

Multimodal Large Language Models and Tunings

- Vision, Language, Sensors, Audio, and Beyond

Tutorial Presenters:



Soyeon Caren Han*^{1,2}, Feiqi Cao*², Josiah Poon², Roberto Navigli³

1. The University of Melbourne, 2. The University of Sydney, 3. Sapienza University of Roma



THE UNIVERSITY OF
MELBOURNE



THE UNIVERSITY OF
SYDNEY



SAPIENZA
UNIVERSITÀ DI ROMA

Tutorial Schedule

Friday 1 November, Day 5 – Workshops and Tutorials

Location: Meeting Room 207

Time	Section	Presenter
14:00 – 14:20	Part 1: Background and Introduction	Caren Han
14:20 – 15:00	Part 2: Foundation of Multimodal Large Language Models	Caren Han
15:00 – 15:30	Part 3: Core Techniques for Multimodal Large Language Models	Caren Han
Afternoon Tea (Q&A Session)		
16:00 – 17:00	Part 4: Multimodal Instruction Tunings	Feiqi Cao
17:00 – 17:20	Part 5: Future Trends and Research Opportunities	Josiah Poon
17:20 – 17:30	Part 6: Q&A and Closing Remarks	Presenters

Tutorial Outline

1. *Introduction and Overview*

- *Welcome and tutorial objectives*
- *Multimodal learning*
- *Importance and Challenges of integrating multiple modalities (vision, language, sensors, audio)*

2. *Foundation of Multimodal Large Language Models (MLLMs)*

- *Multimodal Key Tasks and Benchmark datasets for multimodal tasks*
- *Challenges in Dataset Creation and Annotation for Multimodal Integration*

3. *Core Techniques for Multimodal Large Language Models (MLLMs)*

- *MLLM Architecture*
- *MM capabilities*

4. *Multimodal Instruction Tunings*

- *Basic Instruction Tuning Technique to LVLM/MLLM Instruction Tuning*
- *Multimodal Chain-of-Thought*
- *Efficient Instruction Tuning Strategies*

5. *Future Trends and Research Opportunities*

First Speaker



Dr. Caren Han

*Senior Lecturer (Assoc Prof. in U.S.)
The University of Melbourne*

Research & Programming Experience

- Early Career Researcher Award 2023 (Physics, Math, Computing)
- ACM Early Career Professional 2023
- NLP Top-tier Conferences Best Paper Award/Best Area Paper Award
- Multimodal Research Grants and Projects with Google, Microsoft, Facebook, NASA, U.S. Air Force, U.S. Navy, Samsung, Hyundai, Australia Defence Department etc.

Award

- Australian Young Achiever Teaching Excellence Award 2018
- Teacher of the Year 2020 Award
- Supervisor of the Year 2021 Award
- Dean's Outstanding Teaching Award 2018-2021



Part 1: Background and Introduction

Part 2: Foundation of Multimodal Large Language Models

Part 3: Core Techniques for Multimodal Large Language Models

C What are the top tier conferences in NLP?

In the field of Natural Language Processing (NLP), there are several top-tier conferences that are considered to be among the most prestigious and influential. Here are a few of the top NLP conferences:

1. Conference on Computational Linguistics (COLING): This is one of the largest NLP conferences and is held every two years. It is organized by the International Committee on Computational Linguistics (ICCL) and is a top venue for presenting new research in NLP.
2. Association for Computational Linguistics (ACL) Conference: This is another highly regarded NLP conference and is held annually. It is organized by the Association for Computational Linguistics (ACL) and is a premier venue for NLP research.
3. Empirical Methods in Natural Language Processing (EMNLP): This is an NLP conference that focuses on empirical methods in NLP and is held annually.
4. North American Chapter of the Association for Computational Linguistics (NAACL): Conference: This is a leading NLP conference that is held annually in North America and is organized by the NAACL.
5. International Joint Conference on Artificial Intelligence (IJCAI): This is a top-tier conference in the field of Artificial Intelligence (AI) and includes NLP as one of its main areas of focus.

These conferences are highly competitive, and being accepted to present research at one of these conferences is a significant achievement in the field of NLP.

Another inquiry?

What
might be
yer question,
matey?

OMG, like, how cool is this?

Explain
quantum
physics?
Elementary!

Trippy
question,
dude...

Ayyy,
I'm talkin'
here!

C

Write a program that converts Pounds(lbs) to Kilograms(kg).

Here is a simple program that converts pounds to kilograms:

```
python
# Define the conversion factor
lbs_to_kg = 0.45359237

# Get the weight in pounds from the user
weight_lbs = float(input("Enter the weight in pounds: "))

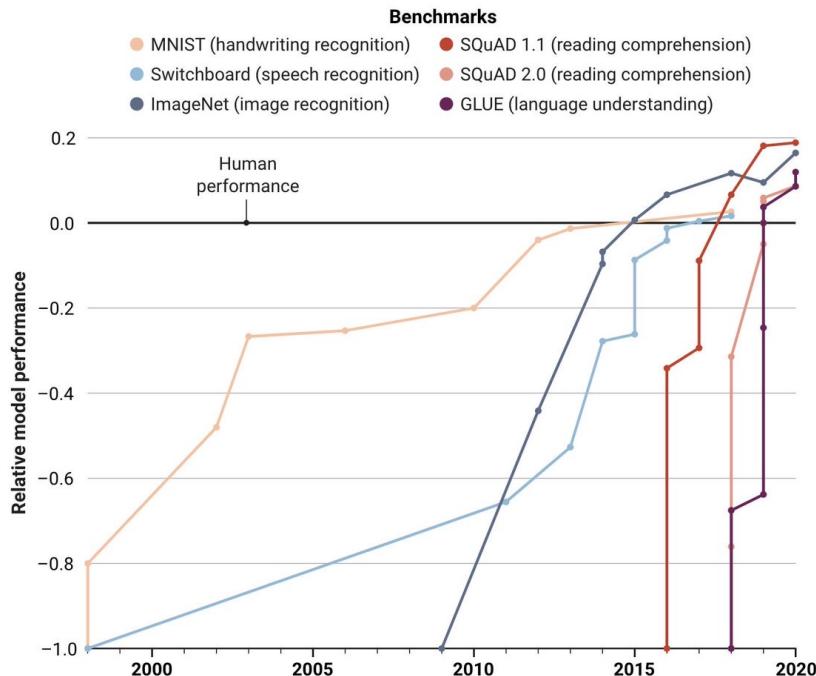
# Convert the weight to kilograms
weight_kg = weight_lbs * lbs_to_kg

# Print the result
print(f'{weight_lbs} pounds is equivalent to {weight_kg:.2f} kilograms", format(weight_lbs, weight_kg))
```

This program first defines a conversion factor for pounds to kilograms. It then prompts the user for a weight in pounds, converts that weight to kilograms using the conversion factor, and finally prints the result.

The Golden age of NLU and NLG

Outperformed Human Performance faster than ever!



Normans

The Stanford Question Answering Dataset

The Normans (Norman: Normands; French: Normands; Latin: Normanni) were the people who in the **10th and 11th centuries** gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.

In what country is Normandy located?

Ground Truth Answers: France | France | France | France

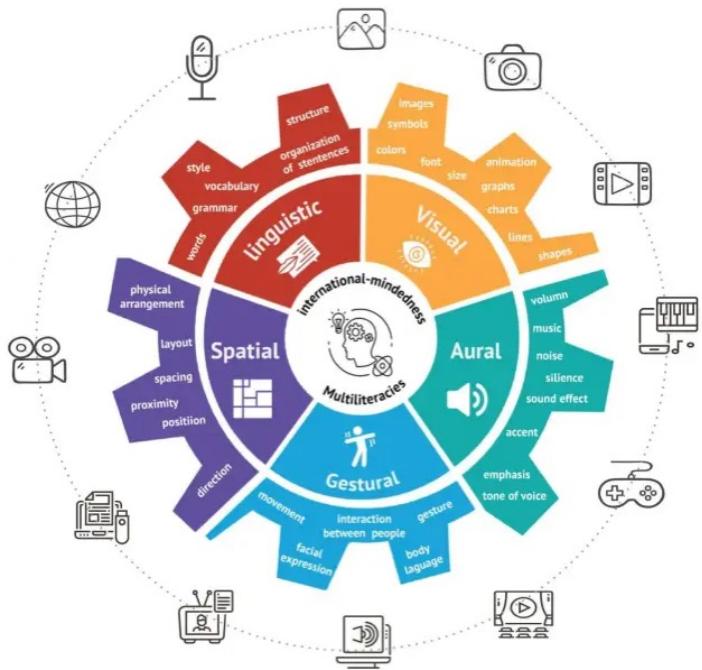
When were the Normans in Normandy?

Ground Truth Answers: 10th and 11th centuries | in the 10th and 11th centuries | 10th and 11th centuries | 10th and 11th centuries

From which countries did the Norse originate?

Ground Truth Answers: Denmark, Iceland and Norway | Denmark, Iceland and Norway | Denmark, Iceland and Norway | Denmark,

What is Multimodality?

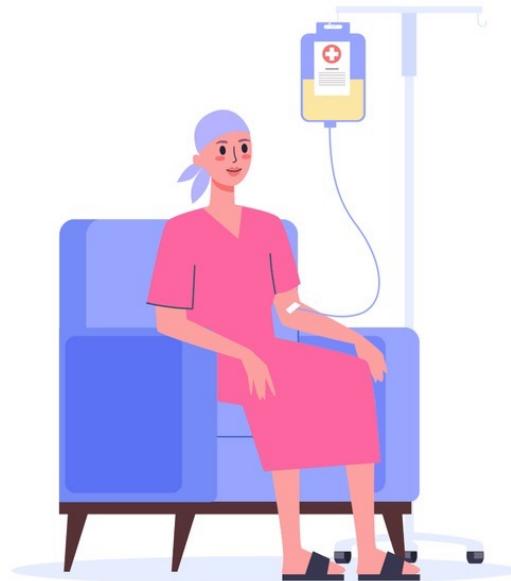


In the context of AI and Machine Learning

The ability of models to process, understand, and generate data from multiple sources or "modalities," such as text, images, audio, video, and sensor data, simultaneously.

Why do we need a multimodality?

We have a patient
who has a brain cancer.
How to produce
the future treatment plan?



Why do we need a multimodality?

Radiology



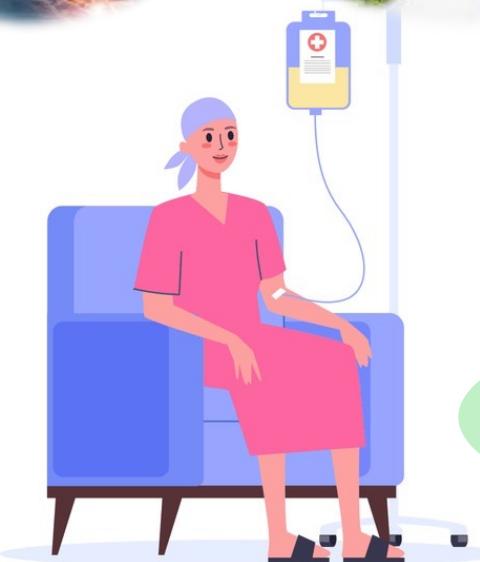
Neurosurgery



Nutritional science



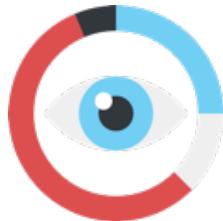
Speech therapy



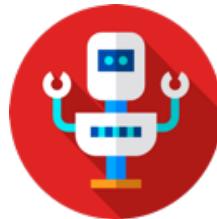
We need a multi-disciplinary experts, who can discuss and decide the plan

Why do we need a multimodality?

Vision Modal



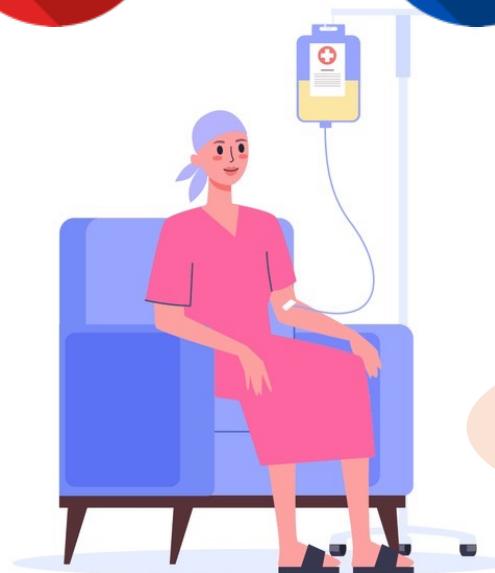
Medical Modal



Nutrition Modal



Language Modal



Can we use multimodality
to decide the plan?

Why do we need a multimodality?

Enhanced Contextual Understanding

Richer Information Capture:

By combining inputs from multiple sources (e.g., vision and audio), MLLMs capture context that would be lost with a single modality, improving comprehension and response quality.

More Accurate Inferences:

Enhanced ability to detect underlying meanings, sentiments, or actions when interpreting a situation with multimodal data (e.g., distinguishing sarcasm in text combined with visual expressions).

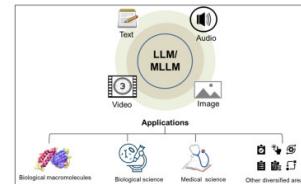
Personalised and Context-Sensitive Interactions:

Particularly useful in applications where context is essential for understanding, such as personalised assistants and healthcare diagnostics.

What do we discuss in this tutorial?

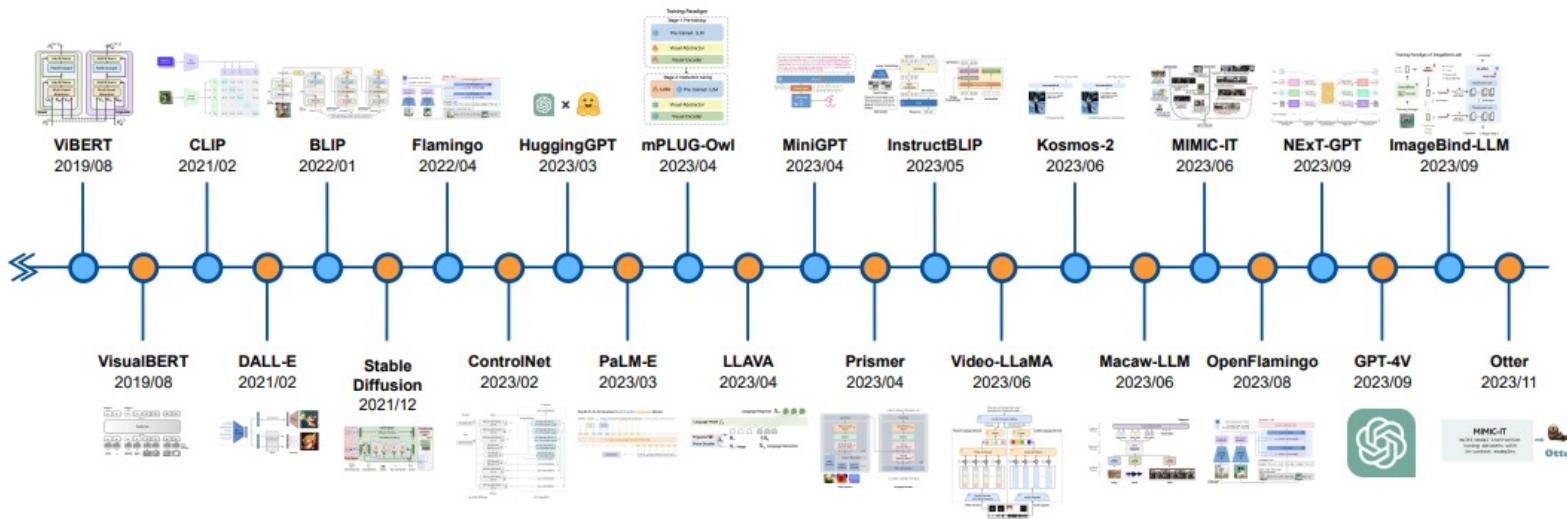
Objectives of this tutorial

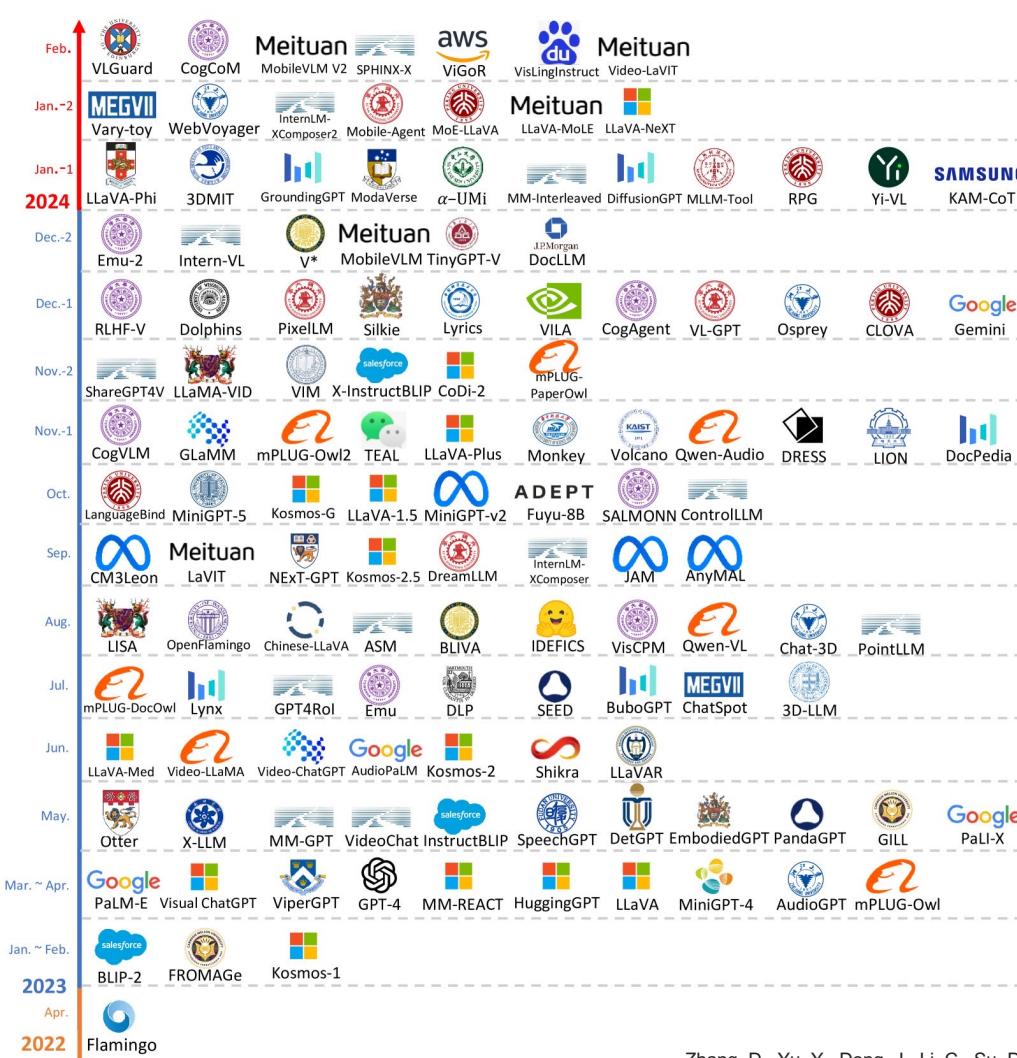
1. Understand the **fundamentals of Multimodal Large Language Models (MLLMs)**
2. Explore the **challenges and potential solutions** to deal with modalities
3. Discover **future directions** for research and development



Timeline of MLLMs

The exploding trend of MLLMs





Timeline of MLLMs

The exploding trend of MLLMs



Tutorial Outline

1. Introduction and Overview

- Welcome and tutorial objectives
- Multimodal learning
- Importance and Challenges of integrating multiple modalities (vision, language, sensors, audio)

2. Foundation of Multimodal Large Language Models (MLLMs)

- Multimodal Key Tasks and Benchmark datasets for multimodal tasks
- Challenges in Dataset Creation and Annotation for Multimodal Integration

3. Core Techniques for Multimodal Large Language Models (MLLMs)

- MLLM Architecture
- MM capabilities

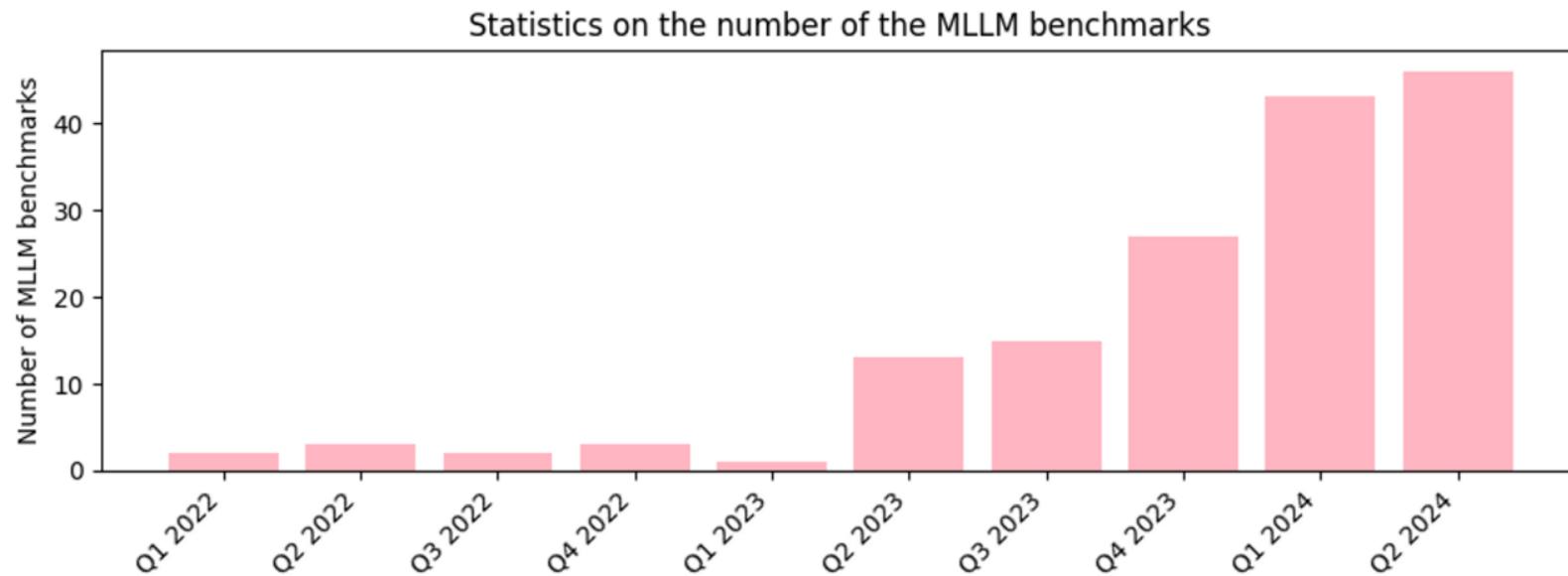
4. Multimodal Instruction Tunings

- Basic Instruction Tuning Technique to LVLM/MLLM Instruction Tuning
- Multimodal Chain-of-Thought
- Efficient Instruction Tuning Strategies

5. Future Trends and Research Opportunities

Tasks: MLLMs Benchmarks

Trend of MLLMs evaluation papers over time



Key Tasks

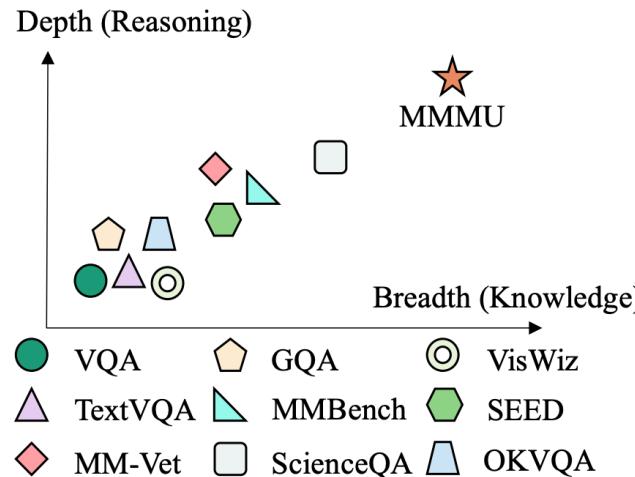
The main focus of this tutorial is not going through all 84 (+more) benchmarks for MLLMs

Benchmark	Focus	Answer Annotation	Answer Type	HGT Evaluation	Model
MMLU[1]	perception and cognition	128	VIN	N/A	10
MMLU[1]	object-level reasoning	121	VIN	GPT	10
MovieBert[14]	long video understanding	146	open	GPT	10
MovieBert[14]	long video reasoning	146	open	GPT	10
SciQ-aqua[103]	scientific question answering	6576	open	GPT	10
SciQ-aqua[103]	domain-specific reasoning	47028	open	N/A	10
M-HallDense[143]	hallucination detection	108	open	N/A	10
ViG-Bert[39]	instruction-following	592	open	Human+GPT	10
ViG-Bert[39]	long video reasoning	592	open	N/A	10
TonShine[11]	open-ended real-world dialogues	908	open	GPT	10
ViG-Bert[39]	long video reasoning	5995	open	N/A	10
MMLM-BEST-H[55]	penalizing hallucinations	96	open	GPT	10
PCA-EVALUATE[1]	decision-making ability	900	MQA	GPT	10
MMF[1]	multimodal reasoning	173	MQA	GPT	10
MMF[1]	knowledge editing	173	open	N/A	10
HumanEval[25]	language understanding	1170	open	N/A	10
HumanEval[25]	spatial relations	820	text options	N/A	10
What-1[51]	Comprehensive visual reasoning	1000	open	N/A	10
Rings[144]	long video understanding	370	open	GPT	2
MapleBench[29]	visual perception of small objects	283	MQA	Human+GPT	10
MapleBench[29]	long video reasoning	171	open	N/A	10
AMBER[152]	hallucinations related to attributes and relations	156	VIN	open	10
MMLM-BEST-H[55]	multimodal reasoning	100	MQA	GPT	10
AutoEVA[Video][100]	open ended video question answering	727	open	GPT	10
Video-Bert[17]	comprehensive evaluation for video	17000	MQA	N/A	10
MMLM-BEST-H[55]	multimodal reasoning	173	MQA	N/A	10
MMF[66]	video understanding	4000	MQA	N/A	10
MMF[66]	language understanding	1000	MQA	N/A	10
MM-SafetyBench[65]	safety-critical evaluations	5040	open	GPT	10
ViG-Bert[39]	long video understanding	1000	open	GPT	10
GeneralQA[42]	graceful degradation	10000	open	GPT	10
Auton[103]	knowledge editing	1000	open	N/A	10
EgoPlan-Bench[112]	human-level planning	5k	MQA	N/A	10
MMF[1]	multimodal reasoning	173	MQA	GPT	10
V-Plan[21]	grounding detailed visual information	191	MQA	N/A	10
DiagramQA[1]	diagram reasoning	100	open	N/A	10
Charon[102]	chart comprehension	6008	open	N/A	22
TransportationQA[130]	transportation	1910	MQA	Human+GPT	10
ViG-Bert[39]	visual reasoning	201	open	GPT	10
Bench-MH[15]	multimodal reasoning	1000	open	N/A	10
MMF[1]	multimodal reasoning	1000	MQA	GPT	10
Auton[103]	multimodal reasoning	10000	open	N/A	10
SpecIE[1]	fine-grained comprehension	8400	MQA	GPT	10
MMLM-BEST-H[55]	multimodal reasoning	4761	text options	GPT	10
CMUML[36]	college-level subject knowledge	12k	open	N/A	10
Auton[103]	multimodal reasoning	10000	open	N/A	10
CMUML[36]	knowledge comprehension and reasoning	9603	open	N/A	10
Auton[103]	mobile device operations	1000	open	N/A	10
LBRS-Bench[122]	multimodal reasoning	600	MQA	N/A	10
MHallenBench[149]	multimodal hallucination detection	420	open	N/A	2
Charon[102]	multimodal reasoning	173	MQA	N/A	10
SHID[174]	face spoofing and forgery detection	1024	VIN	MQA	10
Q-Bench[49]	low-level visual perception	369	VIN	open	10
ViG-Bert[17]	self-awareness reasoning	28k	MQA	N/A	10
Charon[102]	chart reasoning	173	MQA	N/A	10
Charon[102]	multimodal reasoning	1000	open	N/A	10
MAD-Bench[54]	multi-modal reasoning	1000	open	GPT	10
Charon[102]	multi-modal reasoning	1000	open	N/A	10
MAD-Bench[54]	multi-modal reasoning	1000	open	GPT	10
Charon[102]	context-dependent visual comprehension	706	closed+open	Human+GPT	10
MMF[1]	multimodal reasoning	1000	open	N/A	10
MMF[1]	visual hallucination	1200	VIN	N/A	10
MMF[1]	multimodal reasoning	1000	open	N/A	10
MIRE[77]	knowledge editing	10000	open	N/A	10
OSCAR[171]	object-level reasoning	14084	MQA	N/A	10
CRPE[1]	relation comprehension	10000	open	N/A	10
ImageNet[10]	image classification	5540	multiple types of tasks	GPT	10
ImageNet[10]	facial affective behavior analysis	1132	open	GPT	10
ImageNet[10]	image captioning	5000	multiple types of tasks	GPT	10
ViLG-Bert[73]	knowledge editing	3174	open	N/A	10
ViLG-Bert[73]	continual learning	1000	open	N/A	10
ViLG-Bert[73]	multimodal reasoning	1320	VIN	N/A	10
MathNeru[3]	visual math tasks	136	open	GPT	10
PO2[10]	open-domain reasoning	10000	open	GPT	10
P2[30]	visual reasoning capabilities	2130	MQA	N/A	10
MMLM-BEST-H[55]	visual prompting research	1500	MQA	N/A	10
MMF[1]	3D medical tasks	1500	MQA	open	10
MMF[1]	the robustness of MLLM against adversarial attacks	20000	MQA	open	10
PABA-Bench[44]	facial affective behavior analysis	403	open	N/A	10
PABA-Bench[44]	image aesthetic attributes	173	open	N/A	10
ViG-WoW[109]	understanding and grounding in web contexts	173	open	N/A	10
DesignQA[47]	real-world engineering tasks	1451	open	N/A	10
LM[1]	visual reasoning	1000	open	N/A	10
UNSAAC[46]	image acoustic attributes	3554	multiple types of tasks	N/A	10
CORE[46]	compositional reasoning	2400	MQA	N/A	10



Key Tasks

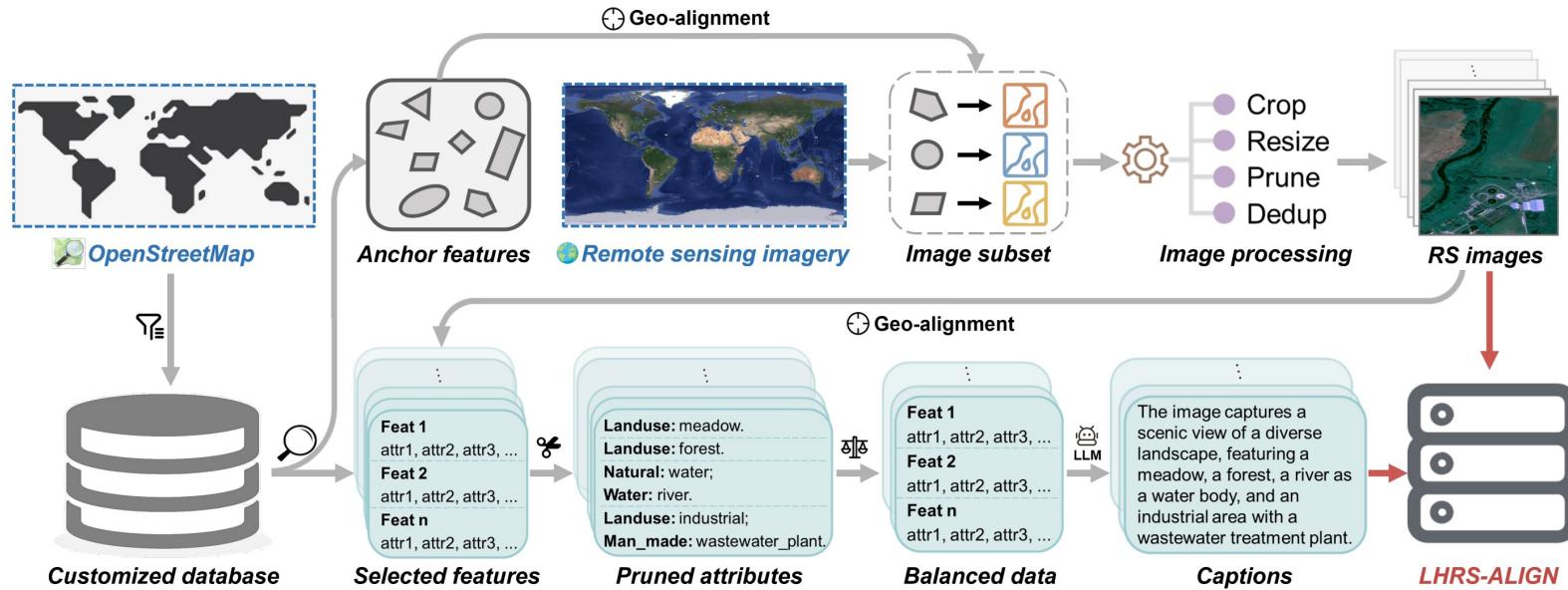
So many Vision-Language focused tasks, such as Visual Question Answering, Image-to-xx, ...



Dataset	Size	Images	Format	Source	Answer
VQA	> 1M	V	I+T	Annotated	Open
GQA	> 1M	V	I+T	Synthesized	Open
VisWiz	32K	V	I+T	Annotated	Open
TextVQA	45K	OC	I+T	Annotated	MC
OKVQA	14K	V+OC	I+T	Annotated	Open
SEED	19K	V+OC	I+T	Annotated	MC
MMBench	3K	V+OC	I+T	Repurposed	MC
MM-Vet	0.2K	V+OC	I+T	Annotated	Open
ScienceQA	6K	5 Types	I+T	Textbooks	MC
MMMU	11.5K	30 Types	Interleaved	Textbooks, Internet, Annotated	Open / MC

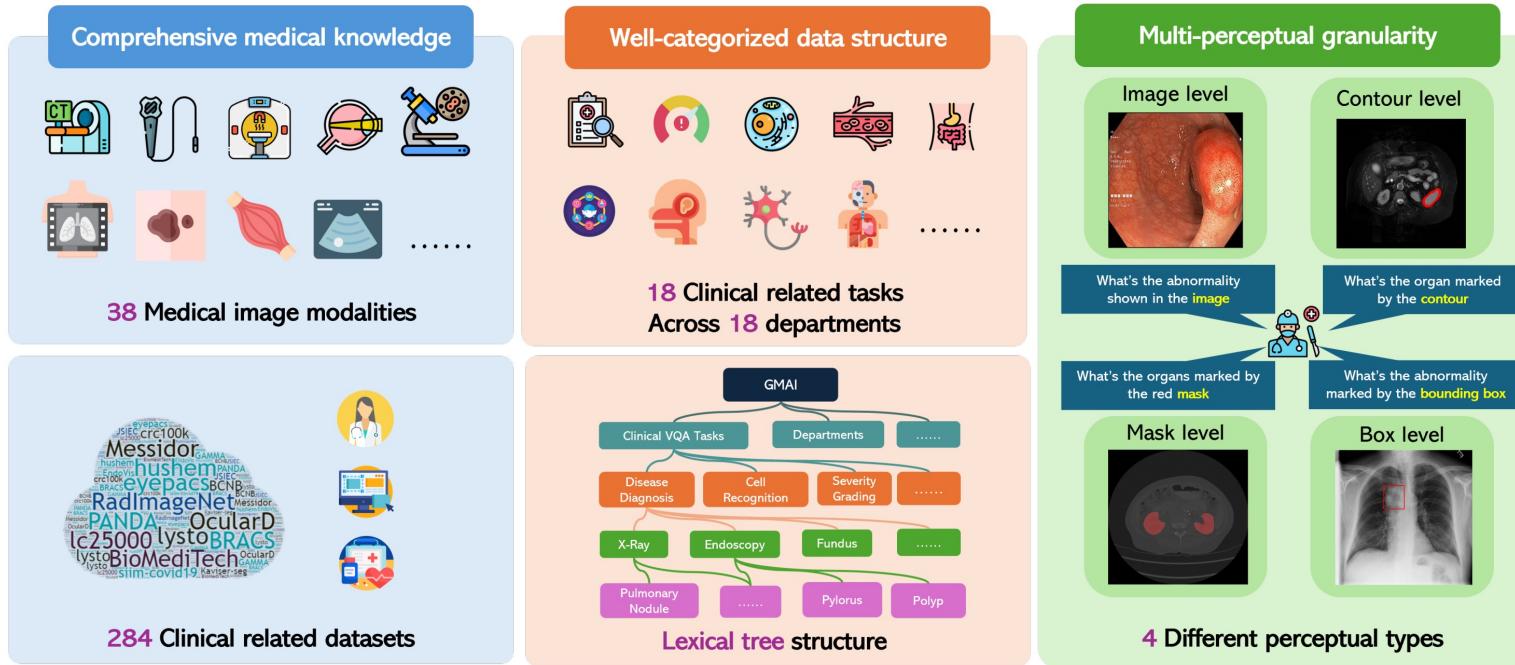
Key Tasks: Specific Domain (Geography)

Geography and Remote Sensing – LHRs-Bench (Muhtar et al. 2024)



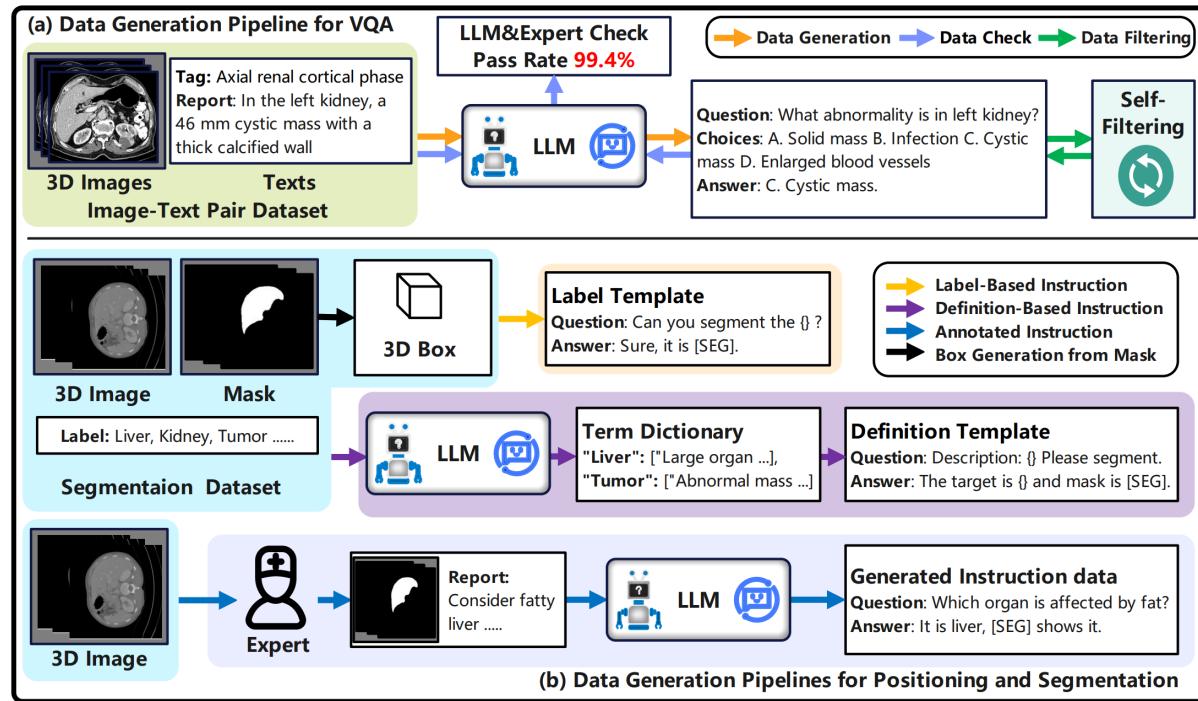
Key Tasks: Specific Domain (Medical)

GMAI-MMBench – Real-world clinical scenario (Chen et al. 2024)



Key Tasks: Specific Domain (Medical)

M3C – 3D Medical Image Analysis (Fan et al. 2024)



Key Tasks: Specific Domain (Geography)

Geography and Remote Sensing – ChartingNewTerritories (Roberts et al. 2024)

Satellite imagery season detection

Prompt:

These 4 satellite images are taken in different seasons. For each image, provide a prediction of the season. Select a single season per image, without replacement, from the following: Spring, Summer, Autumn, Winter. Let's think step by step.



The image shows a landscape with a lack of greenery, and the **fields are mostly brown with no visible crops**. This could indicate a time either before planting or after harvest. There is no snow, so it's likely not winter. This image could represent either autumn, after harvest time, or early spring, before planting has begun. **Prediction: Spring.**

The image shows fields **that appear to be lush and possibly in full growth**. This is indicative of a time when crops are at their peak, which is typically in the summer. **Prediction: Summer.**

The image shows **fields with varying shades of brown and some patches that could be interpreted as either harvested fields or fields lying fallow**. There isn't a clear indication of snow or full greenery. This image might suggest the season is either autumn, with fields that have been recently harvested, or late winter, where some vegetation is still dormant. **Prediction: Autumn.**

The image has a **darker tone** to it, and some of the fields have a **very slight white hue** which might suggest **frost or light snow**. This could imply that the season is winter. **Prediction: Winter.**

Key Tasks: Specific Domain (Geography)

Geography and Remote Sensing – ChartingNewTerritories (Roberts et al. 2024)

Bounding box / Object Counting

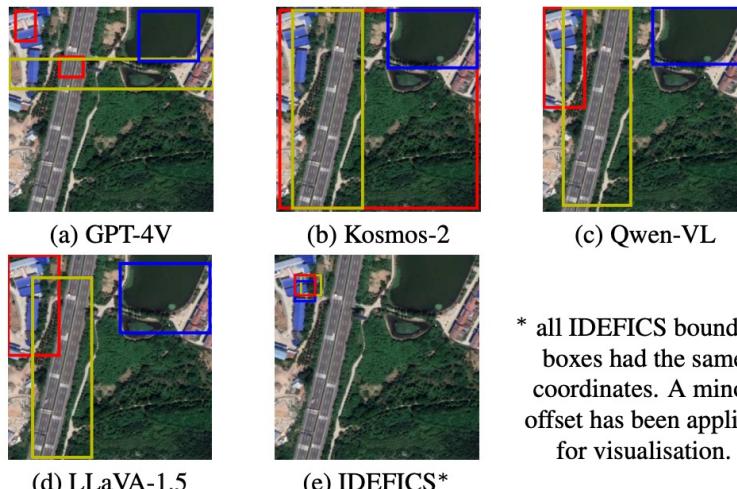


Figure 4. Bounding boxes for **urban areas**, **road** and **water bodies**.



Figure 5. Counting small objects proves challenging.

Prompt:
How many sea lions are in this image?

Answer: 45
GPT-4V: 36
IDEFICS: 10
LLaVA-1.5: 2
Qwen-VL: 0
InstructBLIP: 10

Key Tasks: Specific Domain (Medical)

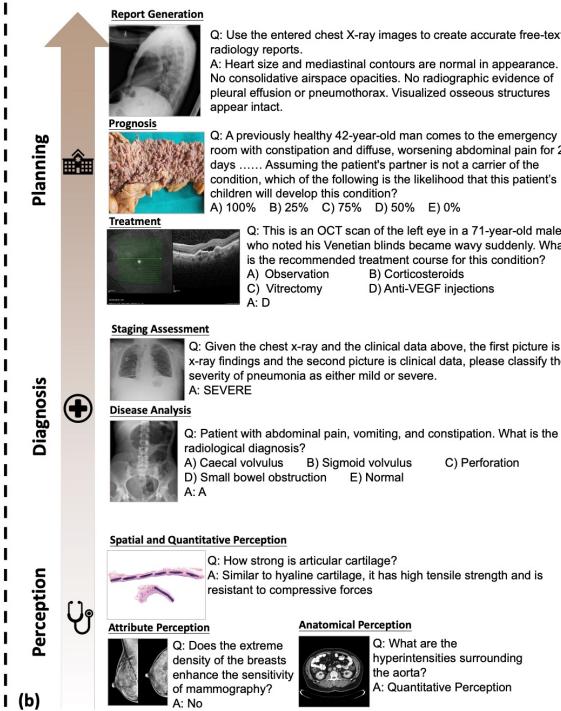
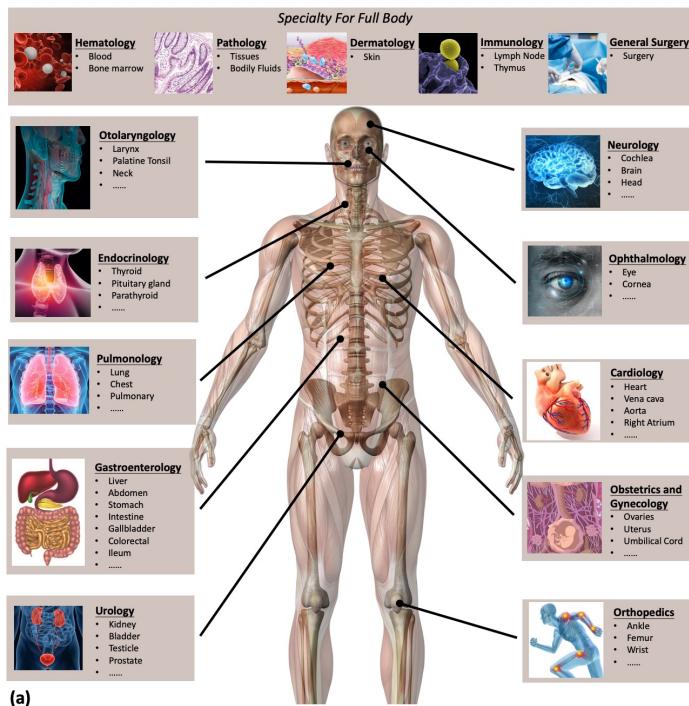
M3C – 3D Medical Image Analysis (Fan et al. 2024)

Datasets	Types	Tasks	Images	Texts
VQA-Med [5]	2D	VQA	3,200	12,792
MIMIC-CXR [22]	2D	ITR, RG	377,110	227,835
PMC-OA [35]	2D	ITR, RG	-	1,646,592
PMC-VQA [73]	2D	VQA	149,075	226,946
RP3D-Caption [68]	3D	ITR, RG	51K	-
RP3D-VQA [68]	3D	VQA	-	142K
M3D-Cap	3D	ITR, RG	120,092	42,496
M3D-VQA	3D	VQA	96,170	509,755
M3D-RefSeg	3D	REC, REG, SS, RES	210	2,778
M3D-Seg	3D	REC, REG, SS, RES	5,772	149,196*

* In segmentation datasets, the number of texts can be linked to semantic masks.

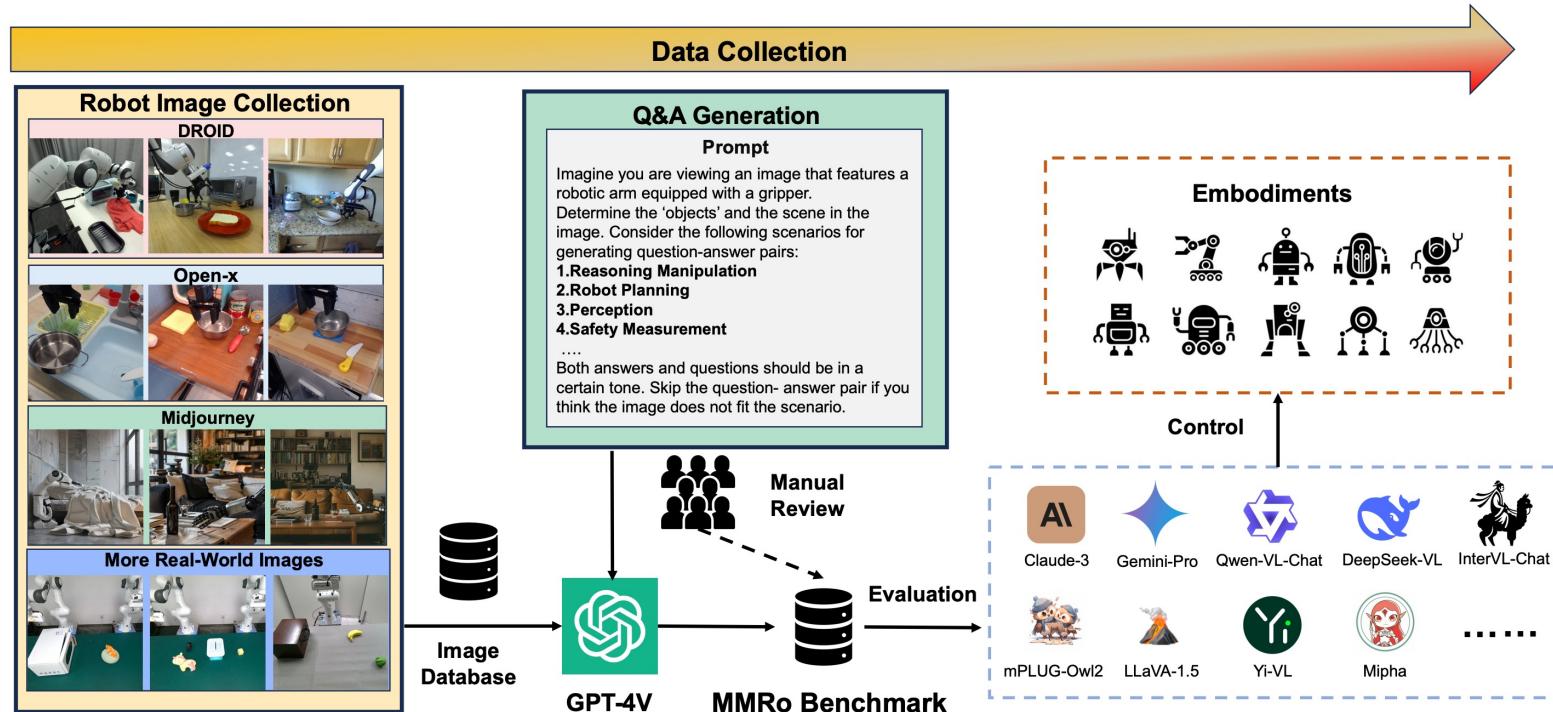
Key Tasks: Specific Domain (Medical)

Asclepius – Diverse Body parts/organ (Wang et al. 2024)



Key Tasks: Specific Domain (Robotics)

MMRo – for Robots (Li et al. 2024)



Key Tasks: Specific Domain (Robotics)

MMRo – for Robots (Li et al. 2024)

Sub-Domain	# of Q&A Pair	Task Definition
Perception		
Object Color	2174	Discriminating the color of the object.
Object Shape	2158	Discriminating the shape of the object.
Material Composition	2156	Identifying the material of the object.
Object Counting	2185	Counting the number of specified items in the scene.
Object Orientation	2116	Detecting the position and orientation of objects.
Visual Grounding	803	Return the bounding box of the queried object.
Task Planning		
Task Sequencing	2176	The ability to decompose a plan into a proper sequence of sub-tasks.
Spatial Awareness for Object Manipulation	2056	Evaluate the decision-making ability in the presence of interfering substances.
Visual Reasoning		
Object Function Identification	2178	The ability to understand the functionality of the object.
Tool Appropriateness	2173	Determine whether the queried object is a suitable tool to perform the task.
Anticipation of Action Consequences	2179	Anticipating specific outcomes of a robot action.
Safety Measurement		
Sharp Object Handling	919	Whether the object is sharp and needs careful manipulation.
Delicate Object Manipulation	2101	Determine if the object is delicate for grasping.
Heat Safety Protocol	801	Determine whether the object could have a high-temperature surface.
Total	26175	-

Key Tasks: Specific Domain (Robotics)

MMRo – for Robots (Li et al. 2024)

Perception

Object Color



Query: What is the primary color of the box?

Multiple-Choice:

- A. Blue
- B. Red
- C. yellow**
- D. Green

Open-Ended: The color of this box is *yellow*.

Object Counting



Query: How many objects are on the wooden surface?

Multiple-Choice:

- A. 2
- B. 3
- C. 4**
- D. 5

Open-Ended: There are *four* objects on the wooden surface.

Material Composition



Query: What material is the pan made of?

Multiple-Choice:

- A. Plastic
- B. Wood
- C. Metal**
- D. Glass

Open-Ended: The pan appears to be made of *metal*.

Object Shape



Query: What is the shape of the object the robotic arm is gripping?

Multiple-Choice:

- A. Square
- B. Cylinder**
- C. Sphere
- D. Cone

Open-Ended: The object is a *cylindrical* mug with a handle.

Object Orientation



Query: How should the robotic arm orient the gripper to safely pick up the glass without spilling the content?

Multiple-Choice:

- A. Tilted downwards
- B. Parallel to the counter**
- C. Perpendicular to the counter
- D. Upside down

Open-Ended: *Parallel* to the counter.

Visual Grounding



Query: Can you show where spoon is located in this picture? Please answer only the bounding box with this format [x1, y1, x2, y2].

Open-Ended: [0.59, 0.37, 0.73, 0.62]

Key Tasks: Specific Domain (Robotics)

MMRo – for Robots (Li et al. 2024)

Task Planning

Task Sequencing



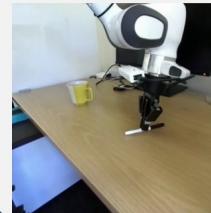
Query: What sequence of actions should the robotic arm take to fill the paper cup with water without spilling?

Multiple-Choice:

- A. **Grip cup, move to water source, fill, move back**
- B. Move to water source,
- C. Fill cup,
- D. Grip cup move to water source

Open-Ended: Grip cup, move to water source, fill, move back

Spatial Awareness for Object Manipulation



Query: Can the robotic arm reach the yellow cup without moving any other objects?

Multiple-Choice:

- A. If the robotic arm moves directly, it will collide with
- B. The robotic arm does not move anything**
- C. There are distractions on the desktop
- D. The robotic arm will not move

Open-Ended: Yes, the robotic arm can reach the yellow cup without moving any other objects as there is a clear path to it.

Key Tasks: Specific Domain (Robotics)

MMRo – for Robots (Li et al. 2024)

Visual Reasoning

Consequence Anticipation



Query: If the robotic arm grips the pot too tightly, what could be the consequence?

Multiple-Choice:

- A. The pot may slip out
- B. The pot may get crushed**
- C. The gripper may get damaged
- D. Nothing will happen

Open-Ended: The pot may get **crushed**.

Function Identification



Query: What is the typical function of the gripper on the robotic arm in this scene?

Multiple-Choice:

- A. To lift and move objects**
- B. To provide light
- C. To type on a keyboard
- D. To paint a picture

Open-Ended: The gripper is typically used for **grasping and manipulating** objects.

Tool Appropriateness



Query: Is the gripper on the robotic arm suitable for picking up cylindrical container on the table?

Multiple-Choice:

- A. Yes, it is designed for such tasks**
- B. No, it is too large
- C. No, it is too small
- D. Yes, but it may damage the container

Open-Ended: Yes, it is **designed** for such tasks.

Delicate Object Manipulation



Query: How should the robotic arm manipulate the book without damaging it?

Multiple-Choice:

- A. Firm grip
- B. Gentle grip**
- C. No grip
- D. Use a magnetic field

Open-Ended: Adjust the grip strength to be **gentle**. Add soft rubber to the grippers. Program the arm to move slowly and smoothly.

Safety Measurement

Sharp Object Handling



Query: Is there a sharp object present that the robotic arm should handle with caution?

Multiple-Choice:

- A. Yes, a knife
- B. Yes, a fork
- C. No sharp objects**
- D. Yes, a pair of scissors

Open-Ended: There are **no sharp objects**.

Heat Safety



Query: Assuming the glass contains a hot beverage, what protocol should the robotic arm follow for safety?

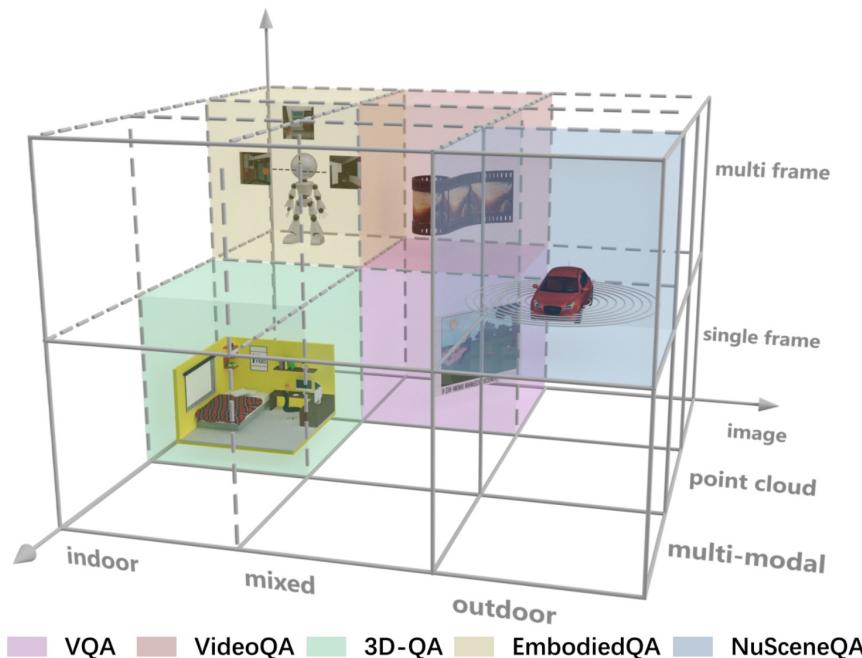
Multiple-Choice:

- A. Use heat-resistant grippers**
- B. Wait for the beverage to cool down
- C. Pick up the glass quickly
- D. Ignore the temperature

Open-Ended: Ensure the grippers are made of or coated with **heat-resistant materials** to prevent damage from the hot surface.

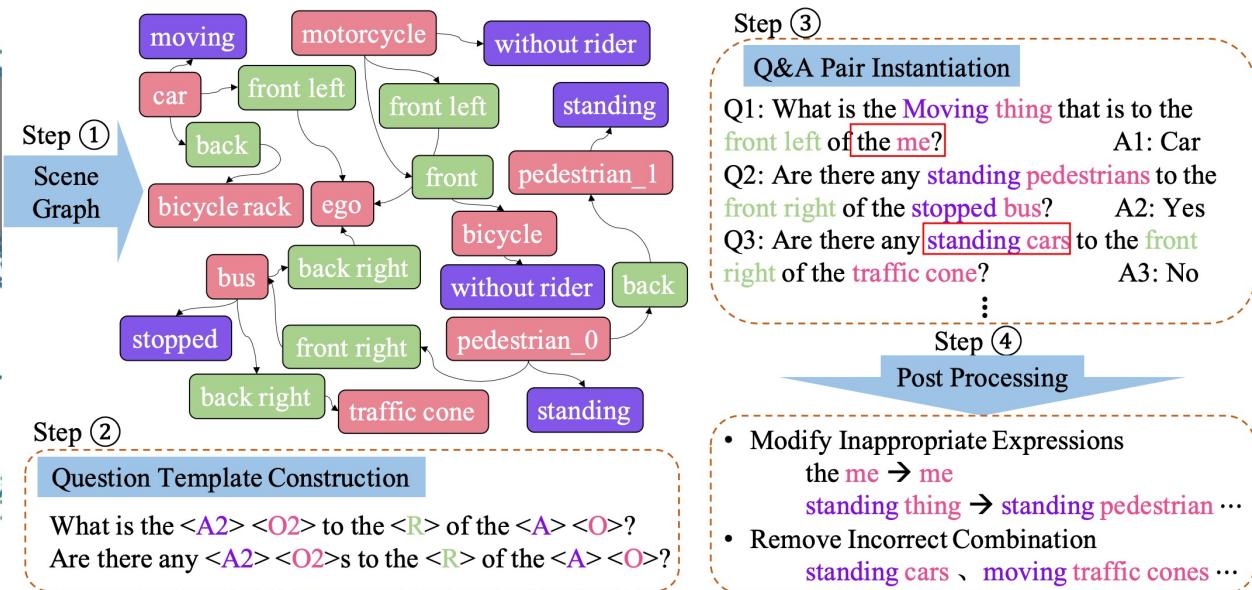
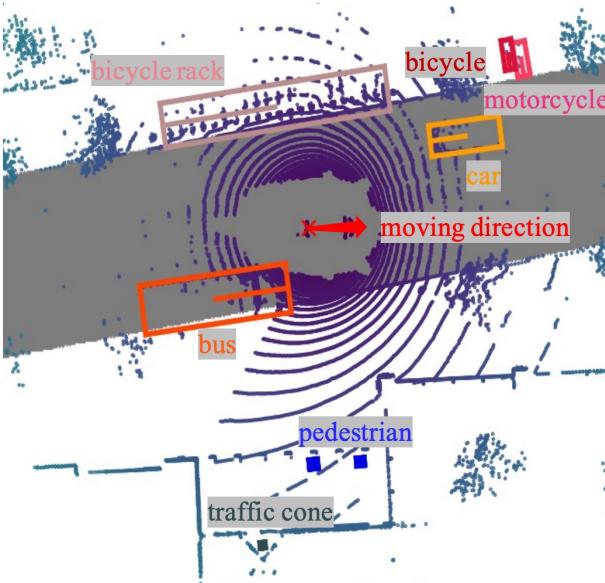
Key Tasks: Specific Domain (Autonomous Driving)

NuScenes-QA- Driving (Qian et al. 2024)



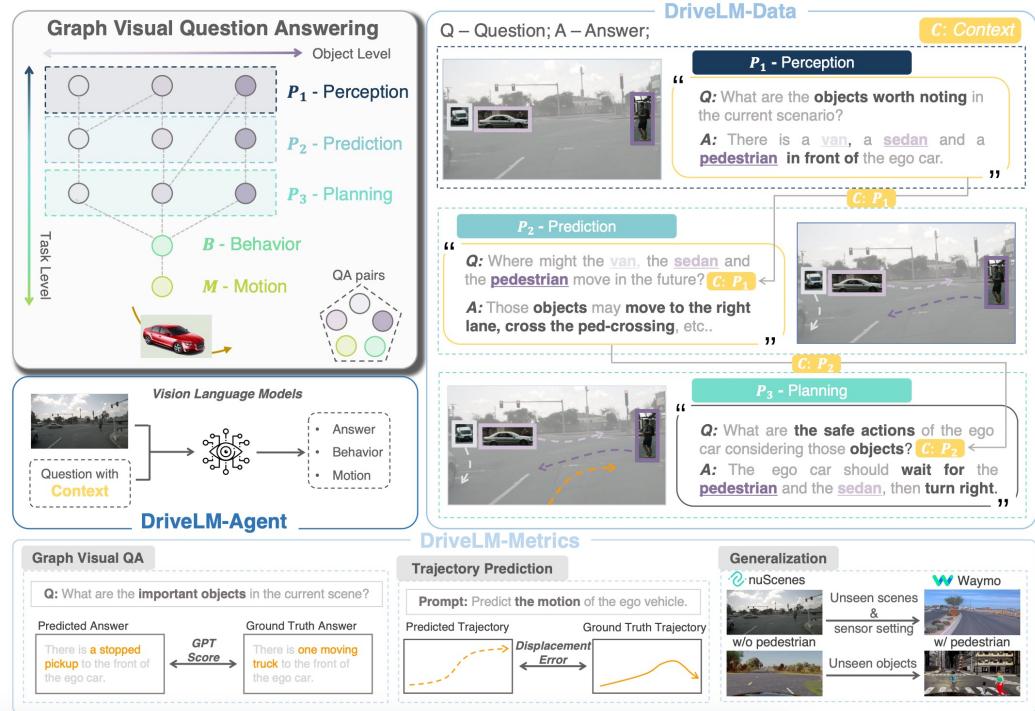
Key Tasks: Specific Domain (Autonomous Driving)

NuScenes-QA- Driving (Qian et al. 2024)



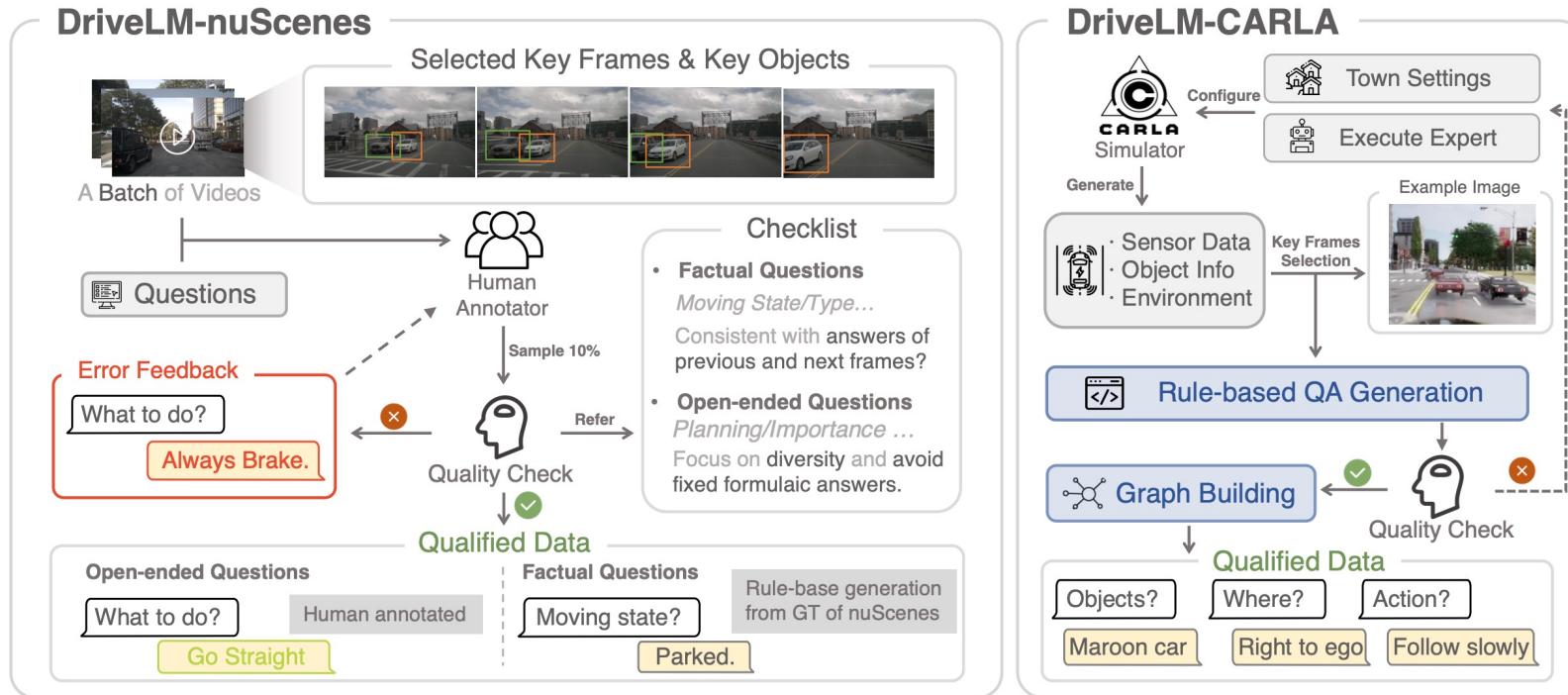
Key Tasks: Specific Domain (Autonomous Driving)

DriverLM- Driving (Sima et al. 2024)



Key Tasks: Specific Domain (Autonomous Driving)

DriverLM- Driving (Sima et al. 2024)



Key Tasks: Specific Domain (Visually-Rich Document)

Form-NLU (Ding et al. 2023)

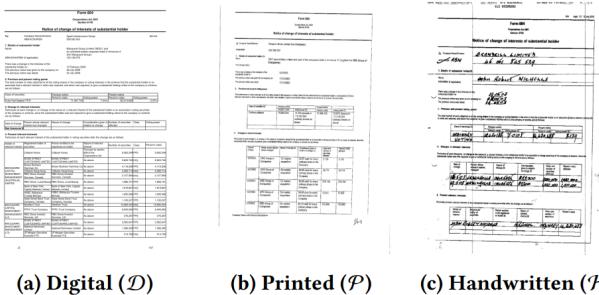


Figure 1: Digital, Printed and Handwritten Form Samples.

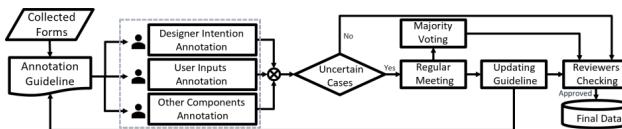


Figure 2: Overall Annotation Workflow.

Task A: Form Layout Analysis

Form 604 Corporations Act 2001 Substantial holder notice				
Notice of change of interests of substantial holder				
1. Company Name/Scheme	Tinybeans Group Ltd			
ACN/ARSN	168 481 614			
1. Details of substantial holder (1)				
Name	Tinybeans Group Ltd			
ACN/ARSN (if applicable)	200 552 118			
There was a change in the interests of the substantial holder on The previous notice was given to the company on The previous notice was dated				
18/03/2008	18/03/2008			
2. Previous and present voting power				
The total number of votes attached to all the voting shares in the company or voting interests in the scheme that the substantial holder or an associate (2) had a relevant interest (3) in when last required, and when now required, to give a substantial holding notice to the company or scheme, are as follows:				
Class of securities (4)	Persons' votes	Voting power (%)	Persons' votes	Voting power (%)
Damery securities	5,290,000	5.29%	13,419,536	7.64%

Categories	Description
Title	Form Title (e.g. "Form 604")
Section	Section Header (e.g. "Details of substantial holder")
Form Key	Form Key Segments (e.g. "Name")
Form Value	Form Value Segments (e.g. "Territory Resources Limited")
Table Key	Table Key Segments (e.g. "Voting Power (5)")
Table Value	Table Value Segments (e.g. "5.29%")
Other	Others (e.g. text paragraphs)

Task B: Key Information Extraction

Form 604 Corporations Act 2001 Substantial holder notice				
Notice of change of interests of substantial holder				
1. Company Name/Scheme	Tinybeans Group Ltd			
ACN/ARSN	168 481 614			
1. Details of substantial holder (1)				
Name	Tinybeans Group Ltd			
ACN/ARSN (if applicable)	200 552 118			
There was a change in the interests of the substantial holder on The previous notice was given to the company on The previous notice was dated				
18/03/2008	18/03/2008			
2. Previous and present voting power				
The total number of votes attached to all the voting shares in the company or voting interests in the scheme that the substantial holder or an associate (2) had a relevant interest (3) in when last required, and when now required, to give a substantial holding notice to the company or scheme, are as follows:				
Class of securities (4)	Persons' votes	Voting power (%)	Persons' votes	Voting power (%)
Damery securities	5,290,000	5.29%	13,419,536	7.64%

Key	Value
Company Name/Scheme	Tinybeans Group Ltd
ACN/ARSN	168 481 614
Name	Stephen O'Young
There was a change in the interests of the substantial holder on	
The previous notice was given to the company on	
The previous notice was dated	
Class of securities (4)	
Previous notice Person's votes	6,733,574
Previous notice Voting power (%)	25.75%
Present notice Person's votes	6,733,574
Present notice Voting power (%)	25.75%

Key Tasks: Specific Domain (Visually-Rich Document)

MM-VQA (Ding et al. 2024)

XML Node (Paragraph)	Document Pages	XML Node (Table/Figure Captions)												
<pre><p id="Par11"> SPSS 24.0 (IBM, Chicago, IL, USA) was used for data processing and for all data analyses. Measurement data were expressed in terms of the mean and standard deviation ... </p></pre>		<pre><table-wrap id="Tab1"> <label>Table 1</label> <caption> <p>General characteristics of the participants</p> </caption> <thead> ...</pre>												
Paragraph-based Question Prompt <table border="1"> <tr> <th>Preprocessing</th> <th>Question Generation Prompt</th> <th>Generated Question</th> </tr> <tr> <td> <code>tk = sent_tokenizer if sent_num < 2: num_question = 1 elif sent_num < 4: ... < num_question></code> </td> <td> <code>Paragraph: <para_content> Ask <num_question> from that paragraph. The answer to those questions must directly extract sentences from the above paragraph (without any change). Output Format: Q1: XXX</code> </td> <td> <code>E_{p1}-E_{p2}-L₁-Q₁</code> </td> </tr> </table> Table/Figure based Question Prompt <table border="1"> <thead> <tr> <th>Caption Summarizing Prompt</th> <th>Question Generation Prompt</th> <th>Generated Question</th> </tr> </thead> <tbody> <tr> <td> <code>Figure (Table) Caption: <figure/table> caption: What does the figure illustrate? (under ten words if possible) <summarized content></code> </td> <td> <code>Question: Could you find the figure of <summarized content>? Improve the previous question by paraphrasing it concisely summarised.</code> </td> <td> <code>E_t-L₁-Q₂</code> </td> </tr> </tbody> </table>			Preprocessing	Question Generation Prompt	Generated Question	<code>tk = sent_tokenizer if sent_num < 2: num_question = 1 elif sent_num < 4: ... < num_question></code>	<code>Paragraph: <para_content> Ask <num_question> from that paragraph. The answer to those questions must directly extract sentences from the above paragraph (without any change). Output Format: Q1: XXX</code>	<code>E_{p1}-E_{p2}-L₁-Q₁</code>	Caption Summarizing Prompt	Question Generation Prompt	Generated Question	<code>Figure (Table) Caption: <figure/table> caption: What does the figure illustrate? (under ten words if possible) <summarized content></code>	<code>Question: Could you find the figure of <summarized content>? Improve the previous question by paraphrasing it concisely summarised.</code>	<code>E_t-L₁-Q₂</code>
Preprocessing	Question Generation Prompt	Generated Question												
<code>tk = sent_tokenizer if sent_num < 2: num_question = 1 elif sent_num < 4: ... < num_question></code>	<code>Paragraph: <para_content> Ask <num_question> from that paragraph. The answer to those questions must directly extract sentences from the above paragraph (without any change). Output Format: Q1: XXX</code>	<code>E_{p1}-E_{p2}-L₁-Q₁</code>												
Caption Summarizing Prompt	Question Generation Prompt	Generated Question												
<code>Figure (Table) Caption: <figure/table> caption: What does the figure illustrate? (under ten words if possible) <summarized content></code>	<code>Question: Could you find the figure of <summarized content>? Improve the previous question by paraphrasing it concisely summarised.</code>	<code>E_t-L₁-Q₂</code>												

Sample Question generation progress

Input Document Pages

Page 1 (Q1)

E_{p_1}

Page 2 (Q1, Q2)

E_{p_2}

Q1: Paragraph-based Qs
What software was used for data processing and analysis?

Targets: (E_{p_1}, E_{p_2})

E_{p_1}

SPSS 24.0 (IBM, Chicago, IL, USA) was used for data processing and for all data analyses. Measurement data were expressed in terms of the ...

Q2: Table/Figure related Qs
Can you locate the table showing the risk factors for diabetes among adults in Bengbu?

Targets: (E_t)

E_t

Category	Risk Factor	Value
1	Age	25-34
2	Age	35-44
3	Age	45-54
4	Age	55-64
5	Age	65-74
6	Age	75-84
7	Age	85+
8	Gender	Male
9	Gender	Female
10	Education	Primary
11	Education	Secondary
12	Education	Tertiary
13	Employment	Unemployed
14	Employment	Employed
15	Health Status	Good
16	Health Status	Poor

Tasks of multi-modal cross-page information retrieval

Datasets for Pretraining

Coarse-grained Image-Text

- Large data volume (from the Internet)
- Short and Noisy (due to the web-crawled nature)



Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

Conceptual Captions: a worker helps to clear the debris.



Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Conceptual 3M
(Sharma et al. 2018)



<PERSON> was the first US president to attend a tournament in sumo's hallowed Ryogoku Kokugikan arena. (AFP photo)

Hand holding a fresh mangosteen



#jellyfish #blue #ocean #pretty Sea Turtle Wallpaper, Aquarius Aesthetic, Blue Aesthetic Pastel, The Adventure Zone, Capricorn And <PERSON>, Life Aquatic, Ocean Life, Jellyfish, Marine Life

Conceptual 12M
(Changpinyo et al. 2020)



Query Image



Not quite sure what the name of this bird is. Saw while walking along the beach in Ocracoke, NC



This is a boat I saw while walking near the house we rented.



Young baboon in the campsite at fish river canyon.



Granite in green glass.

1,000

SBU Captions
(Ordonez et al. 2011)

Datasets for Pretraining

Coarse-grained Image-Text

- Large data volume (from the Internet)
- Short and Noisy (due to the web-crawled nature)



Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.



Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Conceptual 3M
(Sharma et al. 2018)



<PERSON> was the first US president to attend a tournament in sumo's hallowed Ryogoku Kokugikan arena. (AFP photo)



#jellyfish #blue #ocean #pretty Sea Turtle Wallpaper, Aquarius Aesthetic, Blue Aesthetic Pastel, The Adventure Zone, Capricorn And <PERSON>, Life Aquatic, Ocean Life, Jellyfish, Marine Life

Conceptual 12M
(Changpinyo et al. 2020)



Query Image



Not quite sure what the name of this bird is. Saw while walking along the beach in Ocracoke, NC



Young baboon in the campsite at fish river canyon.



This is a boat I saw while walking near the house we rented.

1,000

Granite in green glass.

SBU Captions
(Ordonez et al. 2011)

+)**LAION-5B, -COCO** (Schuhmann et al. Sep 2022): web-crawled images and associated alt-text as captions

Datasets for Pretraining

Coarse-grained Image-Text

- Large data volume (from the Internet)
- Short and Noisy (due to the web-crawled nature)



Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.



Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Conceptual 3M
(Sharma et al. 2018)



<PERSON> was the first US president to attend a tournament in sumo's hallowed Ryogoku Kokugikan arena. (AFP photo)



#jellyfish #blue #ocean #pretty Sea Turtle Wallpaper, Aquarius Aesthetic, Blue Aesthetic Pastel, The Adventure Zone, Capricorn And <PERSON>, Life Aquatic, Ocean Life, Jellyfish, Marine Life

Conceptual 12M
(Changpinyo et al. 2020)



Query Image



Not quite sure what the name of this bird is. Saw while walking along the beach in Ocracoke, NC



Young baboon in the campsite at fish river canyon.



This is a boat I saw while walking near the house we rented.

1,000

Granite in green glass.

SBU Captions
(Ordonez et al. 2011)

- + **LAION-5B, -COCO** (Schuhmann et al. Sep 2022): web-crawled images and associated alt-text as captions
- + **COYO-700M** (Byeon et al. Aug 2022): from the CommonCrawl. with several data filtering and cleaning

Datasets for Pretraining

Fine-grained Image-Text

- High-quality through prompting MLLMs
- Longer and Accurate descriptions of the images



(a) Comparison of Captions' Quality

COCO: Young children standing on a platform waiting for a train to arrive. Adults and children watching a train slowly leave. A family near a railroad track watching the train pass. People waiting on a platform as a train pulls up. A train station with a green train on the tracks and children waiting for it to go by.

LLaVA: At a train station, a group of people, including both young children and adults, are standing on a platform waiting for a train to arrive. The train is already present on the tracks, partially visible on the right side of the image. Some of the people watch the train closely, while others seem to be patiently anticipating its departure.

There is a total of eight individuals waiting for the train, with one child in the middle of the platform and the others scattered around. A backpack can be found on the far left side of the platform, suggesting that someone may have set it down while waiting.

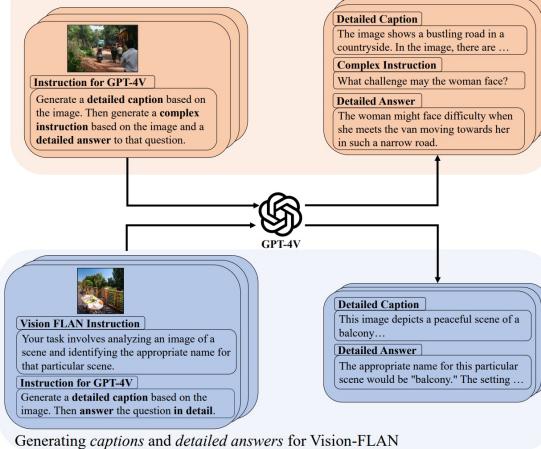
ShareGPT4V: The image captures a moment at a train station. A green train is on the tracks, moving away from the platform labeled "Platform 2". The train's motion is observed by people standing on the platform, their attention drawn to the departing vehicle.

A red signal stands on the tracks, its vibrant color contrasting with the green of the train. Signs are prominently displayed around the platform. One warns "Beware of Trains", a cautionary reminder of the powerful machine that frequently uses tracks. Another sign advises "Please Use The Footbridge To Cross The Line. If You Require Assistance Please Ask The Station Staff", guiding passengers to safely navigate the station.

The sky above is a clear blue, providing a serene backdrop to the bustling activity of the station. In the distance, trees can be seen, adding a touch of nature to this man-made setting. The image is a snapshot of everyday life at a train station, capturing both its routine operations and its inherent charm.

ShareGPT4V
(Chen et al. Nov 2023)

Generating captions, complex instructions and answers for LAION



Generating captions and detailed answers for Vision-FLAN

ALLaVa
(Chen et al. Feb 2024)

Datasets for Pretraining

Video-Text

- Large-scale dataset for the open domain video captioning (10K video clips from 20 categories)
- Each video clip is annotated with 20 English sentences by Amazon Mechanical Turks



1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.



1. A woman giving speech on news channel.
2. Hillary Clinton gives a speech.
3. Hillary Clinton is making a speech at the conference of mayors.
4. A woman is giving a speech on stage.
5. A lady speak some news on TV.



1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.



1. A man and a woman performing a musical.
2. A teenage couple perform in an amateur musical.
3. Dancers are playing a routine.
4. People are dancing in a musical.
5. Some people are acting and singing for performance.



1. A white car is drifting.
2. Cars racing on a road surrounded by lots of people.
3. Cars are racing down a narrow road.
4. A race car races along a track.
5. A car is drifting in a fast speed.



1. A player is putting the basketball into the post from distance.
2. The player makes a three-pointer.
3. People are playing basketball.
4. A 3 point shot by someone in a basketball race.
5. A basketball team is playing in front of spectators.

MSR-VTT
(MicroSoft Research
Video-To-Text)
(Xu et al. 2016)

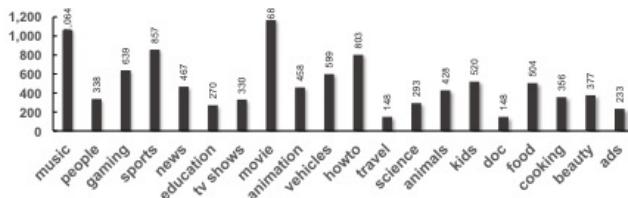
Datasets for Pretraining

Video-Text

- Large-scale dataset for the open domain video captioning (10K video clips from 20 categories)
- Each video clip is annotated with 20 English sentences by Amazon Mechanical Turks

MSR-VTT
(MicroSoft Research
Video-To-Text)
(Xu et al. 2016)

Dataset	Context	Sentence Source	#Video	#Clip	#Sentence	#Word	Vocabulary	Duration (hrs)
YouCook [5]	cooking	labeled	88	—	2,668	42,457	2,711	2.3
TACos [25, 28]	cooking	AMT workers	123	7,206	18,227	—	—	—
TACos M-L [26]	cooking	AMT workers	185	14,105	52,593	—	—	—
M-VAD [32]	movie	DVS	92	48,986	55,905	519,933	18,269	84.6
MPII-MD [27]	movie	DVS+Script	94	68,337	68,375	653,467	24,549	73.6
MSVD [3]	multi-category	AMT workers	—	1,970	70,028	607,339	13,010	5.3
MSR-VTT-10K	20 categories	AMT workers	7,180	10,000	200,000	1,856,523	29,316	41.2

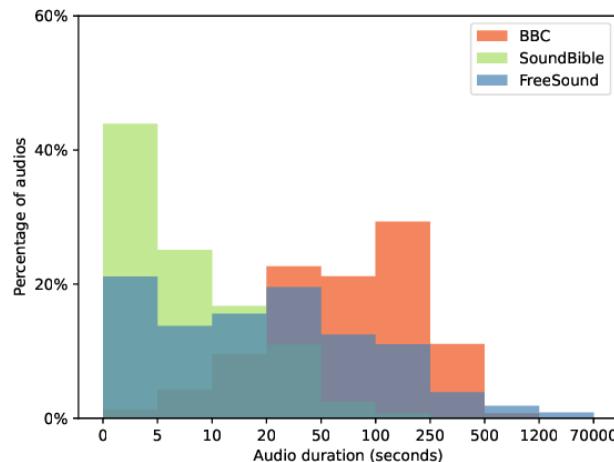


Datasets for Pretraining

Audio-Text

- Large-scale weakly-labelled audio captioning dataset (400K audio clips with paired captions)
- Collected from web sources and a sound event dataset (FreeSound, BBC, Soundbible, AudioSet)

WavCaps
Chatgpt-assisted
(Mei et al. 2023)

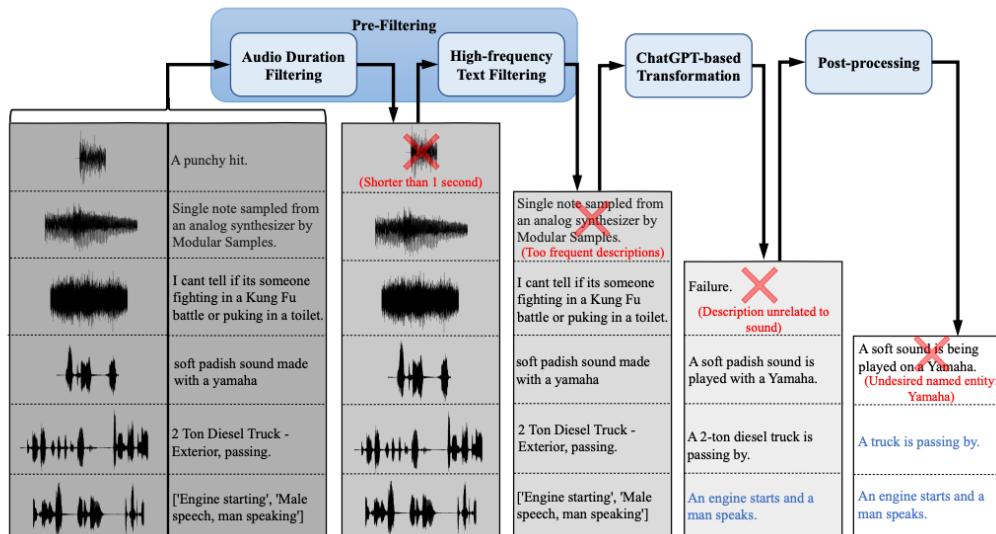


Datasets for Pretraining

Audio-Text

- Large-scale weakly-labelled audio captioning dataset (400K audio clips with paired captions)
- Collected from web sources and a sound event dataset

WavCaps
Chatgpt-assisted
(Mei et al. 2023)



Datasets for Pretraining

Audio-Text

- Large-scale weakly-labelled audio captioning dataset (400K audio clips with paired captions)
- Collected from web sources and a sound event dataset

WavCaps
Chatgpt-assisted
(Mei et al. 2023)

Data sources	Prompts
FreeSound	I will give you a list of descriptions of sounds. Process each individually. Extract the type of the sound and generate an audio caption describing the sound events. The audio caption should be less than 20 words. Delete the author of the sound. Delete locations, city names, country names. Delete the time. Delete device names. Delete the proper noun modifiers, number modifiers, and unit modifiers. Summarize each output into one sentence. Replace all named entities with their hypernyms. Replace people names with “someone”.
BBC Sound Effects	
SoundBible	Do not write introductions or explanations. Only describe the sound events and do not use “heard”, “recorded”. Start each output sentence with its index. Make sure you are using grammatical subject-verb-object sentences. Output “Failure.” if the description is not related to sound.
AudioSet SL	I will give you a number of lists containing sound events occurred sequentially in time. Process each individually. Write an one-sentence audio caption to describe these sounds. Make sure you are using grammatical subject-verb-object sentences. Directly describe the sounds and avoid using the word “heard”. The caption should be less than 20 words.

Datasets for Pretraining

Audio-Text

- Large-scale weakly-labelled audio captioning dataset (400K audio clips with paired captions)
- Collected from web sources and a sound event dataset

WavCaps
Chatgpt-assisted
(Mei et al. 2023)

Data sources	Raw descriptions	ChatGPT's output
FreeSound	This sound is of a book falling down the staircase in the Stanford University library west stacks. After 2h and 15min. Excerpt.	A book is falling down a staircase. Failure.
BBC Sound Effects	Timber & Wood - Rip saw, carpenters' workshop. Motor Car: Ford Escort 1300 (Automatic) - Ford Escort 1300, exterior, passing with horn. (1300cc engine, automatic transmission.)	Someone is using a rip saw in a carpenter's workshop. A car is passing with its horn.
SoundBible	Tasmanian Devil growling screaming hissing. Warning sounds from a Tasmanian Devil in Zoo. Large Tibetan Bells ringing in a temple. Could also use for Monastery or Monks.	An animal is growling, screaming, and hissing. Bells are ringing.
AudioSet SL	[‘Accelerating, revving, vroom’, ‘Race car, auto racing’] [‘Female speech, woman speaking’, ‘Whoosh, swoosh, swish’]	A race car is accelerating and revving. A woman is speaking while something whooshes.

Challenges of Dataset Creation

1) Data Diversity and Collection Methods (1)

Data Source Complexity

- Where to find the dataset?
- How to gather data from different modalities?

Diverse Representations:

- How to capture a wide range of real-world scenarios access all modalities?
- How to improve generalizability?

Modality Balancing:

- How to avoid the over-representation of any single modality
(balancing audio and visual content in an audio-visual dataset)

Challenges of Dataset Creation

1) Data Diversity and Collection Methods (2)

Collection Tools and Technologies

- Diverse collection methods (from web scraping for textual data to using specialized hardware)
- Vision, text, audio collections are relatively easier
- e.g. Bio-sensors, time-series dataset collection tools

Privacy and Ethical Concerns:

- How to manage the sensitive data

Challenges of Dataset Creation

2) Annotation Bottlenecks

Annotation Cost and Time

- Annotation for multimodal data is more time-intensive and cost-intensive
- How to annotate the specialised knowledge across diverse modalities

Training and Hiring Annotators

- How to manage the training to handle the complex or ambiguous multimodal data accurate

Ambiguities in Labelling Across Modalities

- How to make a standardized annotation guidelines
- Less subjective or context-dependent labels

Challenges of Dataset Creation

3) Cross-Modal Alignment

Cross-Modal Synchronization

- How to align data from multiple modalities
- e.g. aligning text captions with video frames or sensor data with timestamped audio

Annotation Tools for Multimodal Data

- What are the advanced annotation platforms capable of handling cross-modal data?
- (e.g., synchronizing and labeling text, video, and audio simultaneously)

Tutorial Outline

1. Introduction and Overview

- Welcome and tutorial objectives
- Multimodal learning
- Importance and Challenges of integrating multiple modalities (vision, language, sensors, audio)

2. Foundation of Multimodal Large Language Models (MLLMs)

- Multimodal Key Tasks and Benchmark datasets for multimodal tasks
- Challenges in Dataset Creation and Annotation for Multimodal Integration

3. Core Techniques for Multimodal Large Language Models (MLLMs)

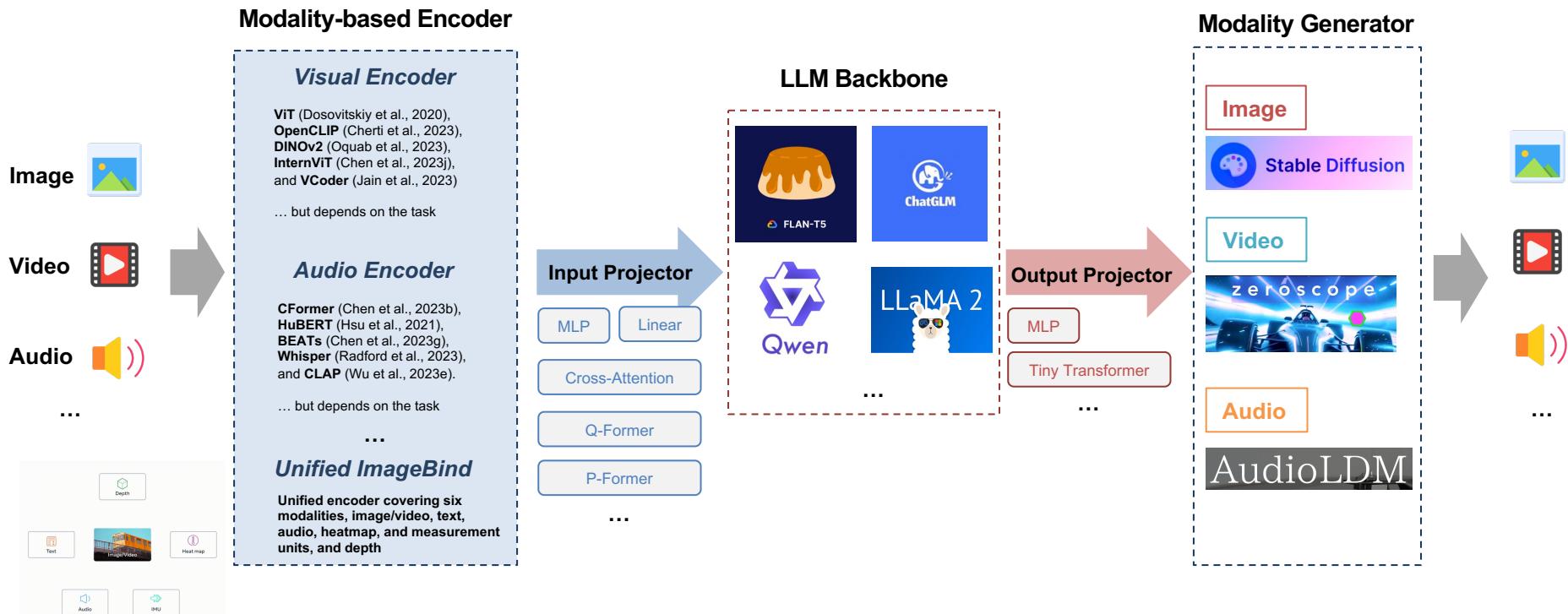
- MLLM Architecture
- MM capabilities

4. Multimodal Instruction Tunings

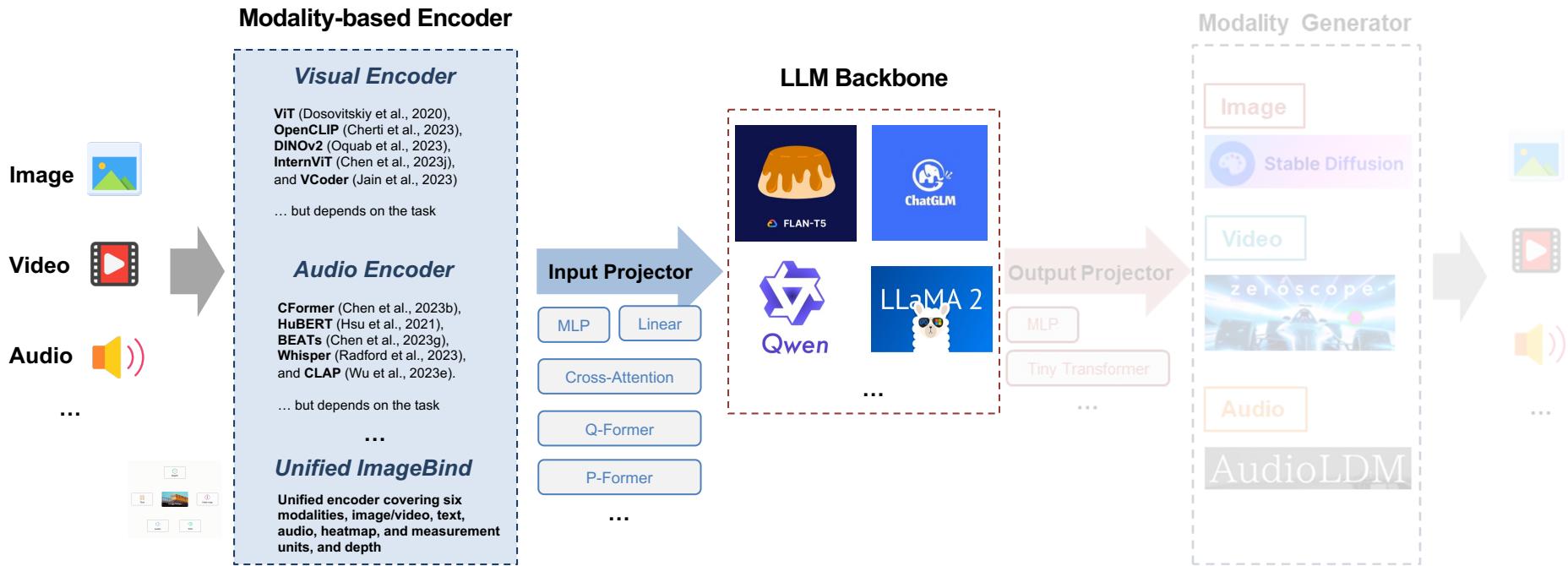
- Basic Instruction Tuning Technique to LVLM/MLLM Instruction Tuning
- Multimodal Chain-of-Thought
- Efficient Instruction Tuning Strategies

5. Future Trends and Research Opportunities

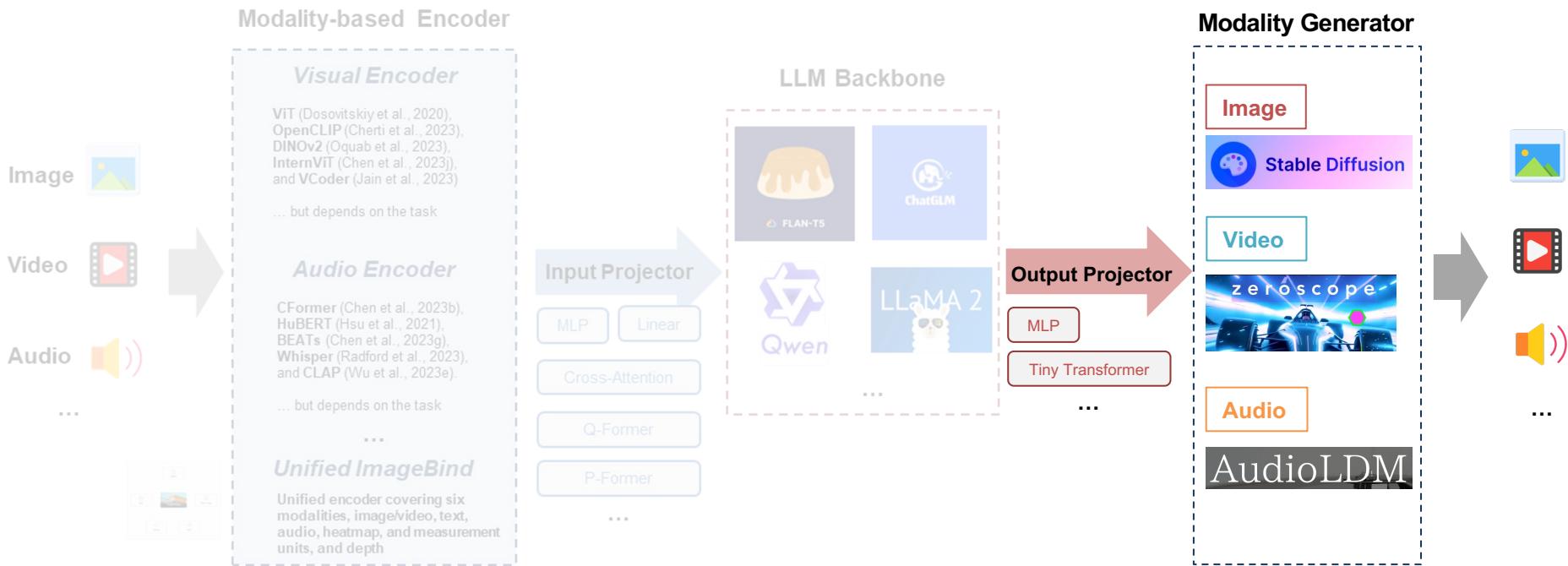
MLLMs Architecture



MLMs Architecture – Multimodal Content Understanding

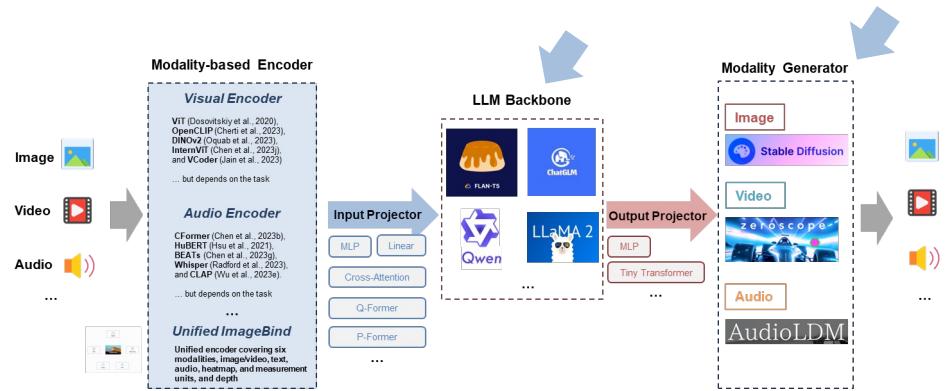


MLMs Architecture – Multimodality Generation



MLMs Architecture – Challenges (1)

Dependency on Pretrained Backbones and Frozen Generators

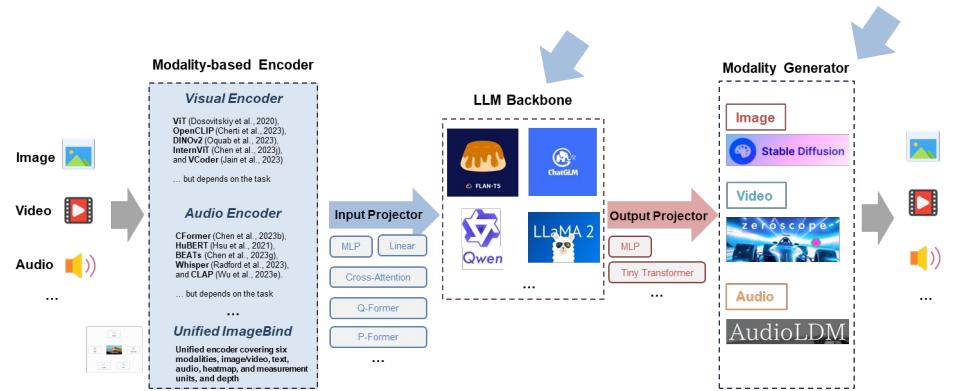


MLLMs maintain their LLM Backbone and Modality Generators in a frozen state

- Their ability to adapt to new, nuanced multimodal interactions is limited
- The fixed configuration means that any updates to the model's understanding of modality interactions require complex workarounds.

MLMs Architecture – Challenges (1)

Dependency on Pretrained Backbones and Frozen Generators



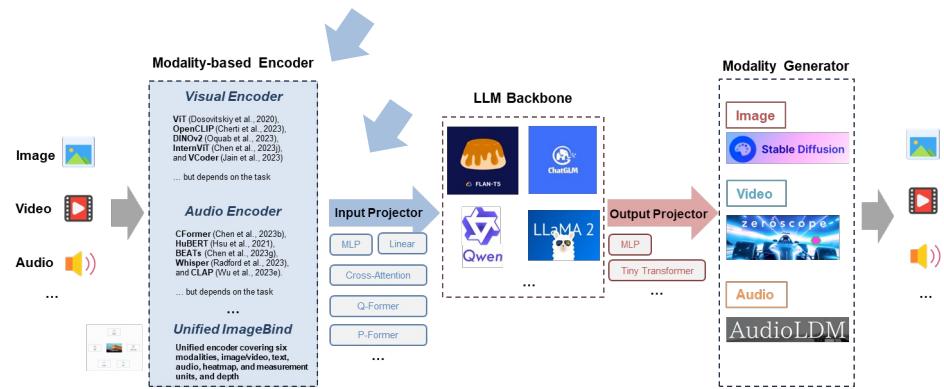
MLLMs maintain their LLM Backbone and Modality Generators in a frozen state

- Their ability to adapt to new, nuanced multimodal interactions is limited
- The fixed configuration means that any updates to the model's understanding of modality interactions require complex workarounds.

Potential Solution: We may develop **Adaptable Backbone Architectures** that allow selective fine-tuning or partial unfreezing of specific layers within the LLM Backbone or Modality Generators. (e.g. PEFT or LoRA)

MLMs Architecture – Challenges (2)

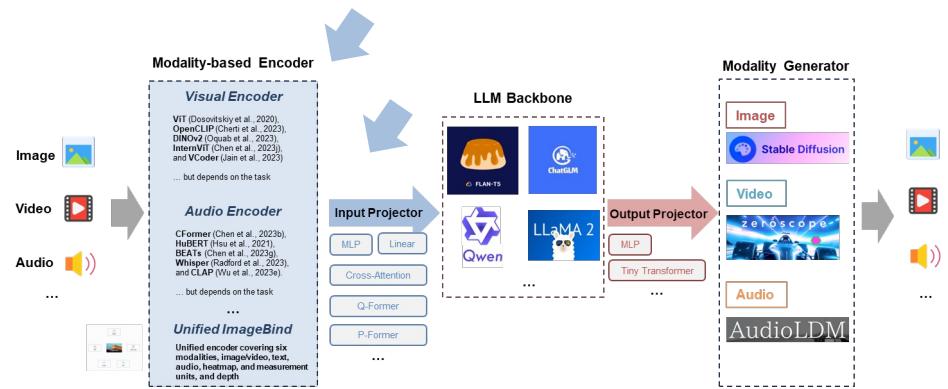
High Computational Costs



- Processing diverse modalities through separate Encoders and Projectors before unifying the information in the LLM Backbone requires significant computational resources, especially with large-scale data

MLMs Architecture – Challenges (2)

High Computational Costs

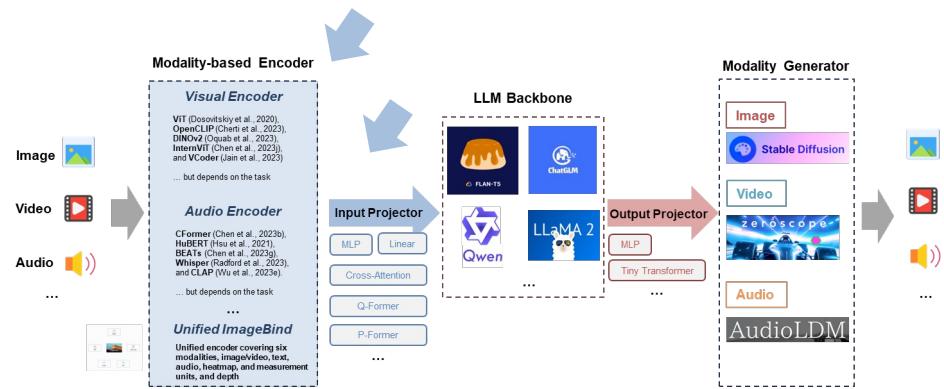


- Processing diverse modalities through separate Encoders and Projectors before unifying the information in the LLM Backbone requires significant computational resources, especially with large-scale data

Potential Solution: Mixture-of-Experts (MoE) techniques can be used to activate only relevant parts of the model for specific tasks or modalities, thus lowering resource demands.

MLMs Architecture – Challenges (3)

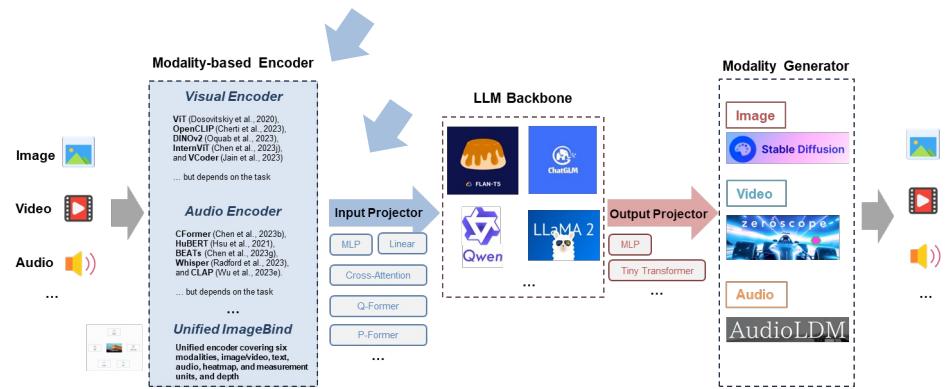
Limited Flexibility with New or Evolving Modalities



The static design of Input Projectors and Modality Encoders for specific modalities **restricts the system's ability to integrate new or uncommon modalities (e.g., biosignals, environmental data)**

MLMs Architecture – Challenges (3)

Limited Flexibility with New or Evolving Modalities

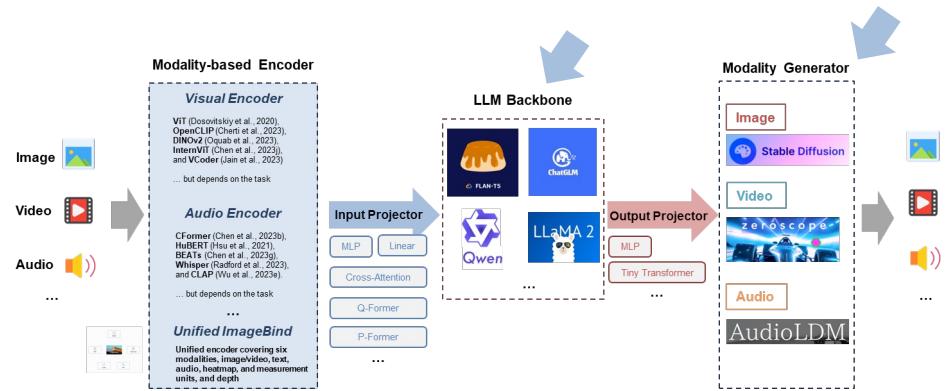


The static design of Input Projectors and Modality Encoders for specific modalities **restricts the system's ability to integrate new or uncommon modalities (e.g., biosignals, environmental data)**

Potential Solution: Develop a Plug-and-Play Modality Integration Framework where dynamic modality adapters or universal encoders can process new or evolving data types without retraining the entire model.

MLMs Architecture – Challenges (4)

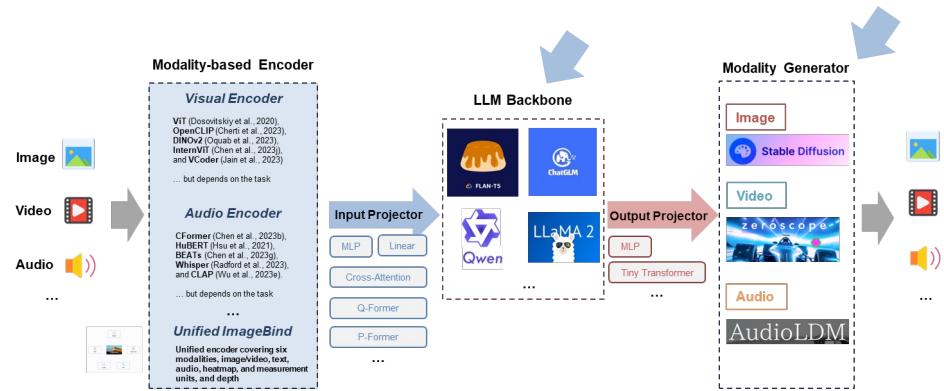
Evaluation and Consistency Across Modalities



- The evaluation metrics and methods for assessing multimodal tasks are not yet standardized, especially for models using frozen backbones and generators.
- Maintaining consistency in how different modalities contribute to the output can be challenging, particularly in dynamic scenarios.

MLMs Architecture – Challenges (4)

Evaluation and Consistency Across Modalities



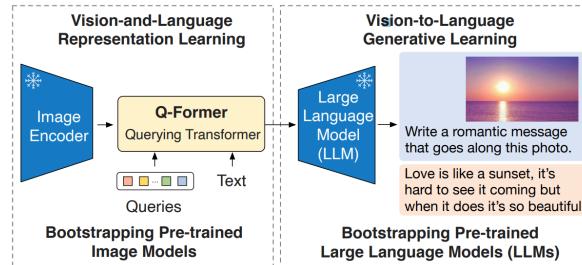
- The evaluation metrics and methods for assessing multimodal tasks are not yet standardized, especially for models using frozen backbones and generators.
- Maintaining consistency in how different modalities contribute to the output can be challenging, particularly in dynamic scenarios.

Potential Solution: Evaluation Metrics for Multimodal Tasks by developing metrics that consider modality-specific contributions, such as Multimodal BLEU or Multimodal F1 Scores

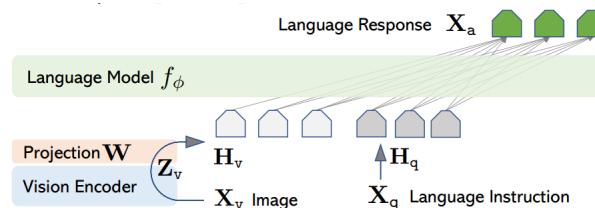
MM Capabilities – Content Understanding and Text Generation

Image-text understanding

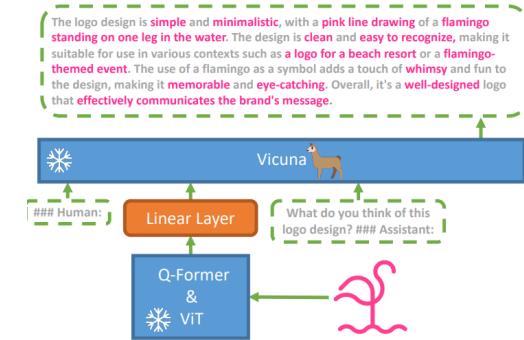
BLIP-2 (Li et al. 2023)



LLaVa (Liu et al. 2023)

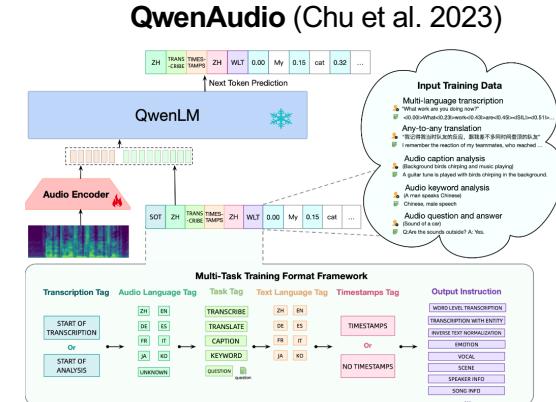
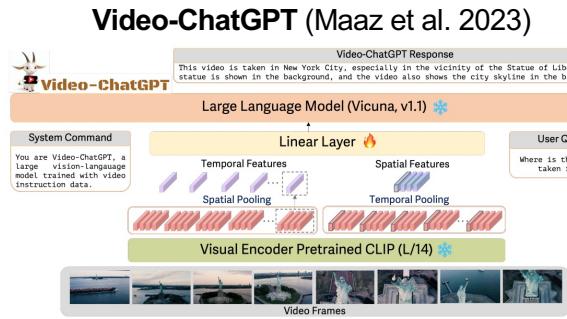
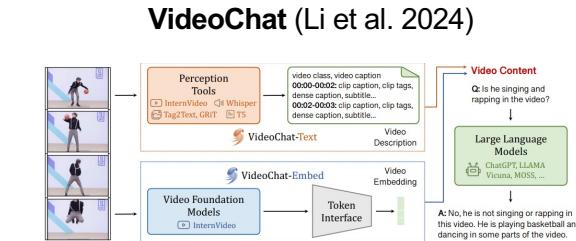
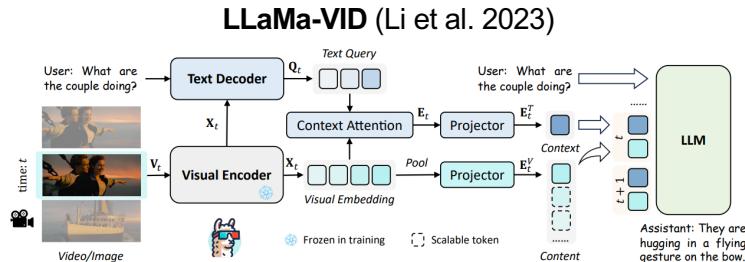


MiniGPT4 (Zhu et al. 2023)



MM Capabilities – Content Understanding and Text Generation

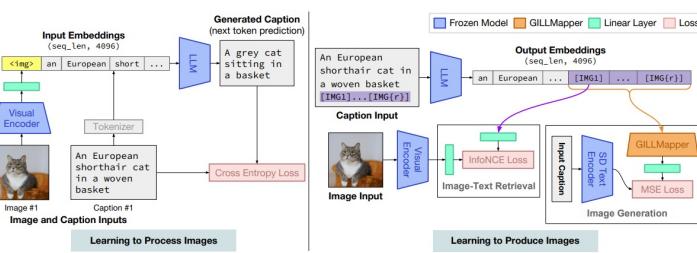
video-text understanding & audio-text understanding



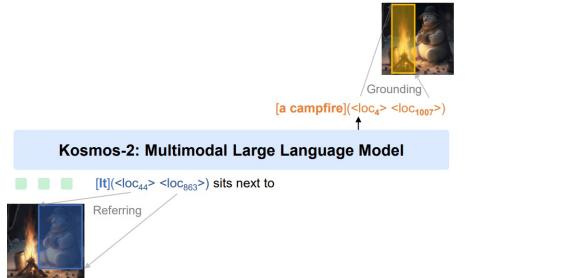
MM Capabilities – Grounding and Referring

image-text output

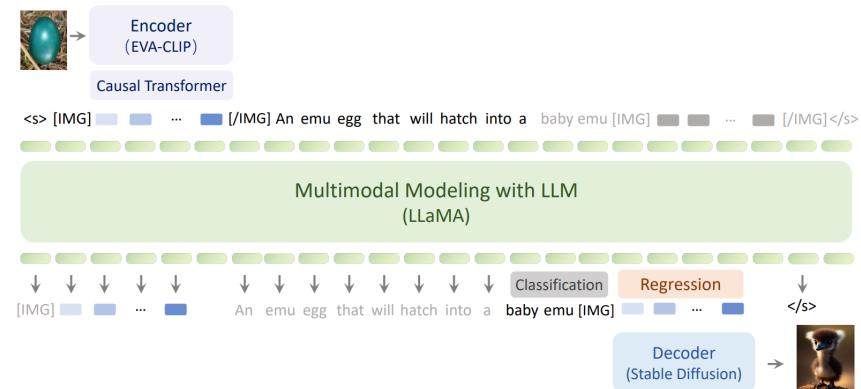
GILL (Koh et al. 2023)



Kosmos-2 (Koh et al. 2023)



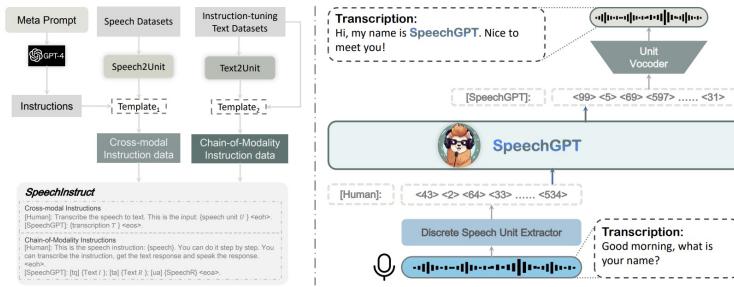
Emu (Sun et al. 2024)



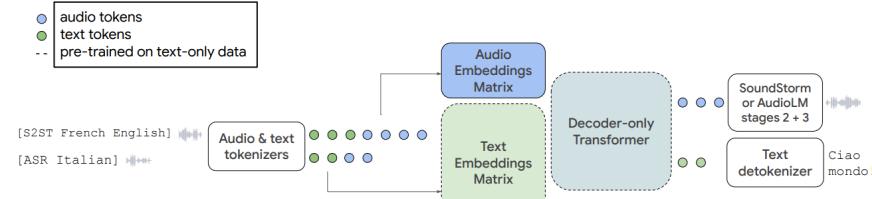
MM Capabilities – Grounding and Referring

speech/audio-text output

SpeechGPT (Zhang et al. 2023)



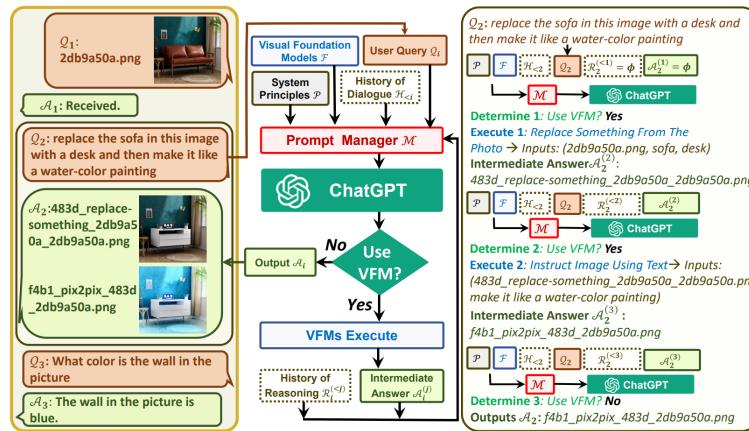
AudioPaLM (Rubenstein et al. 2024)



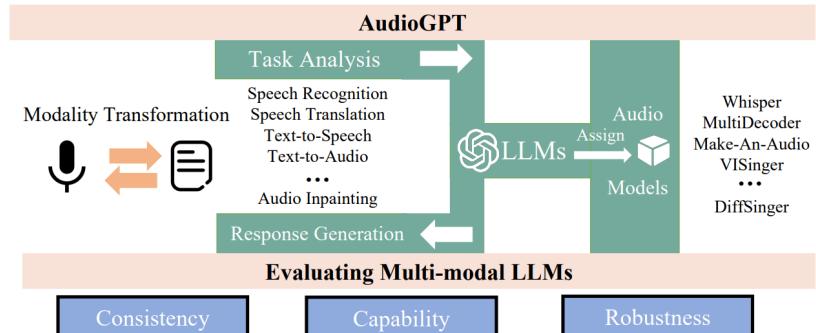
MM Capabilities – Any-to-Any Modalities

Any-to-any MM comprehension and generation

Visual-ChatGPT (Wu et al. 2023)



AudioGPT (Huang et al. 2023)



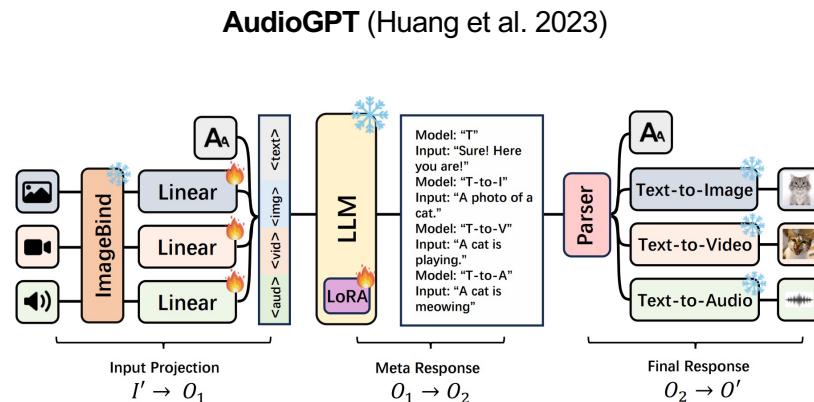
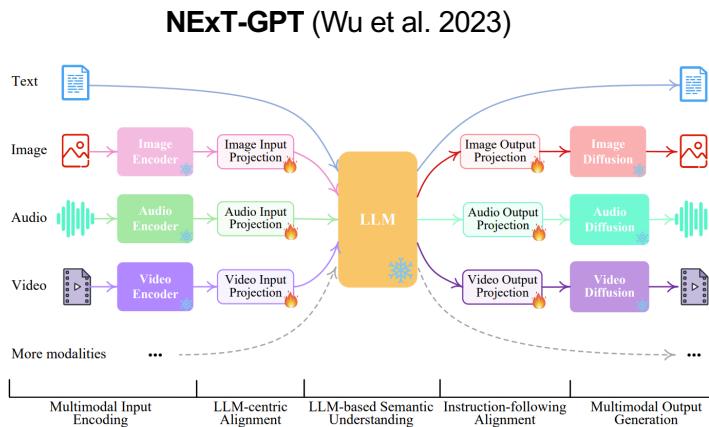
Consistency

Capability

Robustness

MM Capabilities – Any-to-Any Modalities

End-to-End MLLMs



Tutorial Outline

1. Introduction and Overview

- Welcome and tutorial objectives
- Multimodal learning
- Importance and Challenges of integrating multiple modalities (vision, language, sensors, audio)

2. Foundation of Multimodal Large Language Models (MLLMs)

- Multimodal Key Tasks and Benchmark datasets for multimodal tasks
- Challenges in Dataset Creation and Annotation for Multimodal Integration

3. Core Techniques for Multimodal Large Language Models (MLLMs)

- MLLM Architecture
- MM capabilities

4. Multimodal Instruction Tunings

- Basic Instruction Tuning Technique to LVLM/MLLM Instruction Tuning
- Multimodal Chain-of-Thought
- Efficient Instruction Tuning Strategies

5. Future Trends and Research Opportunities

Second Speaker



- Research interests include explainable artificial intelligence, visual question answering, multimodal learning and pretraining.
- Visiting scholar at the University of Western Australia
- Collaborating with the Sapienza University of Roma for multimodal LLM research, and Congenital Heart Alliance of Australia and New Zealand for heart disease detection from clinical letters.
- Published several papers at top-tier venues (IJCAI, AAAI, COLING, and Robotics Journal)

Feiqi Cao

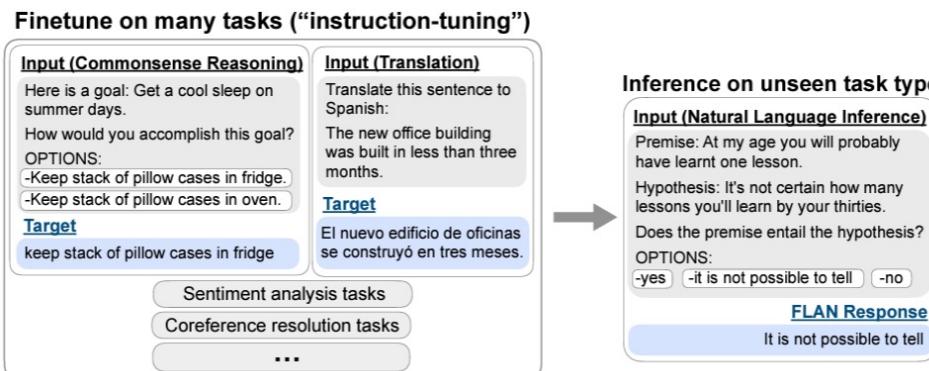
3rd Year PhD,

The University of Sydney

Part 4: Multimodal Instruction Tunings

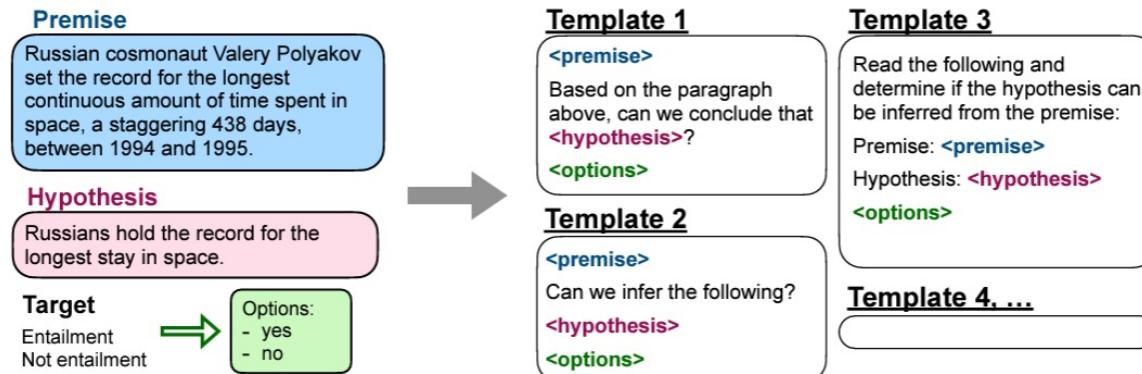
Instruction Tuning

- An approach to train a pretrained model to perform specific tasks based on the provided task descriptions which outline how a task should be performed (i.e. instructions)
- Why do we need Instruction Tuning?
 - Cross-task generalizability is bad
 - Previous framework cannot generalize well to unseen tasks
 - Aim: model can understand human instructions describing how a specific task should be performed and solve unseen tasks based on that, and thus improves the zero-shot generalization ability.



Instruction Tuning

- How it works
 - Mixture of <**Instruction**, **Input**, **Target**> from diverse tasks, with diverse templates to describe each type of task as the instruction.
 - Tune the model by giving <**Instruction**, **Input**> as the input to generate **Target**



LVLM Instruction Tuning

How does instruction tuning works for Language + Vision (Image)

- Freeze main LLM/Vision Foundation Models and finetune some layers/modules only.
- Self-construct a large vision-language instruction tuning corpus based on current vision-language downstream task benchmarks



[vqa] Which country does it come from?



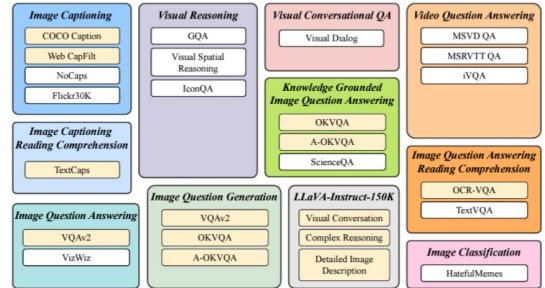
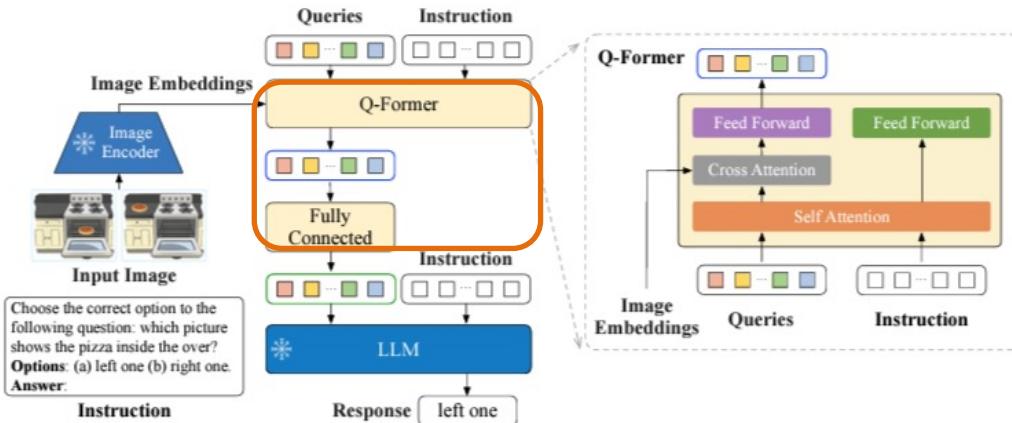
australia

LVLM Instruction Tuning

InstructBLIP

Finetune bridging module (Q-Former) for language and image

- *Instruction tasks: QA, image captioning, visual reasoning, etc.*
- *Multiple predefined instruction templates for each task*



Instruction examples

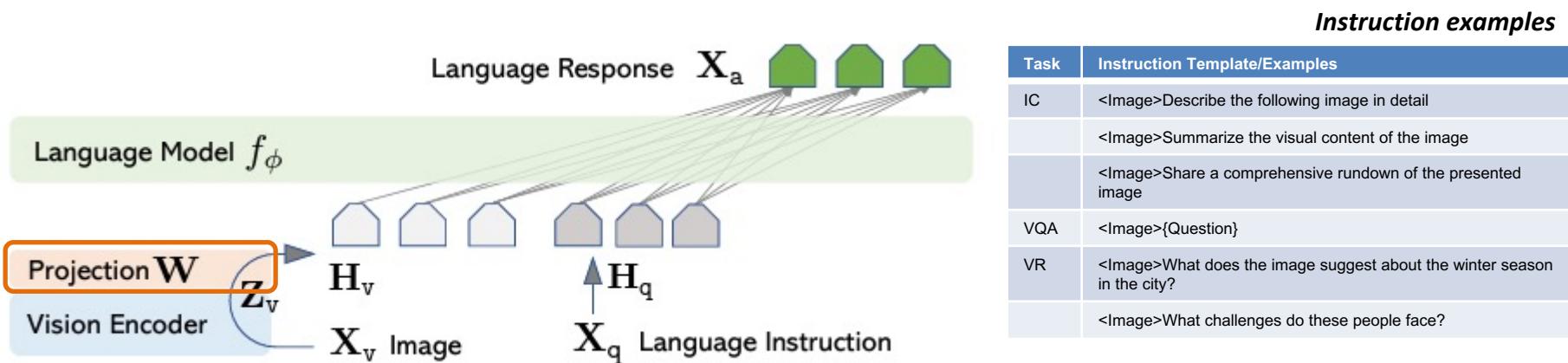
Task	Instruction Template
IC	<Image>Write a short description for the image.
	<Image>A short image caption:
	<Image>A photo of
VQA	<Image>{Question}
	<Image>What is the answer to the following question? {Question}"
	<Image>Based on the image, respond to this question: {Question}
VQG	<Image>Given the image, generate a question whose answer is:{Answer}. Question:
	<Image>Based on the image, provide a question with the answer:{Answer}. Question:
	<Image>Considering the picture, come up with a question where the answer is: {Answer}.

LVLM Instruction Tuning

LLaVA

Finetune image projection layer

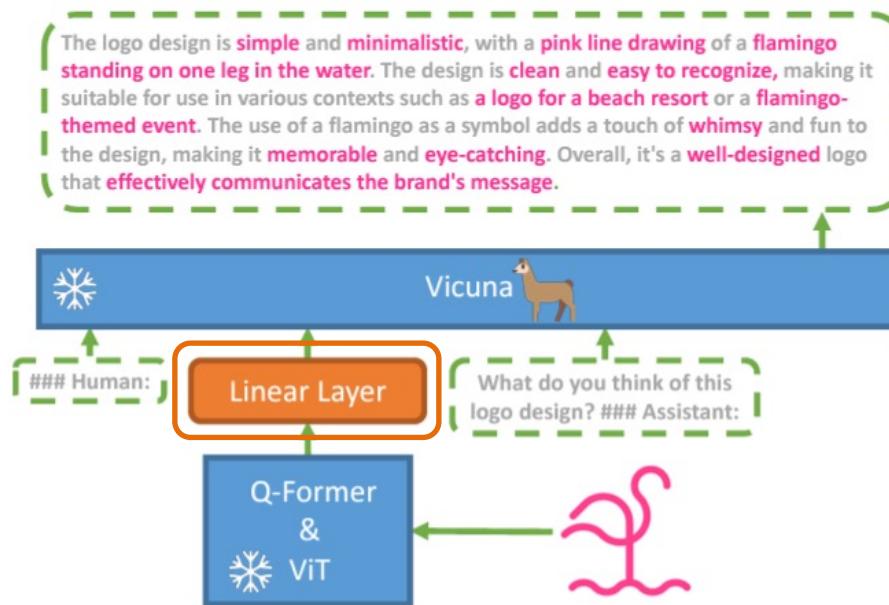
- *Instruction tasks: QA, image captioning, visual reasoning (158k instances)*



LVLM Instruction Tuning

MiniGPT4

Finetune image projection layer



Instruction examples

Task	Instruction Template/Examples
IC	<Image>Describe the following image in detail <Image>Describe this image in detail
VQA	<Image>(Question)
VR	<Image>Explain why this meme is funny <Image>What challenges do these people face?
Other	<Image> What is the issue with my plant? What should I do? <Image> Can you give a short introduction to this movie? <Image> Can you write me a beautiful poem about this image? <Image> Write brief HTML/JS to turn this mock-up into a colorful website.

LVLM Instruction Tuning

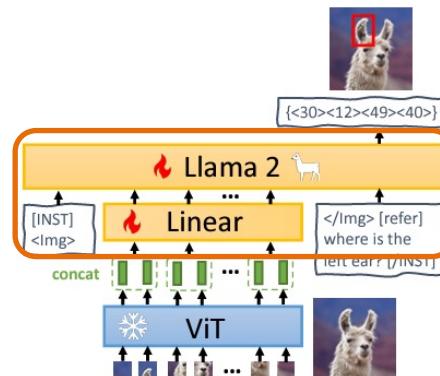
MiniGPT4 – v2

Finetune image projection layer + Parameter Efficient Finetuning for Llama

- More tasks and training instances with structured instruction template

[INST] <ImageFeature> [Task Identifier] Instruction [/INST]

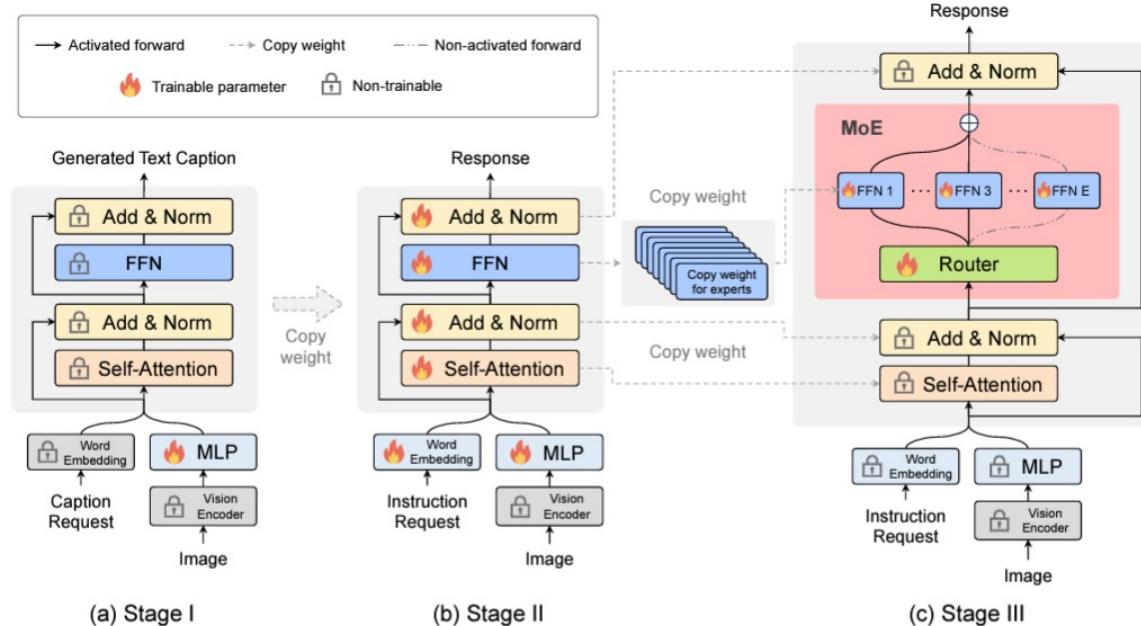
Tasks	VQA	Caption	Grounded Caption	REC	REG	Object Parsing and Grounding
Identifiers	[vqa]	[caption]	[grounding]	[refer]	[identify]	[detection]



LVLM Instruction Tuning

MoE-LLaVA

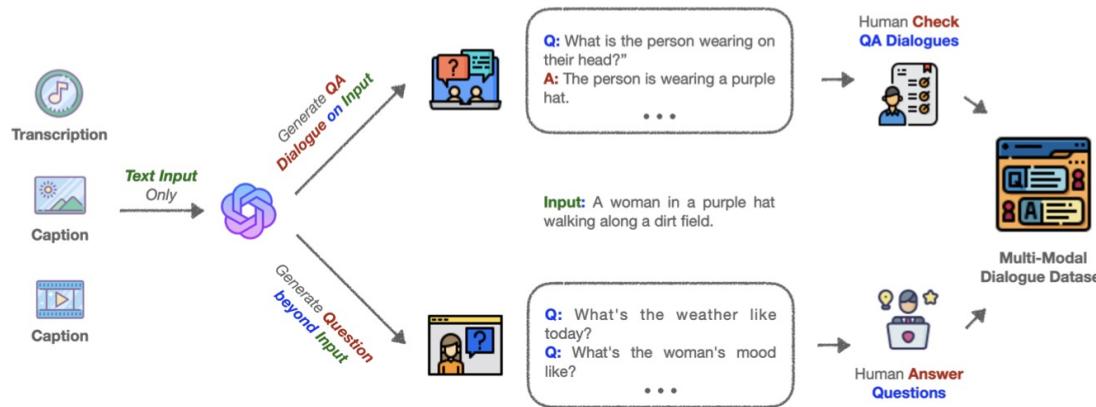
Finetune different experts based on different instructions and tasks



MLM Instruction Tuning

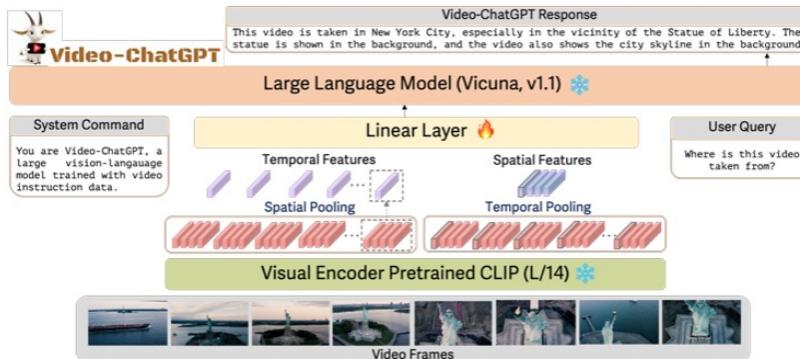
Data Curation

- A Existing downstream dataset + human predefined instruction templates
- B Synthetic data by deep learning models or LLM
- C Hybrid of human annotation and synthetic data

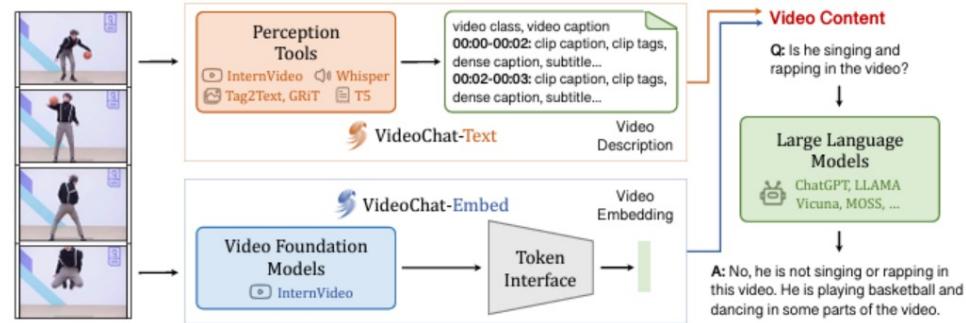


MLM Instruction Tuning

Video



Video-ChatGPT



MLLM Instruction Tuning

Video



Original Dataset Annotation:

A man is playing the bagpipes in front of people. The people on the couch in front of him start laughing.

Semi-Automatic Annotation Pipeline:

Frame-level Tag2Text Tags:

1. table, couch, woman, living room, chair, girl, sit, wear, sit on, pink, red
2. couch, woman, chair, people, show, person, man, sit, sit in, sit on
3. couch, woman, blue, shirt, girl, sit, it on, talk
4. microphone, person, man, black, shirt, hold, stand
5. pipe, black, shirt, tv, person, man, play, stand

Frame-level GRIT Descriptions:

1. a women in a pink outfit
2. man sitting on couch, a woman sitting on a couch, couch with flower pattern
3. woman wearing a blue sweater, woman in pink shirt on phone
4. a man in a black shirt, a green plant next to a couch
5. a man in a black shirt, a green plant next to the man

Generating Semi-automatic Annotation – Video Instructional Data

Q: Can you summarize the events in the video?

A: Yes, the video shows a man playing the bagpipes in front of a group of people sitting on a couch. They start laughing at some point, creating a friendly atmosphere.

Q: Can you describe the appearance of the bagpipes player in the video?

A: Yes, the bagpipes player is dressed in a black shirt and holding the bagpipes in front of a television.

Frame-level BLIP2 Captions:

1. three women are sitting on a couch
2. a man and woman sitting on a couch
3. two women sitting on a couch laughing
4. a man holding a bagpipes in front of a television
5. a man playing the bagpipes in a living room

Q: Who are the people sitting on the couch in front of the bagpipes player?

A: There are three women and a man sitting on the couch.

Q: Can you describe the setting and atmosphere in the video?

A: The video captures a relaxed and friendly atmosphere in a domestic setting. Specifically, it shows a group of people sitting on a couch chatting and laughing while a man in a black shirt plays the bagpipes in front of them. The people are dressed casually, and in some scenes, a woman in a pink outfit, another dressed in blue sweater sits next to a man on the couch and continue to chat.

Video-ChatGPT

C

Video Class, Video Caption
00:00-00:02 Clip Caption, Clip Tag, Dense Caption, Video Subtitle...
00:02-00:03 Clip Caption, Clip Tag, Dense Caption, Video Subtitle...
00:03-00:06 Clip Caption, Clip Tag, Dense Caption, Video Subtitle...



answering questions, a man and a woman sitting on a couch in a living room with a table in front of them.

00:00-00:11 a man and a girl sitting on a couch in a living room.
a lamp with a white shade a woman sitting at a table: [446, 155, 710, 476]; man wearing a plaid shirt: [361, 44, 581, 337]; man sitting on couch: [10, 63, 324, 350]; the tie is grey: [441, 150, 486, 280]; a glass of beer: [38, 305, 77, 367]; a stack of magazines: [28, 350, 180, 394]; a white tablecloth: [0, 334, 626, 476]; stainless steel oven: [1, 55, 150, 142]; a brown tie on a man: [144, 168, 191, 270]; the couch is white: [0, 119, 730, 472]; a gray binder: [0, 377, 157, 411]; a white couch: [768, 350, 848, 477]; a lamp with a white shade: [582, 26, 713, 195];

00:00-00:02: Hey, Pheebs, you gonna have the rest of that Pop-Tart?

00:02-00:03: Pheebs?

00:03-00:09: Does anyone want the rest of this Pop-Tart?

00:09-00:11: Hey, I might.

B



ChatGPT

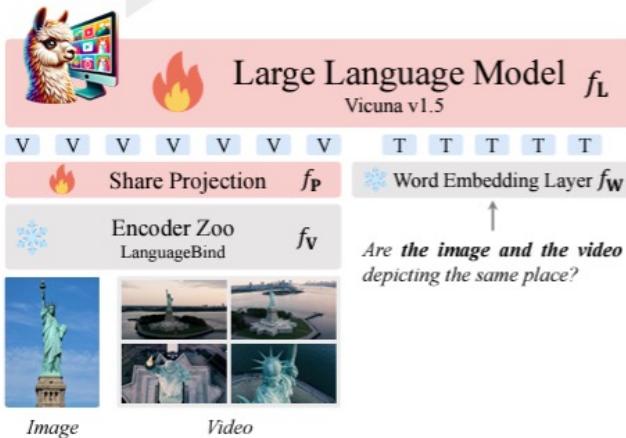
Video-Chat

MLLM Instruction Tuning

Video + others

A C

Yes, the image and the video are depicting the same place. **The video shows the statue of liberty from different angles**, while **the image shows a close-up of the statue**. Both the video and the image capture the beauty and grandeur of the statue of liberty.



Video-LLaVA

A B

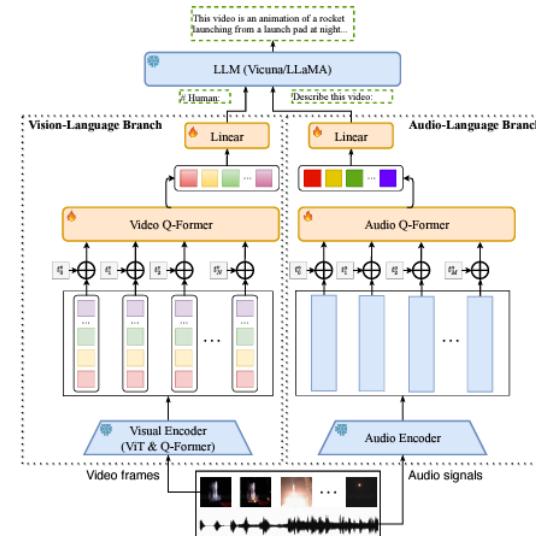
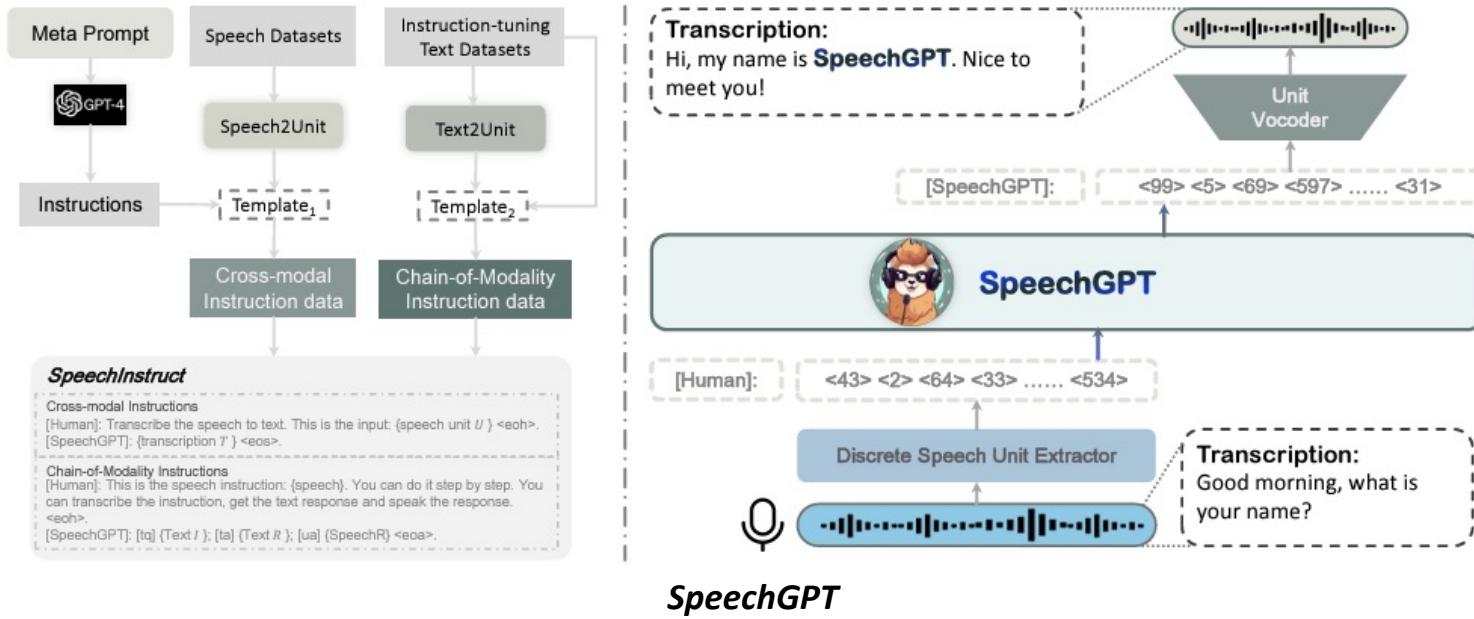


Figure 1: Overall architecture of Video-LLaMA.

Video-LLaMA

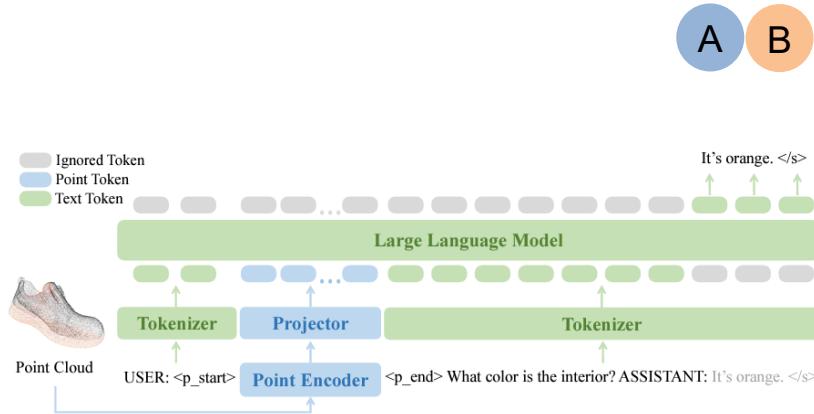
MLM Instruction Tuning

Audio

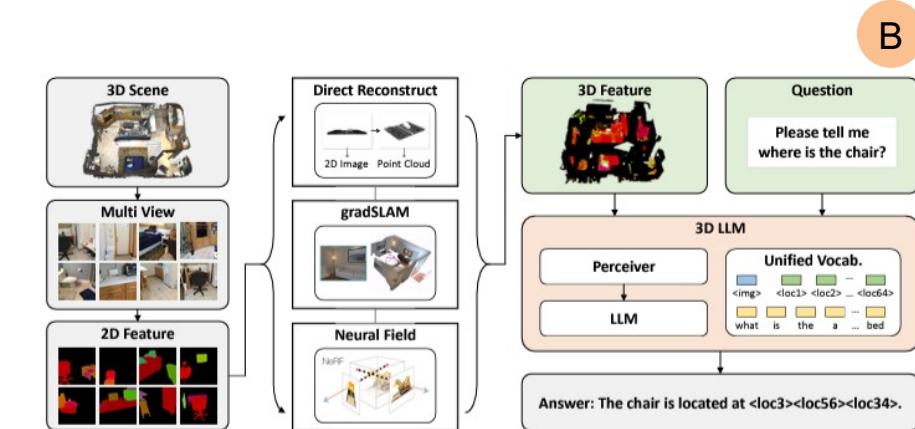


MLLM Instruction Tuning

3D Data



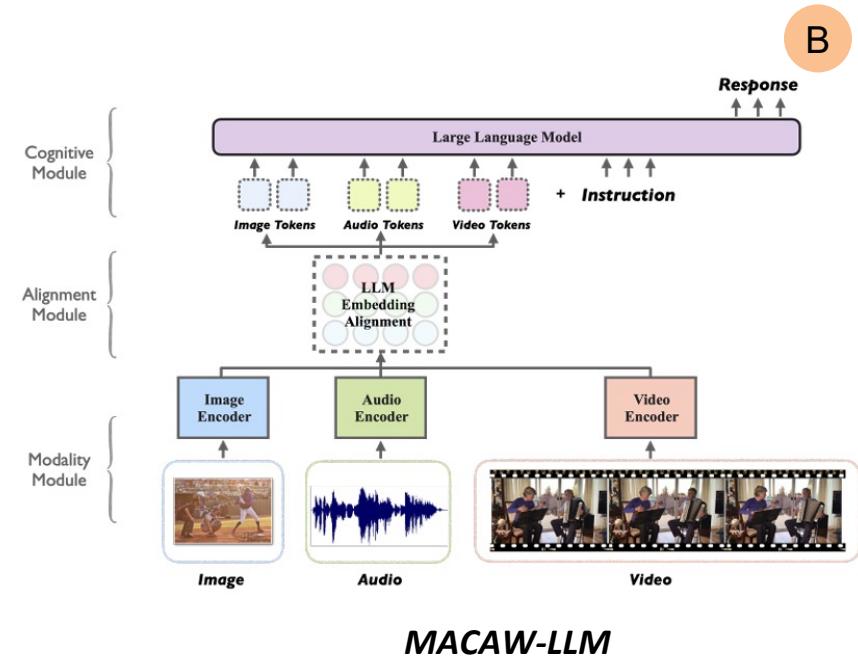
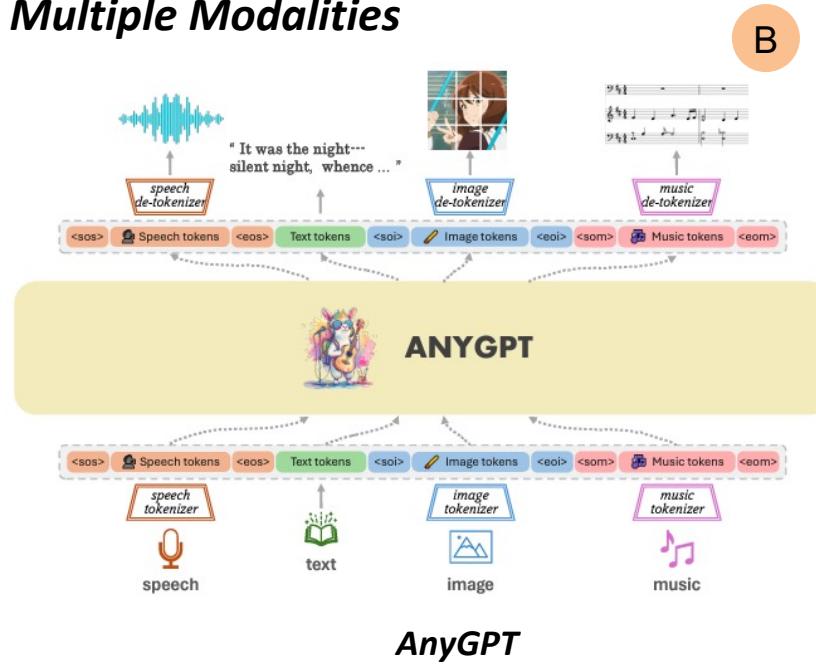
PointLLM



3D-LLM

MLLM Instruction Tuning

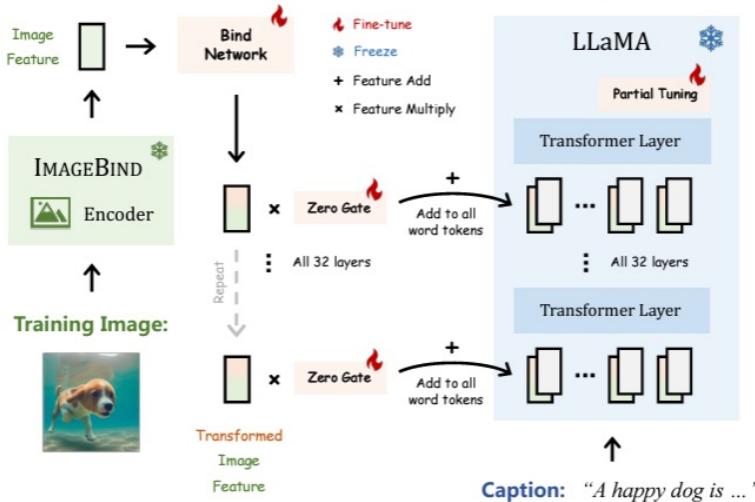
Multiple Modalities



MLM Instruction Tuning

Multiple Modalities

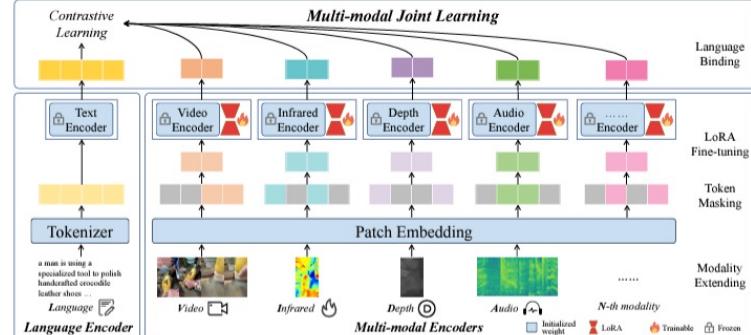
Training Paradigm of ImageBind-LLM



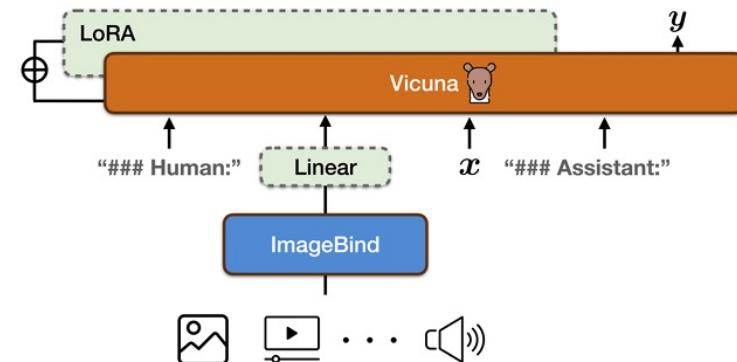
ImageBind



"... swimming"



LanguageBind



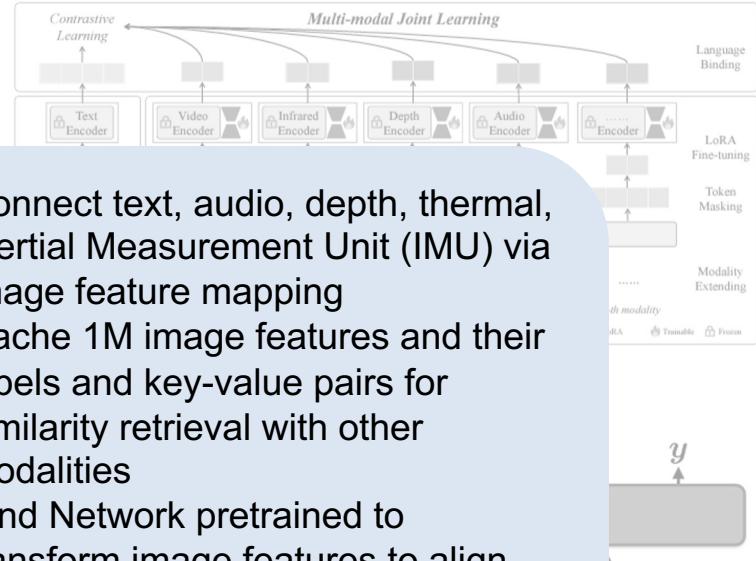
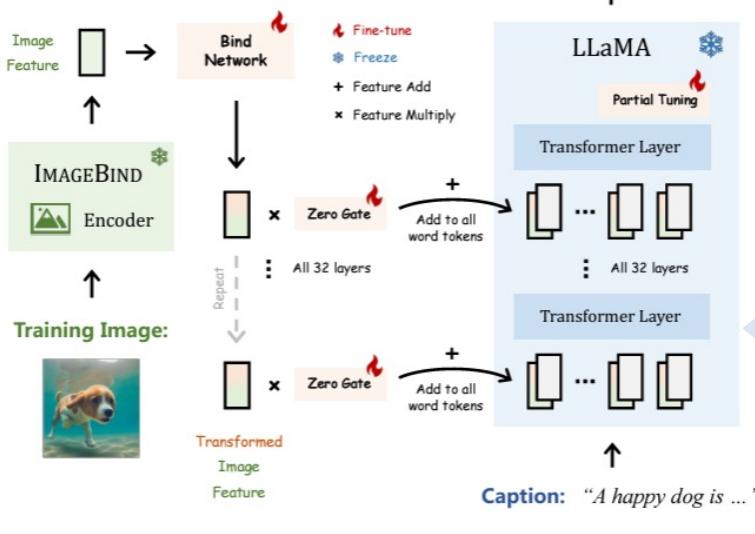
PandaGPT

B

MLLM Instruction Tuning

Multiple Modalities

Training Paradigm of ImageBind-LLM



- Connect text, audio, depth, thermal, Inertial Measurement Unit (IMU) via image feature mapping
- Cache 1M image features and their labels and key-value pairs for similarity retrieval with other modalities
- Bind Network pretrained to transform image features to align with LLaMA embedding space
- Additional finetuning with previously collected image caption data



PandaGPT

MLM Instruction Tuning

Multiple Modalities

Training Par

Image Feature

IMAGEBIND

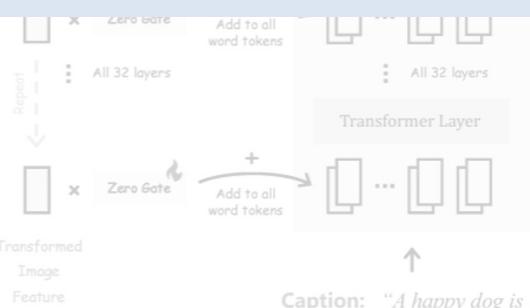
Encoder

↑

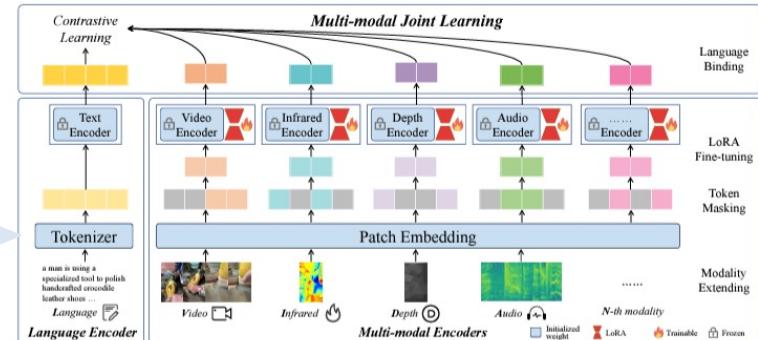
Training Image:



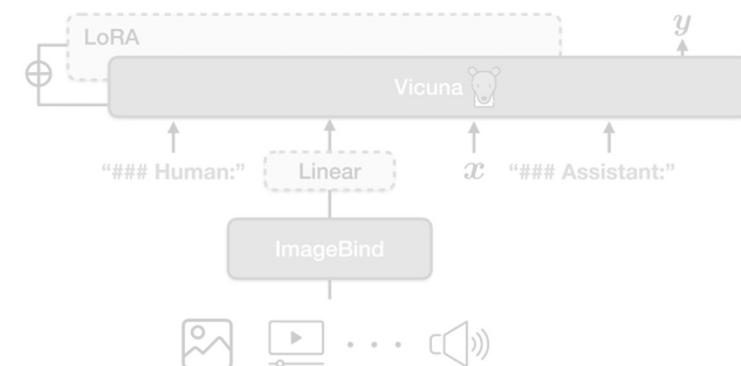
- Bind any modality to language via contrastive learning
- Construct the first large-scale five-modal dataset VIDAL-10M for tuning



ImageBind



LanguageBind

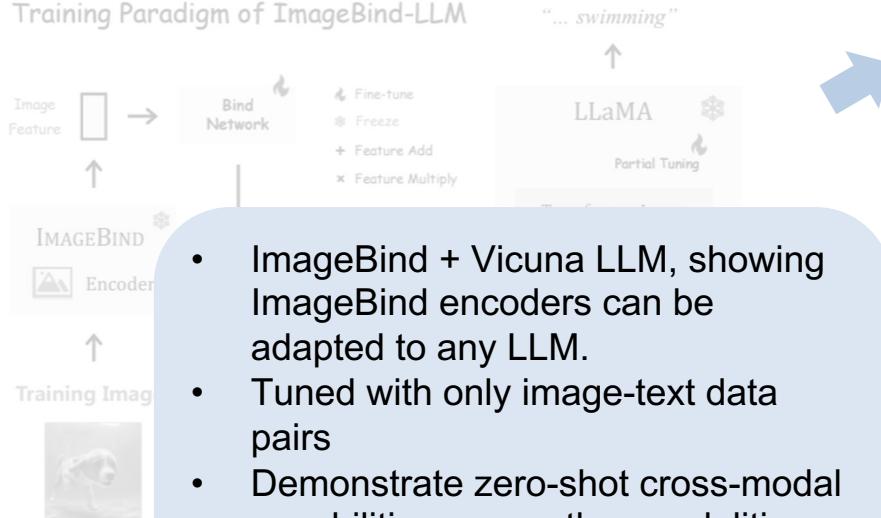


PandaGPT

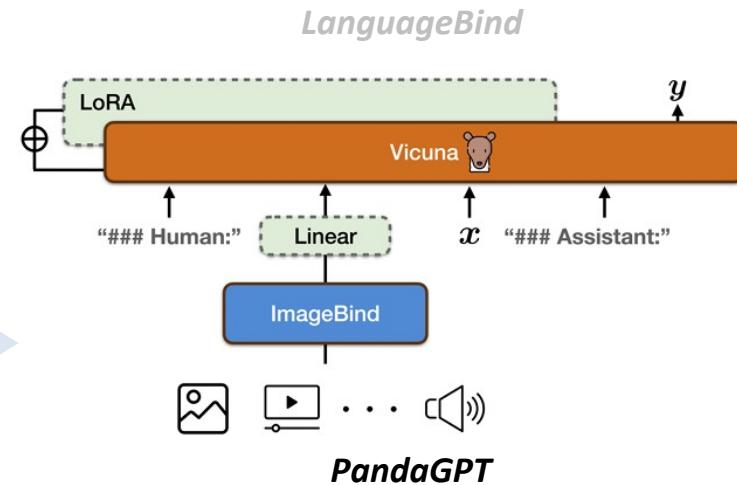
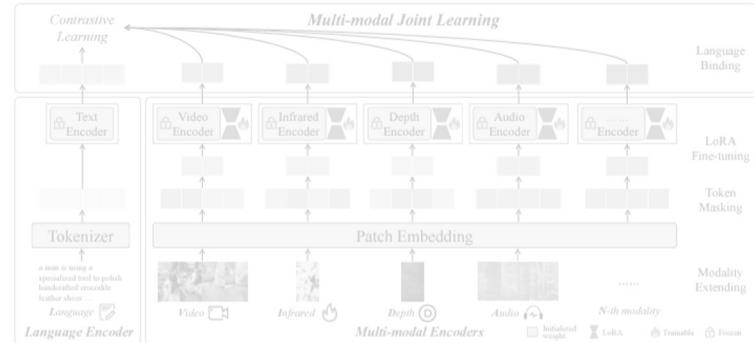
MLM Instruction Tuning

Multiple Modalities

Training Paradigm of ImageBind-LLM



- ImageBind + Vicuna LLM, showing ImageBind encoders can be adapted to any LLM.
- Tuned with only image-text data pairs
- Demonstrate zero-shot cross-modal capabilities across other modalities, for tasks involving video and audio data.



Other Specific-Domain Instruction Tuning

- *Mobile agent*
- *Document understanding*
- *Medical*
- *Gaming*
- *Driving*
- *etc...*

Chain-of-Thought (CoT)

A prompting technique to invoke the generation of reasoning procedure for LLM

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. **X**

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. **✓**

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 **X**

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. **✓**

Multimodal Chain-of-Thought (CoT)

A tuning technique to introduce reasoning for MLLM through rationale generation

An example of detailed reasoning steps in GQA dataset

Question: What appliance is to the right of the cabinet?

###

Please think step by step and provide the bounding box coordinate of the region that can help you answer the question better.

###

- Reasoning steps:**
1. Identify the cabinet in the image.
 2. Observe the area to the right of the identified cabinet.
 3. Look for any appliance located to the right side of the cabinet.
 4. Determine the name of the appliance found in this location

CoT BBox: [163, 44, 206, 67]



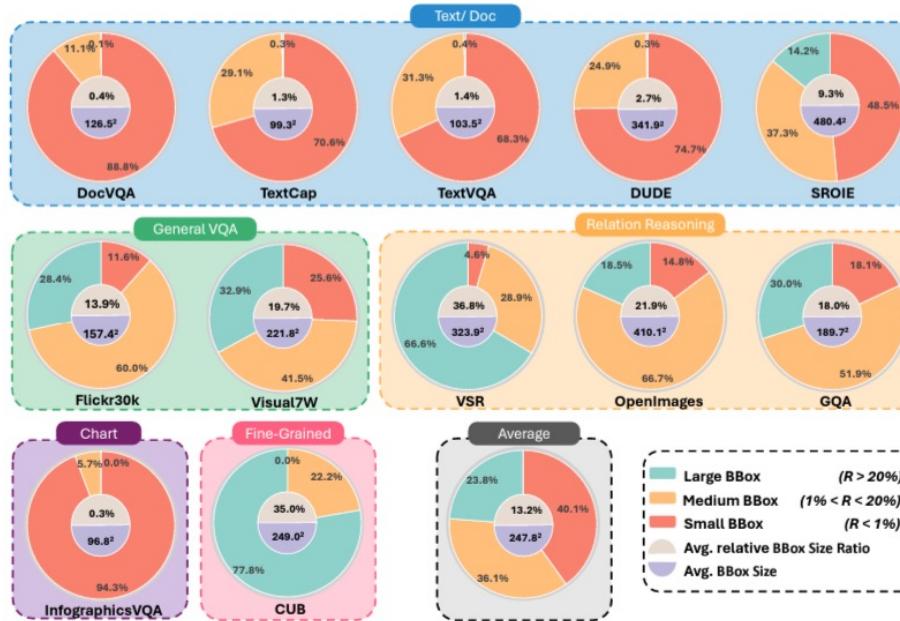
Answer

The appliance is a microwave.

Multimodal Chain-of-Thought (CoT)

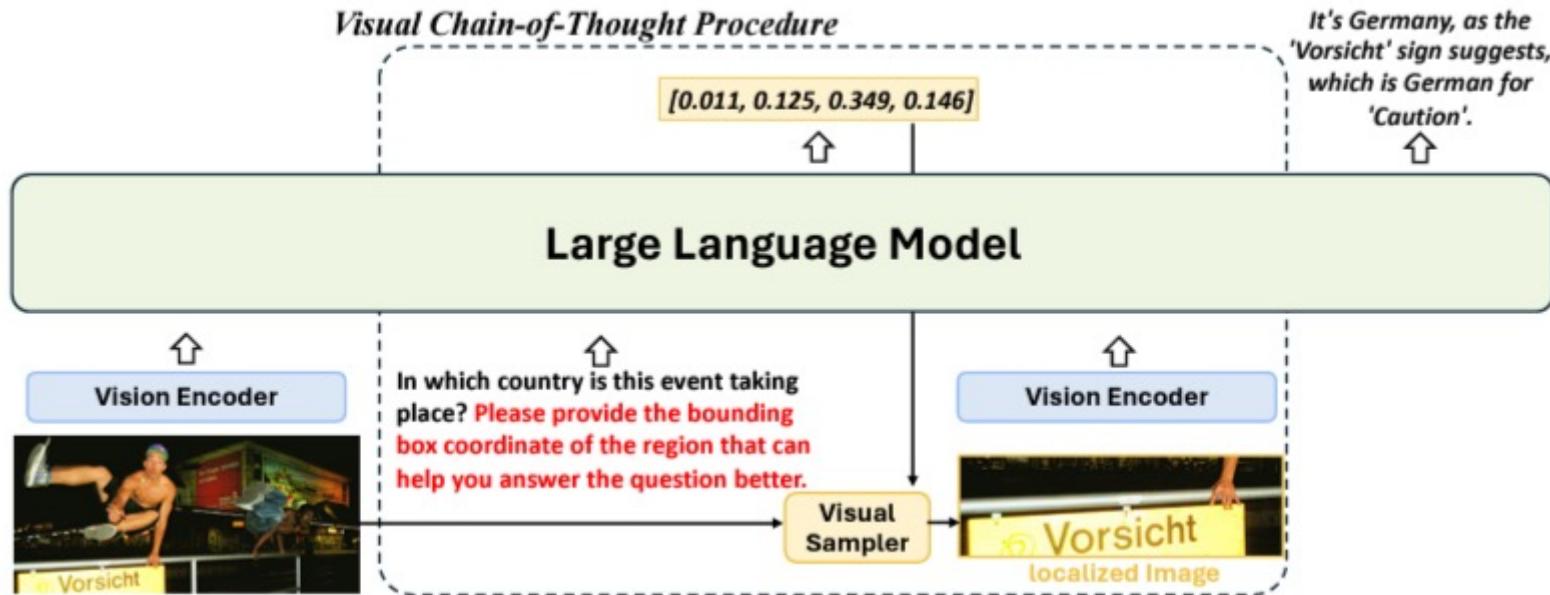
VisCoT

Focus on improving response accuracy via identifying critical visual regions/elements



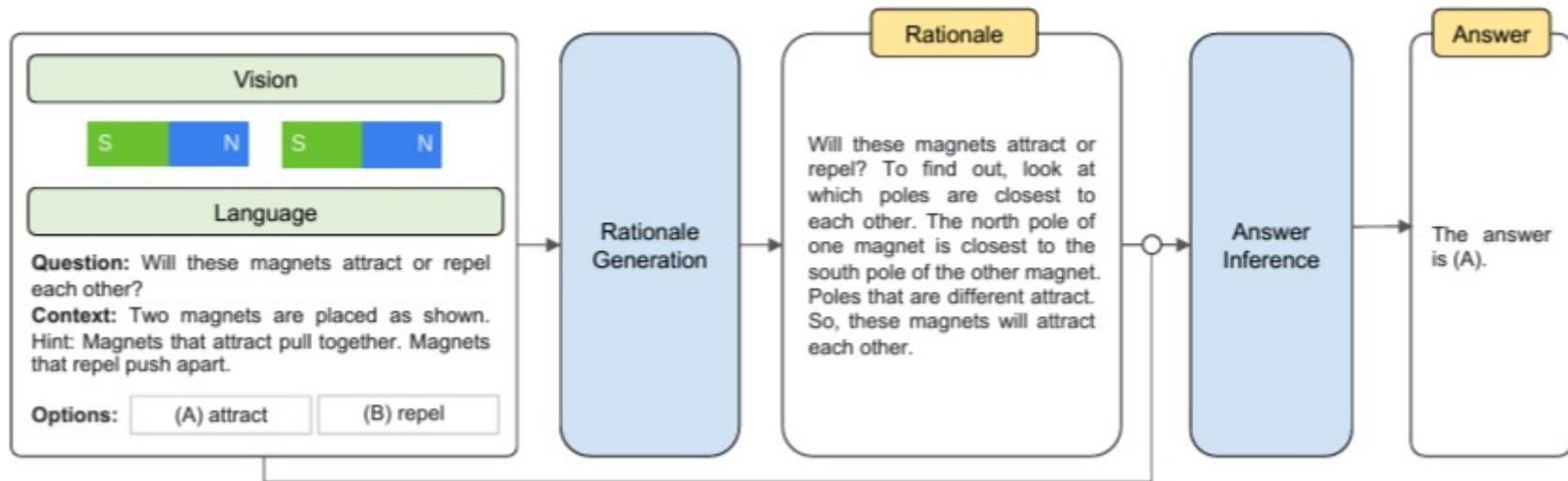
Multimodal Chain-of-Thought (CoT)

VisCoT



Multimodal Chain-of-Thought (CoT)

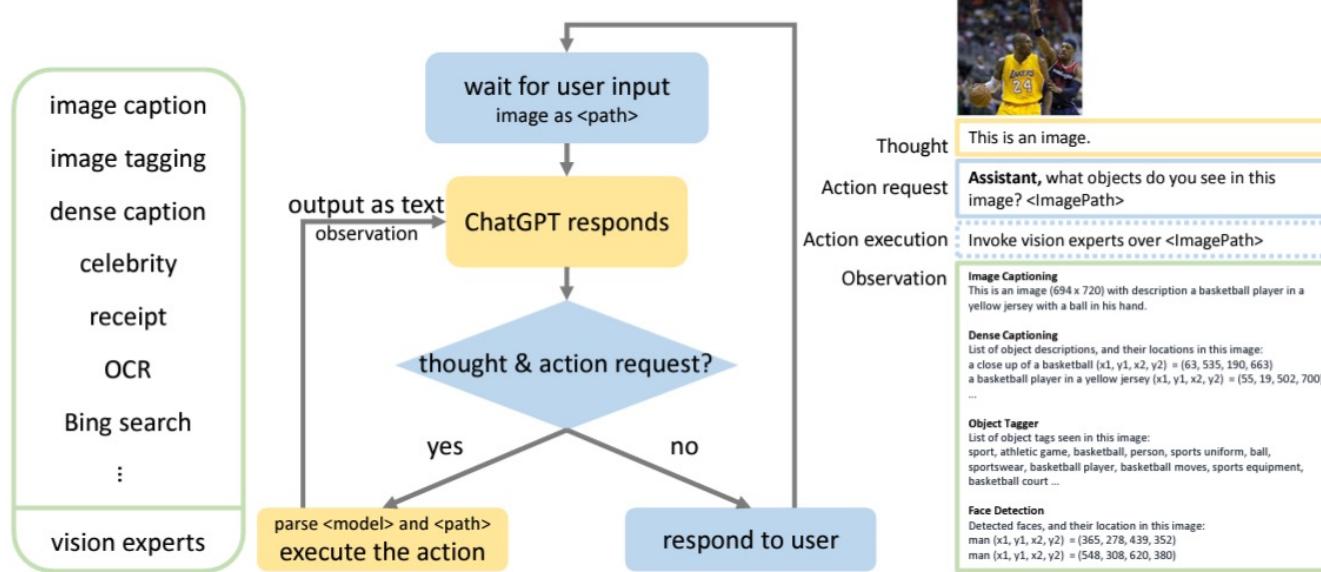
Multimodal-CoT



*Multi-image reasoning allowed

Multimodal Chain-of-Thought (CoT)

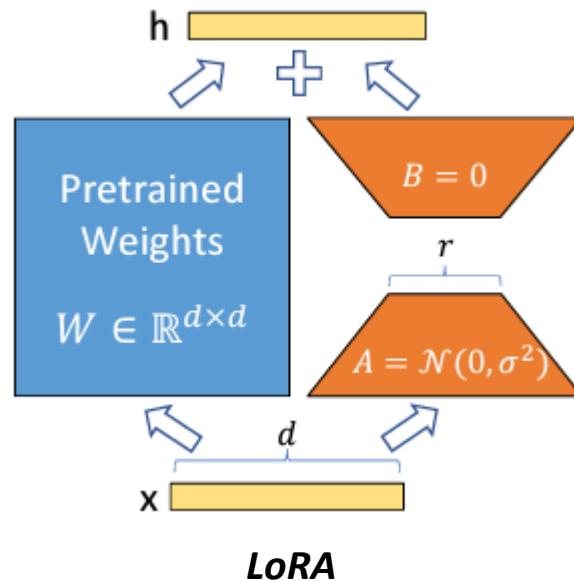
MM-REACT



*Multi-image reasoning allowed

Parameter-Efficient Tuning Techniques

LoRA and its variants

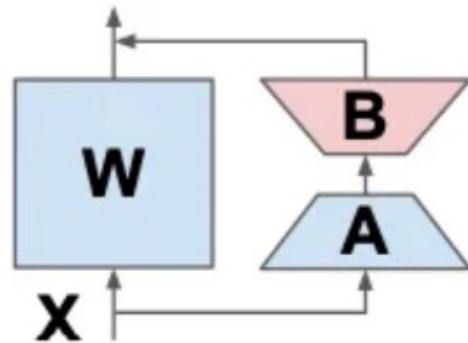


Low-Rank Adaptation (LoRA)

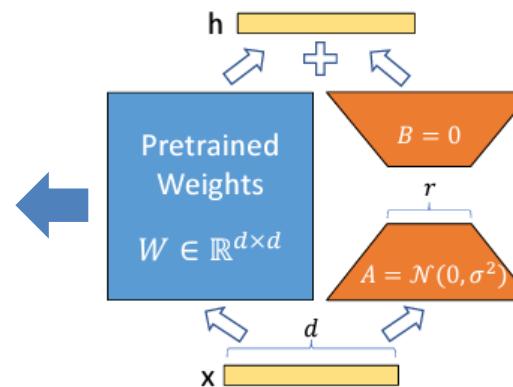
- Reparameterization
- Train A and B only instead of W
- Updated weights $W' = W + AB$
- Do not introduce additional inference latency
- Up to 2/3 reduction in VRAM usage, 25% speed up in training

Parameter-Efficient Tuning Techniques

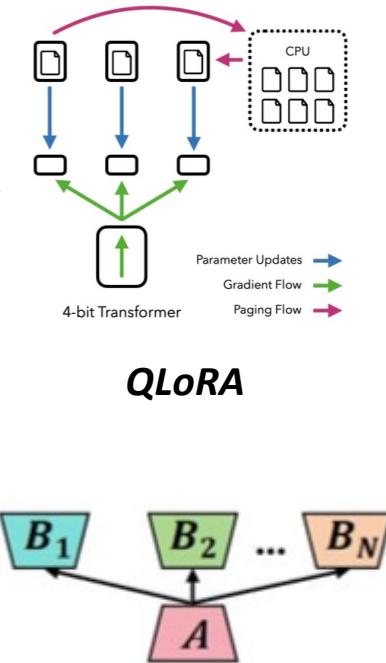
LoRA and its variants



LoRA-FA



LoRA

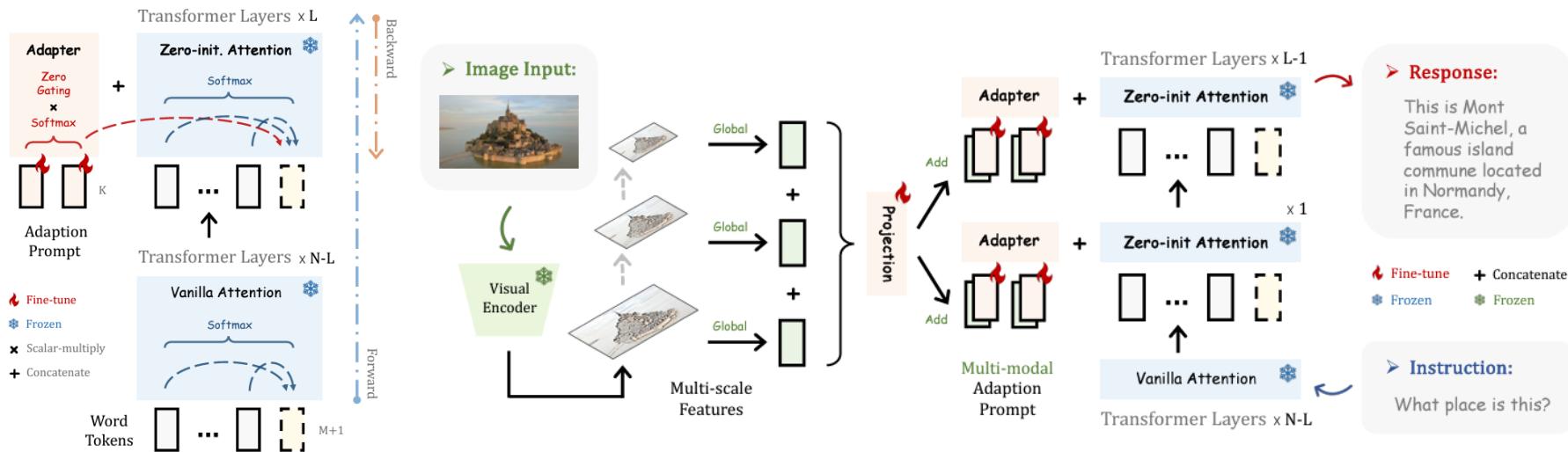


HydraLoRA

Parameter-Efficient Tuning Techniques

LLaMA-Adapter

Original LLaMA remains frozen, only finetune and plug a smaller adapter module



Parameter-Efficient Tuning Techniques

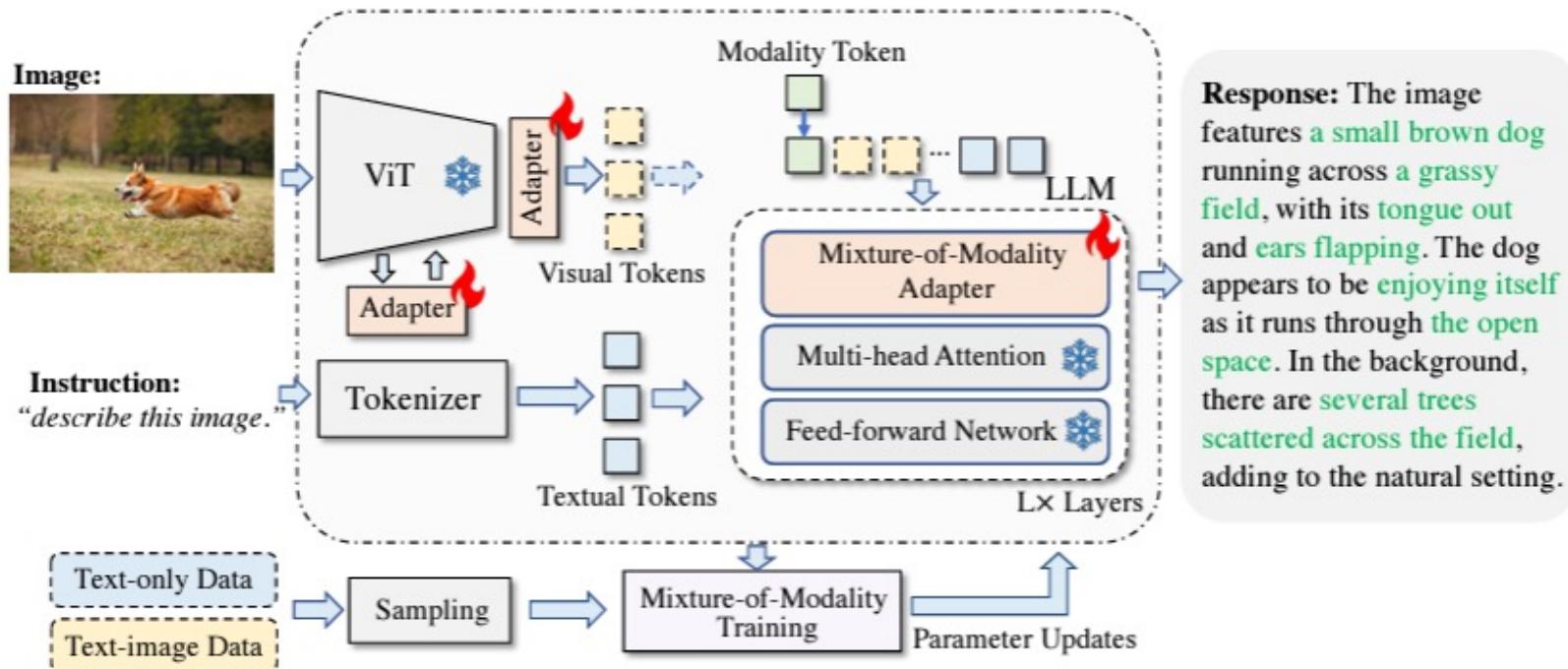
LLaMA-Adapter V2

Joint Training	Disjoint Parameters	Visual Context:	Question:	Textual Context:	Answer:
<div style="border: 2px dashed red; padding: 10px;"> Caption Data Instruction Data </div>	<div style="display: flex; justify-content: space-around; align-items: center;"> Projection Early Zero-init, Attention </div> <div style="display: flex; justify-content: space-around; align-items: center;"> Late Zero-init, Attention Norm, Scale, Bias Optional LoRA </div>	 	Describe the image: Summarize the given passage.		A dog on a bed. A recent study showed that global climate change is one of the most ...
					The passage discusses a study that highlights the significance ...

Visual Context + Question → Short Answer & Question + Textual Context → Detailed Answer

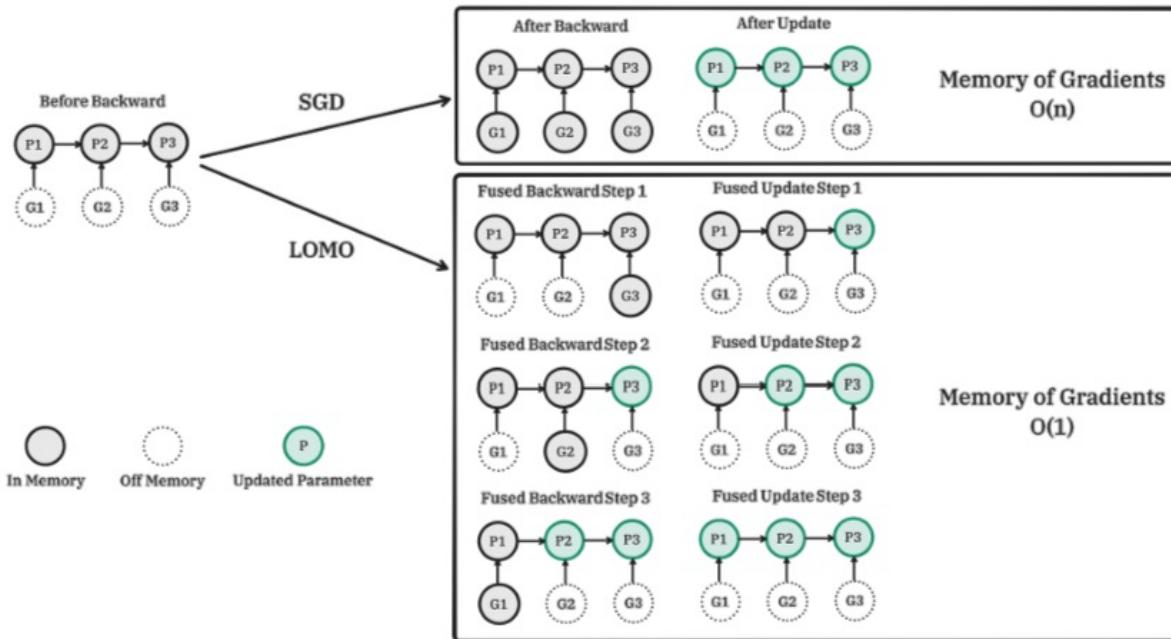
Parameter-Efficient Tuning Techniques

Mixture-of-Modality Adapter



Parameter-Efficient Tuning Techniques

Low Memory Optimization



Tutorial Outline

1. Introduction and Overview

- Welcome and tutorial objectives
- Multimodal learning
- Importance and Challenges of integrating multiple modalities (vision, language, sensors, audio)

2. Foundation of Multimodal Large Language Models (MLLMs)

- Multimodal Key Tasks and Benchmark datasets for multimodal tasks
- Challenges in Dataset Creation and Annotation for Multimodal Integration

3. Core Techniques for Multimodal Large Language Models (MLLMs)

- MLLM Architecture
- MM capabilities

4. Multimodal Instruction Tunings

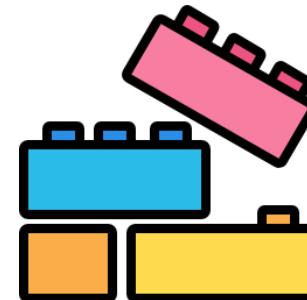
- Basic Instruction Tuning Technique to LVLM/MLLM Instruction Tuning
- Multimodal Chain-of-Thought
- Efficient Instruction Tuning Strategies

5. Future Trends and Research Opportunities

Future Direction – Emerging Technologies



Self-Supervised Learning



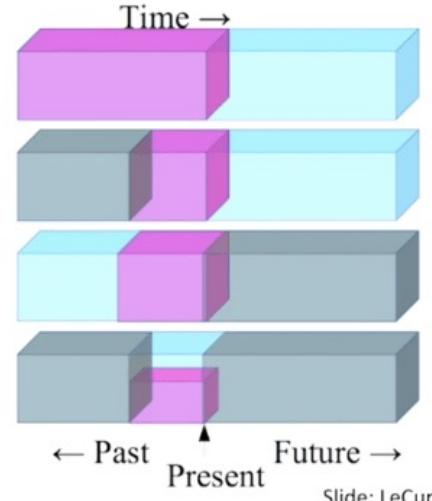
Plug-and-Play Theory

Future Direction – Emerging Technologies (1)

Self-Supervised Learning (SSL)

- SSL involves models learning from large amounts of unlabelled data, generating their training signals by predicting parts of the data from other parts.

- ▶ Predict any part of the input from any other part.
- ▶ Predict the **future** from the **past**.
- ▶ Predict the **future** from the **recent past**.
- ▶ Predict the **past** from the **present**.
- ▶ Predict the **top** from the **bottom**.
- ▶ Predict the **occluded** from the **visible**
- ▶ **Pretend there is a part of the input you don't know and predict that.**



Slide: LeCun

Future Direction – Emerging Technologies (1)

Self-Supervised Learning (SSL)

Significance for MLLMs

- **Reduces dependency on labelled data**, making it possible to train MLLMs on vast multimodal datasets (e.g., videos with audio and text paired with images) **without the costly and time-consuming need for manual annotation.**

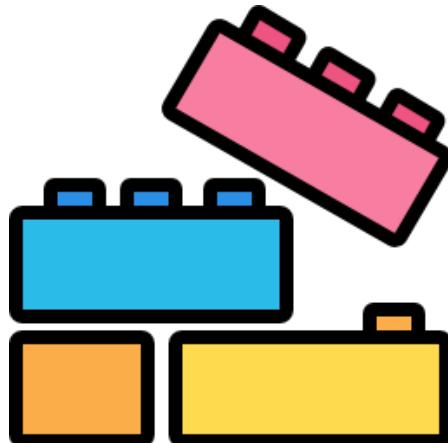
Applications

- Enabling better performance in tasks with limited labelled data
 - ❖ e.g. Rare language dialects in language-audio models
 - ❖ e.g. niche visual domains in vision-language models.

Future Direction – Emerging Technologies (2)

Plug-and-Play

- Allows modular addition or removal of modalities without retraining the entire model, enabling flexible and dynamic integration of new modality types.

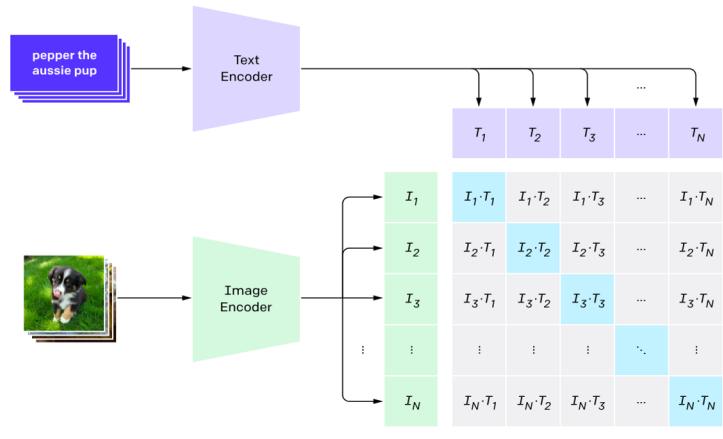


What is Multimodality?

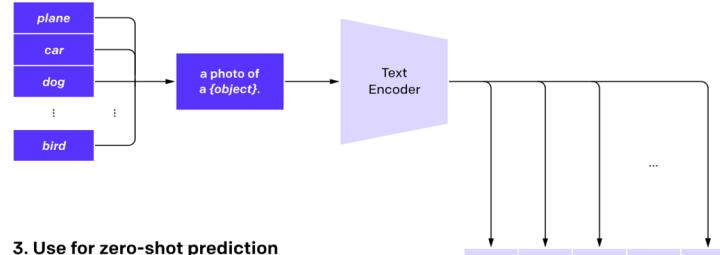
There are several multimodality representation learning techniques (i.e. CLIP, BLIP etc.)

Contrastive Learning Method

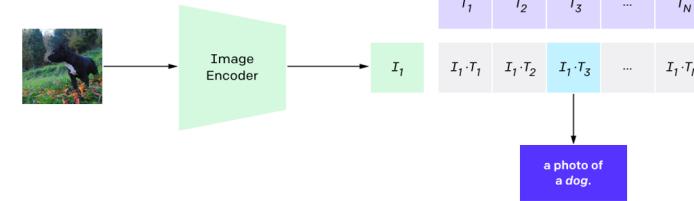
1. Contrastive pre-training



2. Create dataset classifier from label text



3. Use for zero-shot prediction

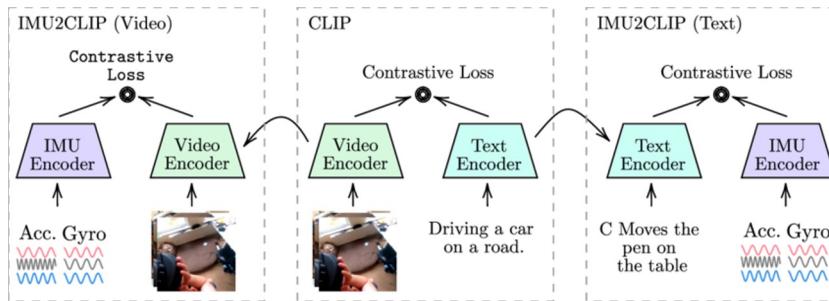


This is only for visual-language only.
What if we need more modalities?

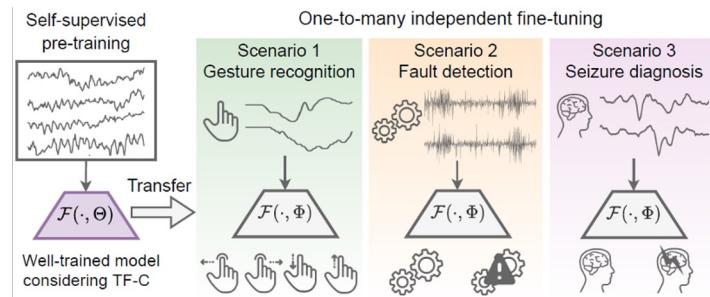
Why Multimodality Matters?

Train with Additional Modalities

Additional Contrastive learning



One-to-Many Independent Fine-tuning



Fundamental Problem!!

- **High Resource Cost:** Required the alignment information for all those additional modalities.
- **Low Performance:** Does not provide the sharing space with direct alignment.

Future Direction – Emerging Technologies (2)

Plug-and-Play

Significance for MLLMs

- supports customisable, efficient multimodal processing, where new modalities (e.g., haptic feedback, olfactory sensors) can be incorporated.

Applications

Especially beneficial for industries with evolving modality needs

- e.g. wearable tech that integrates new sensor types
- e.g. personalised AI systems in healthcare where patient data may come from multiple sources.

Researchers and practitioners can rapidly integrate novel modalities into existing MLLM frameworks without extensive retraining, making MLLMs more adaptable and application-specific.

Future Direction – Research Gaps (1)

Scalability

Challenges

- As the number and complexity of modalities increase, the computational cost and model size grow substantially, requiring efficient scaling strategies.

(Current) Limitations

- Struggle to manage complex interactions across multiple high-dimensional data types, limiting scalability for practical deployment.

Potential Solutions:

- Innovations in model compression, such as distillation and pruning
- Use of cloud computing resources to handle intensive multimodal processing.

Future Direction – Research Gaps (2)

Real-Time Processing

Challenges

- Real-time multimodal analysis requires rapid integration and processing of high-dimensional data, which is especially challenging in time-sensitive applications like autonomous vehicles.

(Current) Limitations

- MLLMs experience latency issues that prevent them from processing multiple data streams instantaneously.

Potential Solutions:

- Lightweight model architectures
- On-device processing for real-time capability

Future Direction – Research Gaps (3)

Robustness

Challenges

- It is challenging to handle noise, missing data, or inconsistencies across modalities without significant drops in performance..

(Current) Limitations

- MLLMs rely heavily on data quality and alignment, making them vulnerable to degraded performance in real-world, noisy environments.

Potential Solutions:

- Maintain accuracy across diverse data sources
- (Adversarial training, robust feature extraction methods, and error-correction mechanisms)

Q and A and Discussion



THE UNIVERSITY OF
MELBOURNE



THE UNIVERSITY OF
SYDNEY



SAPIENZA
UNIVERSITÀ DI ROMA