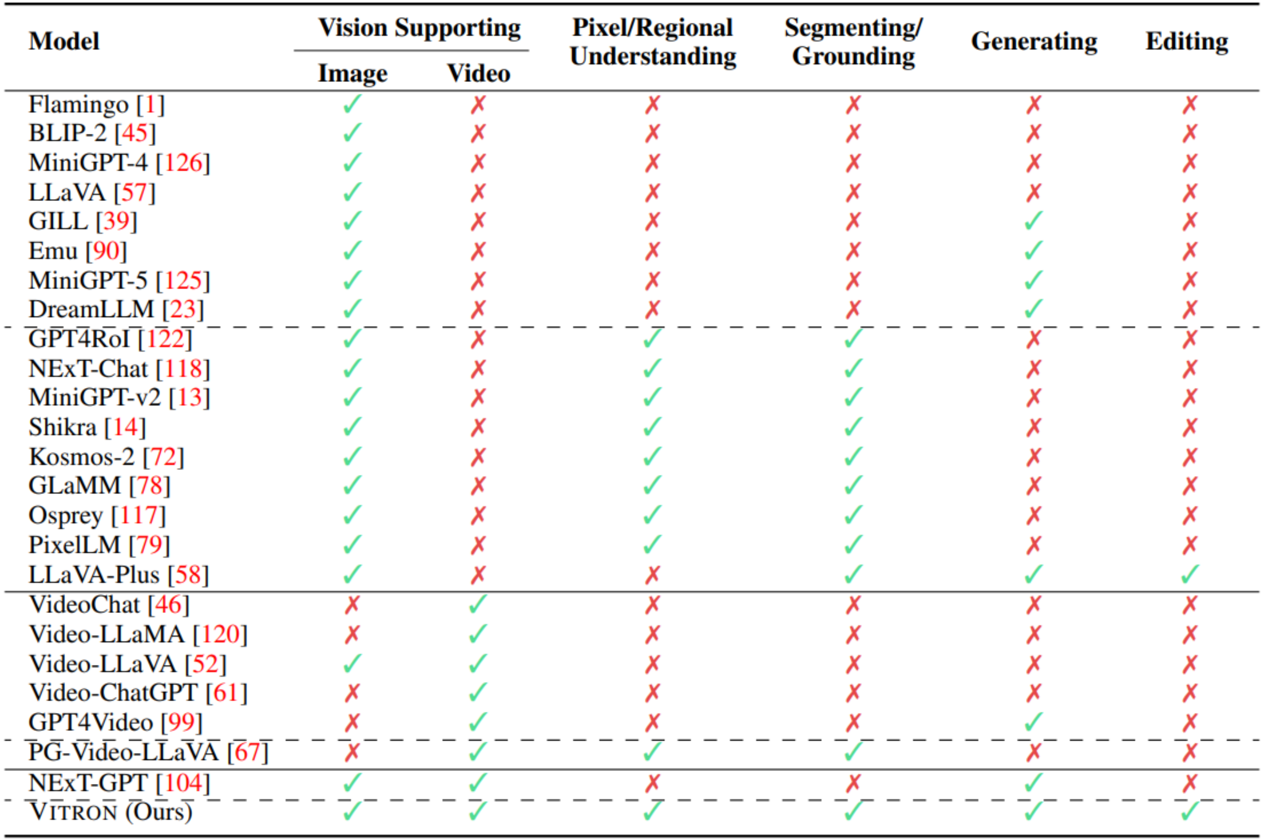
☞ *Users input either an image or video (potentially specifying a region), and the LLM outputs content based on its understanding, generating, grounding or tracking the content to specific pixel-level regions of the image, video.*



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

Fine-grained Capability of MLLM

- Unified
  - Vitron

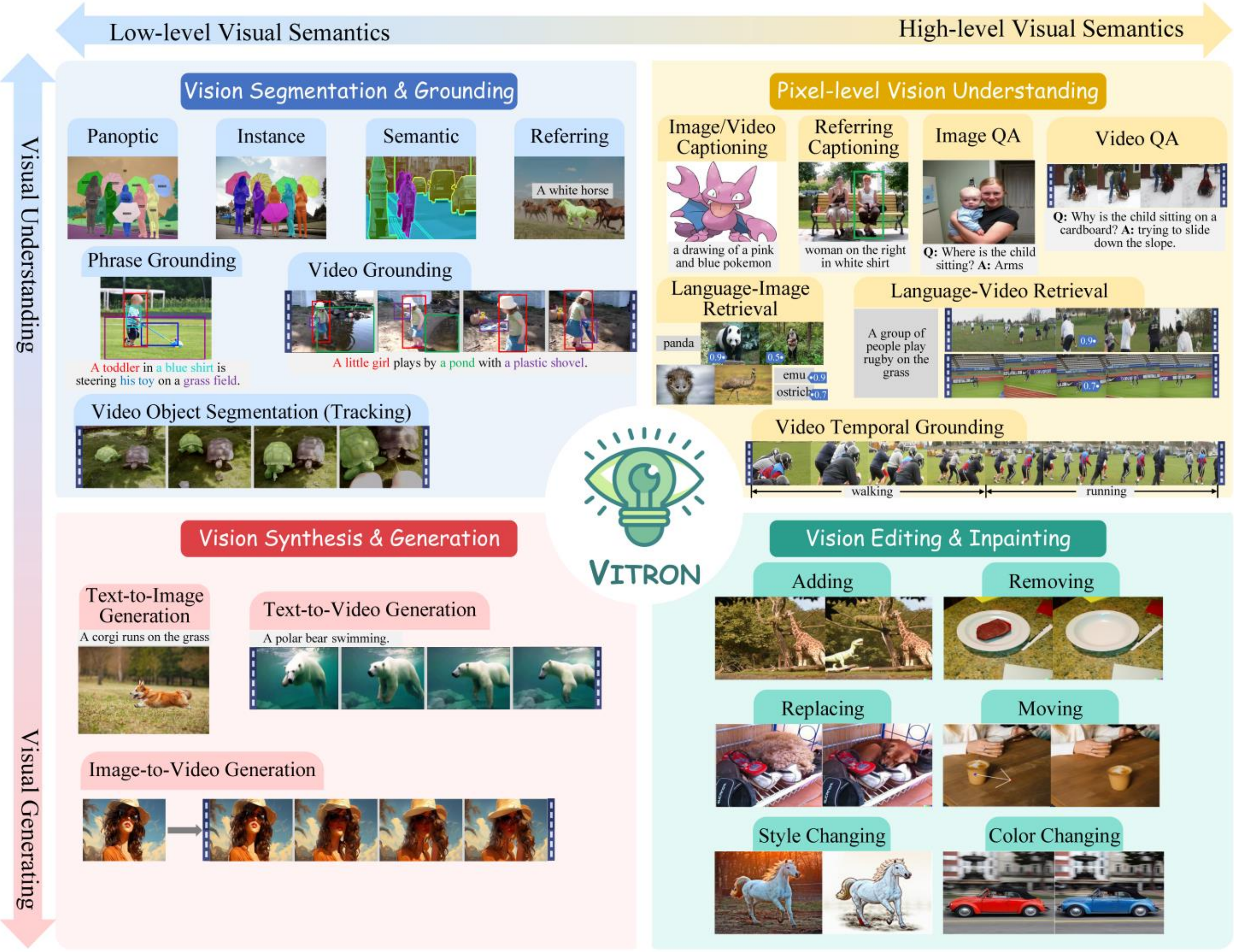


Model	Vision Supporting		Pixel/Regional Understanding	Segmenting/ Grounding	Generating	Editing
	Image	Video				
Flamingo [1]	✓	✗	✗	✗	✗	✗
BLIP-2 [45]	✓	✗	✗	✗	✗	✗
MiniGPT-4 [126]	✓	✗	✗	✗	✗	✗
LLaVA [57]	✓	✗	✗	✗	✗	✗
GILL [39]	✓	✗	✗	✗	✓	✗
Emu [90]	✓	✗	✗	✗	✓	✗
MiniGPT-5 [125]	✓	✗	✗	✗	✓	✗
DreamLLM [23]	✓	✗	✗	✗	✓	✗
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


Low-level Visual Semantics
High-level Visual Semantics
Visual Understanding
Visual Generating
Vision Segmentation & Grounding
Panoptic
Instance
Semantic
Referring
A white horse
Phrase Grounding
A toddler in a blue shirt is steering his toy on a grass field.
Video Grounding
A little girl plays by a pond with a plastic shovel.
Video Object Segmentation (Tracking)
Pixel-level Vision Understanding
Image/Video Captioning
a drawing of a pink and blue pokemon
Referring Captioning
woman on the right in white shirt
Image QA
Q: Where is the child sitting? A: Arms
Video QA
Q: Why is the child sitting on a cardboard? A: trying to slide down the slope.
Language-Image Retrieval
panda
0.9
0.5
emu
0.9
ostrich
0.7
Language-Video Retrieval
A group of people play rugby on the grass
0.9
0.7
Video Temporal Grounding
walking
running
VITRON
Vision Synthesis & Generation
Text-to-Image Generation
A corgi runs on the grass
Text-to-Video Generation
A polar bear swimming.
Image-to-Video Generation
Vision Editing & Inpainting
Adding
Removing
Replacing
Moving
Style Changing
Color Changing

Overview of Modality and Functionality

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

Fine-grained Capability of MLLM

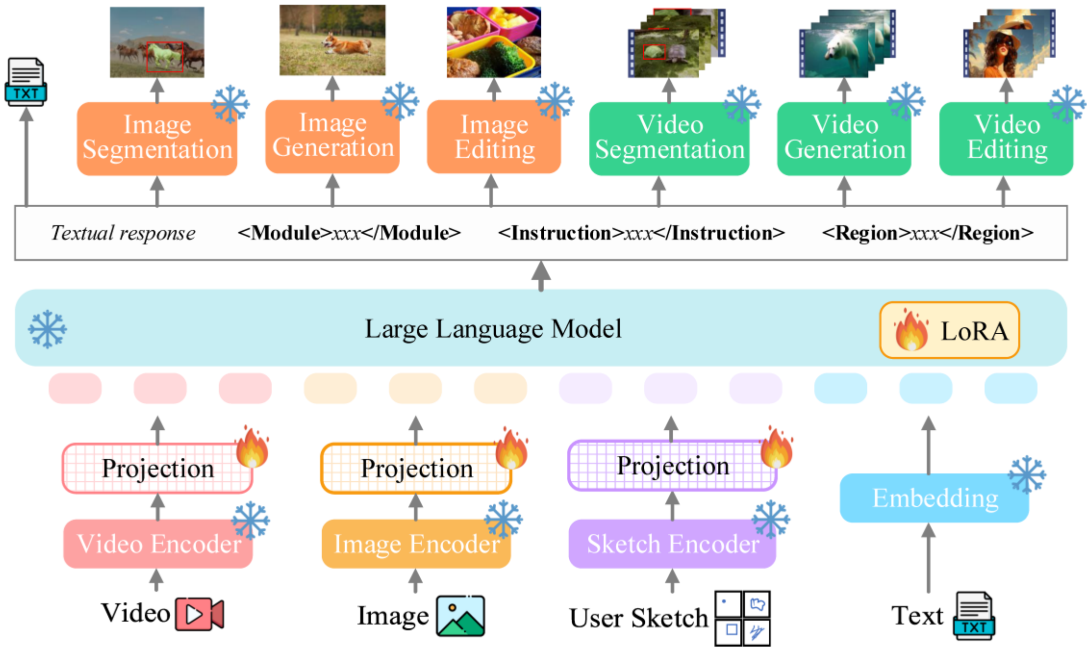
- ## Unified Pixel-wise MLLM

Vitron

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

Paper: https://is.gd/aGu0VV

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



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

□ Image Segmentation

□ Video Segmentation

□ Video Understanding

□ Video Editing

What's Next

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

  - Modality Perspective: Going Broader

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

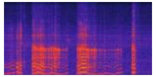
What’s Next from Multimodal LLM to AGI

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

Modality Perspective: Going Broader

- Modalities in current NExT-GPT:
  - language
  - image
  - sound
  - video
- More modalities to go:
  - heat map
  - code
  - time series
  - touch
  - depth&3D

  - infrared/radar
  - document/table
  - spectrogram



  - smell
  - graph

What's Next from Multimodal LLM to AGI

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

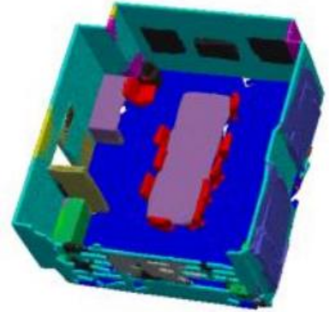
- Task Perspective: Going Deeper

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

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

Referring Segmentation

Panoptic Segmentation



3D Scene Segmentation

What’s Next from Multimodal LLM to AGI

Angle-II: Stronger Generation Ability via Better Tokenization

- Core Idea

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

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

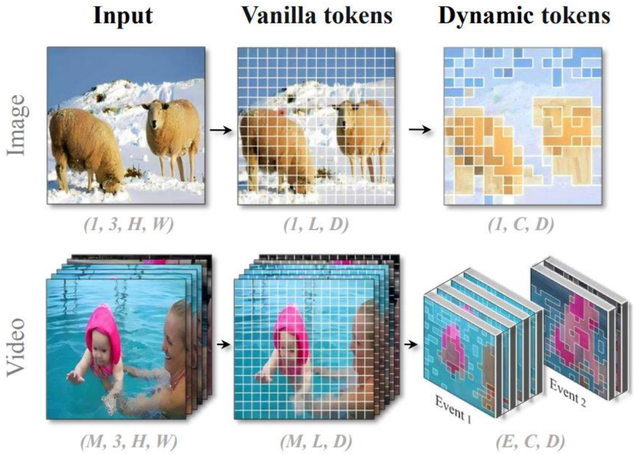
[1] Auto-Encoding Morph-Tokens for Multimodal LLM. 2024

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

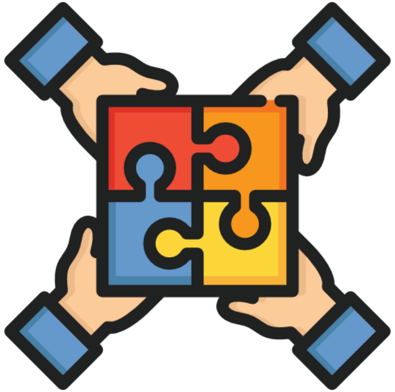
What's Next from Multimodal LLM to AGI

Angle-III: More Multimodality & Multi-Task Synergy

Core Idea

☞ *Achieving a stronger MLLM, and potentially reaching AGI, necessitates enhanced Multimodality & Multi-Task Synergy for the MLLM generalist.*

☞ *Master abductive reasoning to facilitate analogical thinking, allowing different modalities and tasks, as well as the comprehension and generation processes, to mutually assist each other and create synergistic effects.*



[1] Abductive reasoning: Logic, visual thinking, and coherence. 1997.
[2] Reasoning. https://www.butte.edu/departments/cas/tipsheets/thinking/reasoning.html

Thanks!

Any questions?