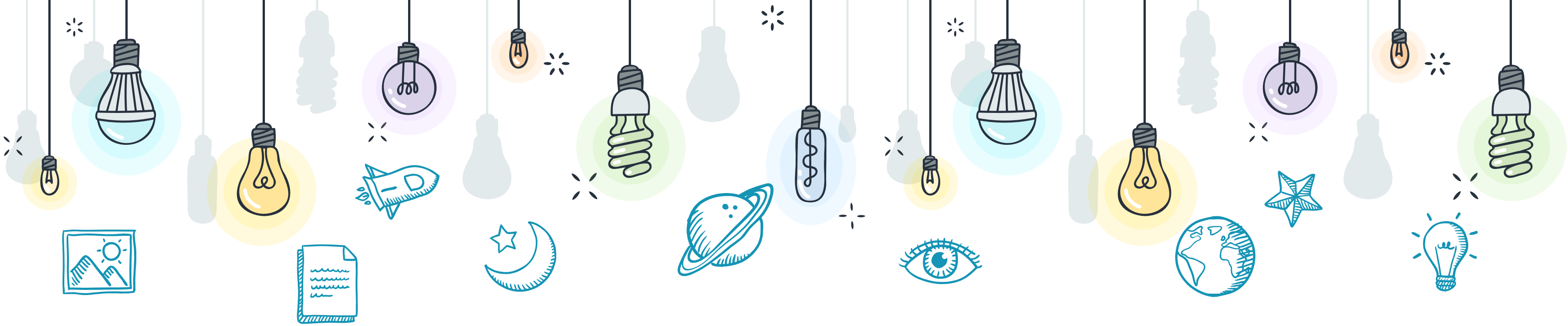


A Brief Intro to Multimodal LLM

Stay tune to MLLM tutorial series:

https://mllm2024.github.io/COLING2024

Hao Fei

Research Fellow

National University of Singapore

http://haofei.vip/

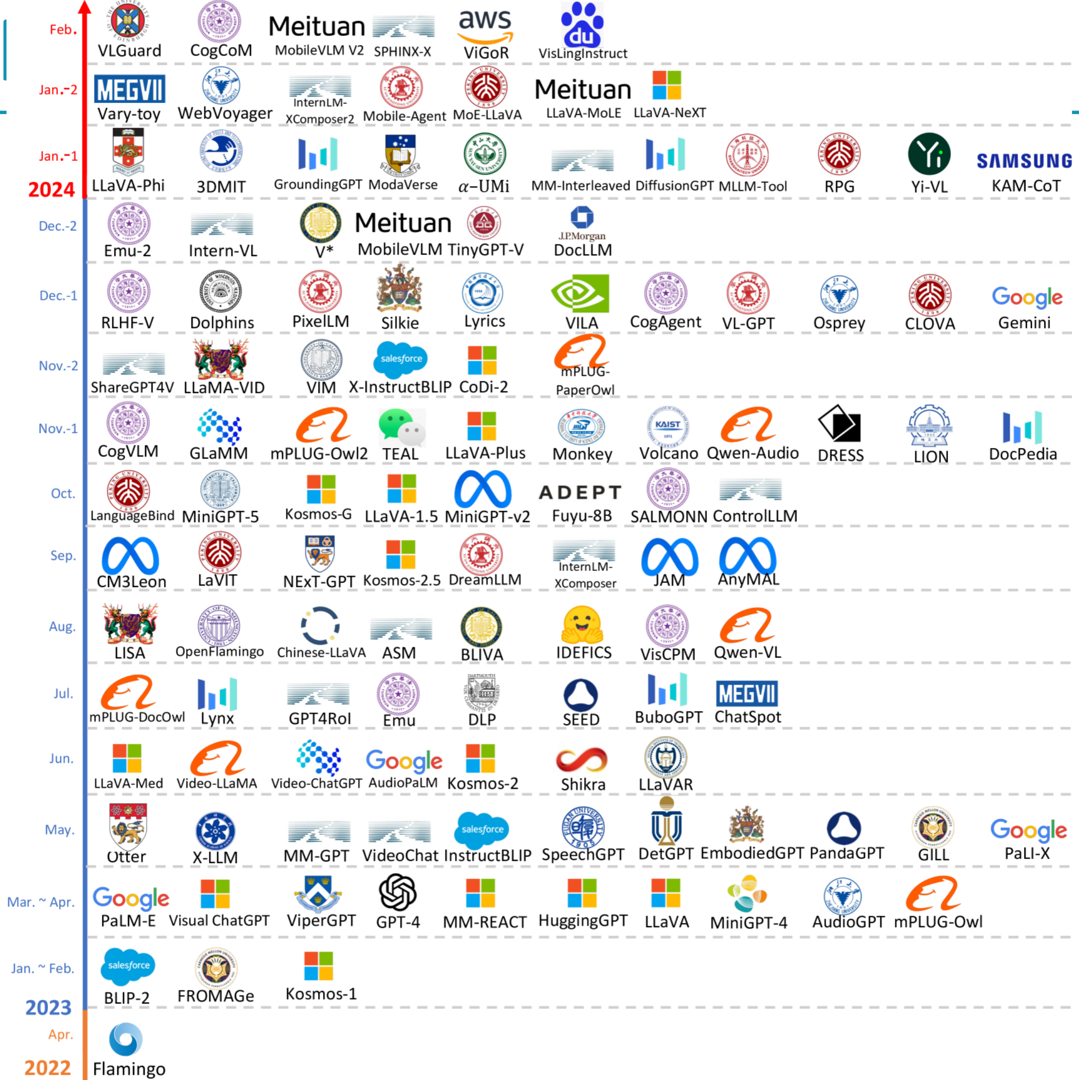
Table of Content

1 Modality
- Overview
- Multimodal Perceiving
- Multimodal Perceiving + Generation
- Unified MLLM
- Fine-grained MLLM

2 Architecture
- Overview
- Multimodal Encoding
- Tokenization
- Input-side Projection
- Backbone LLMs
- Decoding-side Connection
- Multimodal Generation

Intelligence in

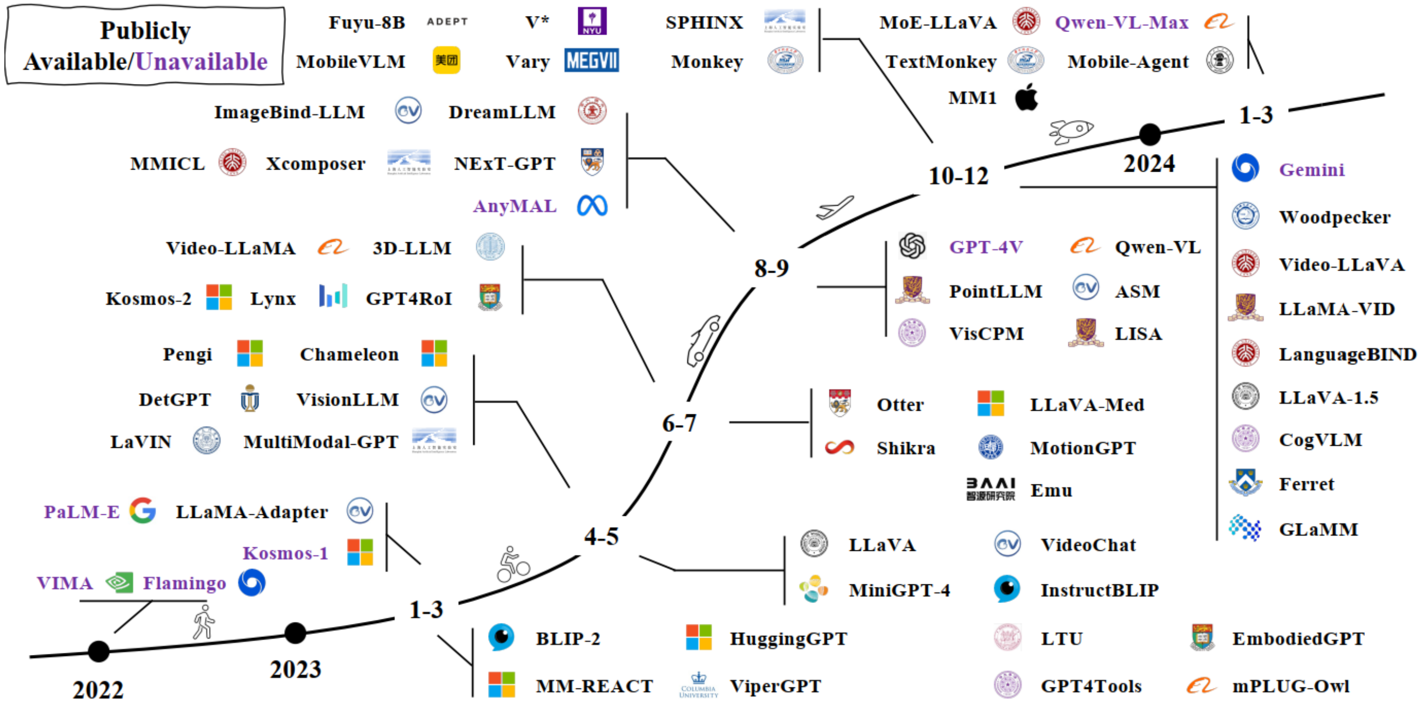
- Trends of MLLMs



[1] MM-LLMs: Recent Advances in MultiModal Large Language Models, 2023.

Intelligence in Multi-Sensory Data

- ## Trends of MLLMs



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

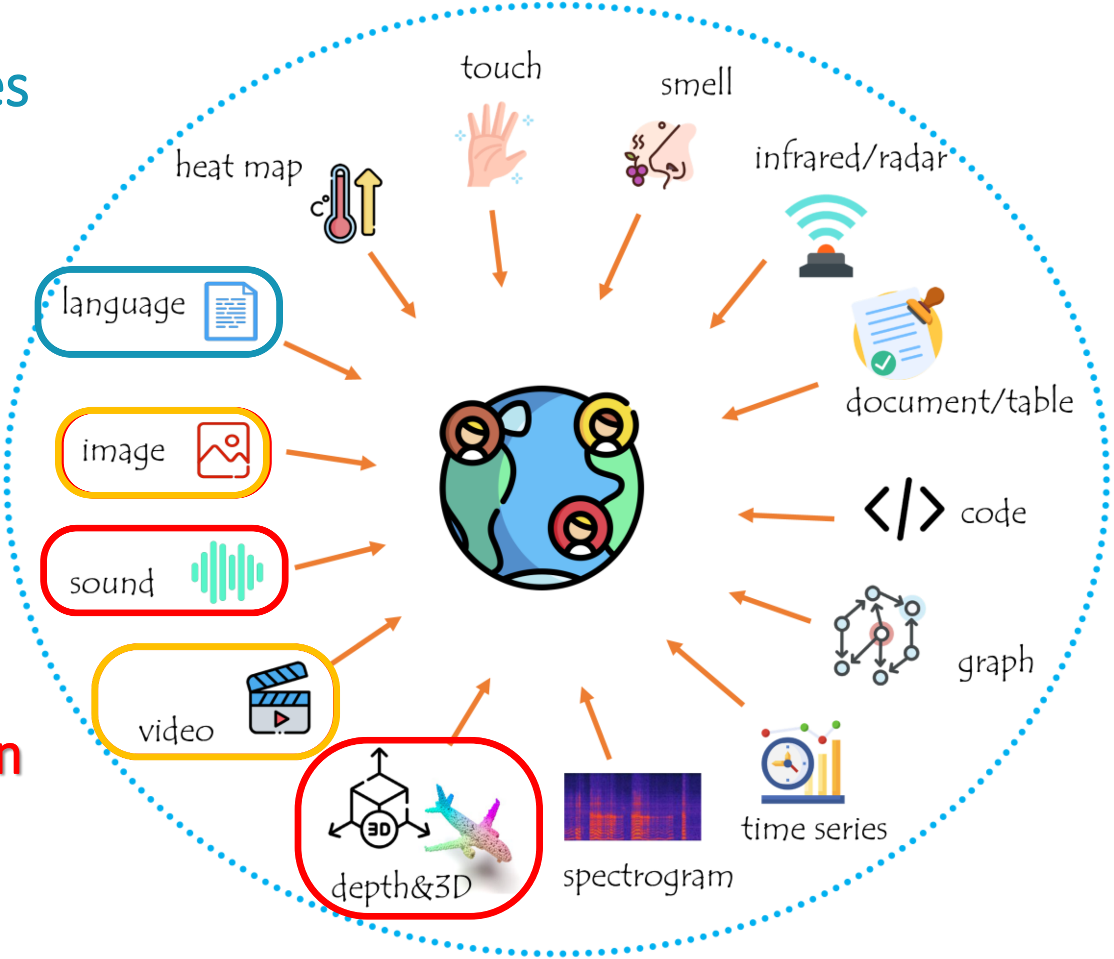
1

Modality and Functionality

What are MLLMs capable of?

Overview of Modality and Functionality

- **Modalities**



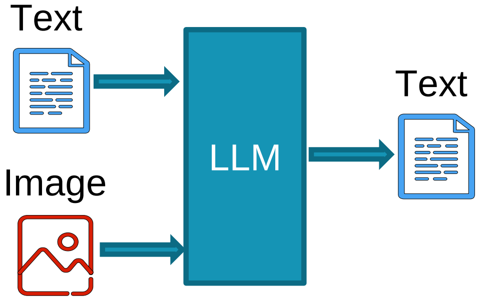
Language + Vision

Overview of Modality and Functionality

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

Multimodal Perceiving

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



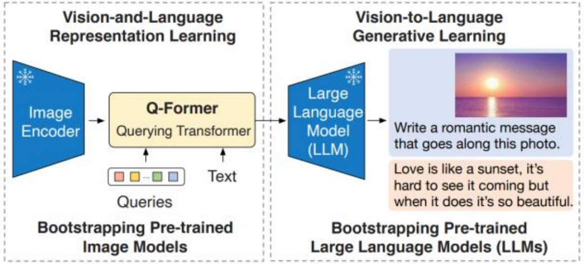
☞ *Encode input images with external image encoders, generating LLM-understandable visual feature, which is then fed into the LLM. LLM then interprets the input images based on the input text instructions and produces a textual response.*

[1] Flamingo: a Visual Language Model for Few-Shot Learning. 2022
[2] Language Is Not All You Need: Aligning Perception with Language Models. 2023
[3] BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. 2023
[4] MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. 2024
…

Multimodal Perceiving

Image-perceiving MLLM

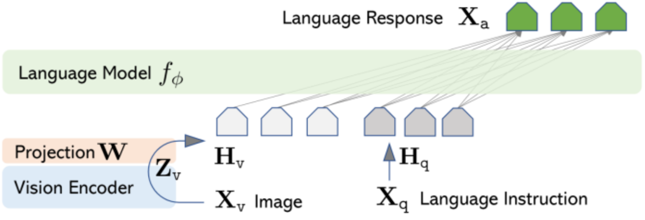
- Blip2


Vision-and-Language Representation Learning
Vision-to-Language Generative Learning
Image Encoder
Q-Former
Querying Transformer
Queries
Text
Large Language Model (LLM)
Bootstrapping Pre-trained Image Models
Bootstrapping Pre-trained Large Language Models (LLMs)

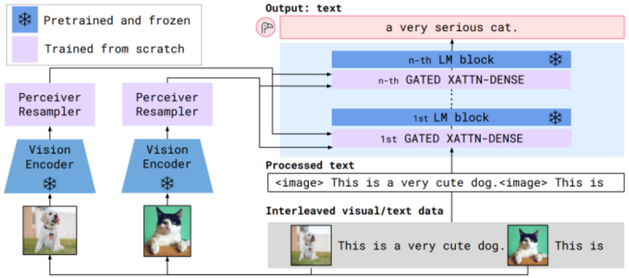
Write a romantic message that goes along this photo.

Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.

- LLaVA



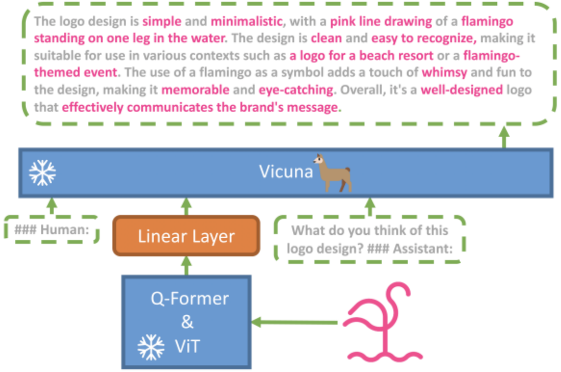
- Flamingo



`<image> This is a very cute dog.<image> This is`

This is a very cute dog. This is

- 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.

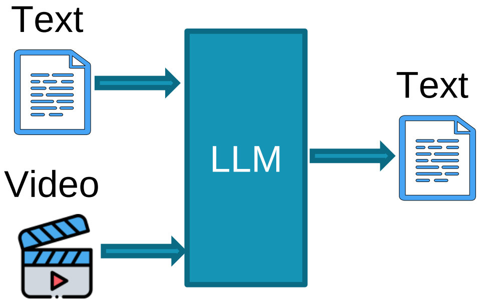
What do you think of this logo design? ### Assistant:

[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
- ...



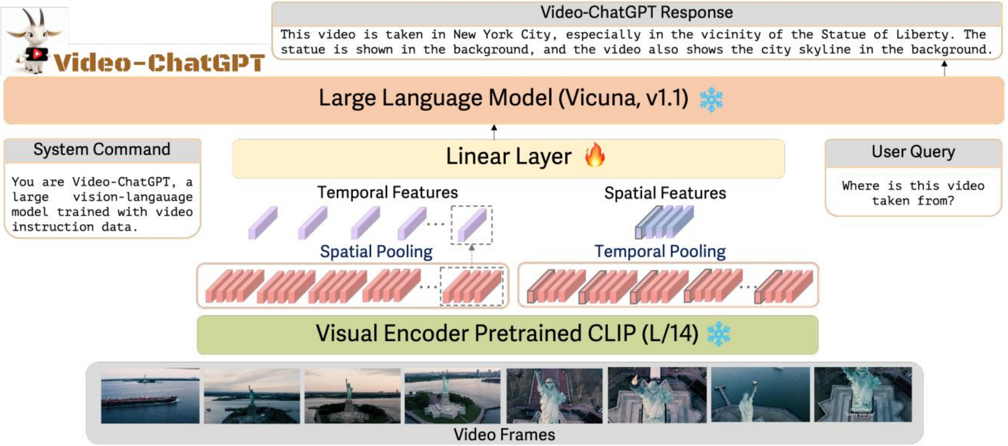
☞ *Encode input videos with external video encoders, generating LLM-understandable visual feature, feeding into LLM, which then interprets the input videos based on the input text instructions and produces a textual response.*

[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

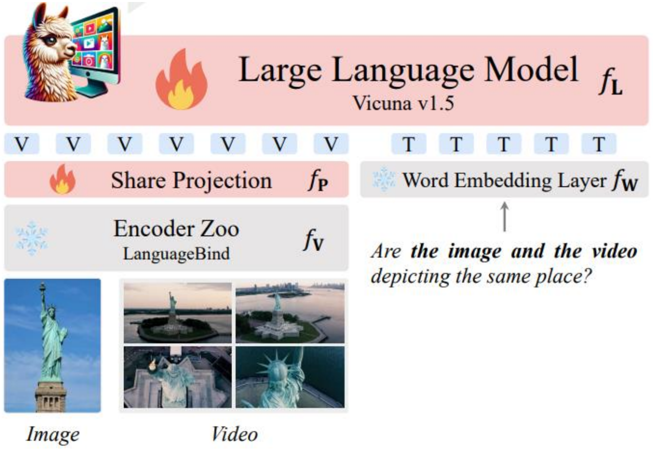


This video is taken in New York City, especially in the vicinity of the Statue of Liberty. The statue is shown in the background, and the video also shows the city skyline in the background.

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

Where is this video taken from?

Video-LLaVA



Are the image and the video depicting the same place?

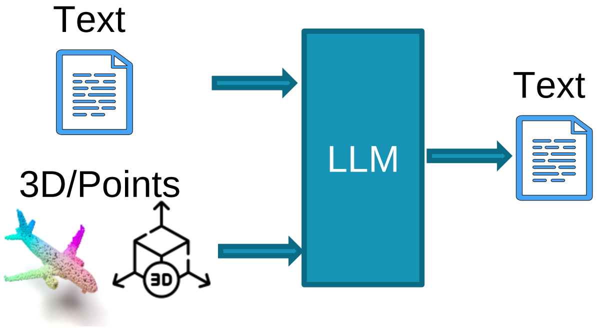
[1] Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. 2023

[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.*

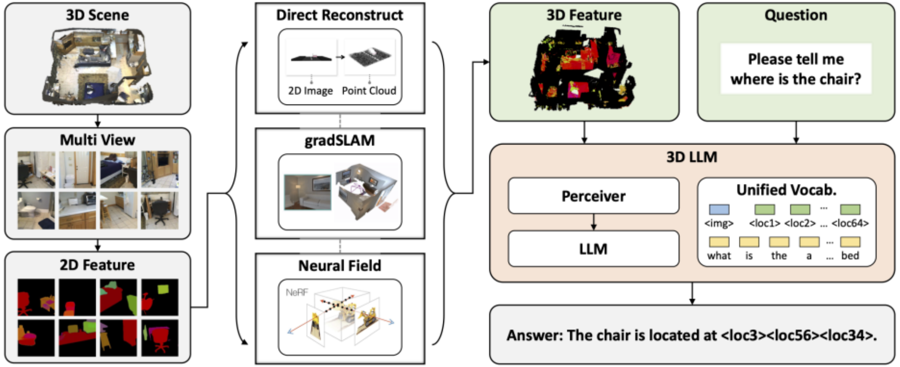
[1] 3D-LLM: Injecting the 3D World into Large Language Models. 2023
[2] 3D-GPT: Procedural 3D Modeling with Large Language Models. 2023
[3] LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning. 2023
[4] PointLLM: Empowering Large Language Models to Understand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

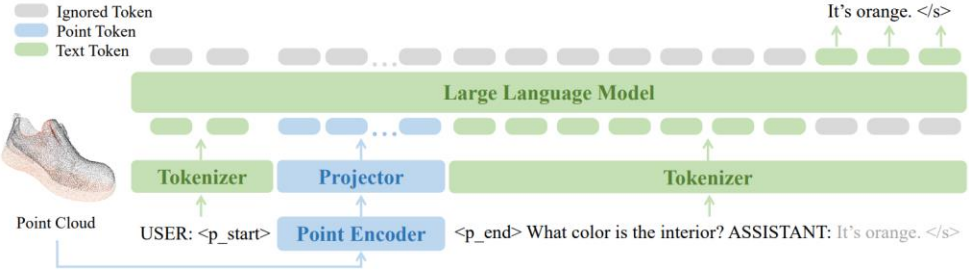
Multimodal Perceiving

- ## 3D-perceiving MLLM

  - 3D-LLM

  - PointLLM



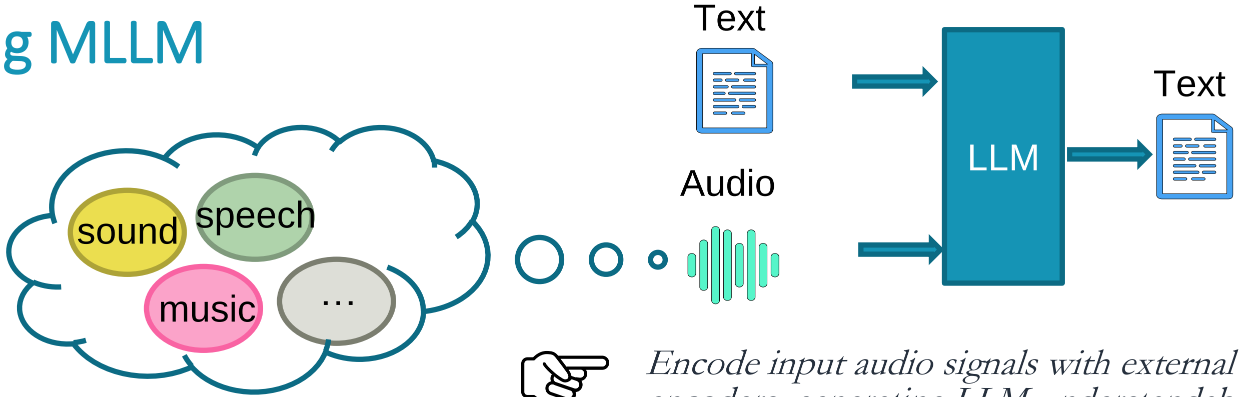


[1] *3D-LLM: Injecting the 3D World into Large Language Models. 2023*

[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.*

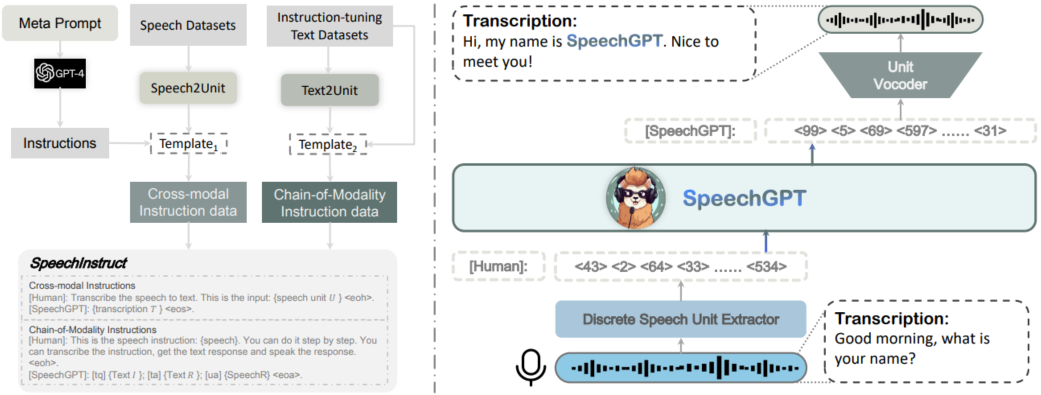
[1] AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. 2023
[2] SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. 2023
[3] VioLA: Unified Codec Language Models for Speech Recognition, Synthesis, and Translation. 2023
[4] AudioPaLM: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

SALMONN



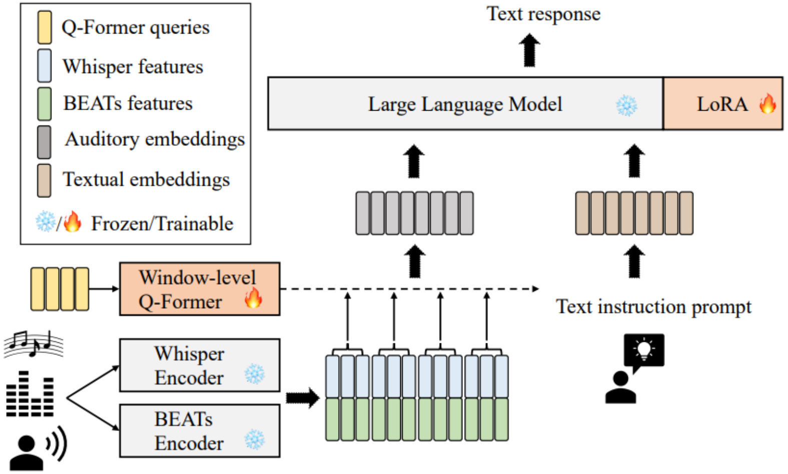
SpeechInstruct

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!

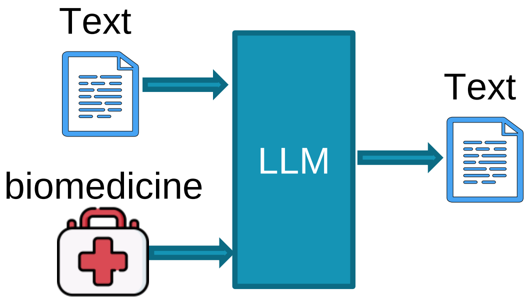
Transcription: Good morning, what is your name?



[1] SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. 2023
[2] SALMONN: Towards Generic Hearing Abilities for Large Language Models. 2023
[3] Sparks of Large Audio Models: A Survey and Outlook. https://github.com/EmulationAI/awesome-large-audio-models, 2023

Multimodal Perceiving

- X-perceiving MLLM
  - Bio-/Medical & Healthcare
    - BioGPT
    - DrugGPT
    - BioMedLM
    - OphGLM
    - GatorTron
    - GatorTronGPT
    - MEDITRON
    - DoctorGLM
    - BianQue
    - ClinicalGPT
    - Qilin-Med
    - ChatDoctor
    - BenTsao
    - HuatuoGPT
    - MedAlpaca
    - AlpaCare
    - Zhongjing
    - PMC-LLaMA
    - CPLLM
    - MedPaLM 2
    - BioMedGPT



[1] *BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. 2022*

[2] *DrugGPT: A GPT-based Strategy for Designing Potential Ligands Targeting Specific Proteins. 2023*

[3] *MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. 2023*

[4] *HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge. 2023*

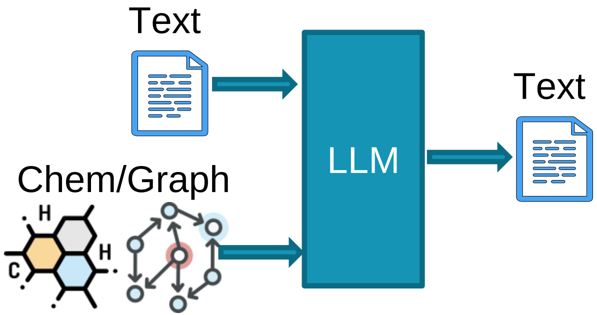
[5] *AlpaCare:Instruction-tuned Large Language Models for Medical Application. 2023*

[6] *A Survey of Large Language Models in Medicine: Progress, Application, and Challenge, https://github.com/AI-in-Health/MedLLMsPracticalGuide. 2023.*

…

Multimodal Perceiving

- X-perceiving MLLM
  - Molecule & Chemistry
    - ChemGPT
    - SPT
    - T5 Chem
    - ChemLLM
    - MolCA
    - MolXPT
    - MolSTM
    - GIMLET
    - …
  - Graph
    - StructGPT
    - GPT4Graph
    - GraphGPT
    - LLaGA
    - HiGPT
    - …
  - Geographical Information System (GIS)
    - GeoGPT



[1] Neural Scaling of Deep Chemical Models. 2022
[2] ChemLLM: A Chemical Large Language Model. 2023
[3] MolCA: Molecular Graph-Language Modeling with Cross-Modal Projector and Uni-Modal Adapter. 2023
[4] StructGPT: A General Framework for Large Language Model to Reason on Structured Data. 2023
[5] LLaGA: Large Language and Graph Assistant. 2023
[6] Awesome-Graph-LLM, https://github.com/XiaoxinHe/Awesome-Graph-LLM. 2023

Unified MLLM: Perceiving + Generation

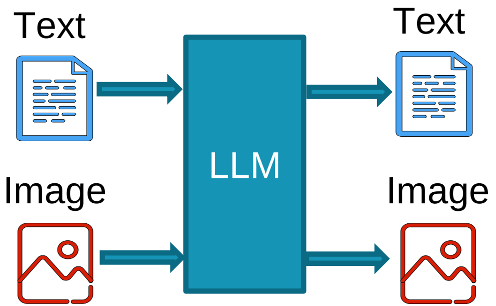
- ## Scenarios

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

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

Unified MLLM: Perceiving + Generation

- Image
  - GILL
  - EMU
  - MiniGPT-5
  - DreamLLM
  - LLaVA-Plus
  - LaVIT
  - …



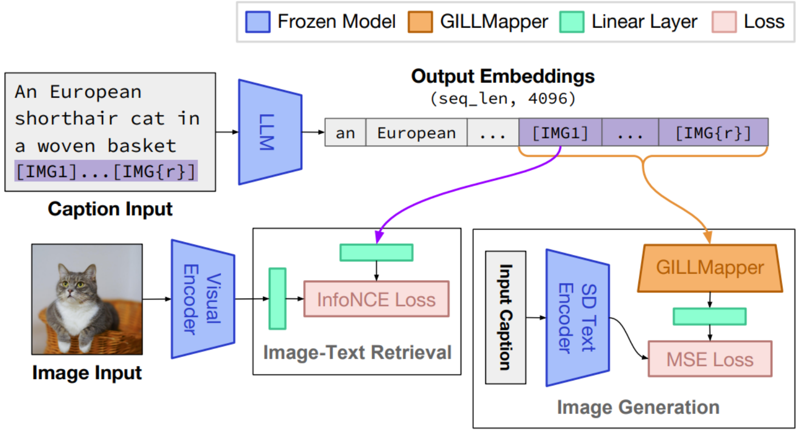
☞ *Central LLMs take as input both texts and images, after semantics comprehension, and generate both texts and images.*

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

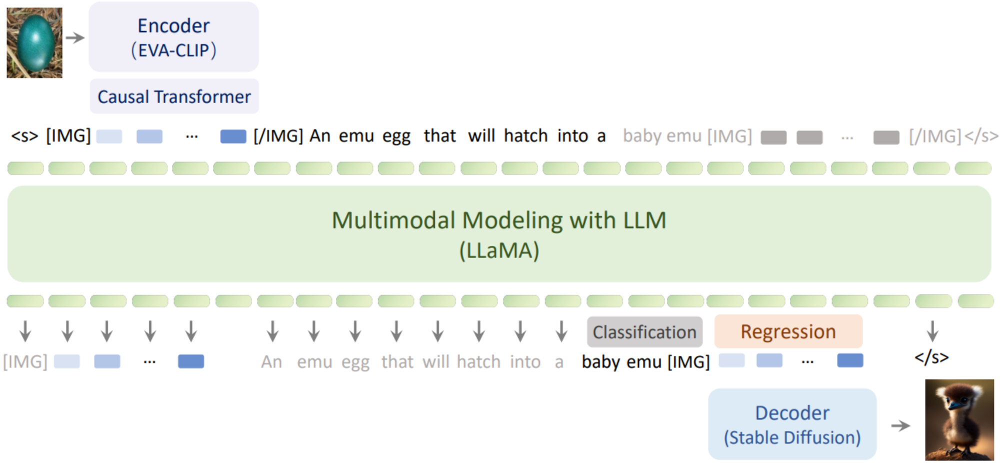
Unified MLLM: Perceiving + Generation

- ## Image

  - ### GILL



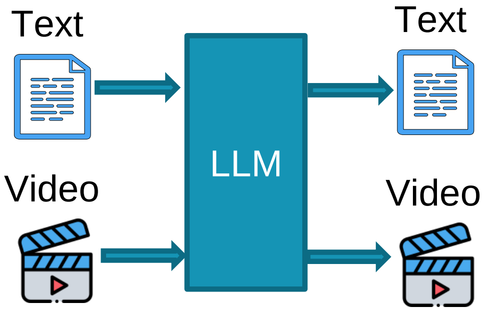
  - ### EMU


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)
[IMG] ... An emu egg that will hatch into a baby emu [IMG] ... </s>
Classification
Regression
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
  - …



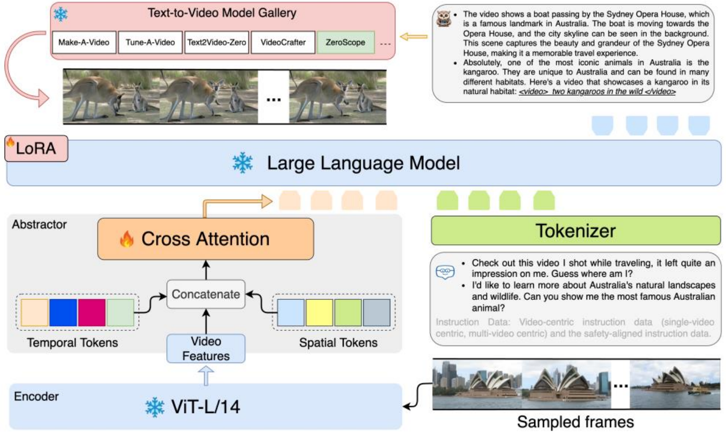
☞ *Central LLMs take as input both texts and videos, after semantics comprehension, and generate both texts and videos.*

[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

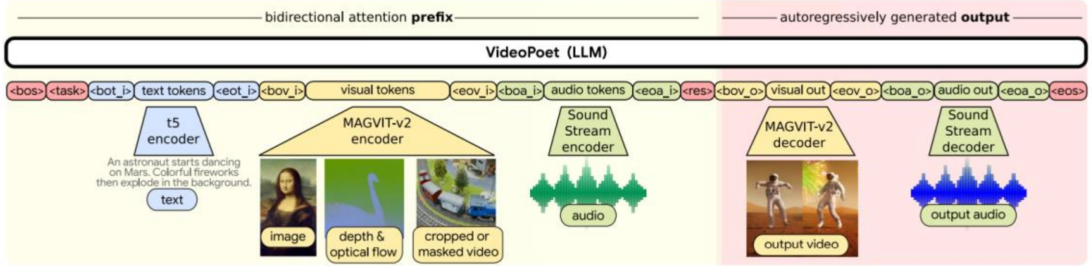


- The video shows a boat passing by the Sydney Opera House, which is a famous landmark in Australia. The boat is moving towards the Opera House, and the city skyline can be seen in the background. This scene captures the beauty and grandeur of the Sydney Opera House, making it a memorable travel experience.
- 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.

  - VideoPoet

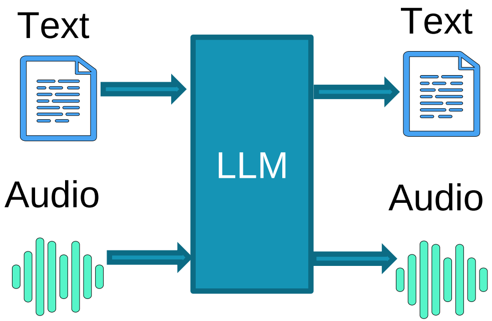


An astronaut starts dancing on Mars. Colorful fireworks then explode in the background.

[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,
  - …



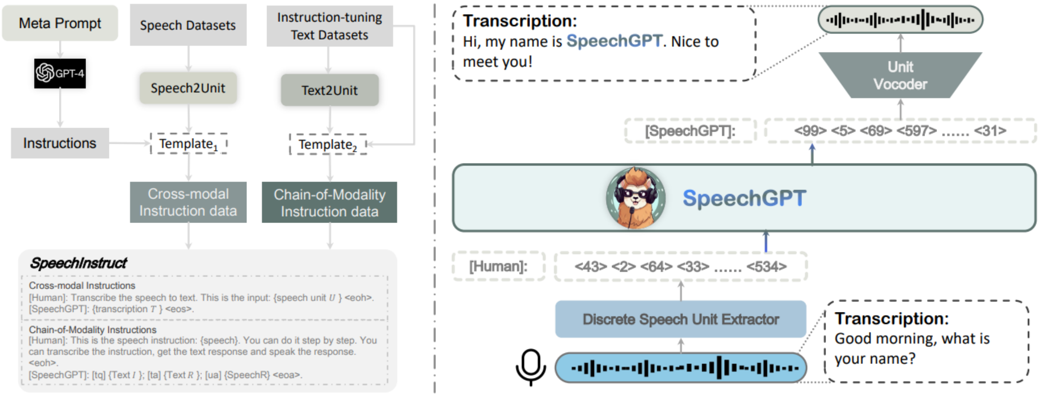
☞ *Central LLMs take as input both texts and audio, after semantics comprehension, and generate both texts and audio.*

[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



SpeechInstruct

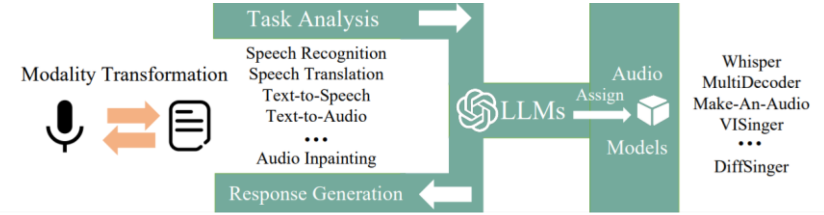
Cross-modal Instructions
[Human]: Transcribe the speech to text. This is the input: {speech unit U } <eoh>.
[SpeechGPT]: {transcription T } <eos>.

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?

AudioGPT



[1] SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. 2023
[2] AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. 2023

Unified MLLM: Harnessing Multi-Modalities

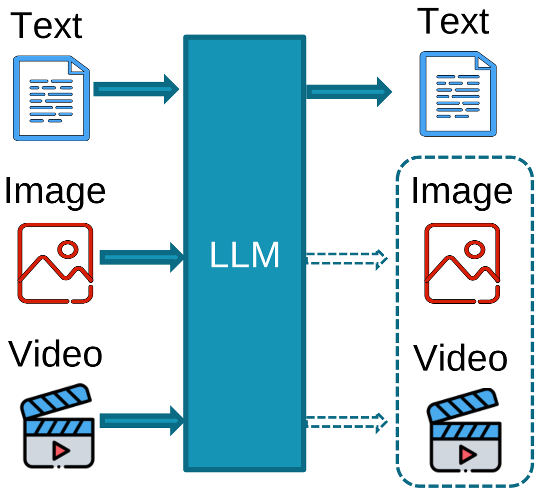
- Scenarios:

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

- Image+Video
- Audio+Video
- Image+Video+Audio
- Any-to-Any
- …

Unified MLLM: Harnessing Multi-Modalities

- Text+Image+Video
  - Video-LLaVA
  - Chat-UniVi
  - LLaMA-VID
  - …

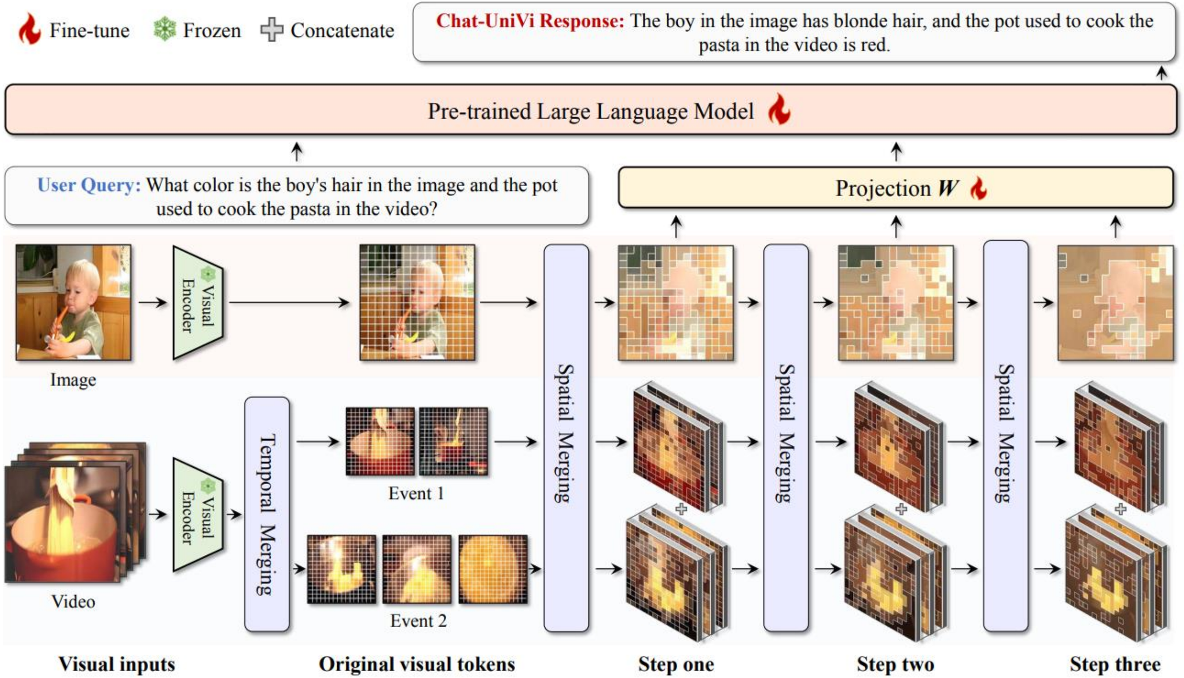


☞ *Central LLMs take as input texts, image and video, after semantics comprehension, and generate texts (maybe also image and video, or combination).*

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

Unified MLLM: Harnessing Multi-Modalities

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



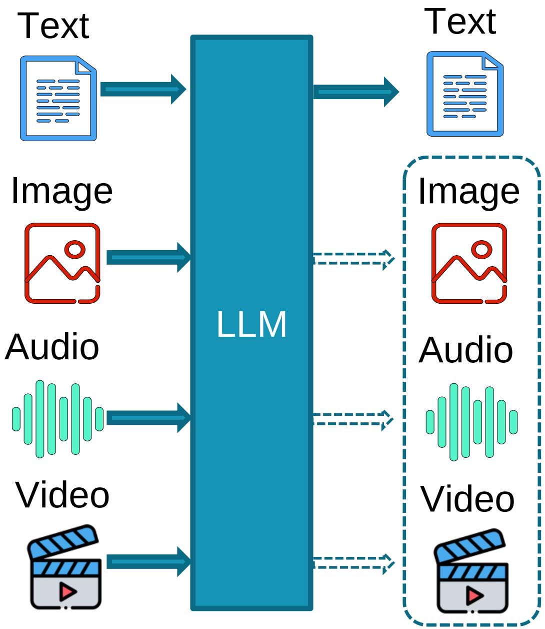
Chat-UniVi Response: The boy in the image has blonde hair, and the pot used to cook the pasta in the video is red.

User Query: What color is the boy's hair in the image and the pot used to cook the pasta in the video?

[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
  - …

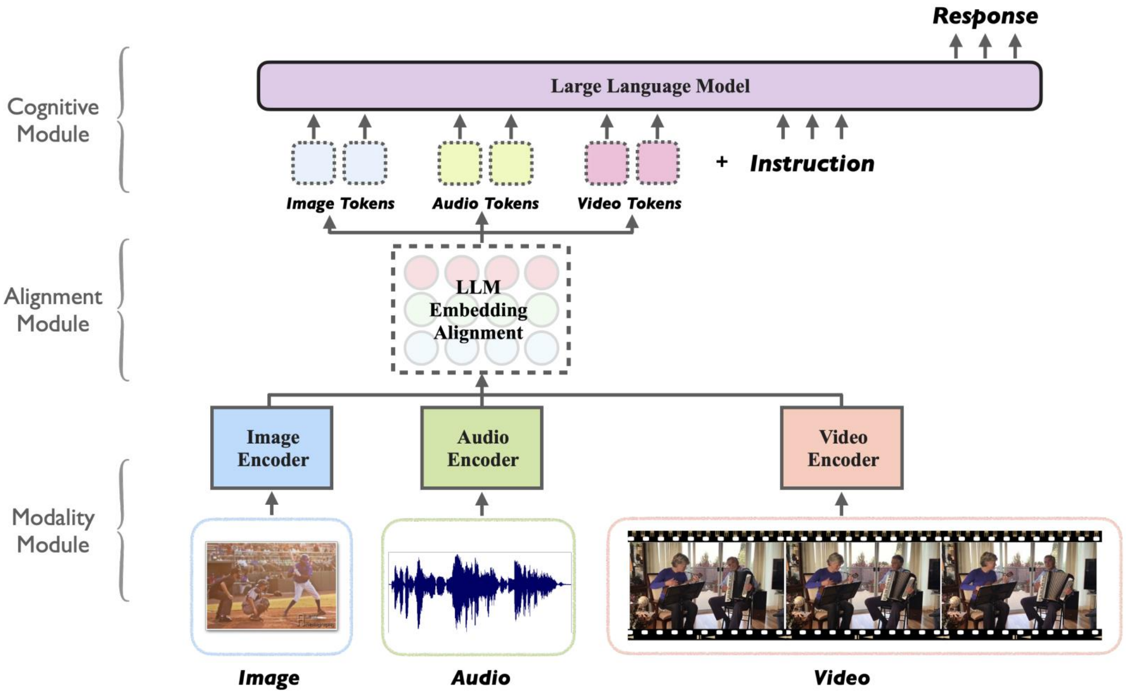


☞ *Central LLMs take as input texts, audio, image and video, and generate texts (maybe also audio, image and video, or combination).*

[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

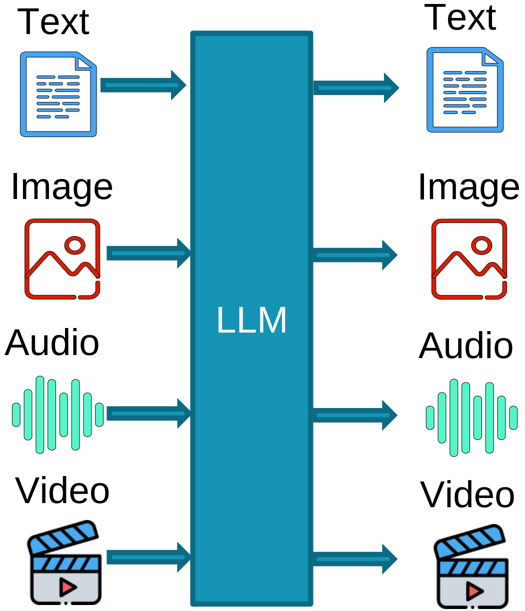
- Text+Image+Video+Audio
  - Macaw-LLM



[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
  - …



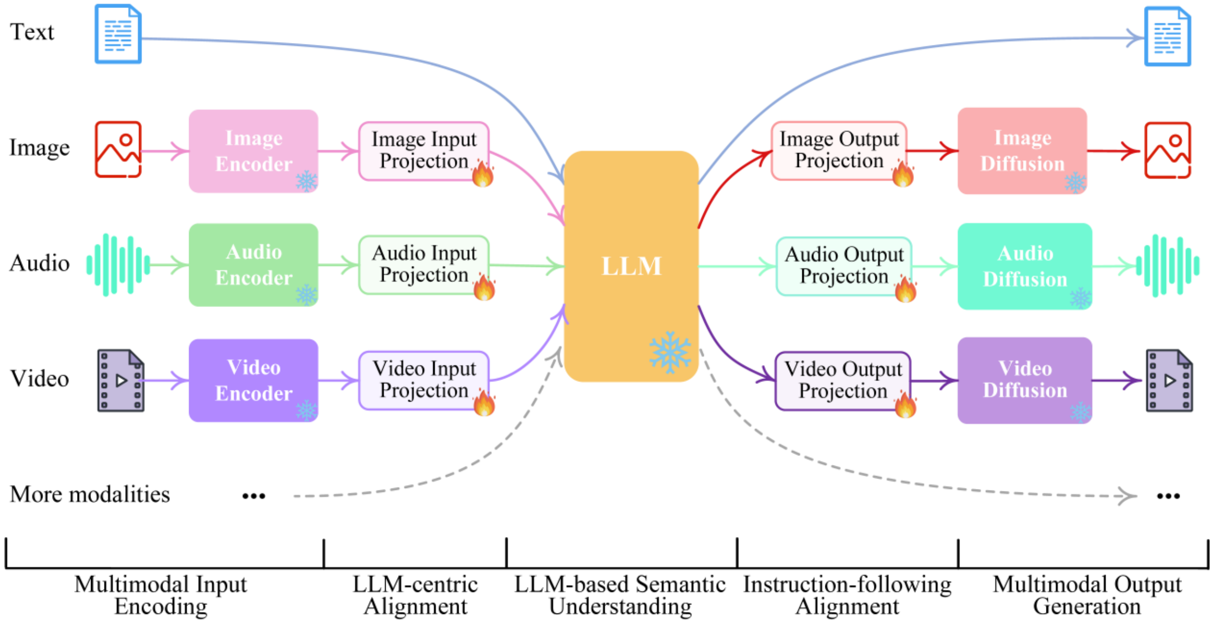
☞ *Central LLMs take as input texts, audio, image and video, and freely generate texts, audio, image and video, or combination.*

[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

Project



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

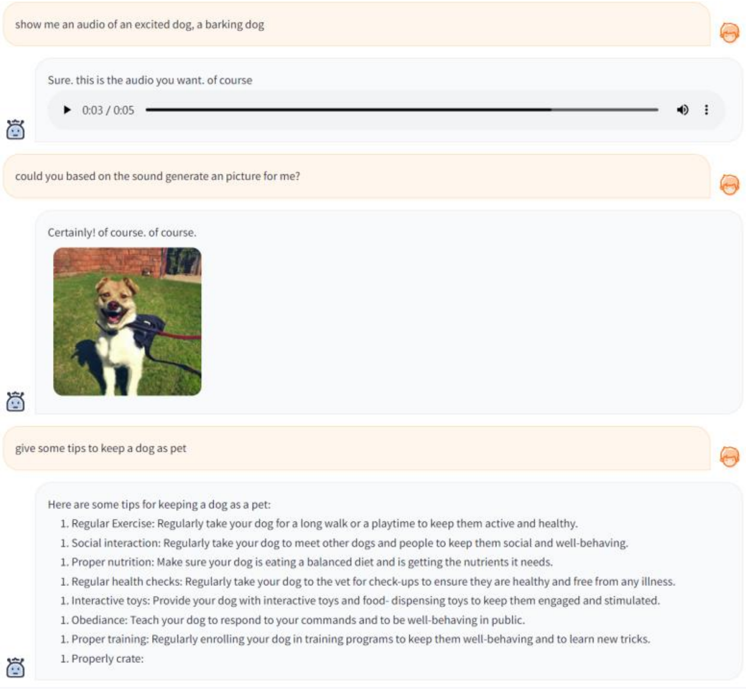
Paper: https://arxiv.org/pdf/2309.05519

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

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

Unified MLLM: Harnessing Multi-Modalities

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



show me an audio of an excited dog, a barking dog

Sure. this is the audio you want. of course

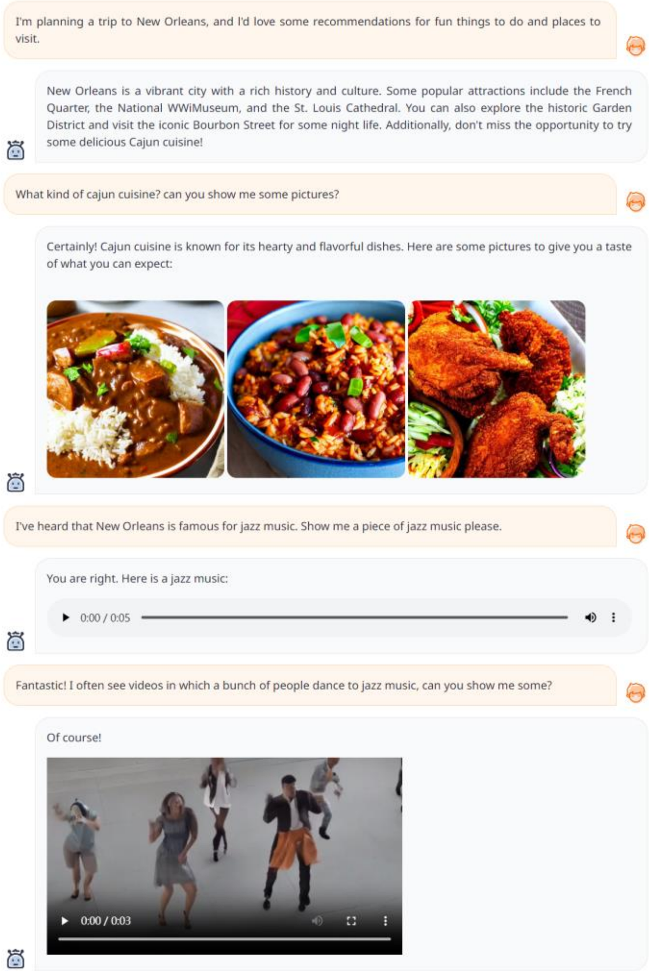
could you based on the sound generate an picture for me?

Certainly! of course. of course.

give some tips to keep a dog as pet

Here are some tips for keeping a dog as a pet:

1. Regular Exercise: Regularly take your dog for a long walk or a playtime to keep them active and healthy.
1. Social interaction: Regularly take your dog to meet other dogs and people to keep them social and well-behaving.
1. Proper nutrition: Make sure your dog is eating a balanced diet and is getting the nutrients it needs.
1. Regular health checks: Regularly take your dog to the vet for check-ups to ensure they are healthy and free from any illness.
1. Interactive toys: Provide your dog with interactive toys and food- dispensing toys to keep them engaged and stimulated.
1. Obediance: Teach your dog to respond to your commands and to be well-behaving in public.
1. Proper training: Regularly enrolling your dog in training programs to keep them well-behaving and to learn new tricks.
1. Properly crate:



I'm planning a trip to New Orleans, and I'd love some recommendations for fun things to do and places to visit.

New Orleans is a vibrant city with a rich history and culture. Some popular attractions include the French Quarter, the National WWiMuseum, and the St. Louis Cathedral. You can also explore the historic Garden District and visit the iconic Bourbon Street for some night life. Additionally, don't miss the opportunity to try some delicious Cajun cuisine!

What kind of cajun cuisine? can you show me some pictures?

Certainly! Cajun cuisine is known for its hearty and flavorful dishes. Here are some pictures to give you a taste of what you can expect:

I've heard that New Orleans is famous for jazz music. Show me a piece of jazz music please.

You are right. Here is a jazz music:

Fantastic! I often see videos in which a bunch of people dance to jazz music, can you show me some?

Of course!

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

Unified MLLM: Harnessing Multi-Modalities

- Any-to-Any MLLM
  - NExT-GPT

NExT-GPT

Text + Audio

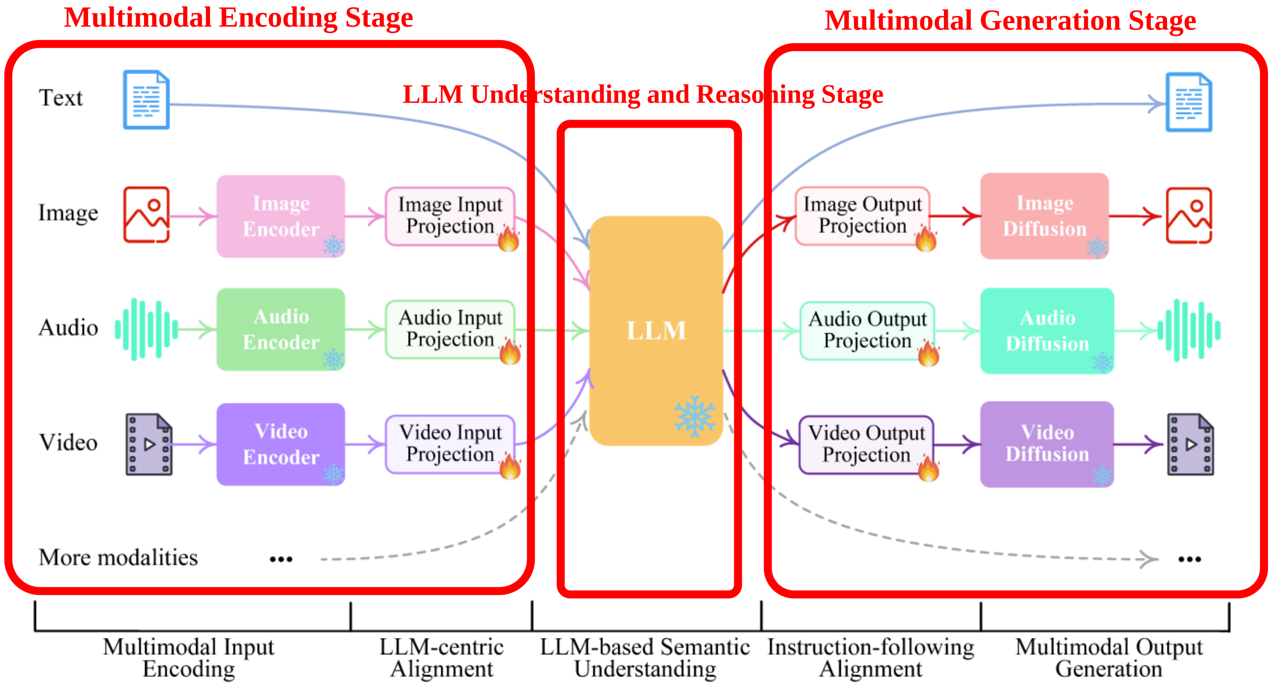
↓

Text + Image + Video

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

Unified MLLM: Harnessing Multi-Modalities

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

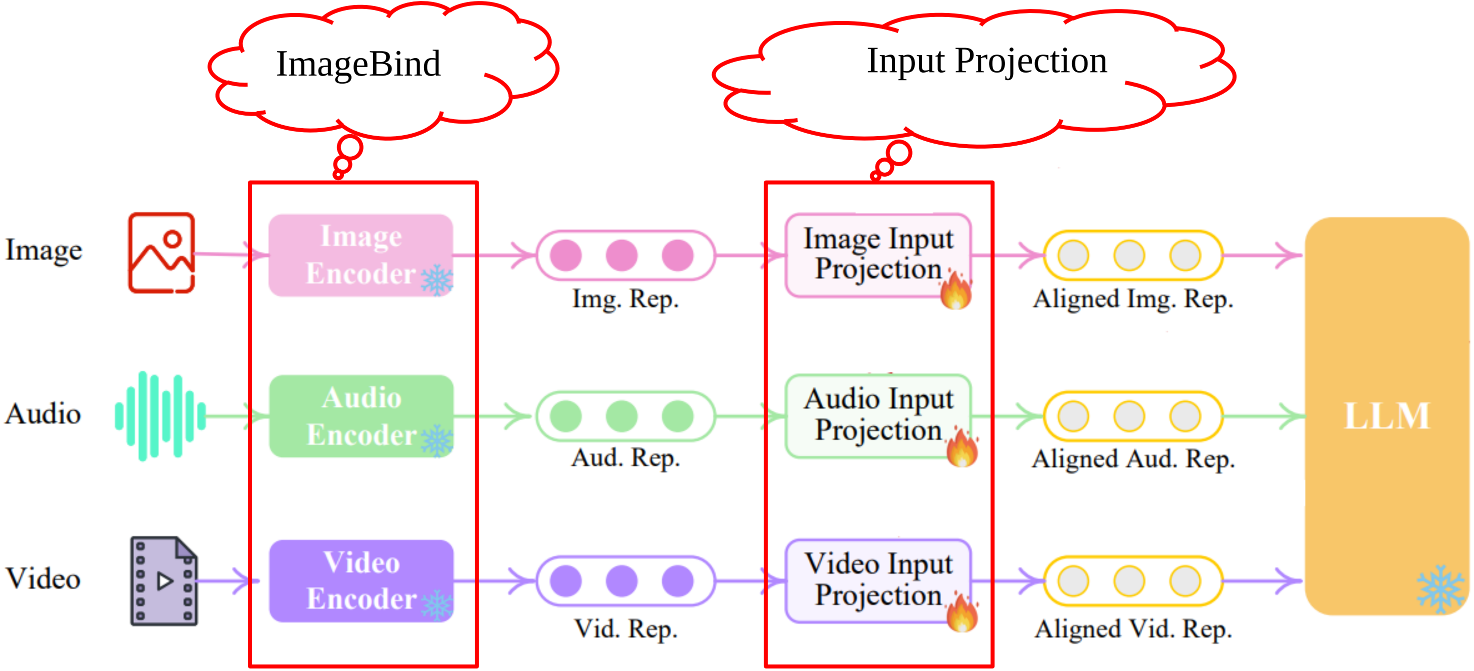


[1] NExT-GPT: Any-to-Any Multimodal LLM. 2023

Unified MLLM: Harnessing Multi-Modalities

- Any-to-Any MLLM
  - NExT-GPT

- Taking ImageBind as the unified multimodal encoder
- An input projection layer to connect multimodal encoder and LLM



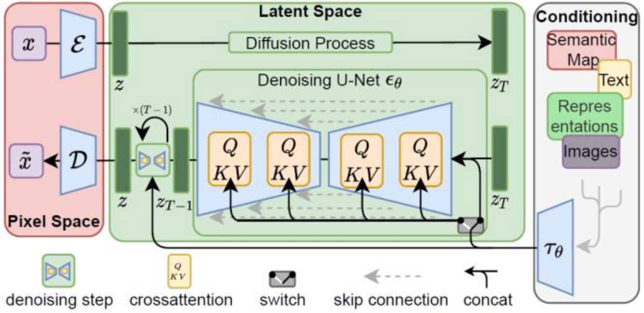
[1] NExT-GPT: Any-to-Any Multimodal LLM. 2023

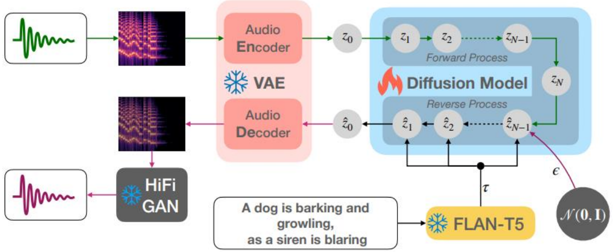
Unified MLLM: Harnessing Multi-Modalities

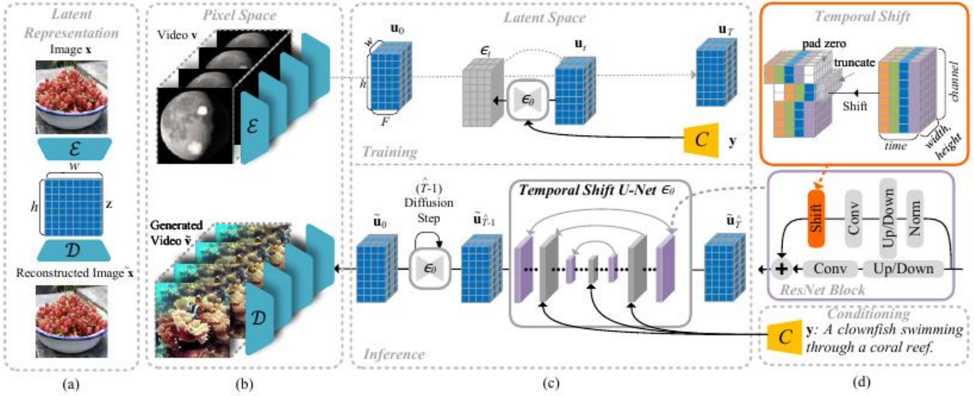
Any-to-Any MLLM

NExT-GPT

- Leveraging the current SoTA diffusion-based Text-to-Image, Video, Audio generation model to generate multimodal content
  - Text encoder – control the generation process
  - VAE
  - UNet



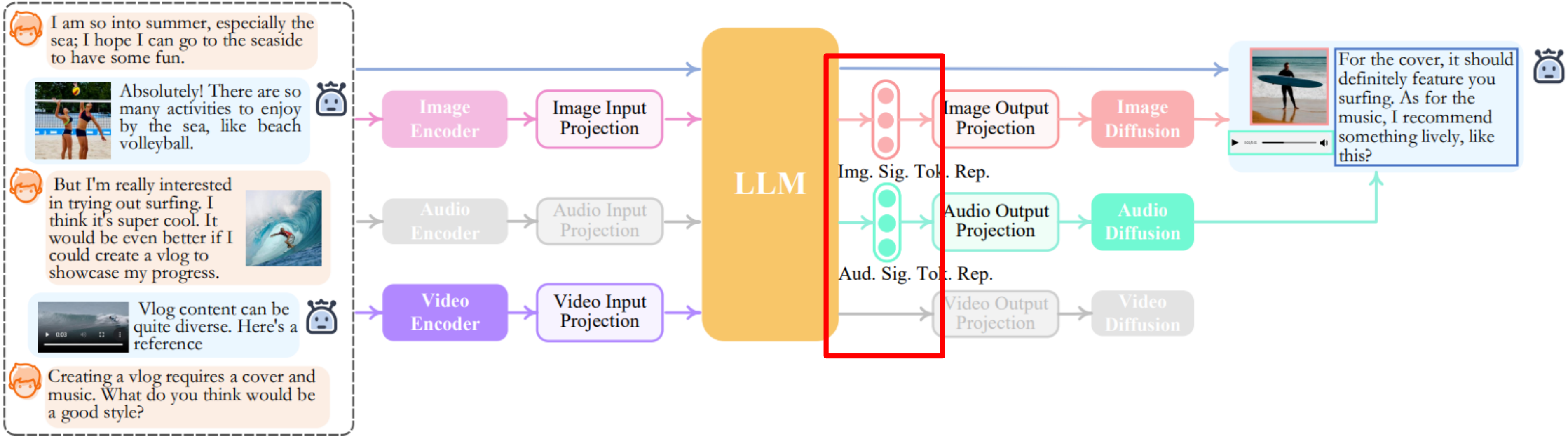




Unified MLLM: Harnessing Multi-Modalities

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

- Harnessing LLM as the core to decide whether & what modal content to output correspondingly.
- Instead of generating textual instructions, LLM produces unique “modality signal” tokens that serve as instructions to guide the generation process.



I am so into summer, especially the sea; I hope I can go to the seaside to have some fun.

Absolutely! There are so many activities to enjoy by the sea, like beach volleyball.

But I'm really interested in trying out surfing. I think it's super cool. It would be even better if I could create a vlog to showcase my progress.

Vlog content can be quite diverse. Here's a reference

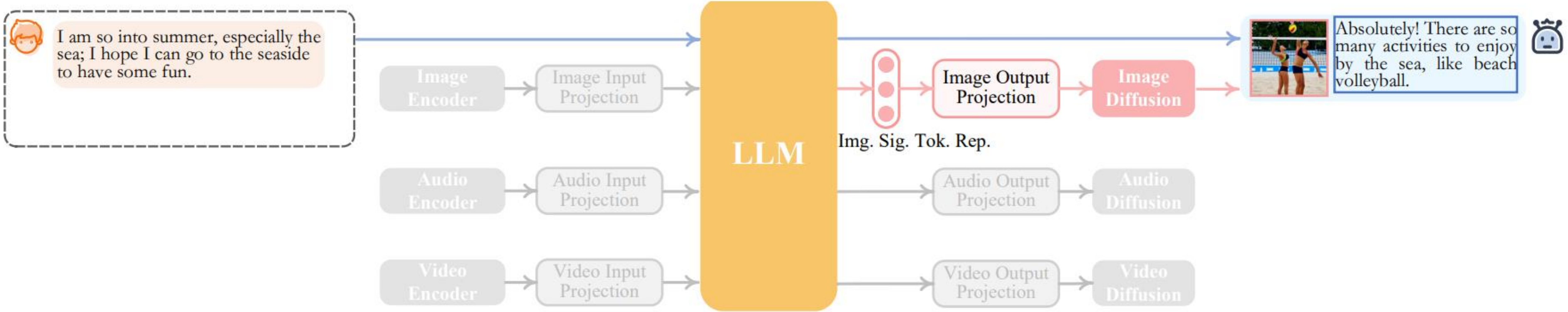
Creating a vlog requires a cover and music. What do you think would be a good style?

For the cover, it should definitely feature you surfing. As for the music, I recommend something lively, like this?

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

Unified MLLM: Harnessing Multi-Modalities

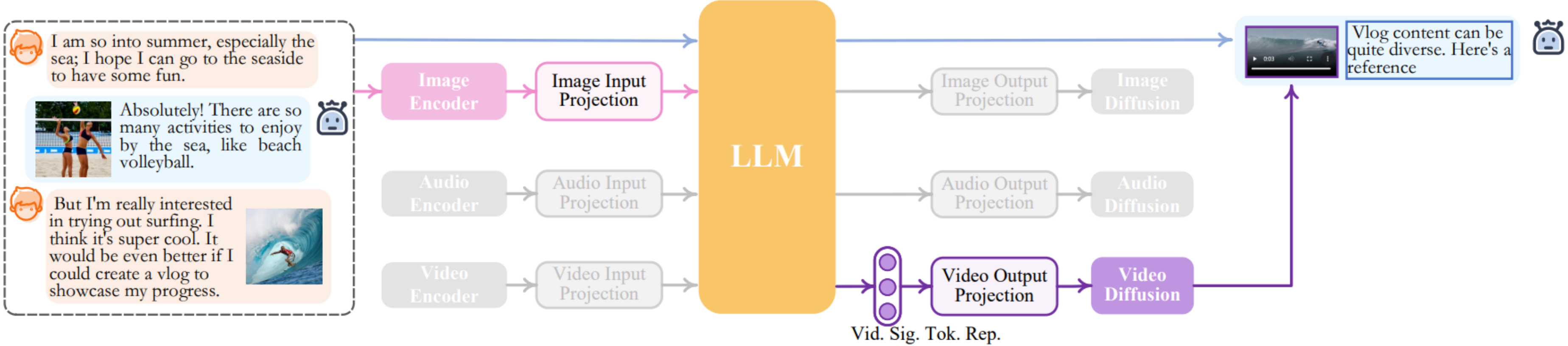
- ## Any-to-Any MLLM
  - NExT-GPT



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

Unified MLLM: Harnessing Multi-Modalities

- Any-to-Any MLLM
  - NExT-GPT



I am so into summer, especially the sea; I hope I can go to the seaside to have some fun.

Absolutely! There are so many activities to enjoy by the sea, like beach volleyball.

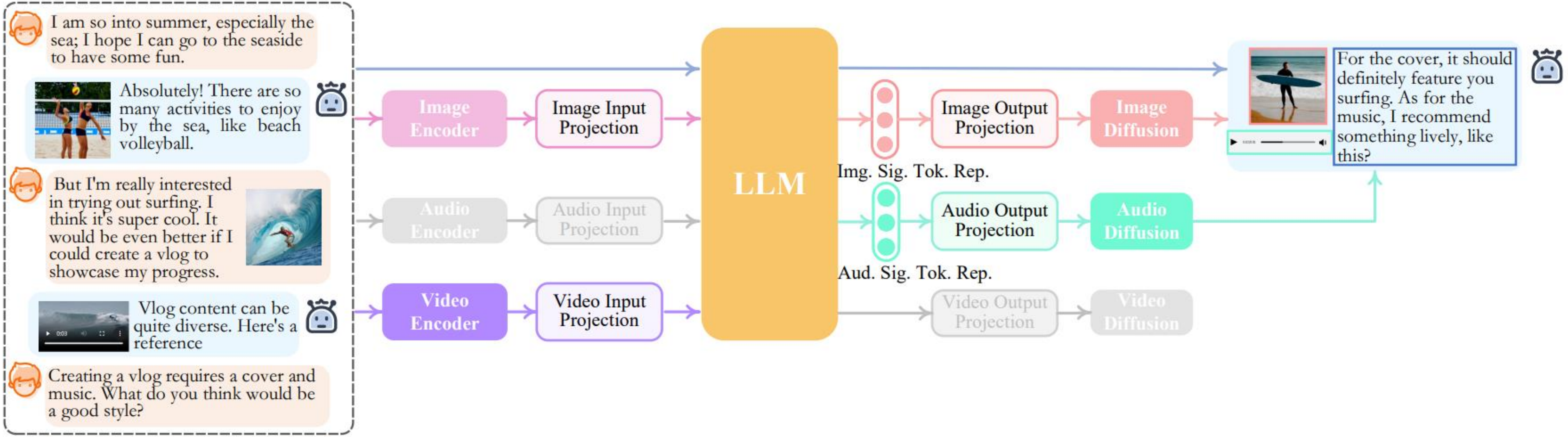
But I'm really interested in trying out surfing. I think it's super cool. It would be even better if I could create a vlog to showcase my progress.

Vlog content can be quite diverse. Here's a reference

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

Unified MLLM: Harnessing Multi-Modalities

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



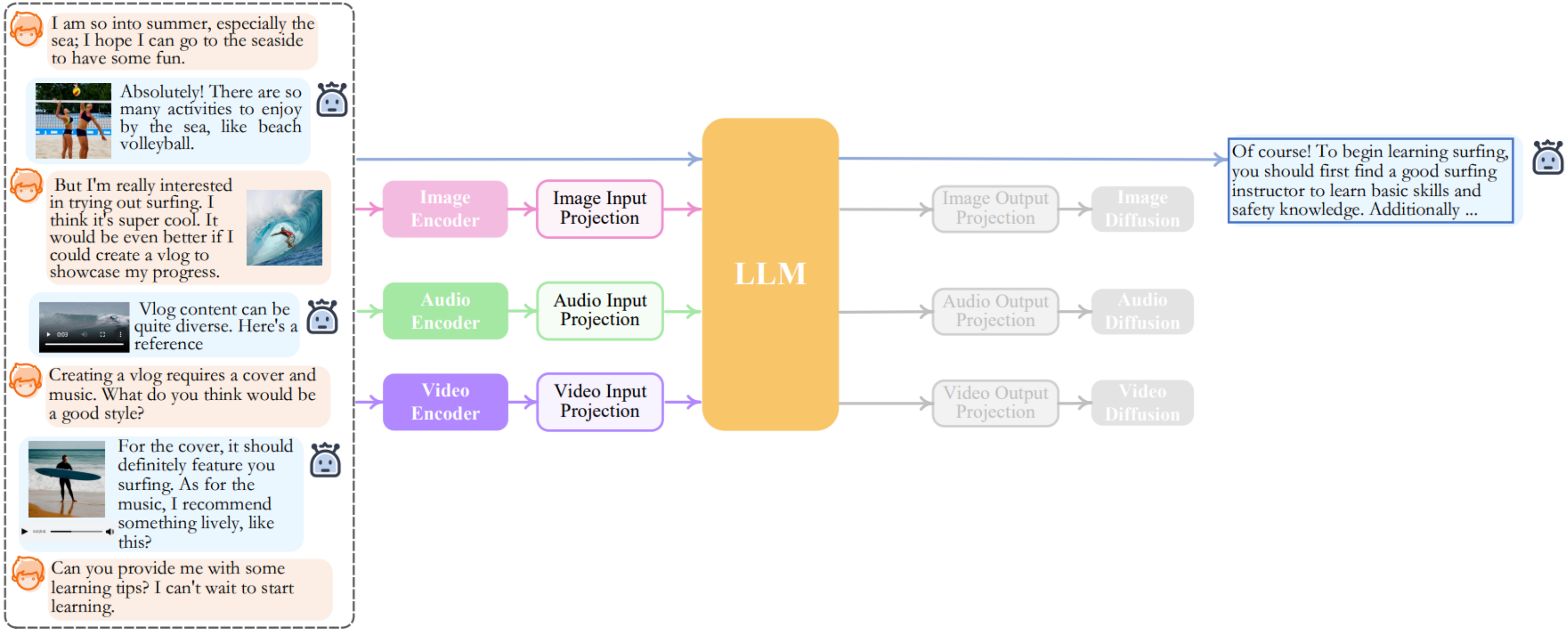
I am so into summer, especially the sea; I hope I can go to the seaside to have some fun.
Absolutely! There are so many activities to enjoy by the sea, like beach volleyball.
But I'm really interested in trying out surfing. I think it's super cool. It would be even better if I could create a vlog to showcase my progress.
Vlog content can be quite diverse. Here's a reference
Creating a vlog requires a cover and music. What do you think would be a good style?
Image Encoder
Image Input Projection
Audio Encoder
Audio Input Projection
Video Encoder
Video Input Projection
LLM
Img. Sig. Tok. Rep.
Aud. Sig. Tok. Rep.
Image Output Projection
Image Diffusion
Audio Output Projection
Audio Diffusion
Video Output Projection
Video Diffusion
For the cover, it should definitely feature you surfing. As for the music, I recommend something lively, like this?

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

Unified MLLM: Harnessing Multi-Modalities

- ## Any-to-Any MLLM

NExT-GPT



I am so into summer, especially the sea; I hope I can go to the seaside to have some fun.

Absolutely! There are so many activities to enjoy by the sea, like beach volleyball.

But I'm really interested in trying out surfing. I think it's super cool. It would be even better if I could create a vlog to showcase my progress.

Vlog content can be quite diverse. Here's a reference

Creating a vlog requires a cover and music. What do you think would be a good style?

For the cover, it should definitely feature you surfing. As for the music, I recommend something lively, like this?

Can you provide me with some learning tips? I can't wait to start learning.

Of course! To begin learning surfing, you should first find a good surfing instructor to learn basic skills and safety knowledge. Additionally ...

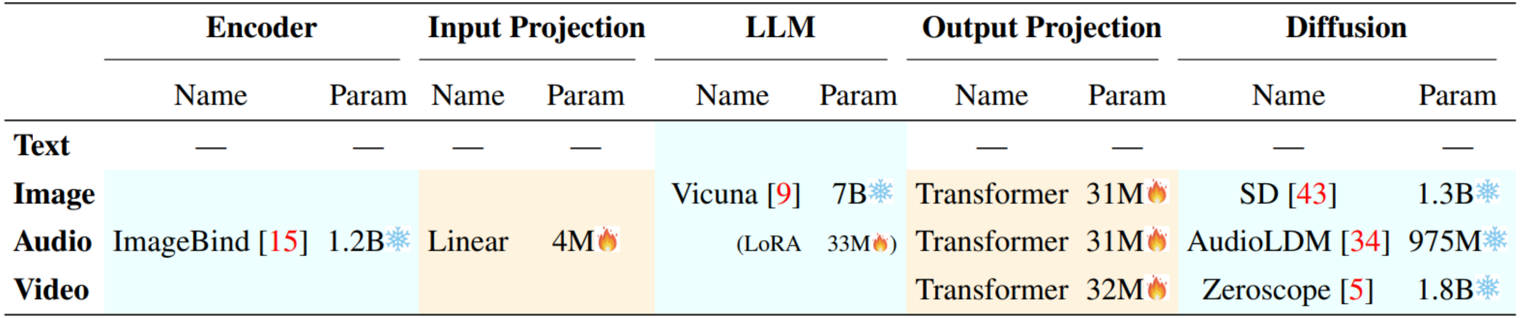
[1] NExT-GPT: Any-to-Any Multimodal LLM. 2023

Unified MLLM: Harnessing Multi-Modalities

- Any-to-Any MLLM

➢ Key Aspect-I: Parameter-efficient Low-cost Training

- NExT-GPT



	Encoder		Input Projection		LLM		Output Projection		Diffusion	
	Name	Param	Name	Param	Name	Param	Name	Param	Name	Param
Text	—	—	—	—	Vicuna [9] (LoRA	7B❄ 33M🔥)	—	—	—	—
Image	ImageBind [15]	1.2B❄	Linear	4M🔥			Transformer	31M🔥	SD [43]	1.3B❄
Audio							Transformer	31M🔥	AudioLDM [34]	975M❄
Video							Transformer	32M🔥	Zeroscope [5]	1.8B❄

Table 1: Summary of system configuration. Only 1% parameters need updating.

$$\frac{Tuned\ Params}{Frozen + Tuned\ Params} = \frac{(4M+33M+31M+31M+32M)}{(4M+33M+31M+31M+32M) + (1.2B+7B+1.3B+1.8B+0.975B)}$$

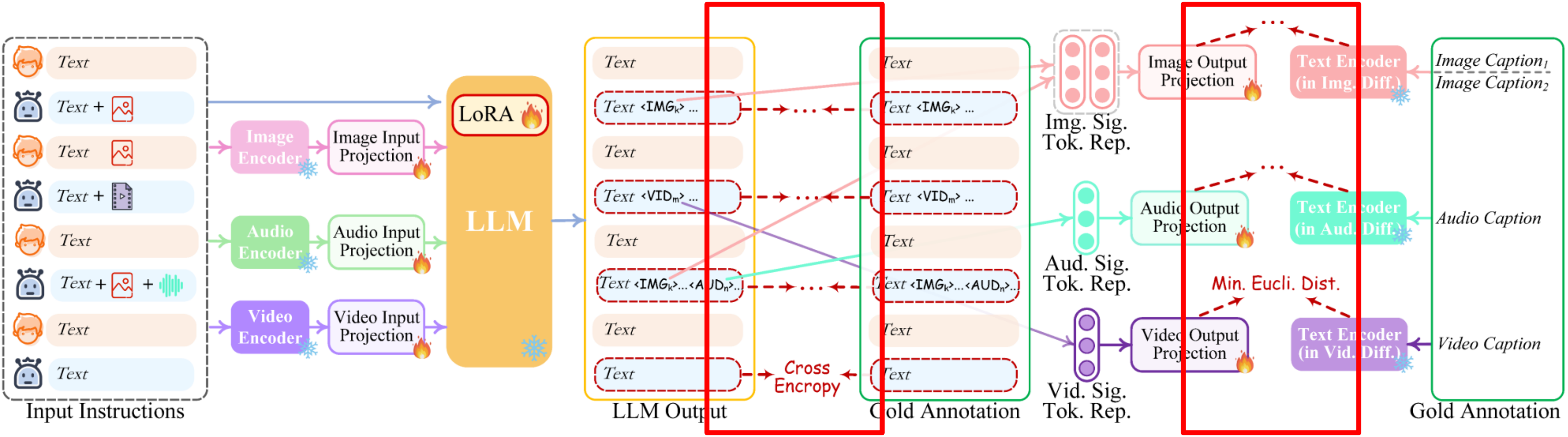
$$= \frac{131M}{131M\ +12.275B} \cong \mathbf{0.01}$$

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

Unified MLLM: Harnessing Multi-Modalities

- Any-to-Any MLLM
  - NExT-GPT

➢ Key Aspect-II： Modality-switching Instruction Tuning



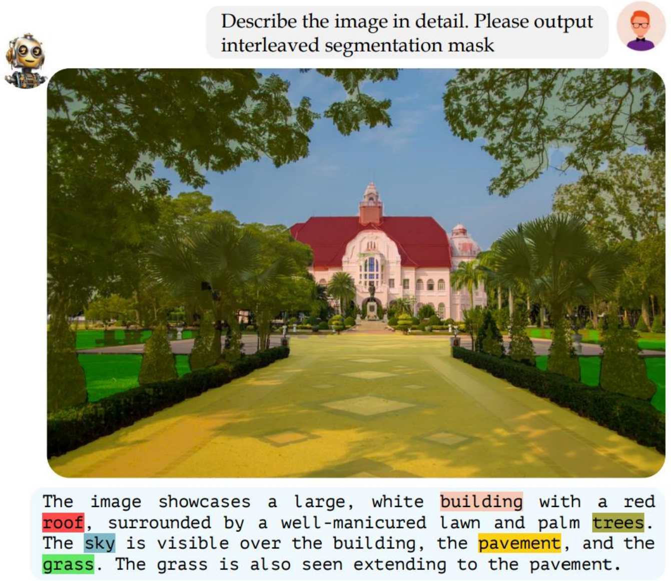
[1] NExT-GPT: Any-to-Any Multimodal LLM. 2023

Fine-grained Capability of MLLM

Pixel-level Vision MLLM

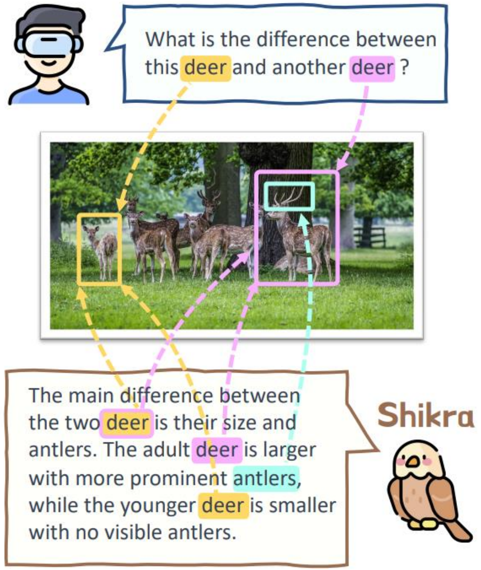
☞ *The vision MLLMs described above generally only support coarse-grained, instance-level visual understanding. This can lead to* ***imprecise visual interpretations****. Also due to the lack of visual grounding, these MLLMs will potentially* ***produce hallucinations****.*

- 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.



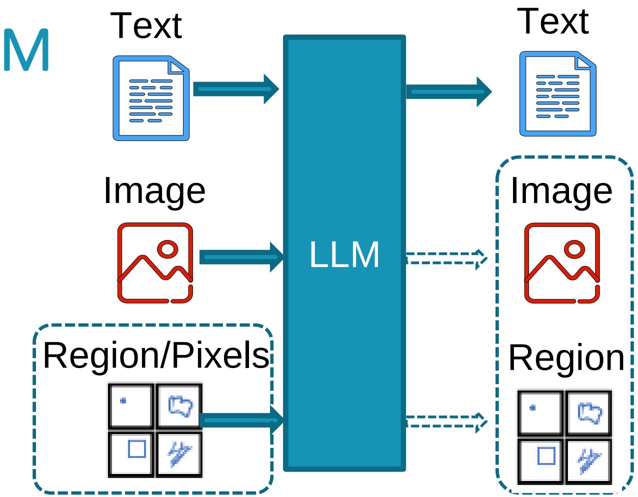
What is the difference between this deer and another deer ?

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
  - …

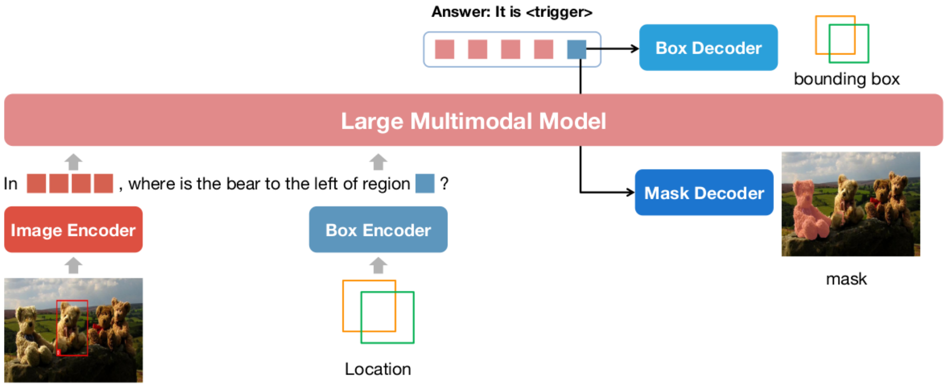
☞ *Users input an image (potentially specifying a region), and the LLM outputs content based on its understanding, grounding the visual content to specific pixel-level regions of the image.*



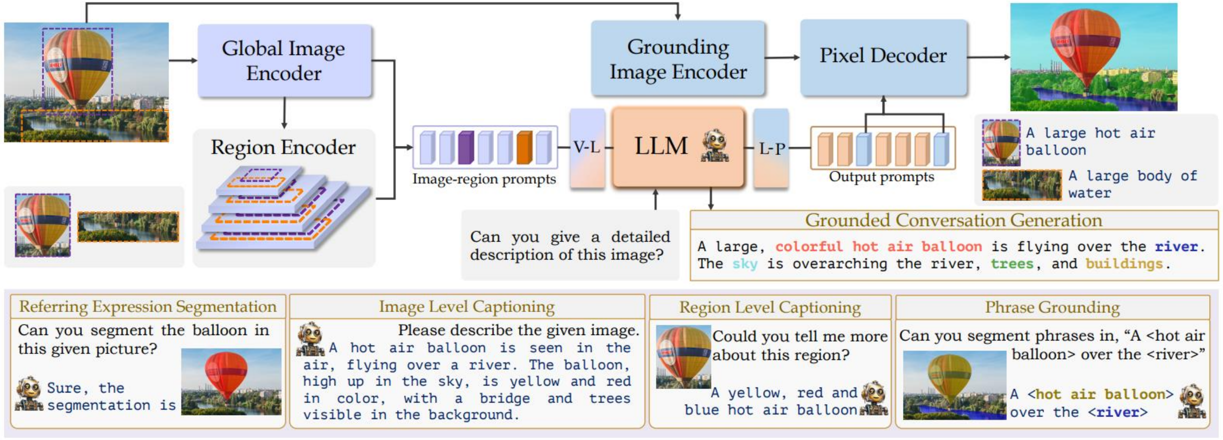
[1] GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. 2023
[2] NExT-Chat: An LMM for Chat, Detection and Segmentation. 2023
[3] MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. 2023
[4] Osprey: Pixel Understanding with Visual Instruction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

Fine-grained Capability of MLLM

- Image-oriented Pixel-wise
  - NExT-Chat



  - GLaMM



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?

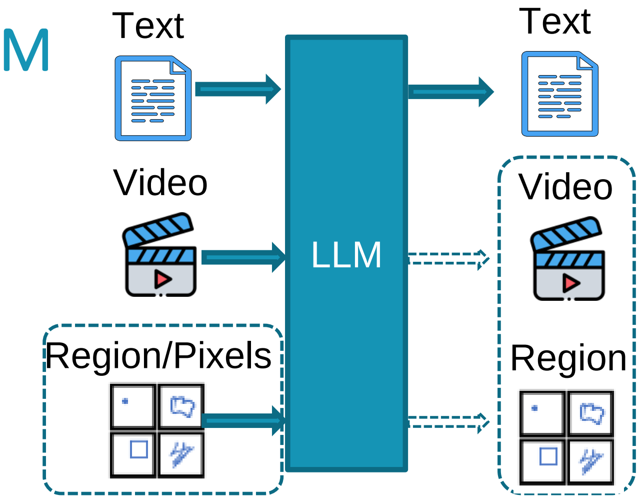
A yellow, red and blue hot air balloon

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

A <hot air balloon> over the <river>

Fine-grained Capability of MLLM

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



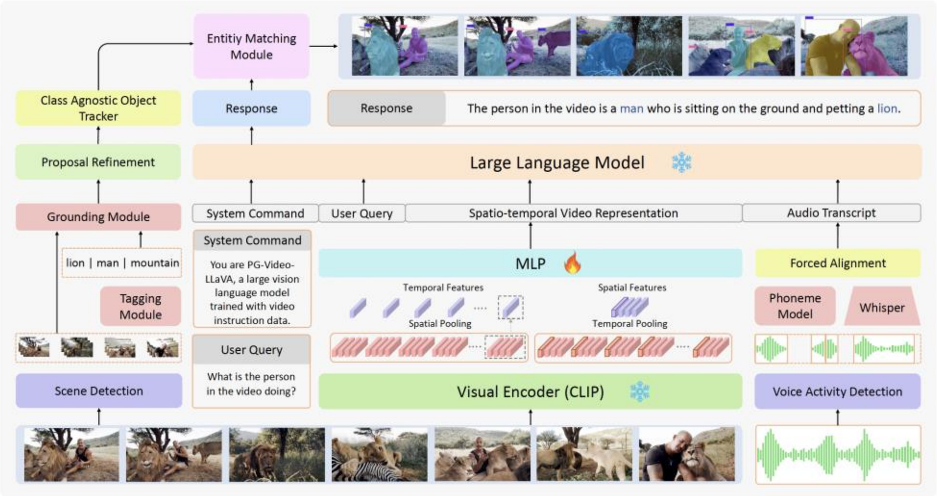
☞ *Users input an video (potentially specifying a region), and the LLM outputs content based on its understanding, grounding or tracking the content to specific pixel-level regions of the video.*

[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

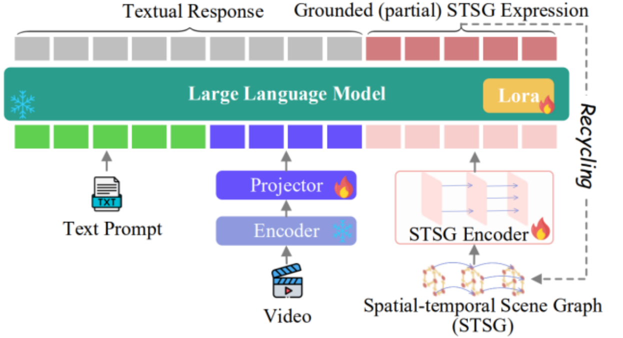


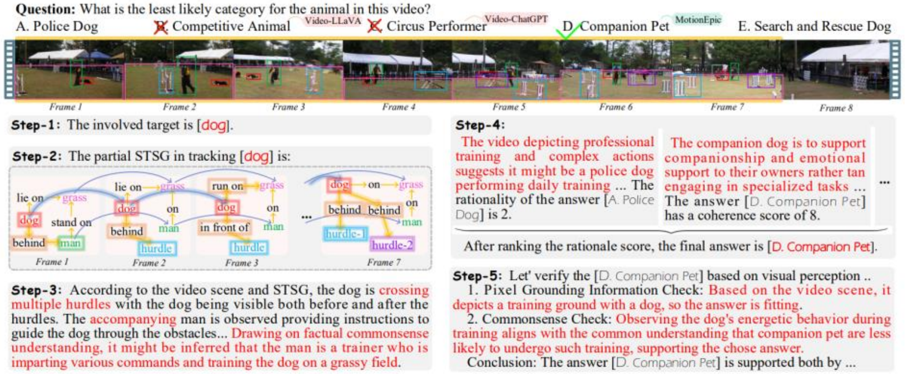
Response: The person in the video is a man who is sitting on the ground and petting a lion.

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

User Query: 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].

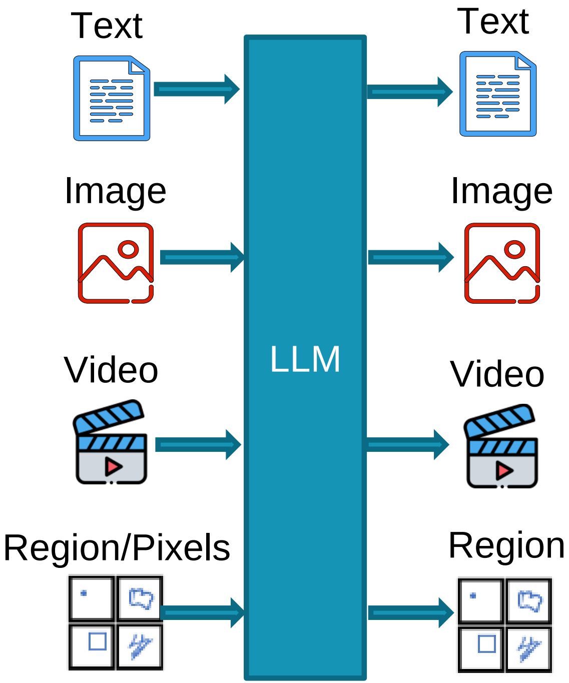
Step-5: Let' verify the [D. Companion Pet] based on visual perception ..
1. Pixel Grounding Information Check: Based on the video scene, it depicts a training ground with a dog, so the answer is fitting.
2. Commonsense Check: Observing the dog's energetic behavior during training aligns with the common 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

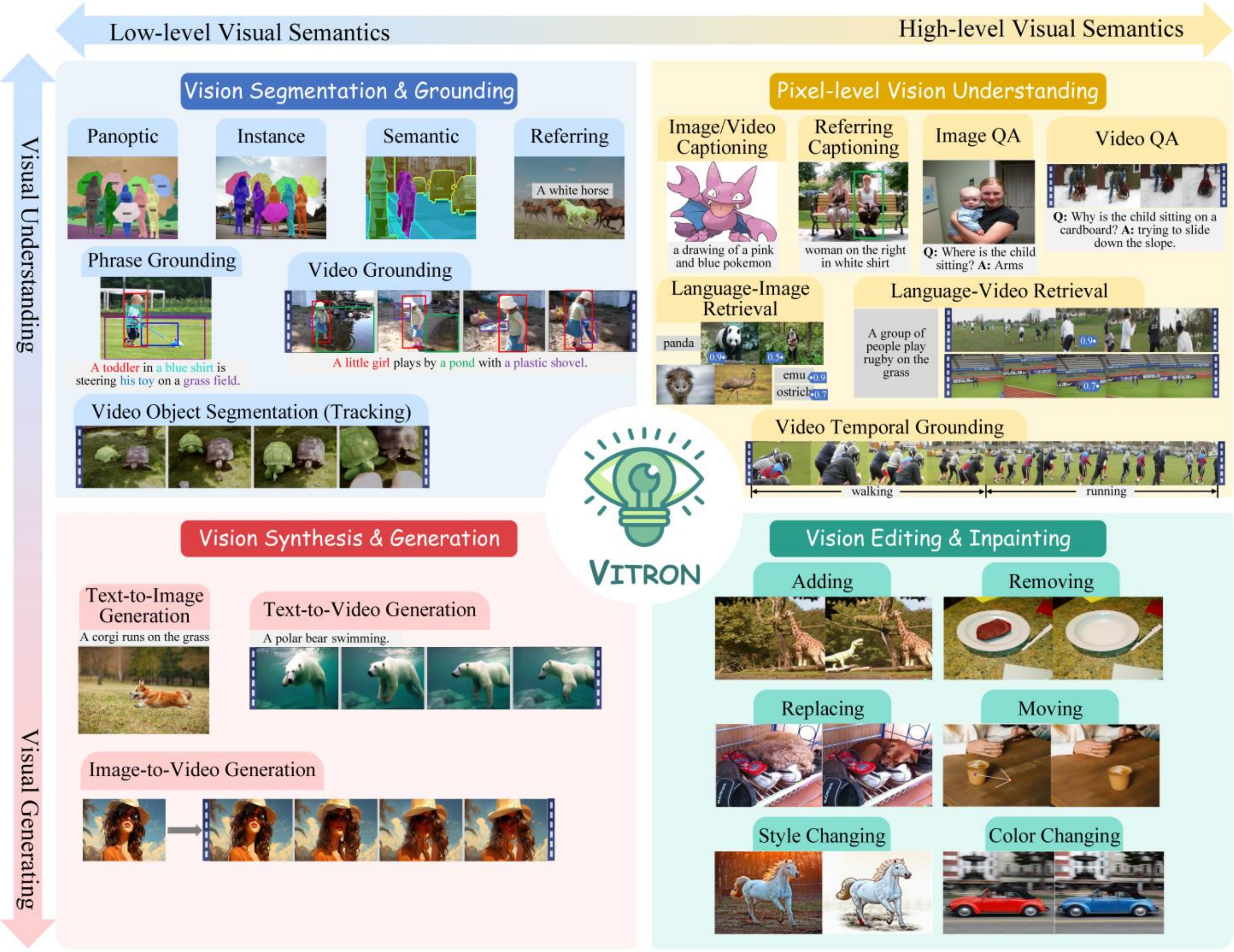
☞ *Users input either an image or video (potentially specifying a region), and the LLM outputs content based on its understanding, generating, grounding or tracking the content to specific pixel-level regions of the image, video.*



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

Fine-

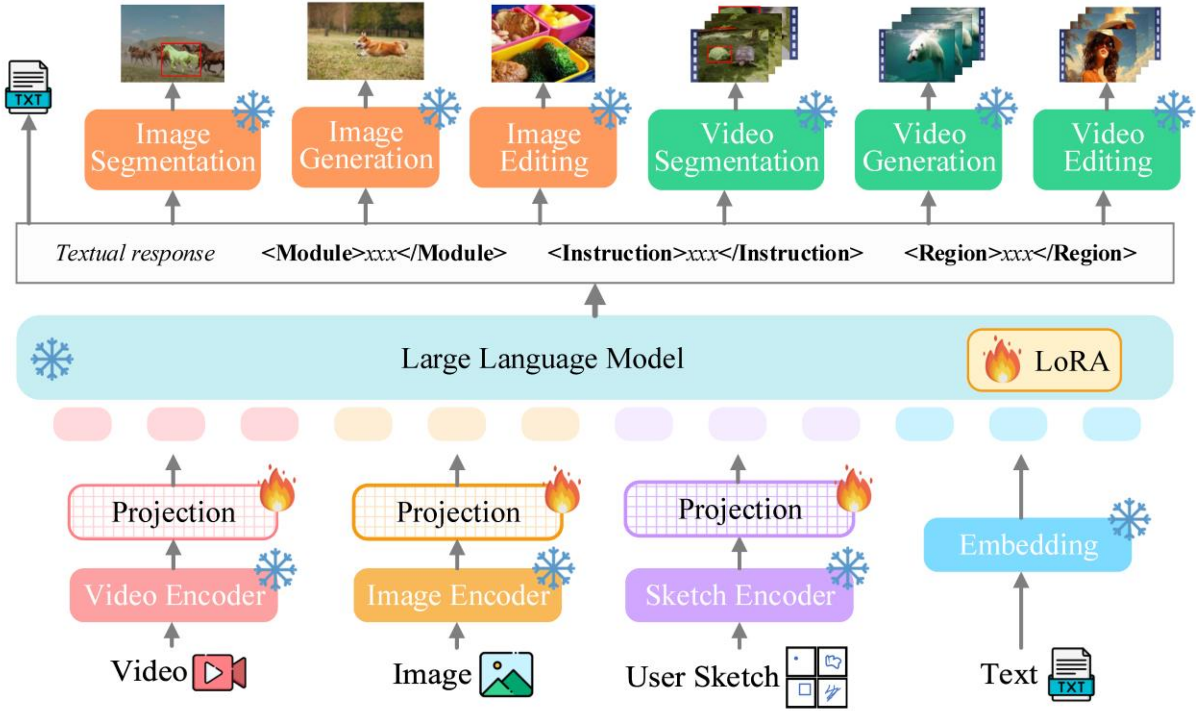
- Unified
  - Vitron


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

Fine-grained Capability of MLLM

- ## Unified Pixel-wise MLLM

  - Vitron

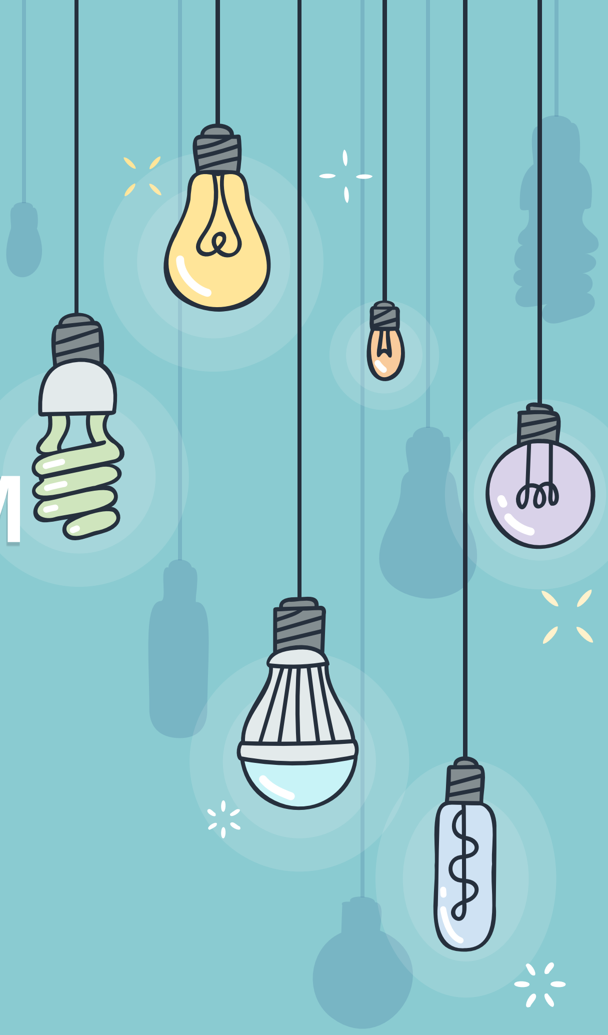


[1] VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing. 2024

2

Architecture of MLLM

How to design an MLLM?



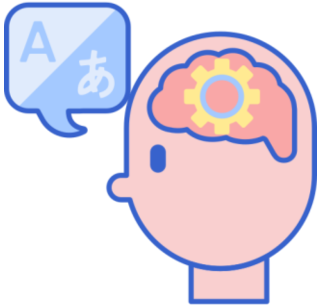
Overview of MLLM Architecture

- ## Preliminary Idea: Intelligence over Language

☞ *Due to the scaling law, emergent phenomena have extensively already occurred in language-based LLMs.*

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

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



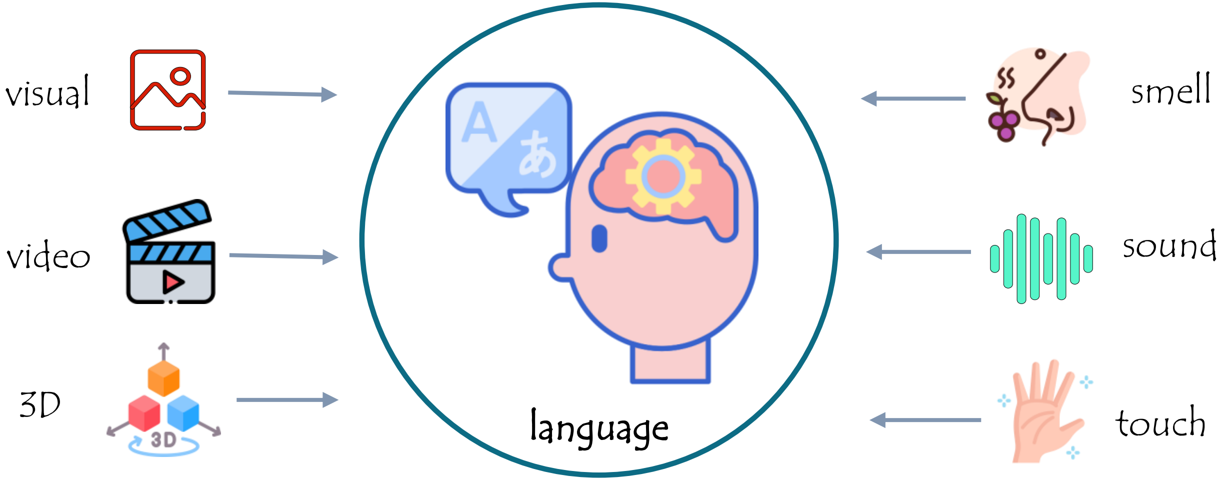
language

Overview of MLLM Architecture

Preliminary Idea: Language Intelligence as Pivot

☞ *Given this premise, nearly all CURRENT MLLMs are built based on language-based LLMs as the core decision-making module (i.e., the brain or central processor).*

☞ *By adding additional external non-textual modality modules or encoders, LLMs are enabled with multimodal perceptual/operation abilities.*



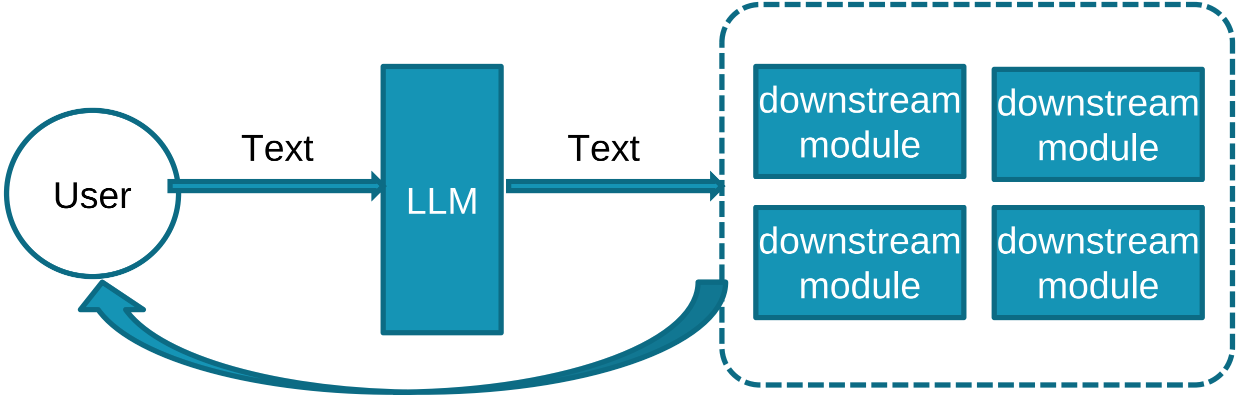
Overview of MLLM Architecture

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

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

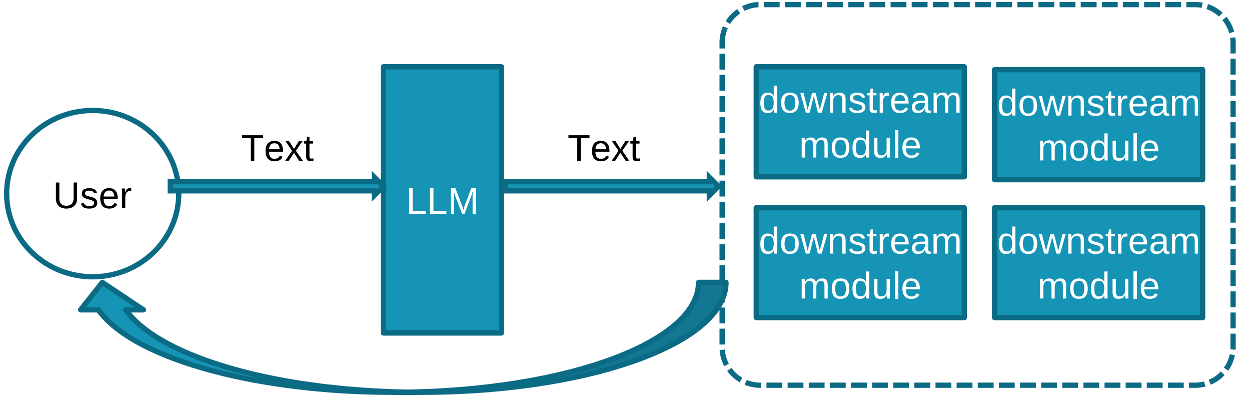
- Key feature:

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



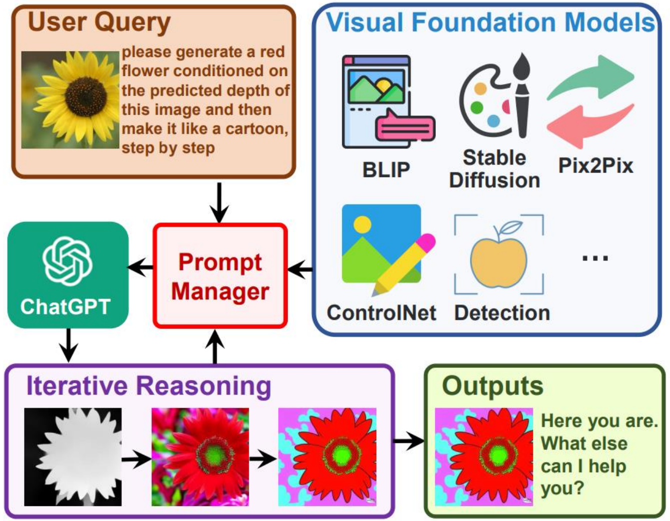
Overview of MLLM Architecture

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



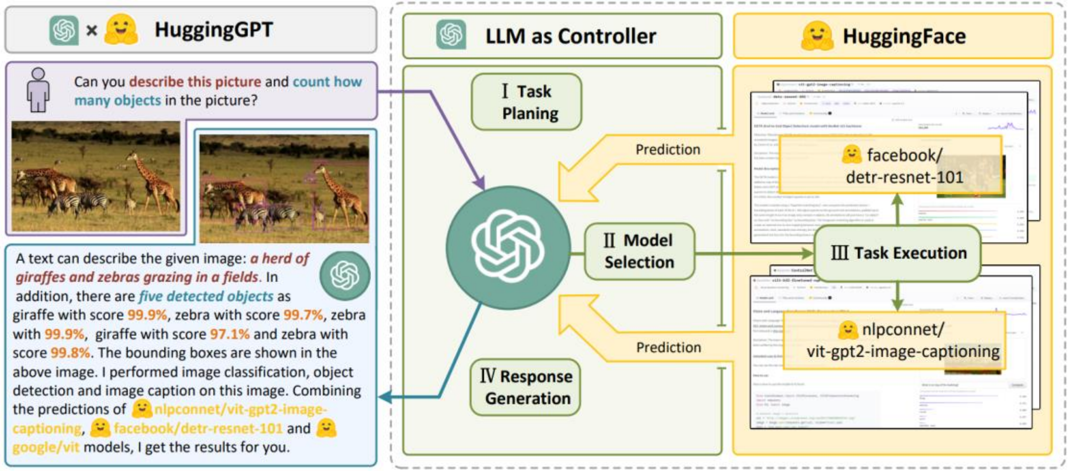
Overview of MLLM Architecture

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



please generate a red flower conditioned on the predicted depth of this image and then make it like a cartoon, step by step

Here you are. What else can I help you?



Can you describe this picture and count how many objects in the picture?

A text can describe the given image: *a herd of giraffes and zebras grazing in a fields*. In addition, there are *five detected objects* as giraffe with score 99.9%, zebra with score 99.7%, zebra with 99.9%, giraffe with score 97.1% and zebra with score 99.8%. The bounding boxes are shown in the above image. I performed image classification, object detection and image caption on this image. Combining the predictions of nlpconnet/vit-gpt2-image-captioning, facebook/detr-resnet-101 and google/vit models, I get the results for you.

[1] Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. 2023
[2] HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. 2023

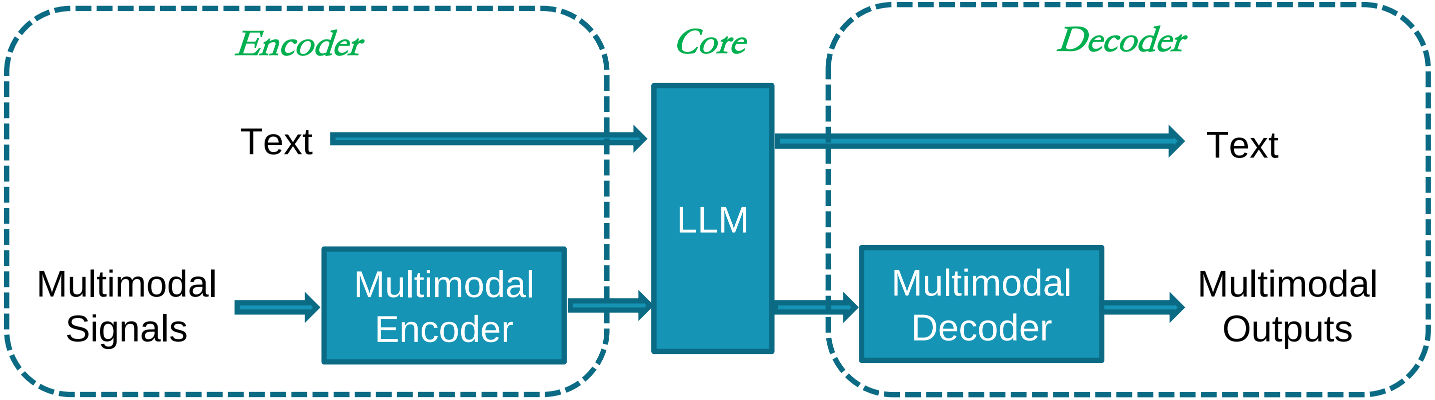
Overview of MLLM Architecture

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

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

- Key feature:

  LLM is the key joint part of the system, receiving multimodal information directly from outside, and delegating instruction to decoders/generators in a more smooth manner.

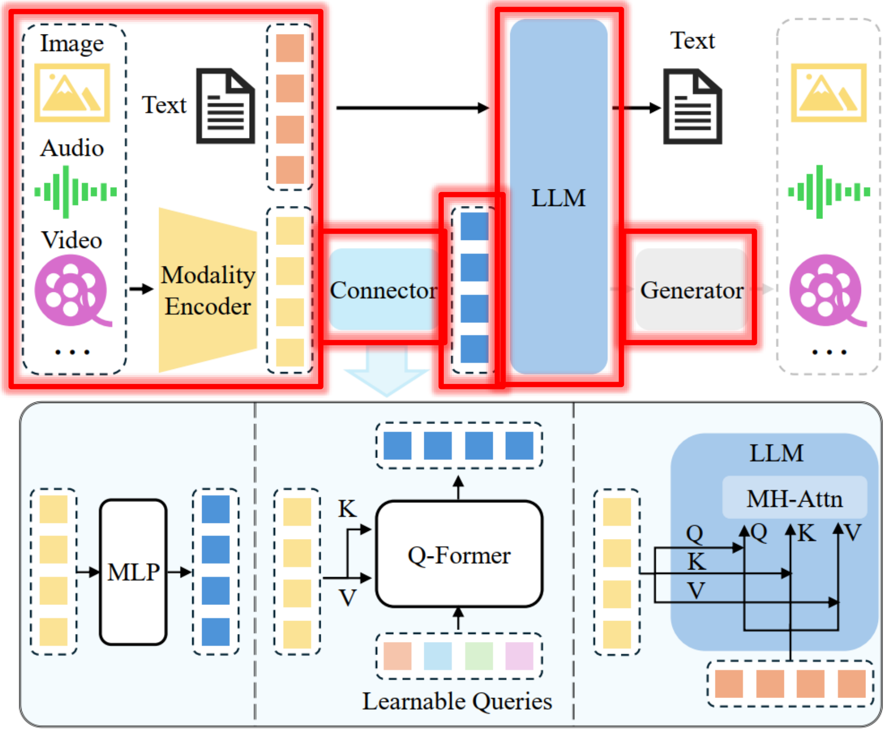


Overview of MLLM Architecture

- Architecture-II: LLM as Joint Part of System

More promising

  - ≈ 96% MLLMs belong to this category.

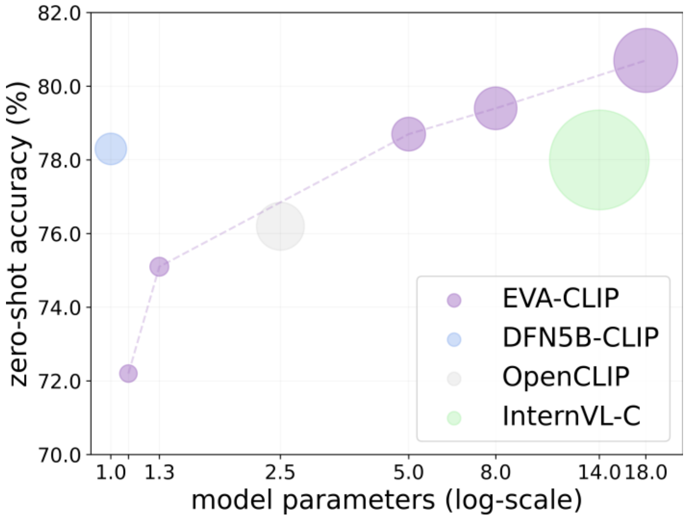


[1] *A Survey on Multimodal Large Language Models. https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models, 2023.*

Multimodal Encoding

Visual (Image&Video) Encoder

- **CLIP-ViT** is the most popular choice for vision-language models.
- Advantages:
  - Providing image representations well aligned with text space.
  - Scale well with respect to parameters and data.



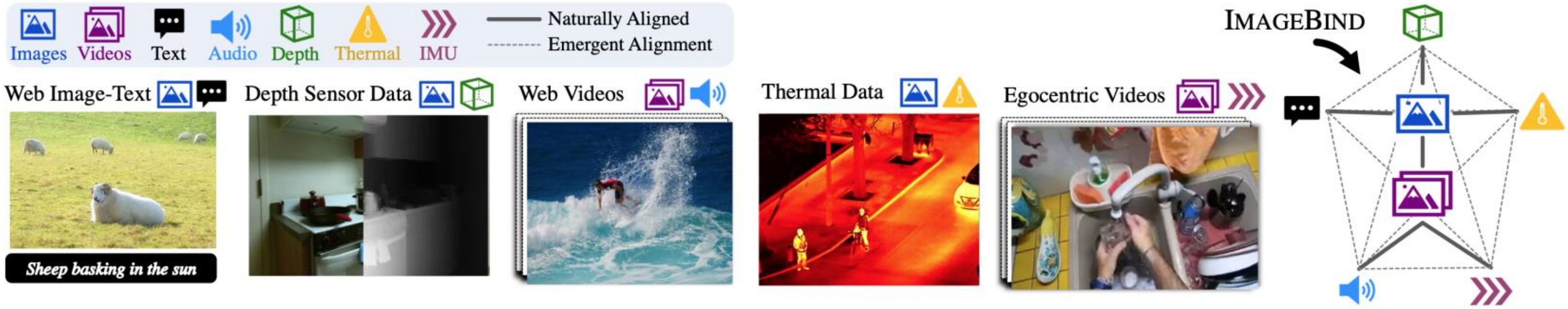
Multimodal Encoding

- ## Non-Visual Encoder
  - Audio:
    - HuBERT
    - Whisper
    - BEATs

  - 3D Point:
    - Point-BERT

Multimodal Encoding

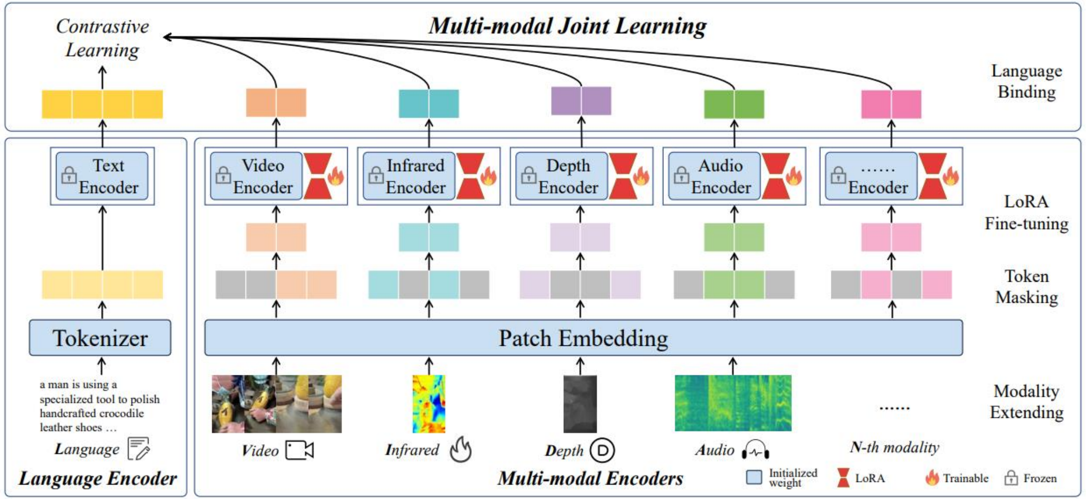
- ## Unified Multimodal Encoder
  - ImageBind:
    - Embedding all modalities into a joint representation space of Image.
    - Well aligned modality representations can benefit LLM understanding



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

Multimodal Encoding

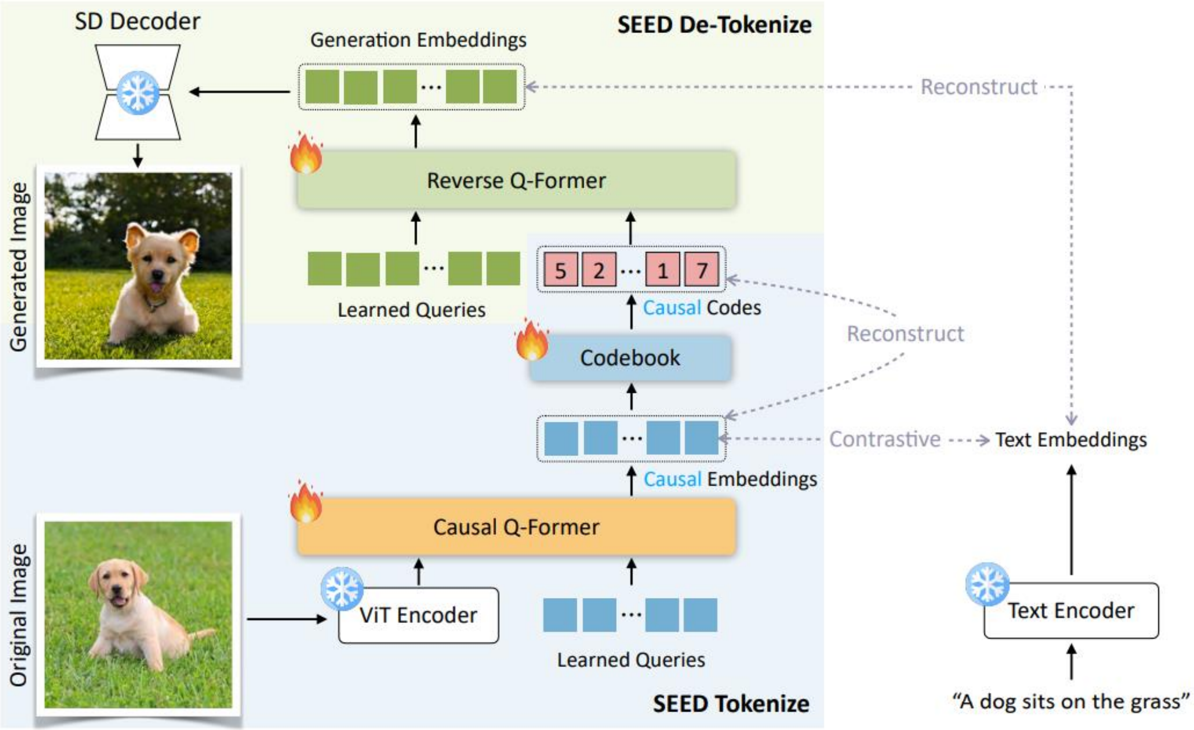
- Unified Multimodal Encoder
  - LanguageBind:
    - Embedding all modalities into a joint representation space of Language.
    - Well aligned modality representations can benefit LLM understanding



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

Multimodal Signal Tokenization

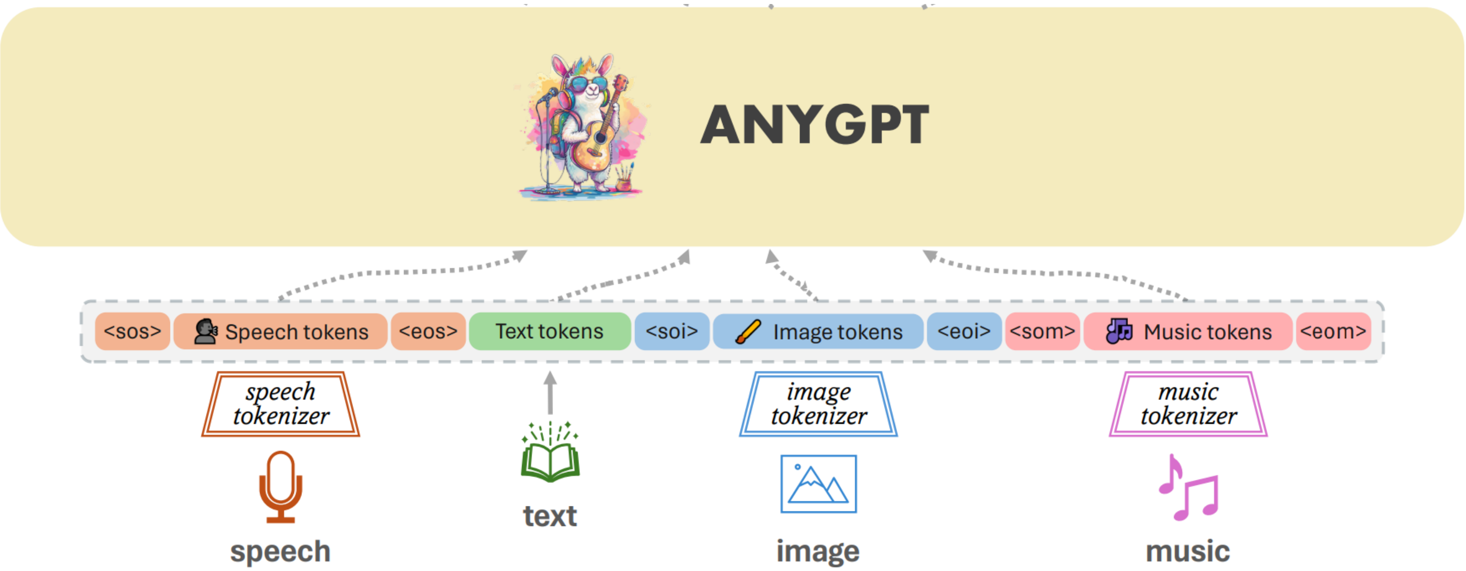
- Tokenization
  - SEED



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

Multimodal Signal Tokenization

- Tokenization
  - AnyGPT



ANYGPT

<sos> Speech tokens <eos> Text tokens <soi> Image tokens <eoi> <som> Music tokens <eom>

speech tokenizer

speech

text

image tokenizer

image

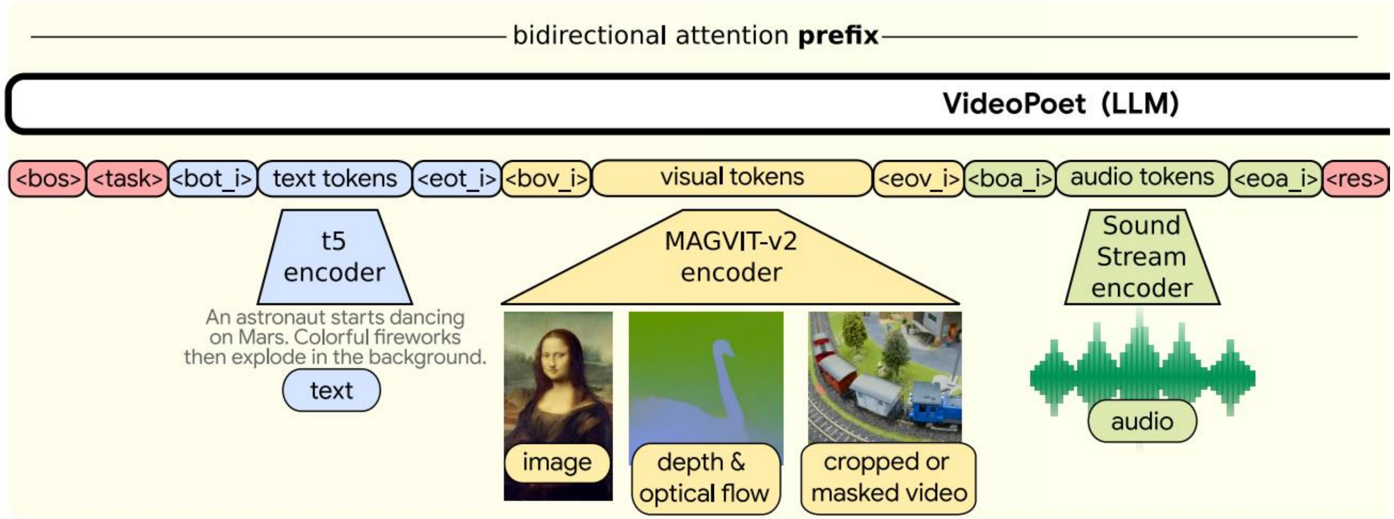
music tokenizer

music

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

Multimodal Signal Tokenization

- ## Tokenization
  - VideoPoet



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

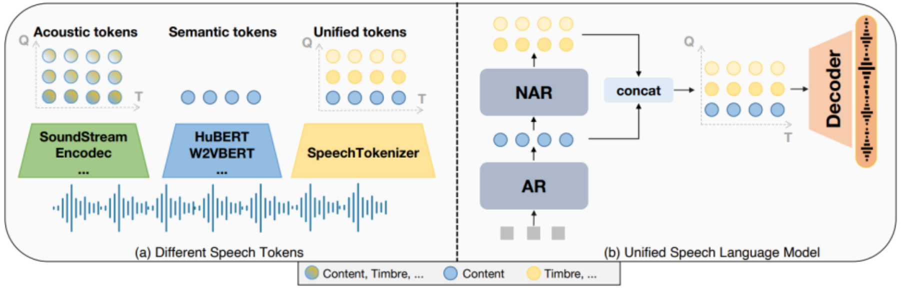
Multimodal Signal Tokenization

Visual (Image&Video) Tokenization in Codebook

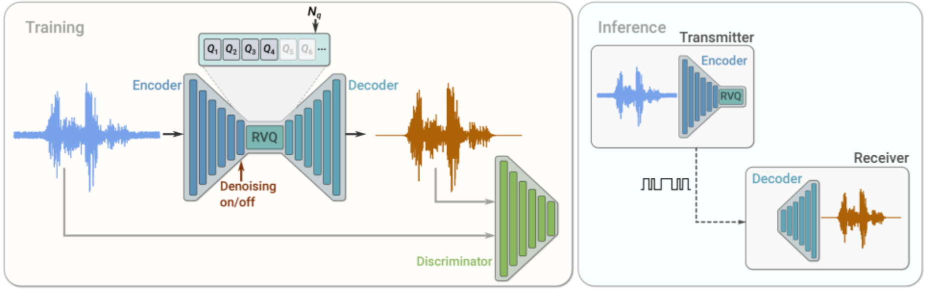
- Represent multimodal signals as discrete tokens in a codebook
  - Advantages: support unified multimodal signal understanding and generation in an auto-regressive next-token prediction framework
  - More commonly used in image synthesize
    - **Parti**
    - **Muse** (parallel)
    - **MaskGIT** (parallel)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

Multimodal Signal Tokenization

- Audio Tokenization
  - SpeechTokenizer +RVQ-VAE



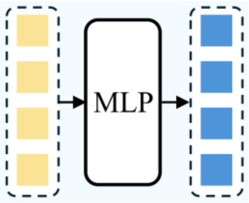
  - SoundStream +RVQ-VAE

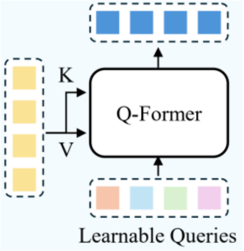


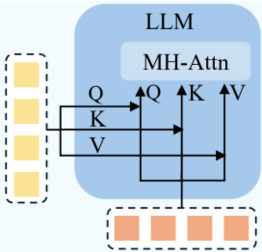
[1] SpeechTokenizer: Unified Speech Tokenizer for Speech Large Language Models. 2023
[2] SoundStream: An End-to-End Neural Audio Codec. 2021

Input-side Projection

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



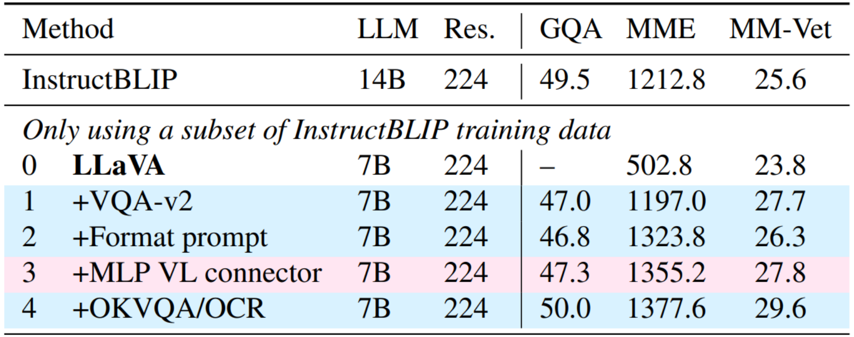




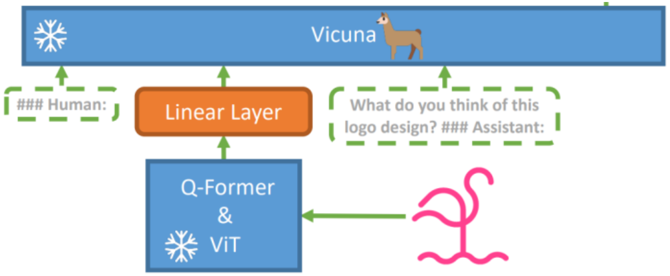
Input-side Projection

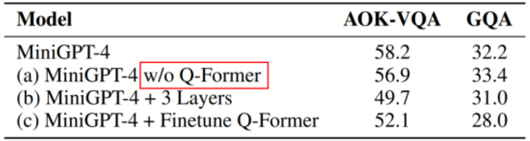
Some Insights

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



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





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

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

Backbone LLMs

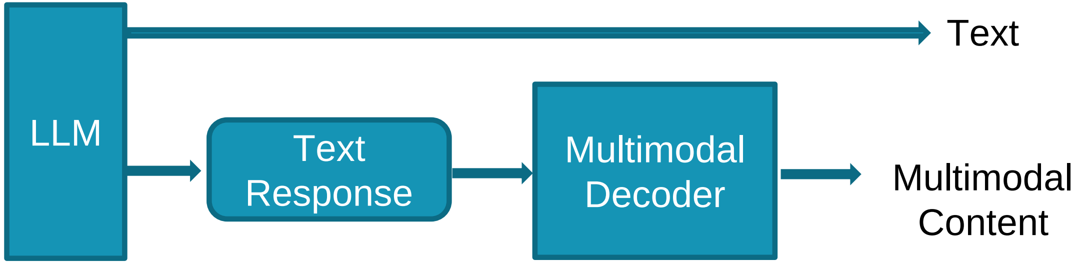
- ## Open-source Language-based LLMs

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

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

Decoding-side Connection

- Message passing via 1) discrete token of language
  - Representative MLLMs:
    - Visual-ChatGPT
    - HuggingGPT
    - GPT4Video
    - MM-REACT
    - ViperGPT
    - ModaVerse
    - Vitron
    - ...



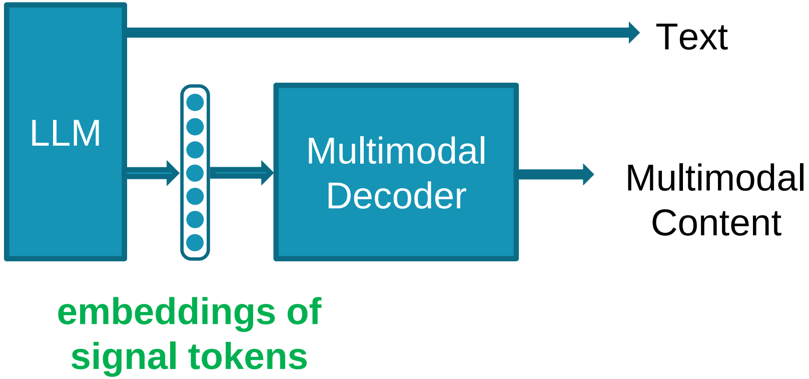
  - Pros:
    - High performance lower-bound
    - More Efficient, i.e., without tuning
  - Cons:
    - Loss of end-to-end tuning capabilities.
    - Performance upper-bound is limited, i.e., some multimodal signals cannot be optimally conveyed through text).

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

Decoding-side Connection

Message passing via 2) continuous embedding

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

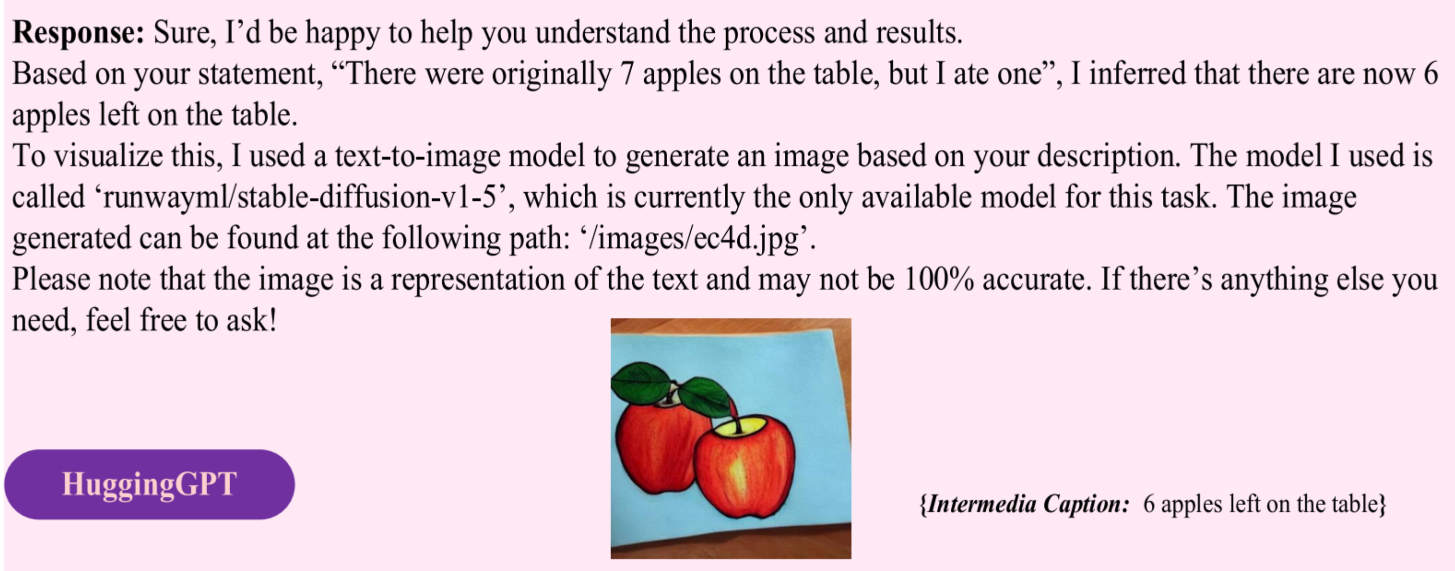


- Merits
  - Capable of end-to-end tuning, resulting in more efficient instruction transmission
  - More able to convey various multimodal signals that text alone cannot express, e.g.,
    - *the numeration of vision*
    - *the visual-spatial relational semantics*

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

Decoding-side Connection

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



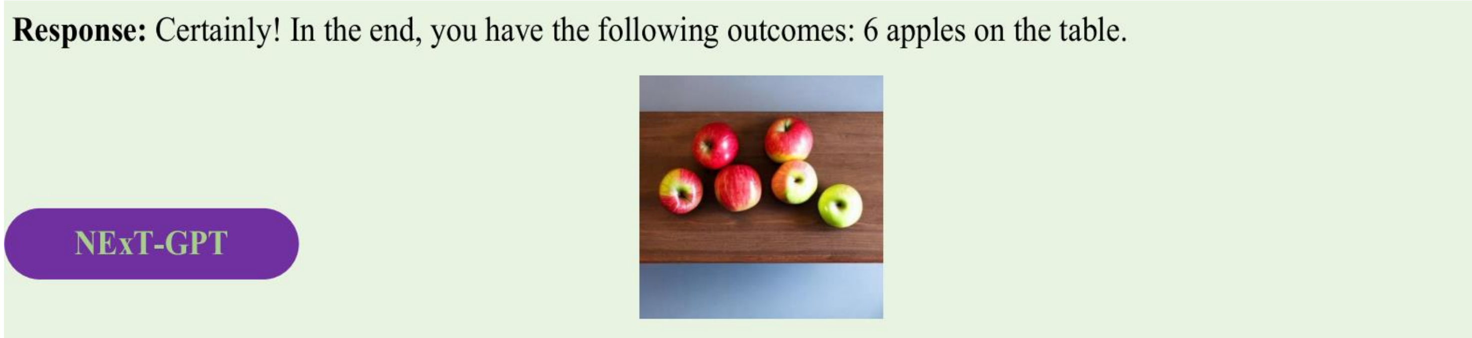
Response: Sure, I'd be happy to help you understand the process and results.
Based on your statement, "There were originally 7 apples on the table, but I ate one", I inferred that there are now 6 apples left on the table.
To visualize this, I used a text-to-image model to generate an image based on your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

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

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

Decoding-side Connection

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



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

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

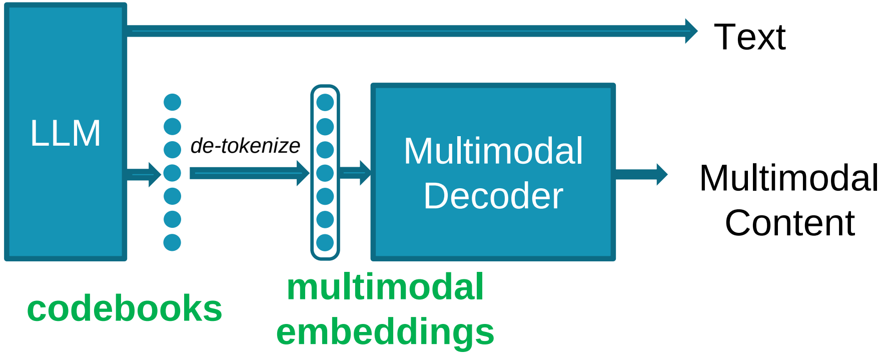
Decoding-side Connection

Message passing via 3) codebooks

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

Merits

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

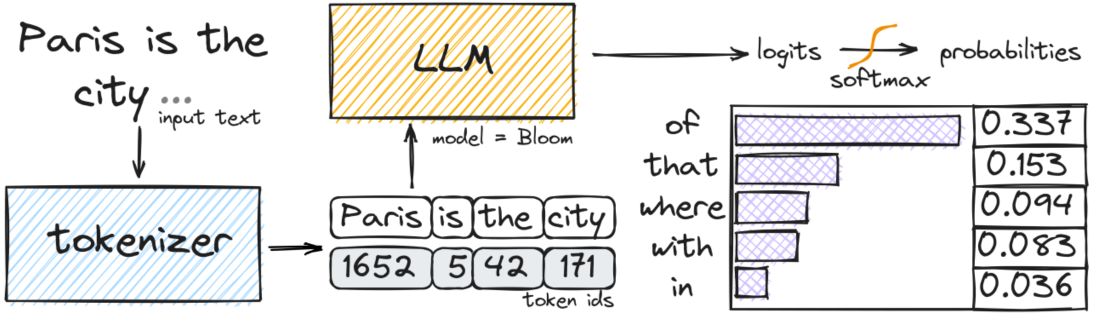


[1] Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action. 2023
[2] LVM: Sequential Modeling Enables Scalable Learning for Large Vision Models. 2023
[3] AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. 2024
[4] VideoPoet: A Large Language Model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

Multimodal Generation

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

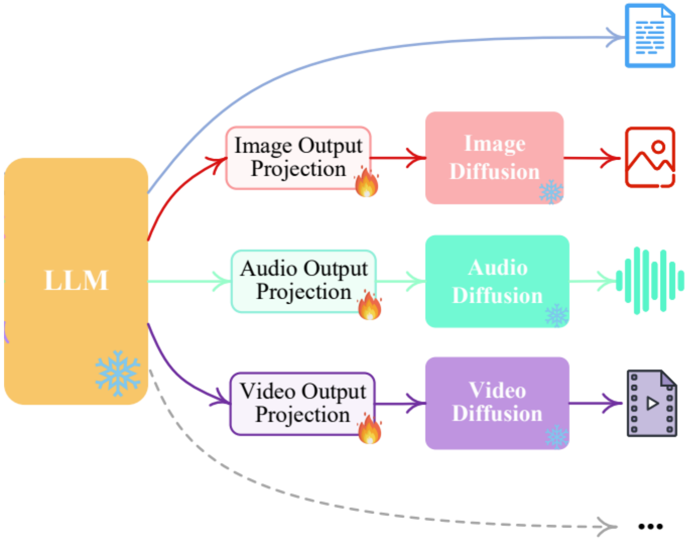
    via e.g., BPE decoding, Beam search, …



Multimodal Generation

- ## Generation via Diffusion Models

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

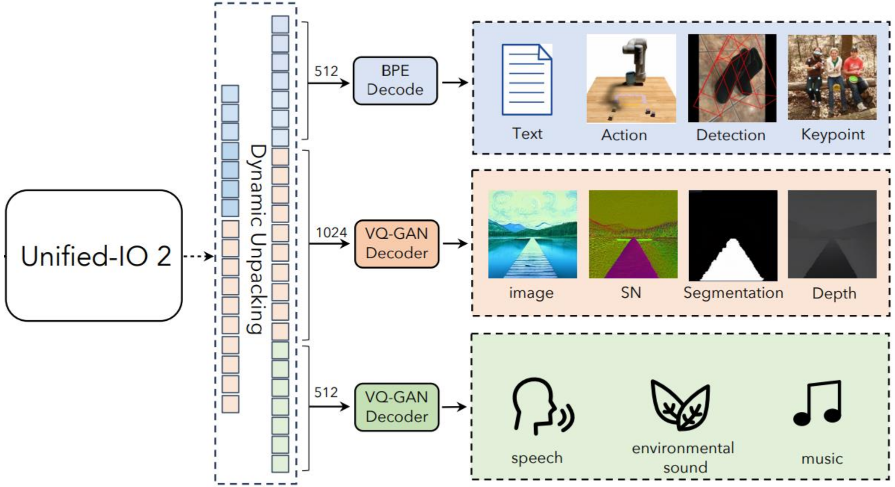


[1] NExT-GPT: Any-to-Any Multimodal LLM. 2023

Multimodal Generation

- ## Generation via Codebooks

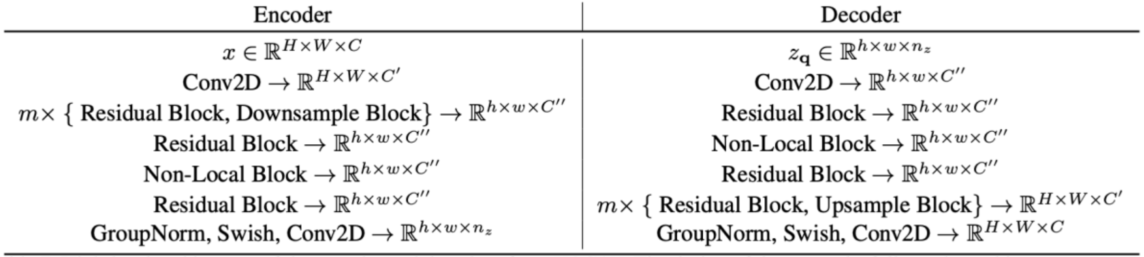
  - Visual (Image/Video) Generator
    - VQ-VAE + Codebooks
    - VQ-GAN + Codebooks
  - Audio Generator
    - SpeechTokenizer + Residual Vector Quantizer
    - SoundStream + Residual Vector Quantizer



[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
    - <u>64 × 64 × 3</u> or <u>32 × 32 × 4</u>



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

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 $f = 2^m$.

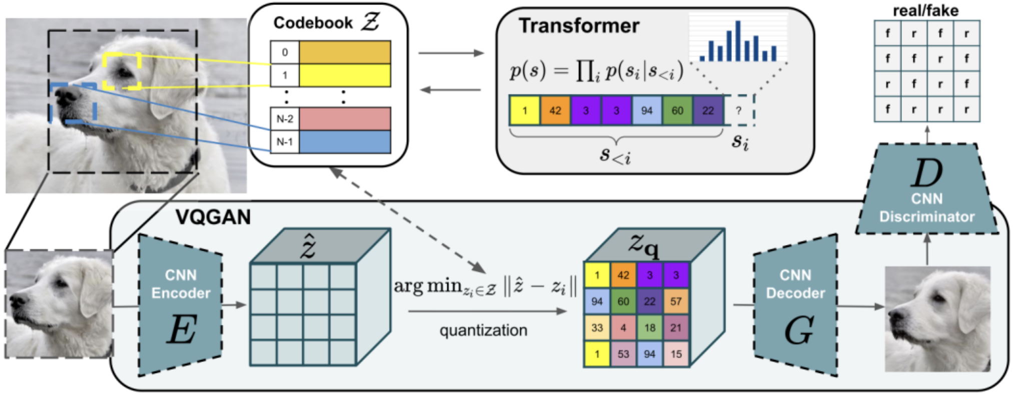
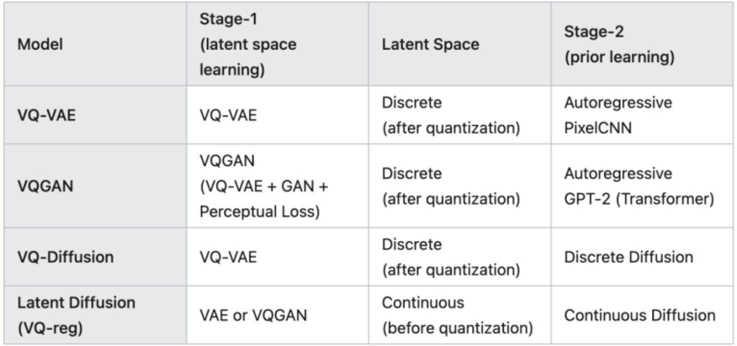


Figure 2. Our approach uses a convolutional *VQGAN* to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables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

Thanks!

Any questions?