MultiModal Large Language Models

INTERPRETING AND INCORPORATING INFORMATION FROM VARIOUS MODALITIES SUCH AS IMAGES, AUDIO, OR VIDEO

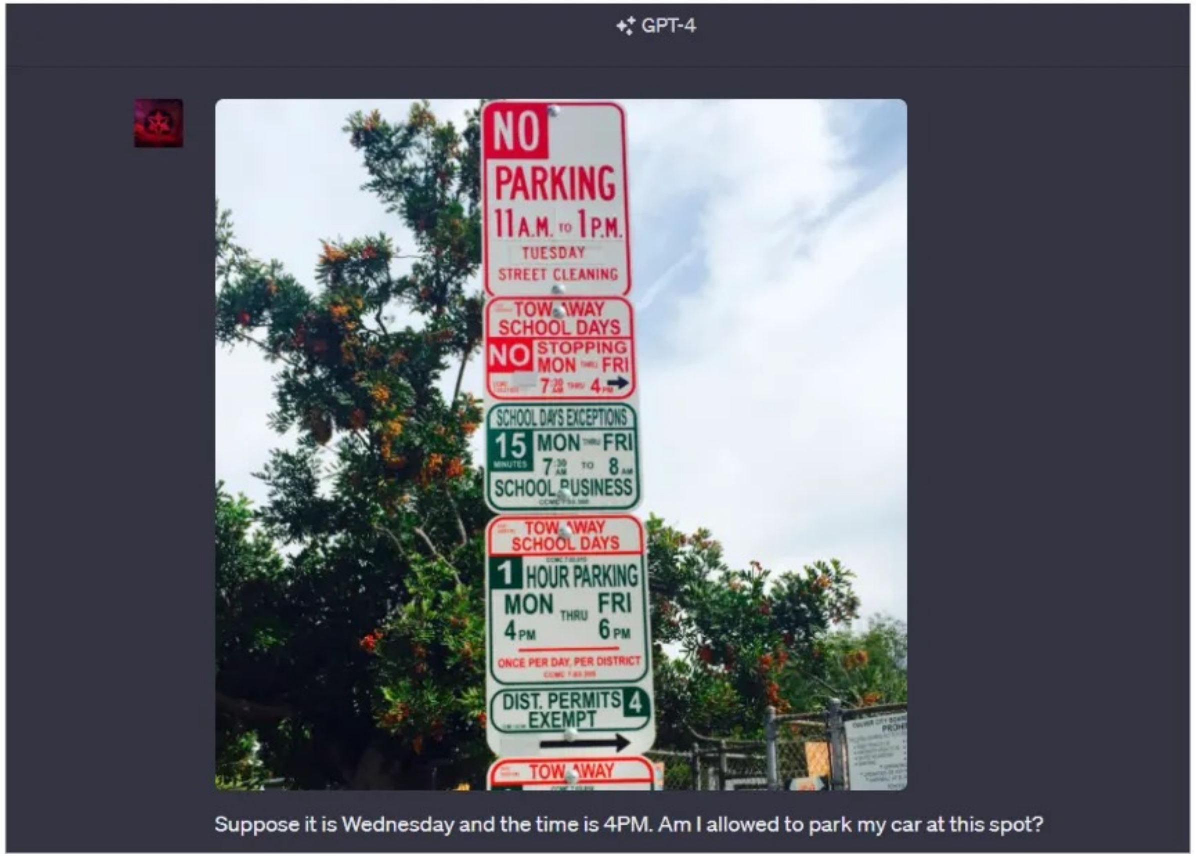
Srikumar Sastry, 02/29/2024
Large Language Models (CSE 561A)

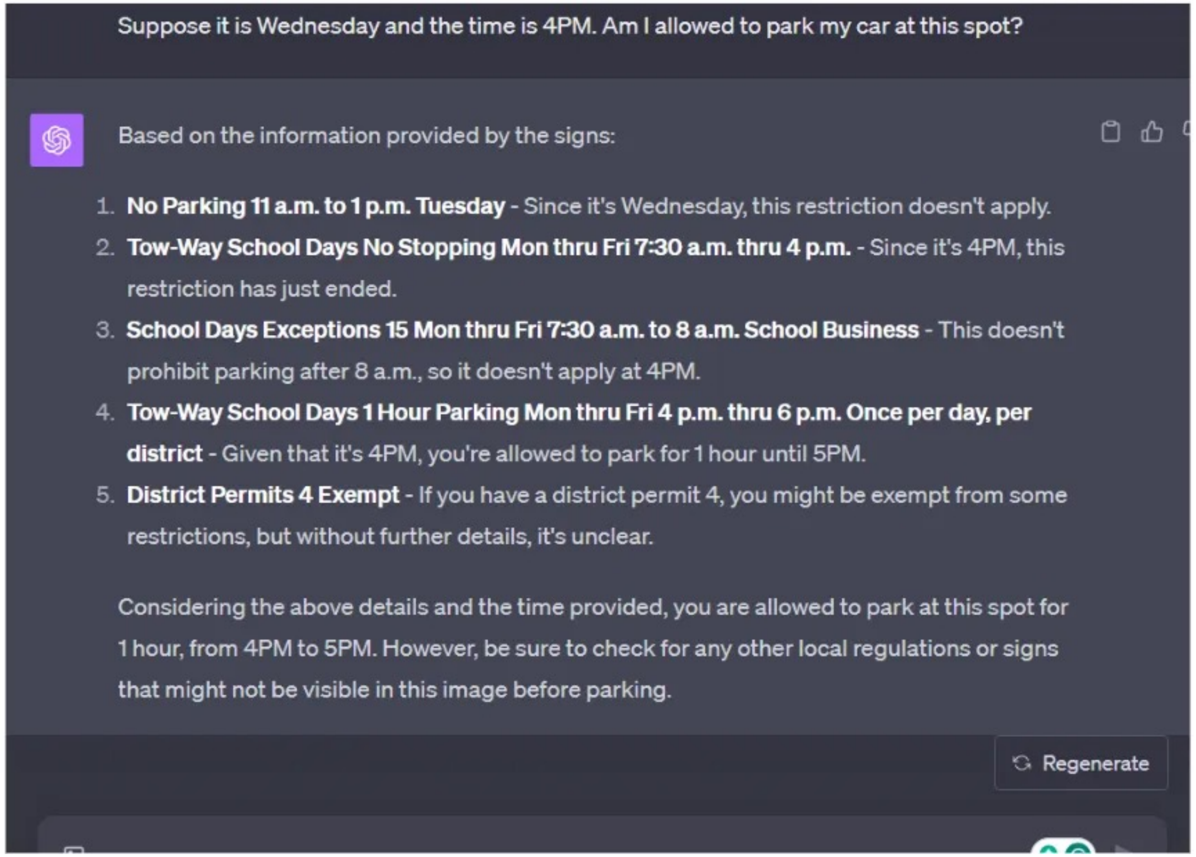
What is MultiModal Large Language Model?

- Multimodal can refer to:
  1. Input and output are of different modalities (e.g. text-to-image, image-to-text).
  2. Inputs are multimodal (e.g. a system that can process both text and images).
  3. Outputs are multimodal (e.g. a system that can generate both text and images)
  4. A combination of all of the above.

- Multimodal Large Language Model (MLLM) is a Large language model that can process different modalities .
- Take them as input or provide them as output after processing.

GPT-4V example





Suppose it is Wednesday and the time is 4PM. Am I allowed to park my car at this spot?

Based on the information provided by the signs:

1. **No Parking 11 a.m. to 1 p.m. Tuesday** - Since it's Wednesday, this restriction doesn't apply.
2. **Tow-Way School Days No Stopping Mon thru Fri 7:30 a.m. thru 4 p.m.** - Since it's 4PM, this restriction has just ended.
3. **School Days Exceptions 15 Mon thru Fri 7:30 a.m. to 8 a.m. School Business** - This doesn't prohibit parking after 8 a.m., so it doesn't apply at 4PM.
4. **Tow-Way School Days 1 Hour Parking Mon thru Fri 4 p.m. thru 6 p.m. Once per day, per district** - Given that it's 4PM, you're allowed to park for 1 hour until 5PM.
5. **District Permits 4 Exempt** - If you have a district permit 4, you might be exempt from some restrictions, but without further details, it's unclear.

Considering the above details and the time provided, you are allowed to park at this spot for 1 hour, from 4PM to 5PM. However, be sure to check for any other local regulations or signs that might not be visible in this image before parking.

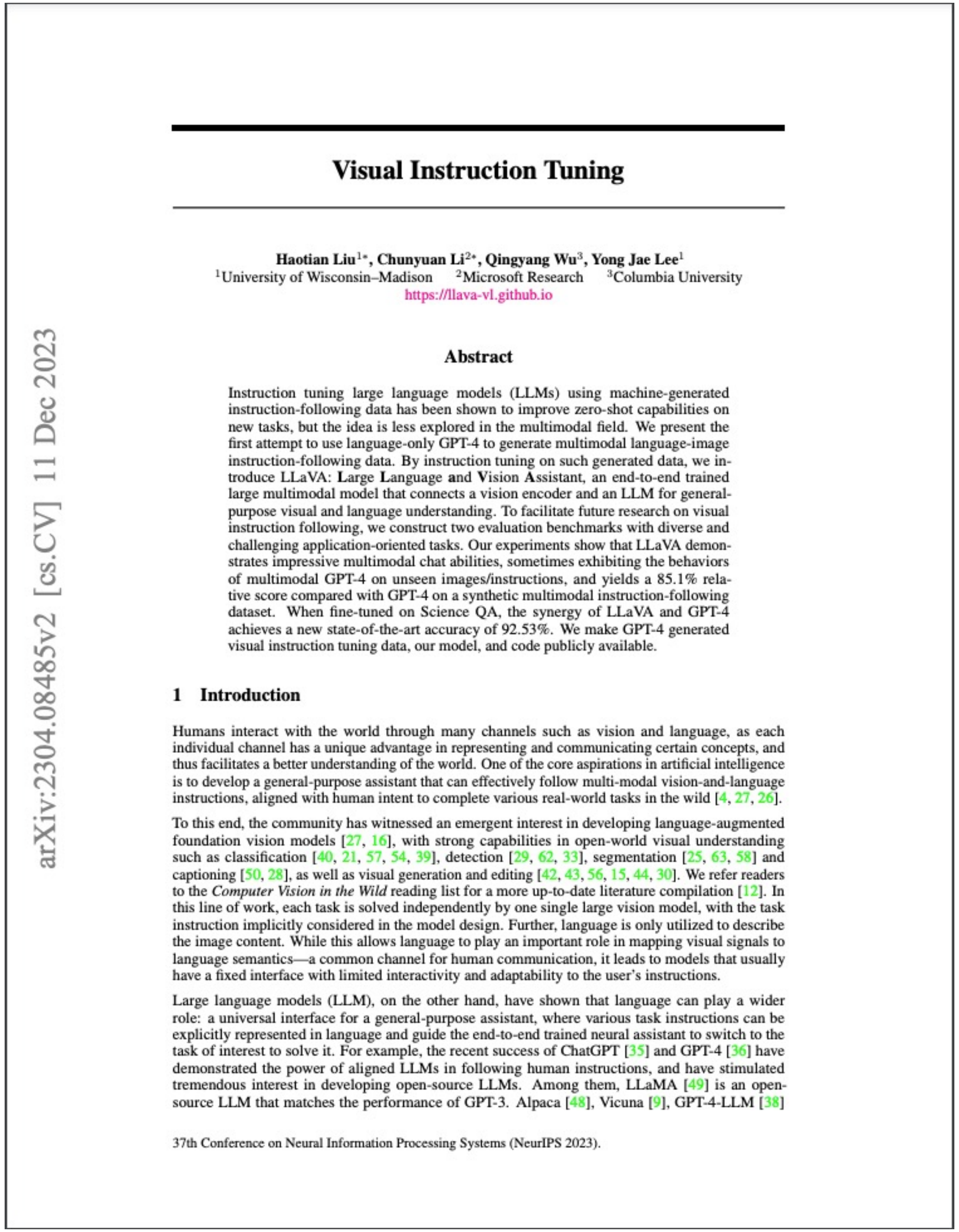
Regenerate

State of the Art

Features/Aspect	GPT-4V	Macaw-LLM	LLaVA	NExT-GPT	CogVLM
Base Model	Transformer-based decoder with 1.7T parameters	LLaMA/Vicuna/Bloom	LLaMA/Vicuna (with modifictaions)	Vicuna	a vision transformer (ViT) encoder, an MLP adapter, a pretrained large language model (GPT), and a visual expert module
Training Data	Not disclosed by OpenAI	Stanford Alpaca Dataset for text data, COCO VQA Dataset for Image data, Charades and Video Dialog dataset for Video Data	Enhanced with academic-task-oriented VQA data and response formatting prompts	LLaVA dataset for images, Alpaca for text and VideoChat for video.	English image-text data from the MiniGPT-4, LLAVA, LRV-Instruction, LLaVAR and Shikra projects, as well as many classic cross-modal work datasets
Vision-Language Connector	Attention mechanism, a neural network technique that allows the model to focus on specific parts of an input sequence or image, improving performance on tasts like image captioning.	Attention function, wherein the multi-modal features serve as the query and the embedding matrix of LLaMA as the key and value.	Fully connected video-language cross-modal connector, which is powerful and data-efficient	LLM-centric Multimodal Alignment	MLP Adapter concatenated with word embeddings from text prompt
Vision Encoder	Vision transformer, converts images into a sequence of tokens to generate text, translate languages, or answer questions.	CLIP	CLIP-VIT-L/14@336px with an MLP projection	ImageBind	ViT Encoder similar to GPT-4V
Performance	Outperforms LLaVA 1.5 in image interpretation, text recognition, and code generation	Video instruction data: Macaw-LLM video instruction dataset	Achieved state-of-the-art across 11 benchmarks.	No information provided	CogVLM-17B achieves state-of-the-art performance on 10 classic cross-modal benchmarks
Training Efficiency	Not publicly disclosed. Likely less efficient than others	No information provided on the efficiency	Final 13B checkpoint uses only 1.2M publicly available data and finishes training in ~1 day on a single 8-A100 node	Encoding-side alignment, 3x GPUs, batch size 18≈30 mins on 40k instances of text-X pairs. Decoding-side alignment, 3x GPUs, batch size 18≈3 hours on 180k instances of text-X pairs.	No information provided on the efficiency
Availability	Closed-Source	Open-Source	Open-Source	Open-Source	Open-Source

Visual Instruction tuning

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2023).



arXiv:2304.08485v2 [cs.CV] 11 Dec 2023

Visual Instruction Tuning

Haotian Liu[1*], **Chunyuan Li**[2*], **Qingyang Wu**[3], **Yong Jae Lee**[1]
[1]University of Wisconsin–Madison [2]Microsoft Research [3]Columbia University
https://llava-vl.github.io

Abstract

Instruction tuning large language models (LLMs) using machine-generated instruction-following data has been shown to improve zero-shot capabilities on new tasks, but the idea is less explored in the multimodal field. We present the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data. By instruction tuning on such generated data, we introduce LLaVA: **L**arge **L**anguage **a**nd **V**ision **A**ssistant, an end-to-end trained large multimodal model that connects a vision encoder and an LLM for general-purpose visual and language understanding. To facilitate future research on visual instruction following, we construct two evaluation benchmarks with diverse and challenging application-oriented tasks. Our experiments show that LLaVA demonstrates impressive multimodal chat abilities, sometimes exhibiting the behaviors of multimodal GPT-4 on unseen images/instructions, and yields a 85.1% relative score compared with GPT-4 on a synthetic multimodal instruction-following dataset. When fine-tuned on Science QA, the synergy of LLaVA and GPT-4 achieves a new state-of-the-art accuracy of 92.53%. We make GPT-4 generated visual instruction tuning data, our model, and code publicly available.

1 Introduction

Humans interact with the world through many channels such as vision and language, as each individual channel has a unique advantage in representing and communicating certain concepts, and thus facilitates a better understanding of the world. One of the core aspirations in artificial intelligence is to develop a general-purpose assistant that can effectively follow multi-modal vision-and-language instructions, aligned with human intent to complete various real-world tasks in the wild [4, 27, 26].

To this end, the community has witnessed an emergent interest in developing language-augmented foundation vision models [27, 16], with strong capabilities in open-world visual understanding such as classification [40, 21, 57, 54, 39], detection [29, 62, 33], segmentation [25, 63, 58] and captioning [50, 28], as well as visual generation and editing [42, 43, 56, 15, 44, 30]. We refer readers to the *Computer Vision in the Wild* reading list for a more up-to-date literature compilation [12]. In this line of work, each task is solved independently by one single large vision model, with the task instruction implicitly considered in the model design. Further, language is only utilized to describe the image content. While this allows language to play an important role in mapping visual signals to language semantics—a common channel for human communication, it leads to models that usually have a fixed interface with limited interactivity and adaptability to the user's instructions.

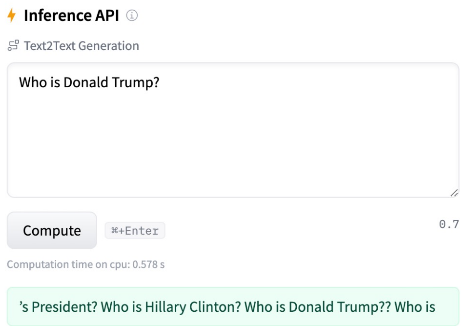
Large language models (LLM), on the other hand, have shown that language can play a wider role: a universal interface for a general-purpose assistant, where various task instructions can be explicitly represented in language and guide the end-to-end trained neural assistant to switch to the task of interest to solve it. For example, the recent success of ChatGPT [35] and GPT-4 [36] have demonstrated the power of aligned LLMs in following human instructions, and have stimulated tremendous interest in developing open-source LLMs. Among them, LLaMA [49] is an open-source LLM that matches the performance of GPT-3. Alpaca [48], Vicuna [9], GPT-4-LLM [38]

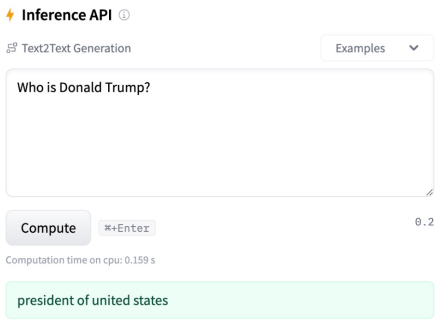
37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Instruction Tuning

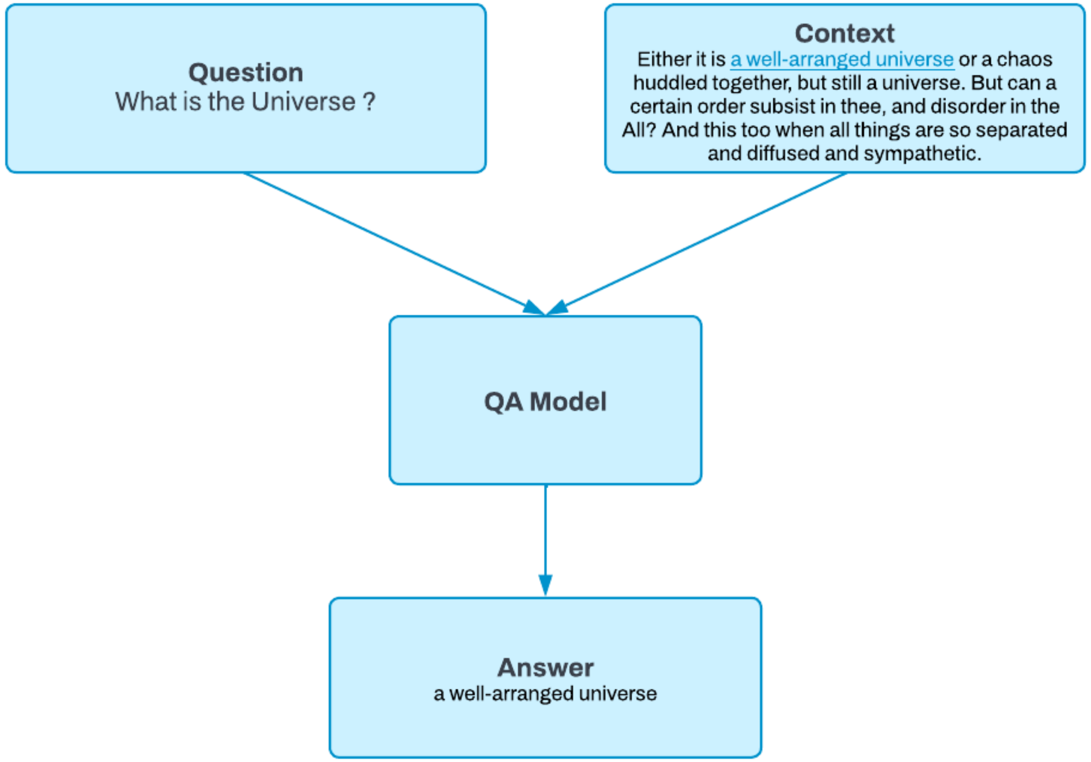
Pre-Training Features	Fine-Tuning Features
Broad language understanding	Task-specific adaptation
Large, diverse dataset training	Smaller, targeted dataset training
General knowledge base development	Rapid specialization
Facilitates transfer learning	Quick learning from few examples
High initial computational cost	Lower computational cost
Scalable with continual learning	Customizable to current data
Sets performance benchmarks	Enhances specific task performance
Flexible across various applications	Efficient for niche applications

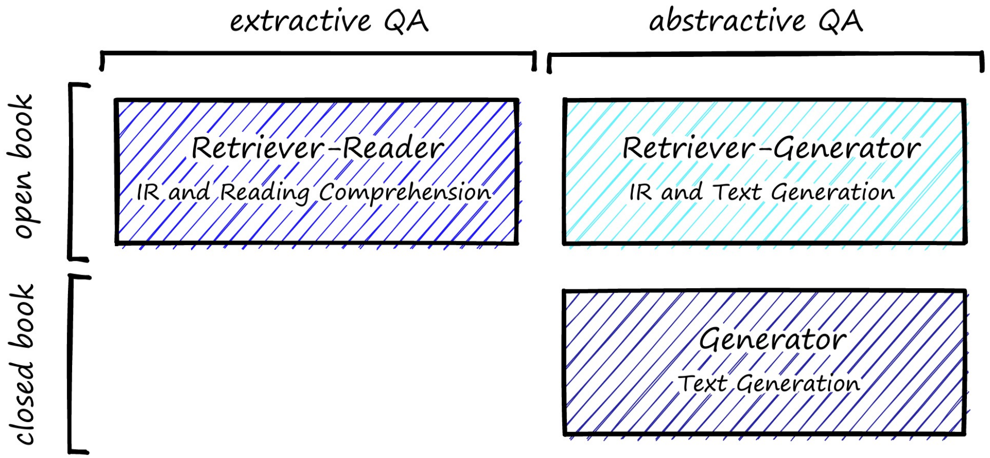
- Pre-trained LLM learn a generic representation.
- Might not answer specific queries a user has.
- Finetuning LLM with input-output pairs.





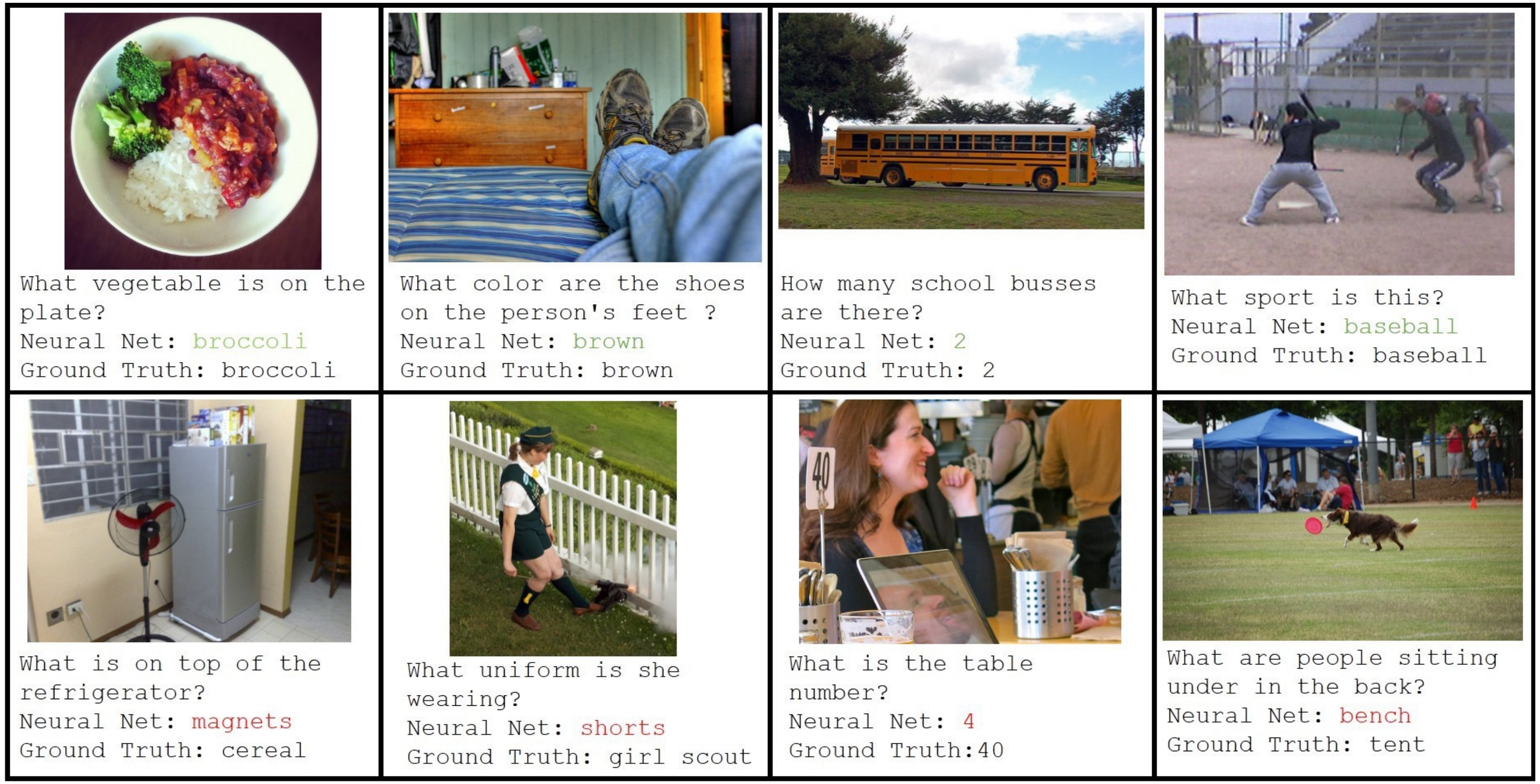
Instruction Tuning





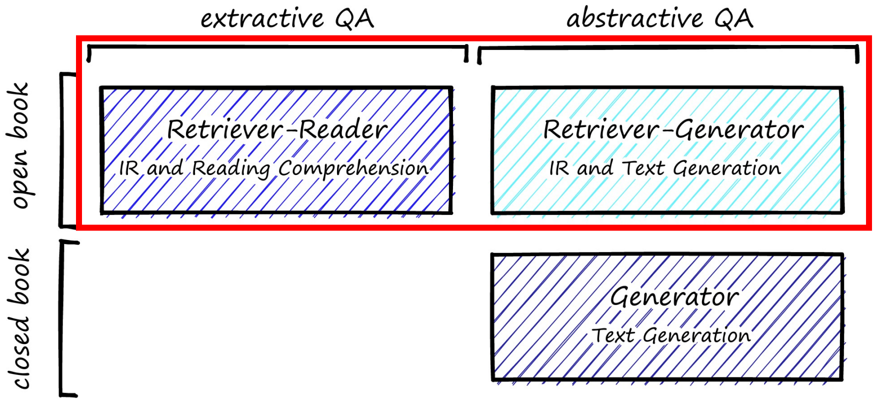
Visual Instruction tuning

- Visual instruction tuning to training an LLM to solve Visual Question Answering.

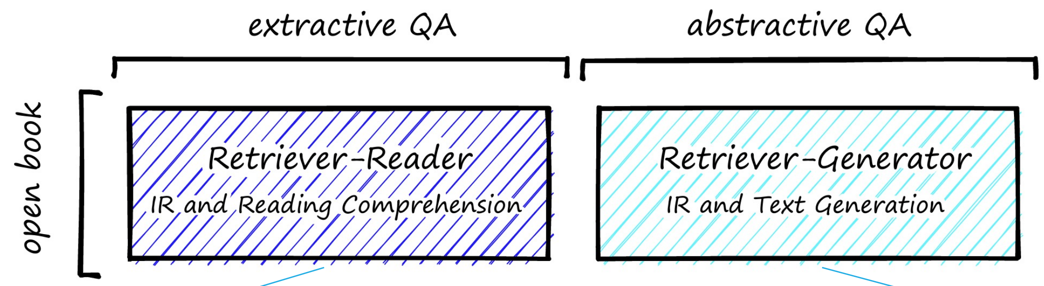


Visual Instruction tuning

- Visual Question Answering falls in the domain of open book Question Answering.
- Image or video act as the context and LLM has to reason based on that.
- Question can be open-ended (answer not present in image) or closed-domain (answer is present in the image).



Visual Instruction tuning



Contributions

- Multimodal instruction-following data – Creation of a large-scale image-text dataset in instruction-following format using GPT-4.
- Large multimodal models – First paper to build a LLM for visual question answering.
- Multimodal instruction-following benchmark – They present LLaVA-bench, that contains two benchmarks to evaluate LLMs on visual-language task.
- Open-source – Model, datasets and code is open-sourced.

Vision-Text Data Generation

Use GPT-4 to compile a dataset of question and answers

$\mathbf{X}_{\mathrm{v}}$ - Image $\mathbf{X}_{\mathrm{c}}$ - Caption $\mathbf{X}_{\mathrm{q}}$ - Question

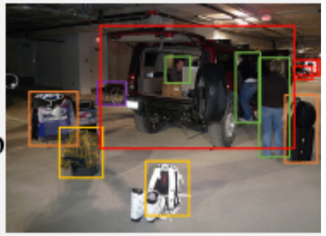
$$\texttt{Human}: \mathbf{X}_{\mathrm{q}}\ \mathbf{X}_{\mathrm{v}}\texttt{<STOP>}\ \texttt{Assistant}: \mathbf{X}_{\mathrm{c}}\texttt{<STOP>}$$

This setup would lack diversity and in-depth reasoning.

1. Multi-Turn Conversation using GPT-4.
2. Created a list of question for extracting detailed description from the image.
3. Create in-depth reasoning questions

Vision-Text Data Generation

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

- Two context types:
  1. Captions
  2. Bounding Boxes
- Compiled the data set using in-context learning, i.e. few shot learning.
- The first few examples are generated by human annotators.
- Total Dataset : 158k instructions
  - 58k – conversations
  - 23k – detailed description
  - 77k - complex reasoning

Architecture

- Simple architecture. Use Pretrained Vicuna.
- Two stage training framework:
  1. Visual feature alignment with the LLM.
  2. Fine-tuning End-to-End.
- First stage: Freeze everything except the visual projection layer W.
- Second stage: Freeze only the vision encoder and train end-to-end.

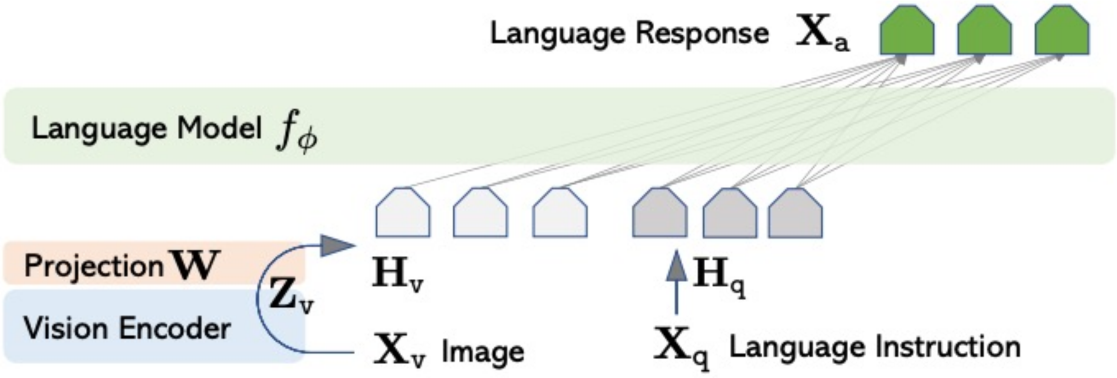


Figure 1: LLaVA network architecture.

Training: First Stage – Multimodal projection alignment

- Filter CC3M to 595K image-text pairs.
- Create single turn conversation using GPT-4.
- Train only the projection layer on next token prediction task.

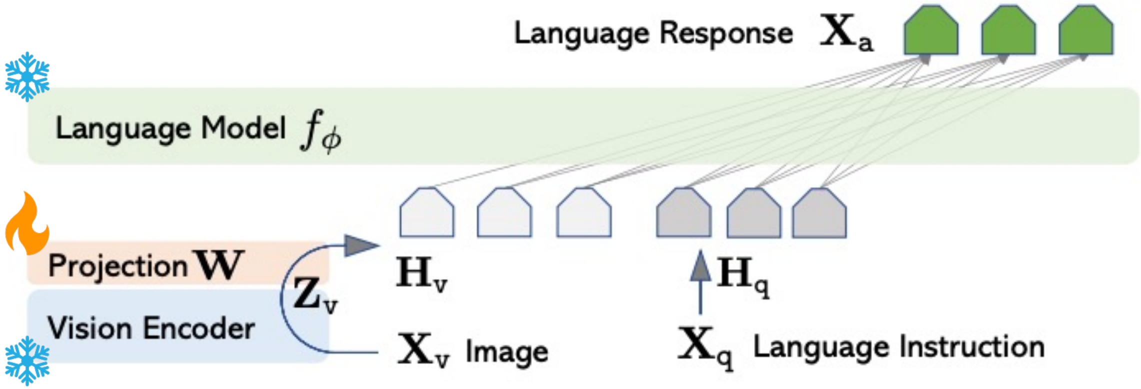


Figure 1: LLaVA network architecture.

Training: Second Stage – end-to-end finetuning

- Created multi-turn conversation data.

$$(\mathbf{X}_q^1, \mathbf{X}_a^1, \cdots, \mathbf{X}_q^T, \mathbf{X}_a^T)$$

$$\mathbf{X}_{\text{instruct}}^t = \begin{cases} \text{Randomly choose } [\mathbf{X}_q^1, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}_q^1], & \text{the first turn } t = 1 \\ \mathbf{X}_q^t, & \text{the remaining turns } t > 1 \end{cases}$$

- Training LLM to maximize likelihood:

$$p(\mathbf{X}_a | \mathbf{X}_v, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i | \mathbf{X}_v, \mathbf{X}_{\text{instruct},<i}, \mathbf{X}_{a,<i})$$

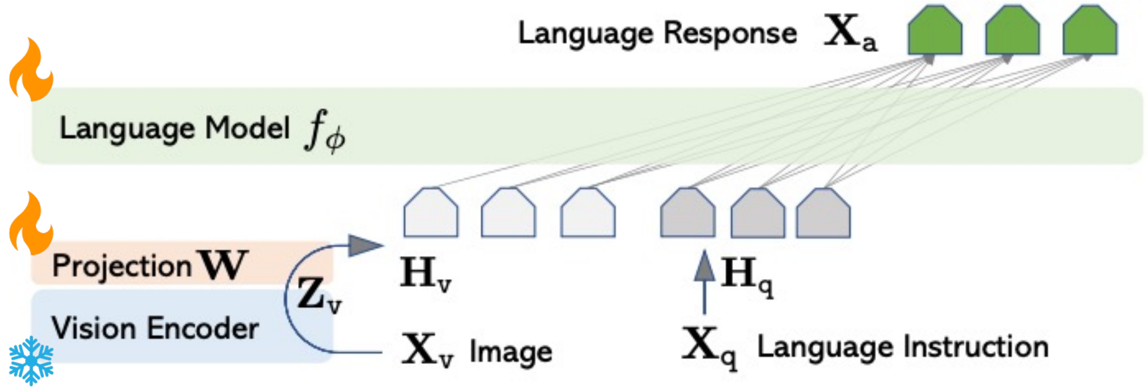


Figure 1: LLaVA network architecture.

Example

Visual input example, Extreme Ironing:

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [36]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Table 3: Example prompt from GPT-4 paper [36] to compare visual reasoning and chat capabilities. Compared to BLIP-2 [28] and OpenFlamingo [5], LLaVA accurately follows the user's instructions, instead of simply describing the scene. LLaVA offers a more comprehensive response than GPT-4. Even when merely asked to describe the image, LLaVA identifies atypical aspects of the image.

Evaluation

- Create 2 benchmarks:
  - LLaVA-Bench (COCO) – 30 images from COCO dataset (val) and generated a instruction following dataset. Total 90 questions
  - LLaVA-Bench (in-the-Wild) – More complex reasoning. 24 images with 60 questions.

- For scoring, use GPT-4 text-only model to rate the responses from the chatbot.

Results

	Conversation	Detail description	Complex reasoning	**All**
Full data	83.1	75.3	96.5	85.1
Detail + Complex	81.5 (-1.6)	73.3 (-2.0)	90.8 (-5.7)	81.9 (-3.2)
Conv + 5% Detail + 10% Complex	81.0 (-2.1)	68.4 (-7.1)	91.5 (-5.0)	80.5 (-4.4)
Conversation	76.5 (-6.6)	59.8 (-16.2)	84.9 (-12.4)	73.8 (-11.3)
No Instruction Tuning	22.0 (-61.1)	24.0 (-51.3)	18.5 (-78.0)	21.5 (-63.6)

Table 4: Ablation on LLaVA-Bench (COCO) with different training data. We report relative scores *w.r.t.* a text-only GPT-4 model that uses ground truth image captions and bounding boxes as visual input. We prompt GPT-4 with the answers from our model outputs and the answers by GPT-4 (text-only), and let it compare between both responses and give a rating with an explanation.

Results

	Conversation	Detail description	Complex reasoning	All
OpenFlamingo [5]	19.3 ± 0.5	19.0 ± 0.5	19.1 ± 0.7	19.1 ± 0.4
BLIP-2 [28]	54.6 ± 1.4	29.1 ± 1.2	32.9 ± 0.7	38.1 ± 1.0
LLaVA	57.3 ± 1.9	52.5 ± 6.3	81.7 ± 1.8	67.3 ± 2.0
LLaVA†	58.8 ± 0.6	49.2 ± 0.8	81.4 ± 0.3	66.7 ± 0.3

Table 5: Instruction-following capability comparison using relative scores on LLaVA-Bench (In-the-Wild). The results are reported in the format of *mean* $\pm$ *std*. For the first three rows, we report three inference runs. LLaVA performs significantly better than others. † For a given set of LLaVA decoding sequences, we evaluate by querying GPT-4 three times; GPT-4 gives a consistent evaluation.

Results

- ScienceQA dataset
- Contains 21k multiple choice questions.

Method	Subject			Context Modality			Grade		Average
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Representative & SoTA methods with numbers reported in the literature									
Human [34]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [34]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [34]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [59]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [61]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [61]	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
Results with our own experiment runs									
GPT-4†	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
LLaVA	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVA+GPT-4† (complement)	90.36	95.50	88.55	89.05	87.80	91.08	92.22	88.73	90.97
LLaVA+GPT-4† (judge)	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	**92.53**

Table 7: Accuracy (%) on Science QA dataset. Question categories: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. †Text-only GPT-4, our eval. Our novel model ensembling with the text-only GPT-4 consistently improves the model's performance under all categories, setting the new SoTA performance.

Limitations

- Some inconsistency in interpreting visual information.
  - Responds with yes when asked if strawberry-flavored yogurt is present, even though the fridge contains only yogurt and strawberries.
- Vicuna-7B might not be the best. Could be switched with larger models.
- Could be expanded to multilingual understanding.

NeXT-GPT

Wu, Shengqiong, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. "Next-gpt: Any-to-any multimodal llm." *arXiv preprint arXiv:2309.05519* (2023).

NExT-GPT: Any-to-Any Multimodal LLM

Shengqiong Wu Hao Fei* Leigang Qu Wei Ji Tat-Seng Chua
NExT++, School of Computing, National University of Singapore

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

arXiv:2309.05519v2 [cs.AI] 13 Sep 2023

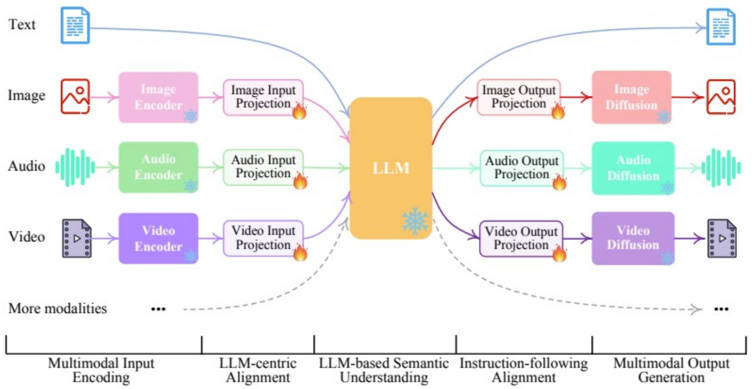


Figure 1: By connecting LLM with multimodal adaptors and diffusion decoders, NExT-GPT achieves universal multimodal understanding and any-to-any modality input and output.

Abstract

While recently Multimodal Large Language Models (MM-LLMs) have made exciting strides, they mostly fall prey to the limitation of only input-side multimodal understanding, without the ability to produce content in multiple modalities. As we humans always perceive the world and communicate with people through various modalities, developing any-to-any MM-LLMs capable of accepting and delivering content in any modality becomes essential to human-level AI. To fill the gap, we present an end-to-end general-purpose any-to-any MM-LLM system, **NExT-GPT**. We connect an LLM with multimodal adaptors and different diffusion decoders, enabling NExT-GPT to perceive inputs and generate outputs in arbitrary combinations of text, images, videos, and audio. By leveraging the existing well-trained highly-performing encoders and decoders, NExT-GPT is tuned with only a small amount of parameter (1%) of certain projection layers, which not only benefits low-cost training and also facilitates convenient expansion to more potential modalities. Moreover, we introduce a modality-switching instruction tuning (MosIT) and manually curate a high-quality dataset for MosIT, based on which NExT-GPT is empowered with complex cross-modal semantic understanding and content generation. Overall, our research showcases the promising possibility of building a unified AI agent capable of modeling universal modalities, paving the way for more human-like AI research in the community.

*Hao Fei is the corresponding author: haofei37@nus.edu.sg

Contributions

- First work on any-to-any MLLM: text, audio, image, videos as input and/or output.
- Lightweight alignment technique, requiring only 1% parameters.
- Present a multi modality instruction following dataset.

Architecture

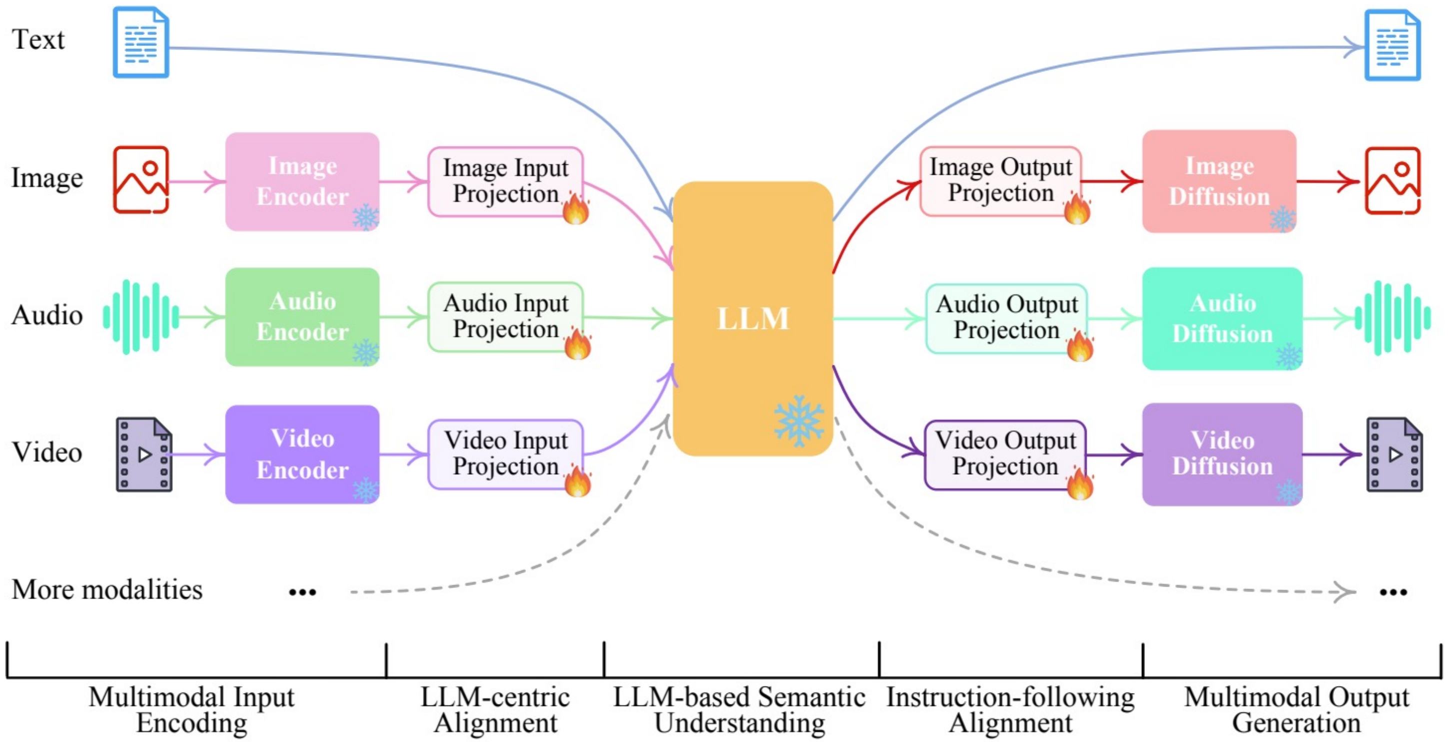


Figure 1: By connecting LLM with multimodal adaptors and diffusion decoders, NExT-GPT achieves universal multimodal understanding and any-to-any modality input and output.

	Encoder		Input Projection		LLM		Output Projection		Diffusion	
	Name	Param	Name	Param	Name	Param	Name	Param	Name	Param
Text	—	—	—	—			—	—	—	—
Image					Vicuna [12]	7B❄	Transformer	31M🔥	SD [68]	1.3B❄
Audio	ImageBind [25]	1.2B❄	Linear	4M🔥	(LoRA	33M🔥)	Transformer	31M🔥	AudioLDM [51]	975M❄
Video							Transformer	32M🔥	Zeroscope [8]	1.8B❄

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

Architecture Details

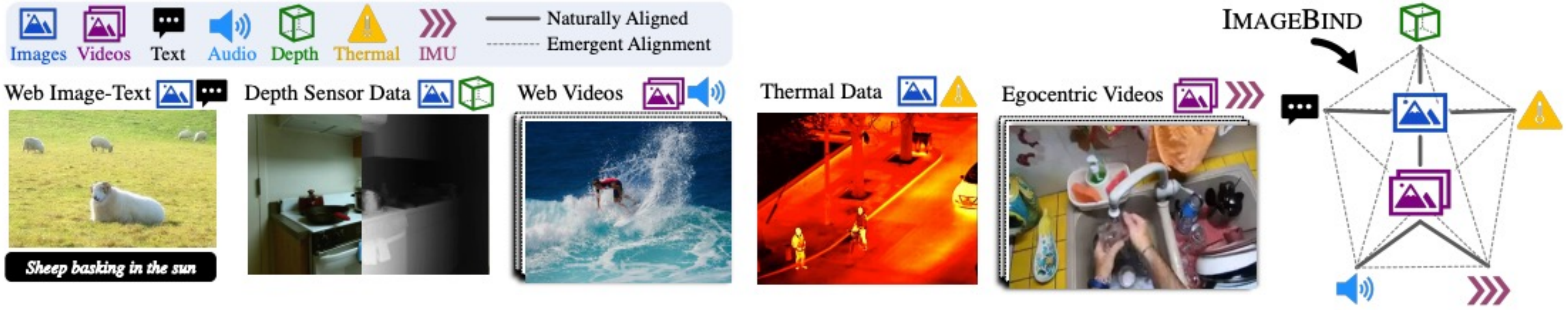


Figure 2. IMAGEBIND overview. Different modalities occur naturally aligned in different data sources, for instance images+text and video+audio in web data, depth or thermal information with images, IMU data in videos captured with egocentric cameras, *etc*. IMAGEBIND links all these modalities in a common embedding space, enabling new emergent alignments and capabilities.

ImageBind

Girdhar, Rohit, et al. "Imagebind: One embedding space to bind them all." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Stage 1: Multi Modal Projection Alignment

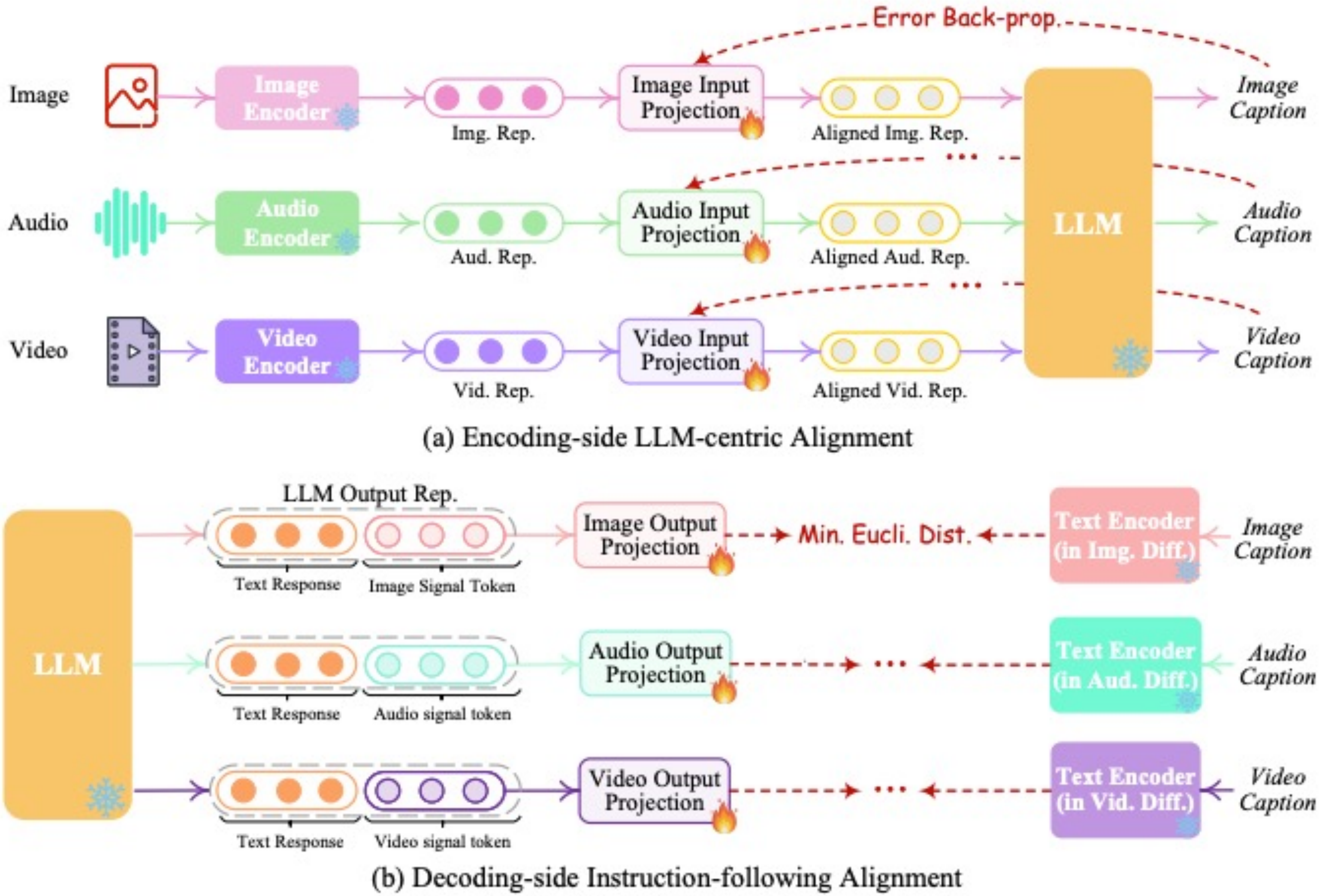


Figure 3: Illustration of the lightweight multimodal alignment learning of encoding and decoding.

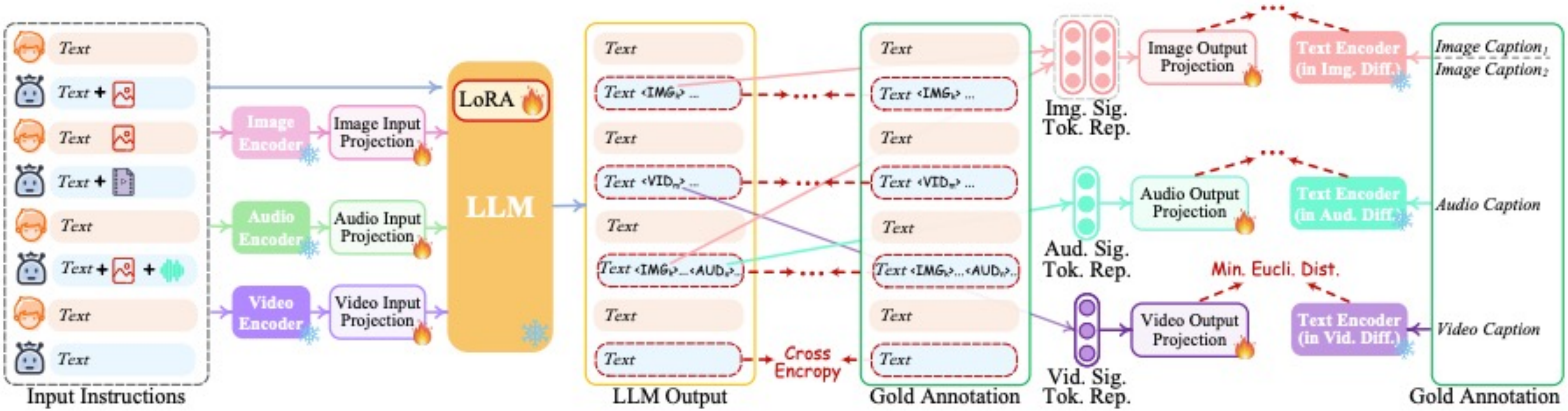


Figure 4: Illustration of modality-switching instruction tuning.

Stage 2: Instruction Tuning

Dataset	Data Source	In→Out Modality	Approach	Multi-turn Reason	#Img/Vid/Aud	#Dialog Turn.	#Instance
► *Existing data*							
MiniGPT-4 [109]	CC [10], CC3M [71]	T+I→T	Auto	✗	134M/-/-	1	5K
StableLLaVA [47]	SD [68]	T+I→T	Auto+Manu.	✗	126K/-/-	1	126K
LLaVA [104]	COCO [50]	T+I→T	Auto	✓	81K/-/-	2.29	150K
SVIT [106]	MS-COCO [50], VG [41]	T+I→T	Auto	✓	108K/-/-	5	3.2M
LLaVAR [104]	COCO [50], CC3M [71], LAION [70]	T+I→T	LLaVA+Auto	✓	20K/-/-	2.27	174K
VideoChat [44]	WebVid [5]	T+V→T	Auto	✓	-/8K/-	1.82	11K
Video-ChatGPT [54]	ActivityNet [28]	T+V→T	Inherit	✗	-/100K/-	1	100K
Video-LLaMA [103]	MiniGPT-4, LLaVA, VideoChat	T+I/V→T	Auto	✓	81K/8K/-	2.22	171K
InstructBLIP [15]	Multiple	T+I/V→T	Auto	✗	-	-	~ 1.6M
MIMIC-IT [42]	Multiple	T+I/V→T	Auto	✗	8.1M/502K/-	1	2.8M
PandaGPT [77]	MiniGPT-4, LLaVA	T+I→T	Inherit	✓	81K/-/-	2.29	160K
MGVLID [107]	Multiple	T+I+B→T	Auto+Manu.	✗	108K/-/-	-	108K
M^{3}IT [45]	Multiple	T+I/V/B→T	Auto+Manu.	✗	-/-/-	1	2.4M
LAMM [97]	Multiple	T+I+PC→T	Auto+Manu.	✓	91K/-/-	3.27	196k
BuboGPT [108]	Clotho [20], VGGSS [11]	T+A/(I+A)→T	Auto	✗	5k/-/9K	-	9K
mPLUG-DocOwl [96]	Multiple	T+I/Tab/Web→T	Inherit	✗	-	-	-
► *In this work*							
T2M	Webvid [5], CC3M [71], AudioCap [38]	T→T+I/A/V	Auto	✗	4.9K/4.9K/4.9K	1	14.7K
MosIT	Youtube, Google, Flickr, Midjourney, etc.	T+I+A+V→T+I+A+V	Auto+Manu.	✓	4K/4K/4K	4.8	5K

Table 2: Summary and comparison of existing datasets for multimodal instruction tuning. T: text, I: image, V: video, A: audio, B: bounding box, PC: point cloud, Tab: table, Web: web page.

Dataset Compilation

Results

Method	FID (↓)
CogVideo [17]	27.10
GLIDE [58]	12.24
CoDi [78]	11.26
SD [68]	**11.21**
NExT-GPT	11.28

Table 3: Text-to-image generation results on COCO-caption data [50].

Method	FD (↓)	IS (↑)
DiffSound [95]	47.68	4.01
AudioLDM-S [51]	29.48	6.90
AudioLDM-L [51]	23.31	8.13
CoDi [78]	**22.90**	**8.77**
NExT-GPT	23.58	8.35

Table 4: Text-to-audio generation results on AudioCaps [38].

Method	FID (↓)	CLIPSIM (↑)
CogVideo [30]	23.59	0.2631
MakeVideo [74]	13.17	0.3049
Latent-VDM [68]	14.25	0.2756
Latent-Shift [2]	15.23	0.2773
CoDi [78]	—	0.2890
NExT-GPT	**13.04**	**0.3085**

Table 5: Text-to-video generation results (zero-shot) on MSR-VTT [92].

Method	B@4	METEOR	CIDEr
Oscar [46]	36.58	30.4	124.12
BLIP-2 [43]	43.7	—	145.8
OFA [86]	44.9	32.5	154.9
CoDi [78]	40.2	31.0	149.9
NExT-GPT	44.3	**32.9**	**156.7**

Table 6: Image-to-text generation (image captioning) results on COCO-caption data [50].

Method	SPIDEr	CIDEr
AudioCaps [38]	0.369	0.593
BART [26]	0.465	0.753
AL-MixGen [39]	0.466	0.755
CoDi [78]	0.480	0.789
NExT-GPT	**0.521**	**0.802**

Table 7: Audio-to-text generation (audio captioning) results on AudioCaps [38].

Method	B@4	METEOR
ORG-TRL [105]	43.6	28.8
GIT [85]	54.8	33.1
mPLUG-2 [91]	57.8	34.9
CoDi [78]	52.1	32.5
NExT-GPT	**58.4**	**38.5**

Table 8: Video-to-text generation (video captioning) results on MSR-VTT [92].

Method	Object		Background	
	CLIP (↑)	FID (↓)	CLIP (↑)	FID (↓)
PTP [29]	30.33	9.58	31.55	13.92
BLDM [4]	29.95	6.14	30.38	20.44
DiffEdit [14]	29.30	**3.78**	26.92	**1.74**
PFB-Diff [36]	**30.81**	5.93	**32.25**	13.77
NExT-GPT	29.31	6.52	27.29	15.20

Table 9: Text+image-to-image generation (text-conditioned image editing) results on COCO data [50].

Method	MCD (↓)
CampNet [87]	0.380
MakeAudio [33]	0.375
AudioLDM-L [51]	0.349
NExT-GPT	**0.302**

Table 10: Text+audio-to-audio generation (text-conditioned speech editing) results on VCTK data [83].

Method	CLIP-T (↑)	CLIP-I (↑)
CogVideo [30]	0.2391	0.9064
TuneVideo [89]	0.2758	0.9240
SDEdit [55]	0.2775	0.8731
Pix2Video [9]	**0.2891**	**0.9767**
NExT-GPT	0.2683	0.9645

Table 11: Text+video-to-video generation (text-conditioned video editing) results on DAVIS data [62].

Examples

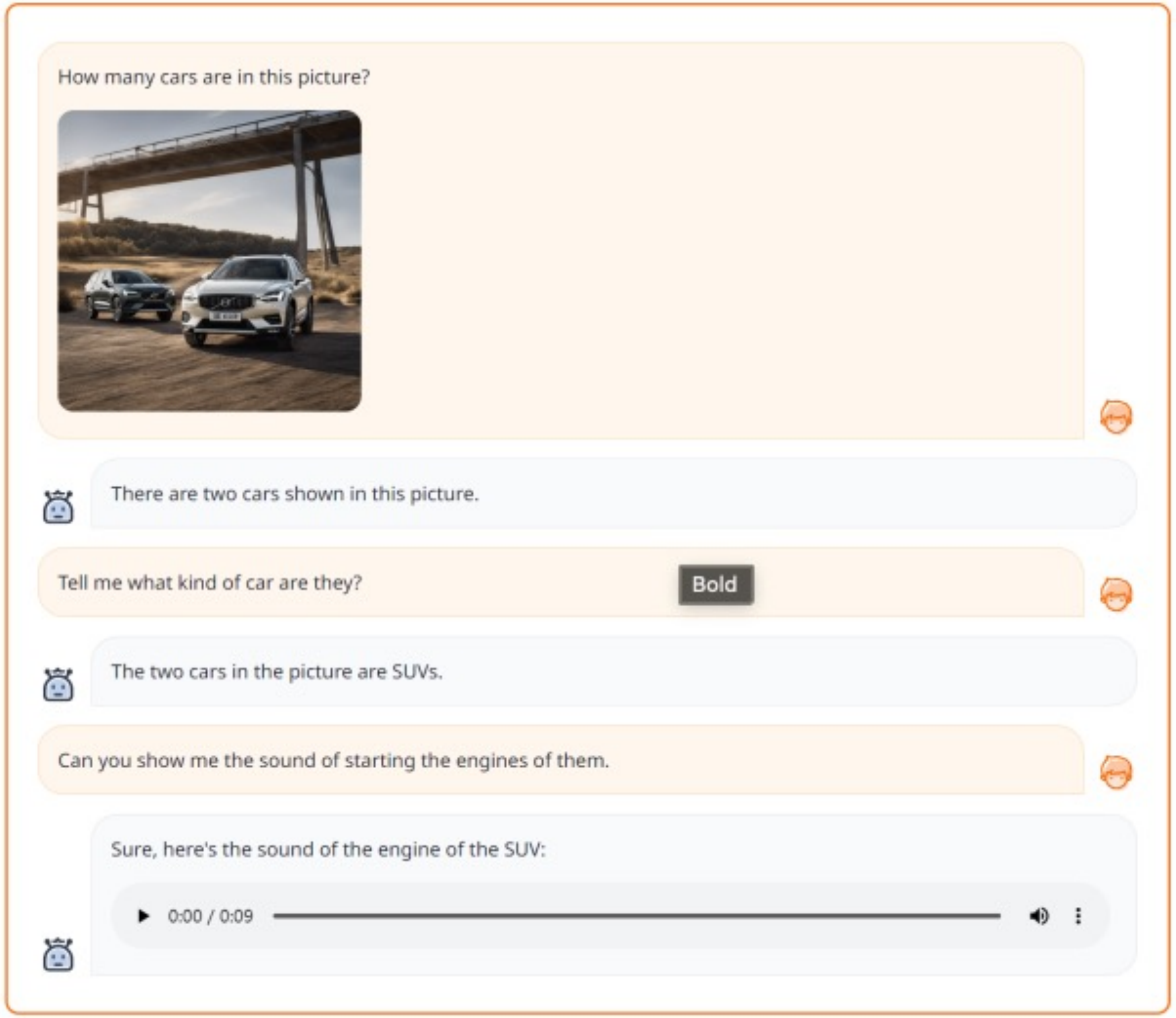


Figure 6: Example of Text+Image → Text+Audio.

Examples

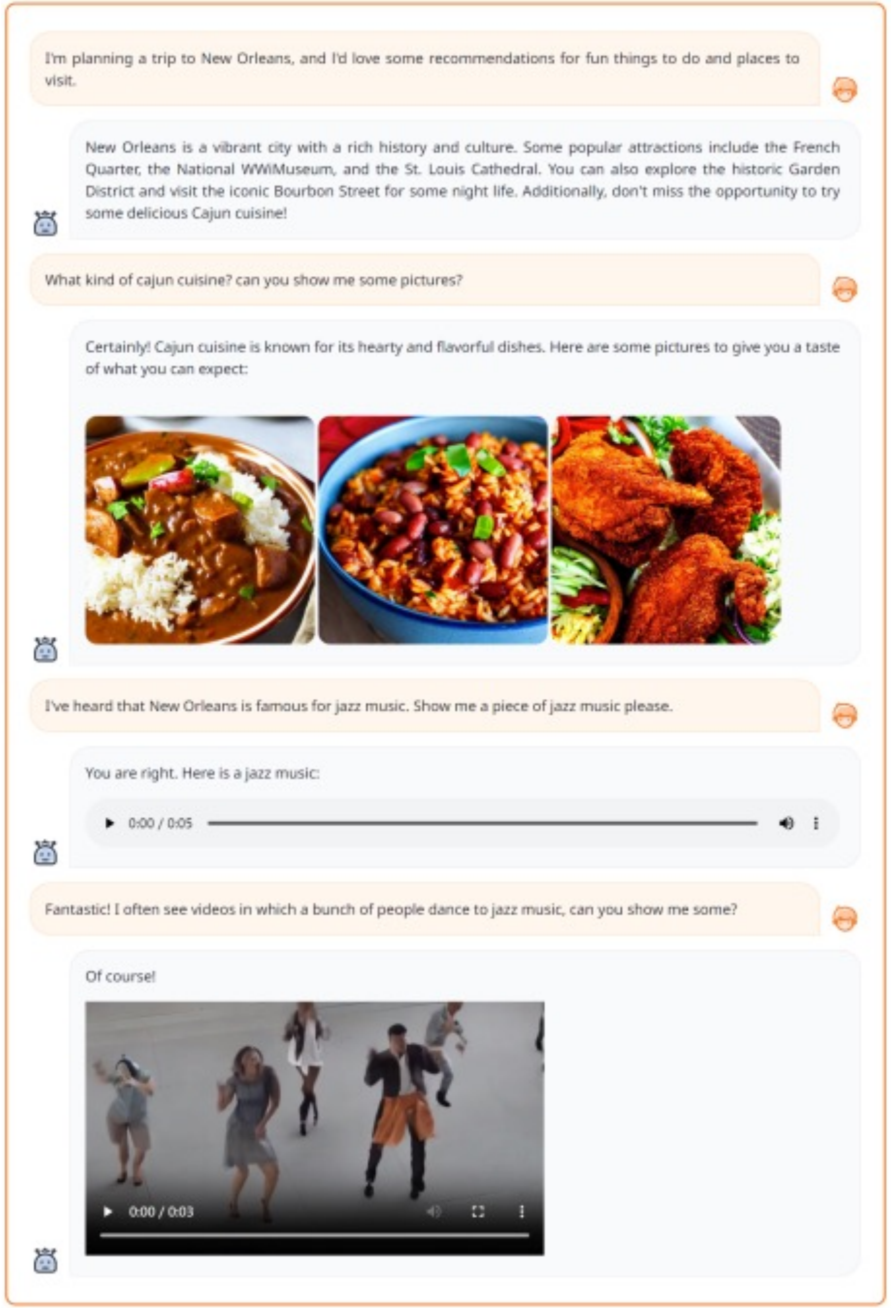


Figure 7: Example of Text → Text+Image+Video+Audio.

Limitations

- Only 4 supported modalities. Future work is to expand to 3D vision, tables etc.
- LLM variants: Use Vicuna-7b which might not be the best.
- Output strategy: Apart from generation, use add retrieval capabilities.
- Dataset is limited and could be expanded.

Thank you

ANY QUESTIONS?