

From Multimodal LLM to Human-level AI

Modality, Instruction, Reasoning, Efficiency and Beyond

CVPR 2024 tutorial

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

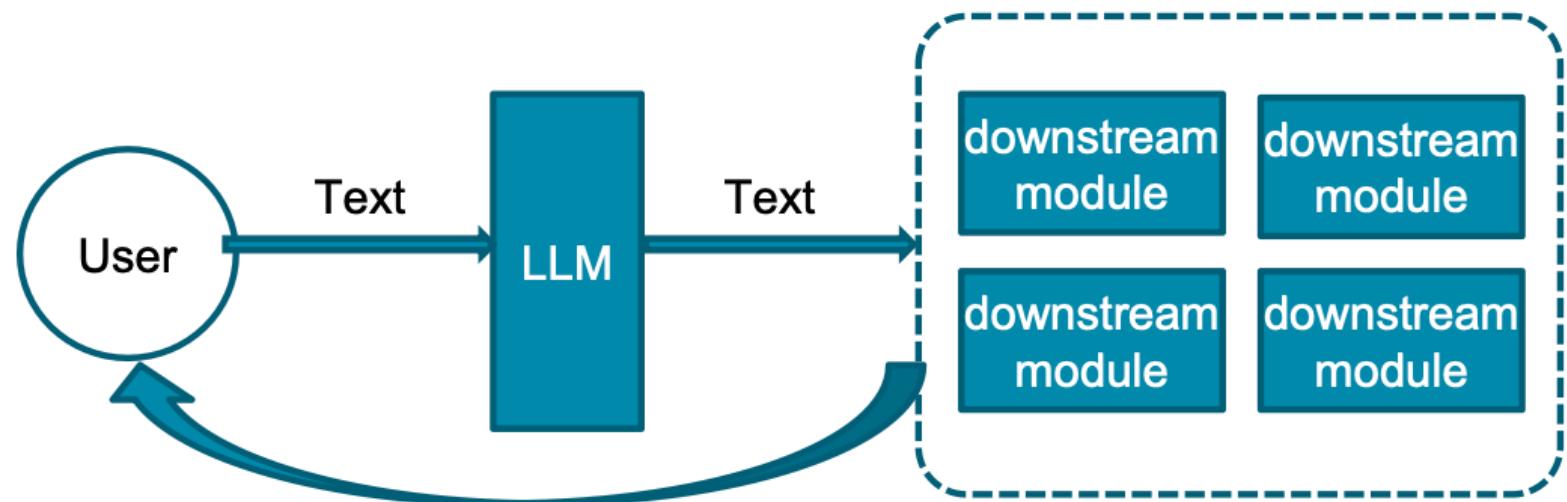
Outline

- Tutorial
 - Architecture
 - Modality
 - Multimodal Instruction Tuning
- LLaVA
- LION
- GILL
- GLaMM
- SmartEdit
- Conclusion

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

- + Representative MLLMs:

- + Visual-ChatGPT
 - + HuggingGPT
 - + MM-REACT
 - + ViperGPT
 - + AudioGPT
 - + LLaVA-Plus
 - + ...



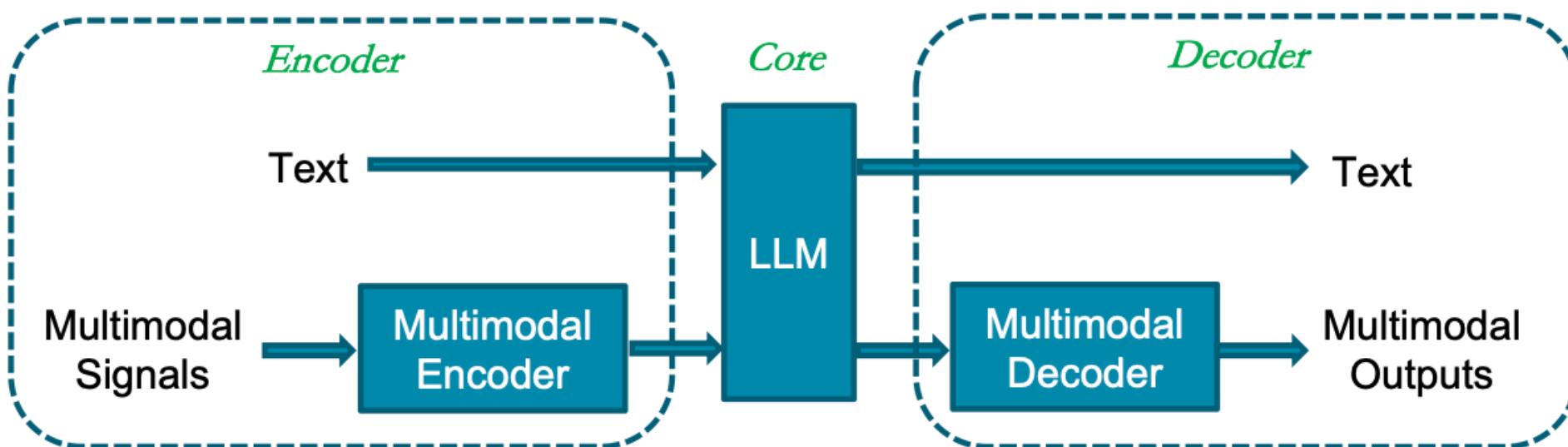
• Architecture-II: LLM as Joint Part of System



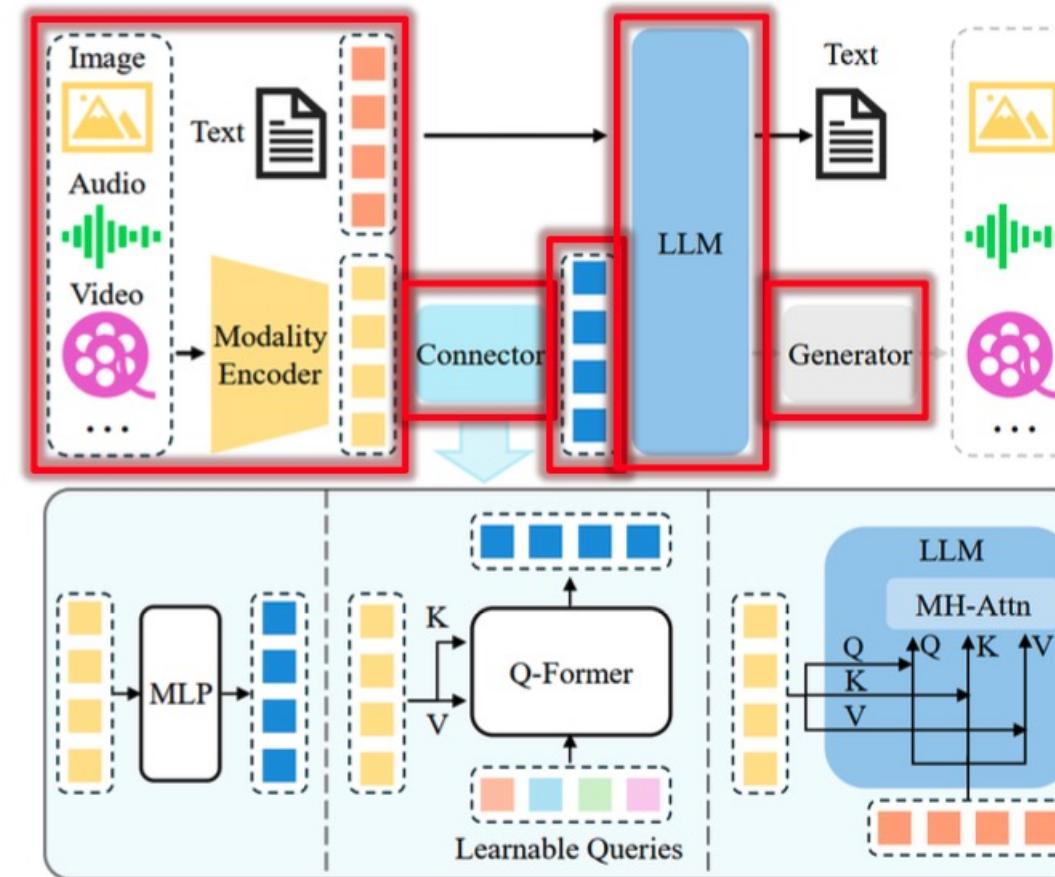
The role of the LLM is to perceive multimodal information, and **react by itself**, in a structure of **Encoder-LLM-Decoder**.

- + Key feature:

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



- **Architecture-II: LLM as Joint Part of System** •◦◦ More promising
- + > 90% MLLMs belong to this category.
- + Higher upper-bound, better integrated into a unified model

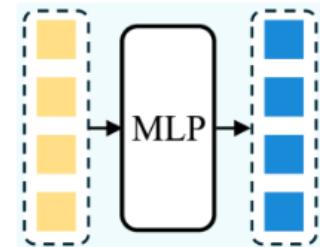


[1] A Survey on Multimodal Large Language Models.
<https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models>, 2023.

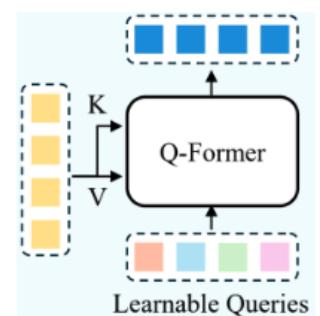
• Methods to Connect Multimodal Representation with LLM

- + Projecting multimodal (e.g., image) representations into LLM semantic space

- ✗ Q-Former: **BLIP-2, InstructBLIP, VisCPM, VisualGLM**

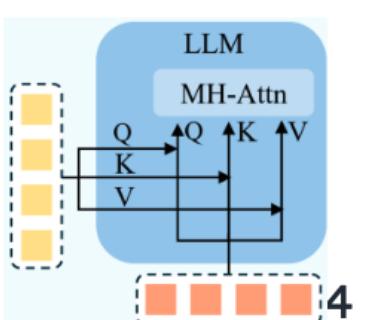


- ✗ Linear projection: **LLaVA, MiniGPT-4, NExT-GPT**



- ✗ Two-layer MLP: **LLaVA-1.5/NeXT, CogVLM, DeepSeek-VL, Yi-VL**

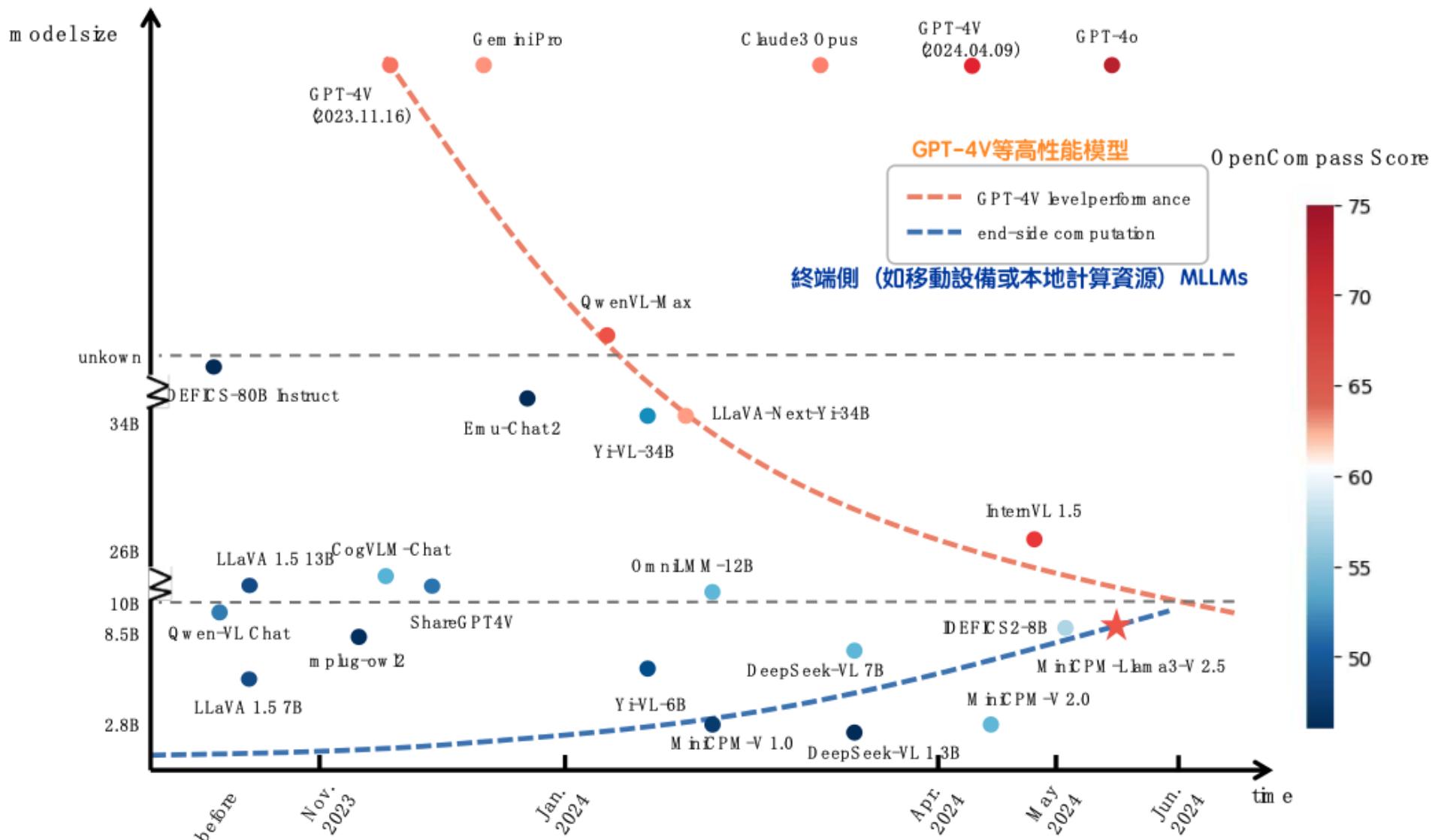
- ✗ Perceiver Resampler: **Flamingo, Qwen-VL, MiniCPM-V, LLaVA-UHD**



- ✗ C-Abstractor: **HoneyBee, MM1**

End-side MLLMs show promising potentials

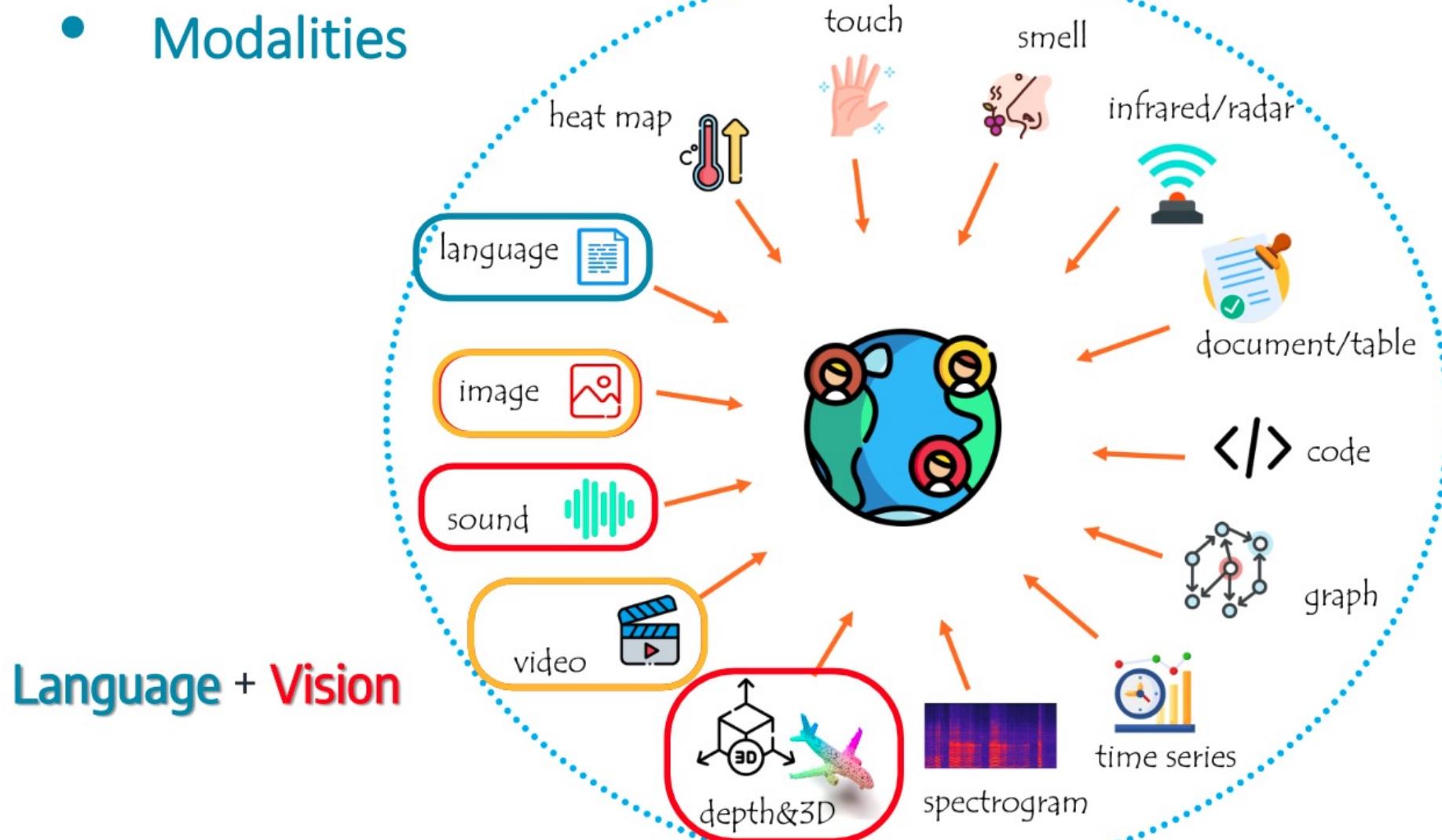
- + Promising performance: Matching larger LLMs



Outline

- Tutorial
 - Architecture
 - Modality
 - Multimodal Instruction Tuning
- LLaVA
- LION
- GILL
- GLaMM
- SmartEdit
- Conclusion

- Modalities



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

- SmartEdit

- **Image-oriented Pixel-wise Regional MLLM**



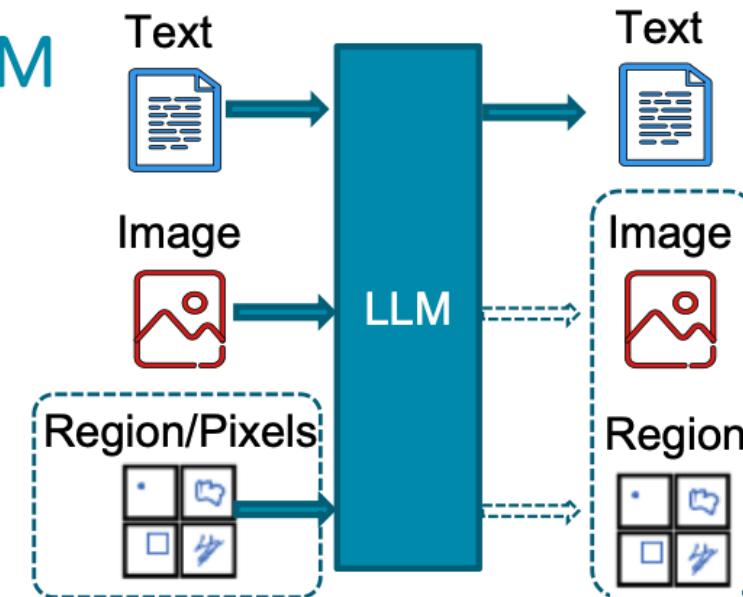
Pixel-level Awareness at Input/Output

+ Output-side Only Pixel-wise Awareness

LISA, PixelLM, DetGPT, MiniGPT-v2, LION

+ Input-&Output-side Pixel-wise Awareness

NExT-Chat, GPT4RoT, Shikra,
KOSMOS-2, GLaMM, Osprey



- **Image-oriented Pixel-wise Regional MLLM**

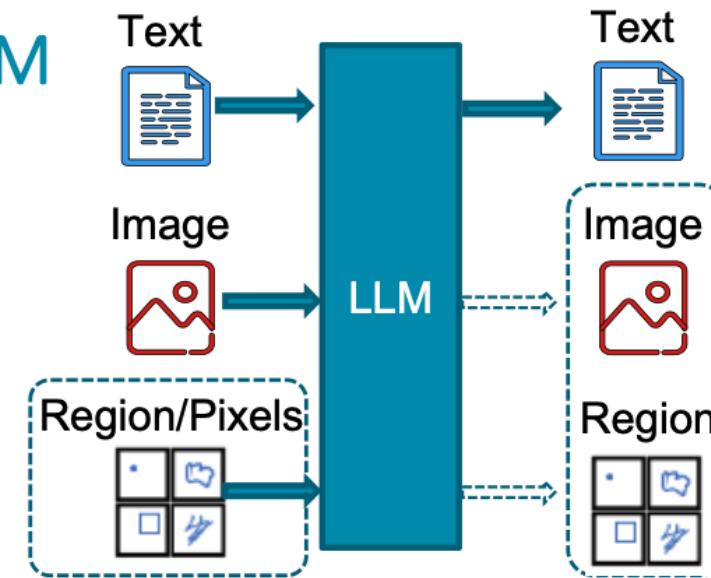
👉 **Pixel Granularity**

+ **Bounding-box Coordinates**

NExT-Chat, GPT4RoI, Shikra, **LION**,
KOSMOS-2, DetGPT, MiniGPT-v2

+ **Finer-grained Mask-based Segments**

NExT-Chat, LISA, PixelLM,
GLaMM, Osprey



- **Image-oriented Pixel-wise Regional MLLM**

👉 **User Input Interaction**

+ **No Image User Interaction**

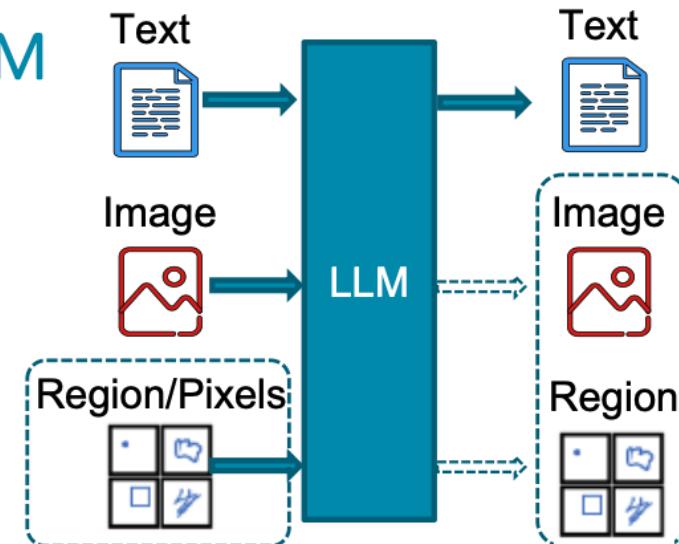
LISA, PixelLM, DetGPT, MiniGPT-v2, **LION**

+ **Bounding-box Coordinates**

GPT4RoI, Shikra, KOSMOS-2, **GLaMM**

+ **User Sketches**

NExT-Chat, Osprey,



Outline

- Tutorial
 - Architecture
 - Modality
 - Multimodal Instruction Tuning
- LLaVA
- LION
- GILL
- GLaMM
- SmartEdit
- Conclusion

- Multitask learning (with task tokens)

Training



INPUT: <image><tok_task_1=short_cap>
OUTPUT: <generated short descriptions>

INPUT: <image><tok_task_2=yes_no>
OUTPUT: yes/no

Testing

Only with <tok_task_1>, <tok_task_2>...

Does not work with <new_task=long_cap>

- Instruction tuning (with natural language task instructions)

Training



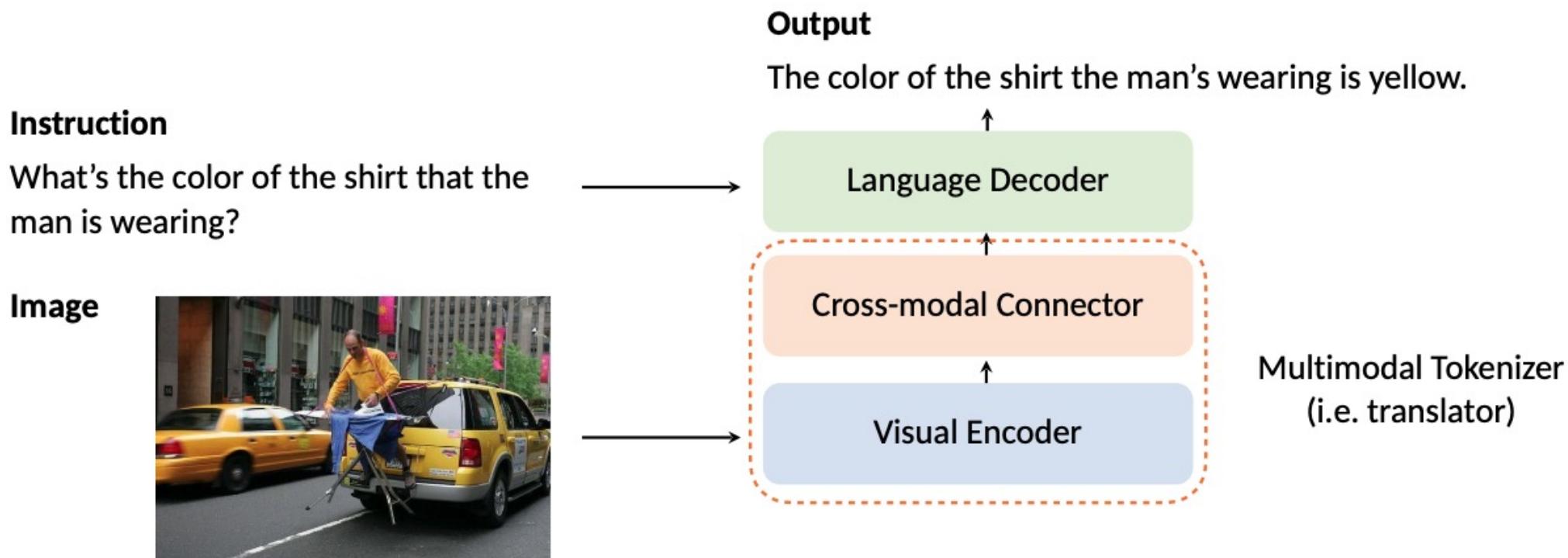
INPUT: <image>Describe this image
briefly.
OUTPUT: <generated short descriptions>

INPUT: <image>Is this xxx?
OUTPUT: yes/no

Testing

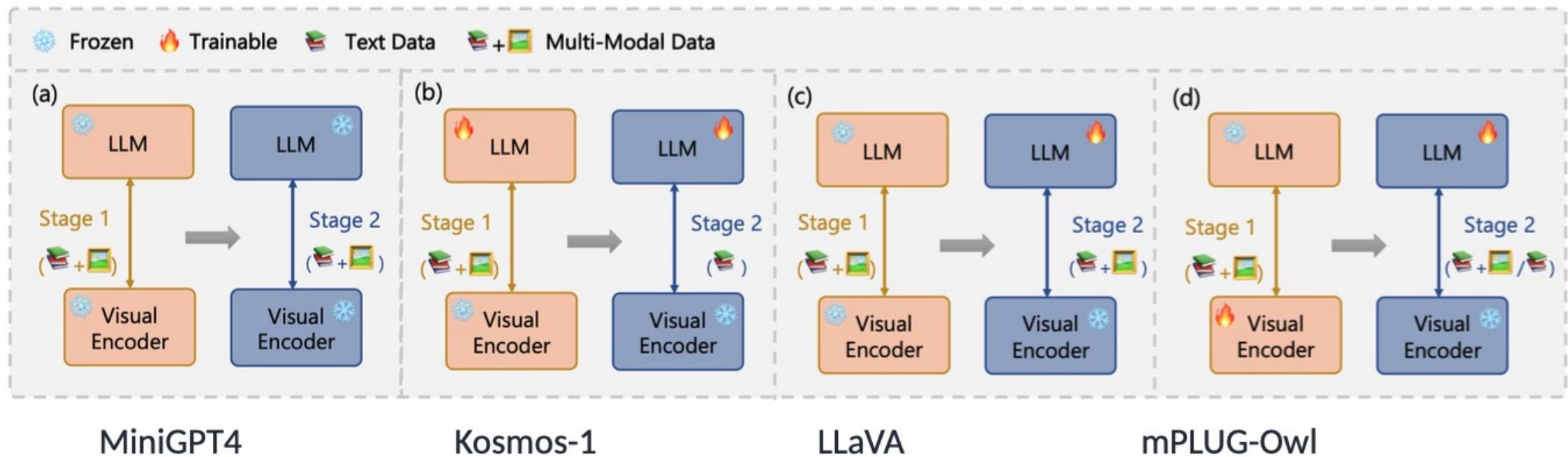
INPUT: <image>Describe this image in detail.
OUTPUT: <long descriptions>

Generalizes to new instructions zero-shot.



LLM “learns” a **visual** foreign language efficiently.

👉 Training paradigms of popular multimodal large language models.



Existing Instruction Tuning Dataset

Dataset	Size	Modalities	Constructions
LLaVA-Instruct-158k	158k	Image, Text	ChatGPT-generated
LRV-Instruction	400k	Image, Text	GPT4-generated
MMC-Instruction	600k	Chart, Text	GPT4-generated/adapted
Clotho-Detail	3.9k	Text, Audio	GPT4-generated
MACAW-LLM	119k	Image, Video, Text	GPT-3.5-turbo-generated
MIMIC-IT	2.8M	Image, Video, Text	ChatGPT-generated
StableLLaVA	126k	Image, Text	StableDiffusion & ChatGPT-generated
LAMM	196k	Image, PointCloud, Text	GPT4-generated
VIGC-LLaVA	1.8M	Image, Text	Model-generated
X-LLM	10k	Image, Video, Text	ChatGPT-generated

LLaVA: Large Language and Vision Assistant

Visual Instruction Tuning

NeurIPS 2023 (Oral)

Haotian Liu*, Chunyuan Li*, Qingyang Wu, Yong Jae Lee

► University of Wisconsin-Madison ► Microsoft Research ► Columbia University

*Equal Contribution

NeurIPS 2023

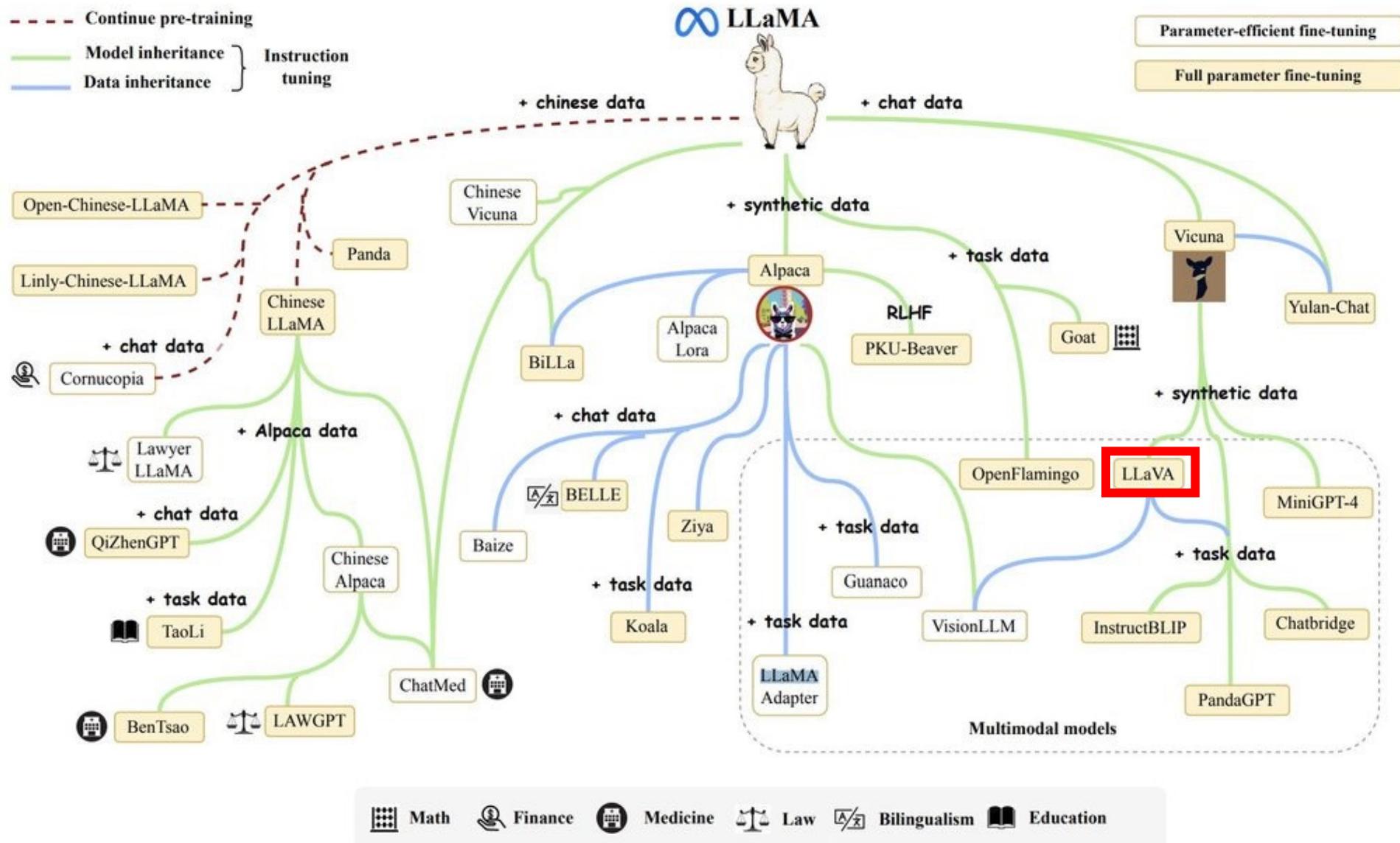
Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Outline

- Tutorial
- LLaVA
 - Framework
 - Experiments
- LION
- GILL
- GLaMM
- SmartEdit
- Conclusion

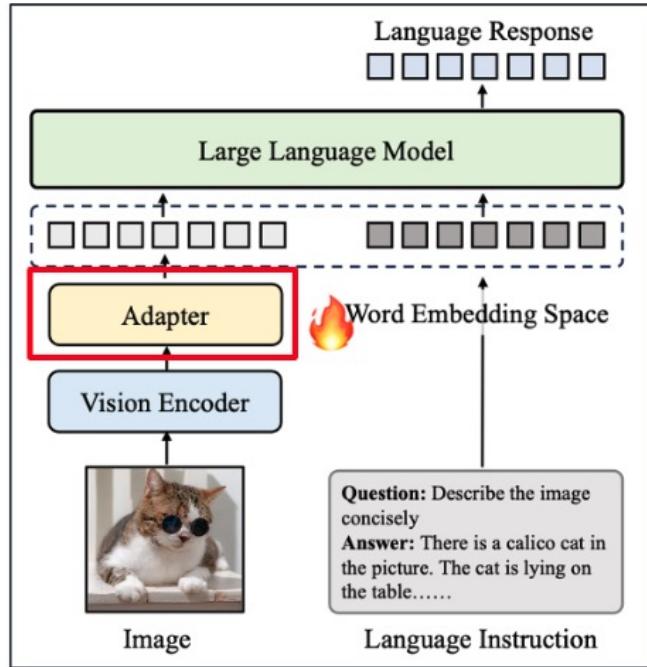
	Modality (w/ Language)			
	Image	Video	Audio	3D
Input-side Perceiving	Flamingo, Kosmos-1, Blip2, mPLUG-Owl, Mini-GPT4, LLaVA , InstructBLIP, VPGTrans, CogVLM, Monkey, Chameleon, Otter, Qwen-VL, GPT-4v, SPHINX, Yi-VL, Fuyu, ...	VideoChat, Video-ChatGPT, Video-LLaMA, PandaGPT, MovieChat, Video-LLaVA, LLaMA-VID,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





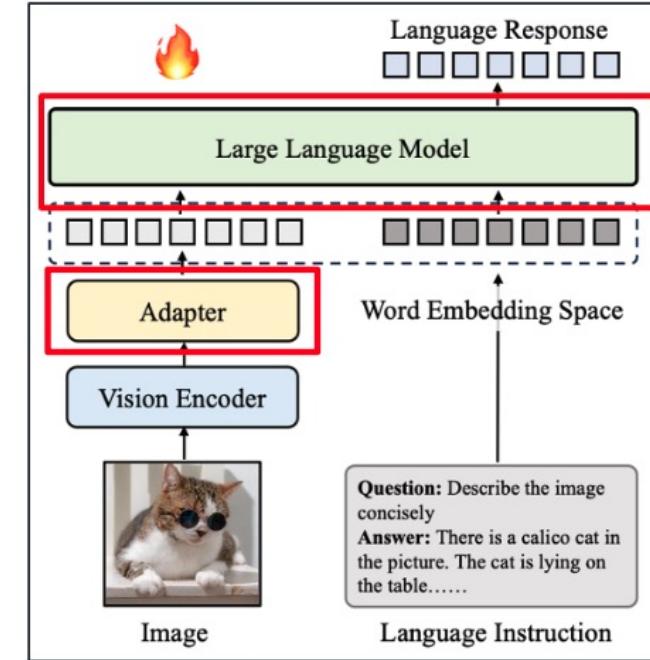
Stage1: Pretraining Stage

- Align different modalities, provide world knowledge



Stage2: Instruction Tuning Stage

- Teach models to better understand the instructions from users and fulfill the demanded tasks.



- end-to-end trained large multimodal model that connects a vision encoder and an LLM for general-purpose visual and language understanding.
- open-set visual encoder of **CLIP** with the language decoder **Vicuna**

Context type 1: Captions

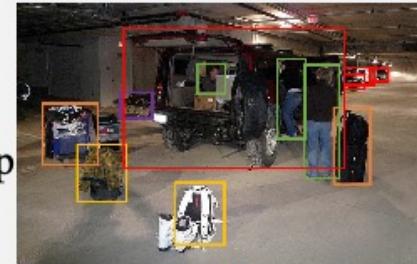
A group of people standing outside of a black vehicle with various luggage.

Luggage surrounds a vehicle in an underground parking area

People try to fit all of their luggage in an SUV.

The sport utility vehicle is parked in the public garage, being packed for a trip

Some people with luggage near a van that is transporting it.



Context type 2: Boxes

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

Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

- The top block used to prompt GPT, and the bottom block shows **the three types of responses**.
- Visual image is not used to prompt GPT, we only show it here as a reference.

$\mathbf{X}_{\text{system-message}} \text{ <STOP>}$
 Human : $\mathbf{X}_{\text{instruct}}^1 \text{ <STOP>} \text{ Assistant: } \mathbf{X}_a^1 \text{ <STOP>}$
 Human : $\mathbf{X}_{\text{instruct}}^2 \text{ <STOP>} \text{ Assistant: } \mathbf{X}_a^2 \text{ <STOP>} \dots$

For each image \mathbf{X}_v , we generate multi-turn conversation data $(\mathbf{X}_a^1, \mathbf{X}_a^1, \dots, \mathbf{X}_a^T, \mathbf{X}_a^T)$

$$\mathbf{X}_{\text{instruct}}^t = \begin{cases} \text{Randomly choose } [\mathbf{X}_q^1, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}_q^1], & \text{the first turn } t = 1 \\ \mathbf{X}_q^t, & \text{the remaining turns } t > 1 \end{cases} \quad (2)$$

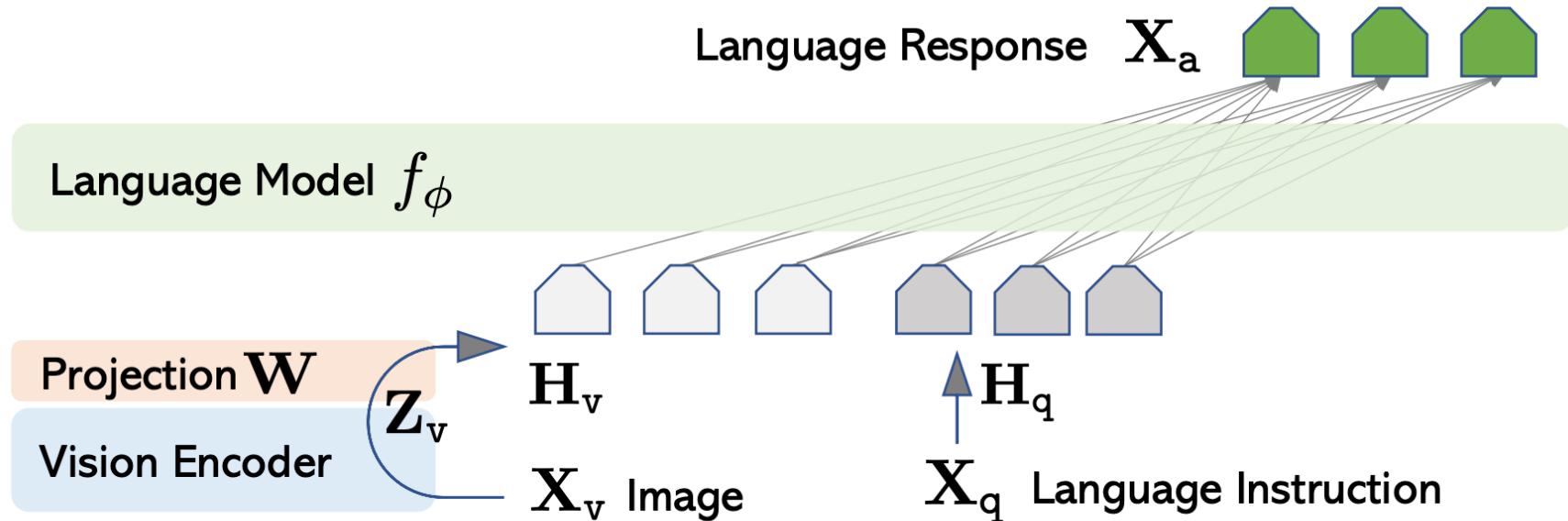


Figure 1: LLaVA network architecture.

$$\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v, \text{ with } \mathbf{Z}_v = g(\mathbf{X}_v) \quad (1)$$

$$p(\mathbf{X}_a | \mathbf{X}_v, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^L p_\theta(\mathbf{x}_i | \mathbf{X}_v, \mathbf{X}_{\text{instruct}, < i}, \mathbf{X}_{a, < i}), \quad (3)$$

Outline

- Tutorial
- LLaVA
 - Framework
 - Experiments
- LION
- GILL
- GLaMM
- SmartEdit
- Conclusion

Visual input example, Extreme Ironing:



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

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

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

- Metrics:

- Reference prediction based on the question and the ground-truth textual descriptions, using the text-only GPT-4.
- Obtaining the responses from both models, we feed the question, visual information (in the format of textual descriptions), and the generated responses from both assistants, to the judge (i.e., text-only GPT-4).

Challenging examples from LLaVA-Bench (In-the-Wild):



ICHIRAN Ramen [[source](#)]



Filled fridge [[source](#)]

Annotation

A close-up photo of a meal at **ICHI-RAN**. The chashu ramen bowl with a spoon is placed in the center. The ramen is seasoned with **chili sauce**, **chopped scallions**, and served with **two pieces of chashu**. Chopsticks are placed to the right of the bowl, still in their paper wrap, not yet opened. The ramen is also served with nori on the left. On top, from left to right, the following sides are served: a bowl of **orange spice** (possibly garlic sauce), a plate of **smoke-flavored stewed pork with chopped scallions**, and a cup of **matcha green tea**.

An open refrigerator filled with a variety of food items. In the left part of the compartment, towards the front, there is a **plastic box of strawberries** with a small bag of baby carrots on top. Towards the back, there is a stack of sauce containers. In the middle part of the compartment, towards the front, there is a green plastic box, and there is an unidentified plastic bag placed on it. Towards the back, there is a carton of milk. In the right part of the compartment, towards the front, there is a box of blueberries with three yogurts stacked on top. The large bottle of yogurt is **Fage non-fat yogurt**, and **one of the smaller cups is Fage blueberry yogurt**. The brand and flavor of the other smaller cup are unknown. Towards the back, there is a container with an unknown content.

Question 1

What's the name of the restaurant?

What is the brand of the blueberry-flavored yogurt?

Question 2

Describe this photo in detail.

Is there strawberry-flavored yogurt in the fridge?

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

- LLaVA model ensembling with the text-only GPT-4 consistently improves the model's performance under all categories, setting the new SoTA performance.

LION : Empowering Multimodal Large Language Model with Dual-Level Visual Knowledge

Gongwei Chen, Leyang Shen, Rui Shao[†], Xiang Deng, Liqiang Nie[†]

School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen

{chengongwei, shaorui, dengxiang, nieliqiang}@hit.edu.cn

<https://github.com/rshaojimmy/JiuTian>

CVPR 2024

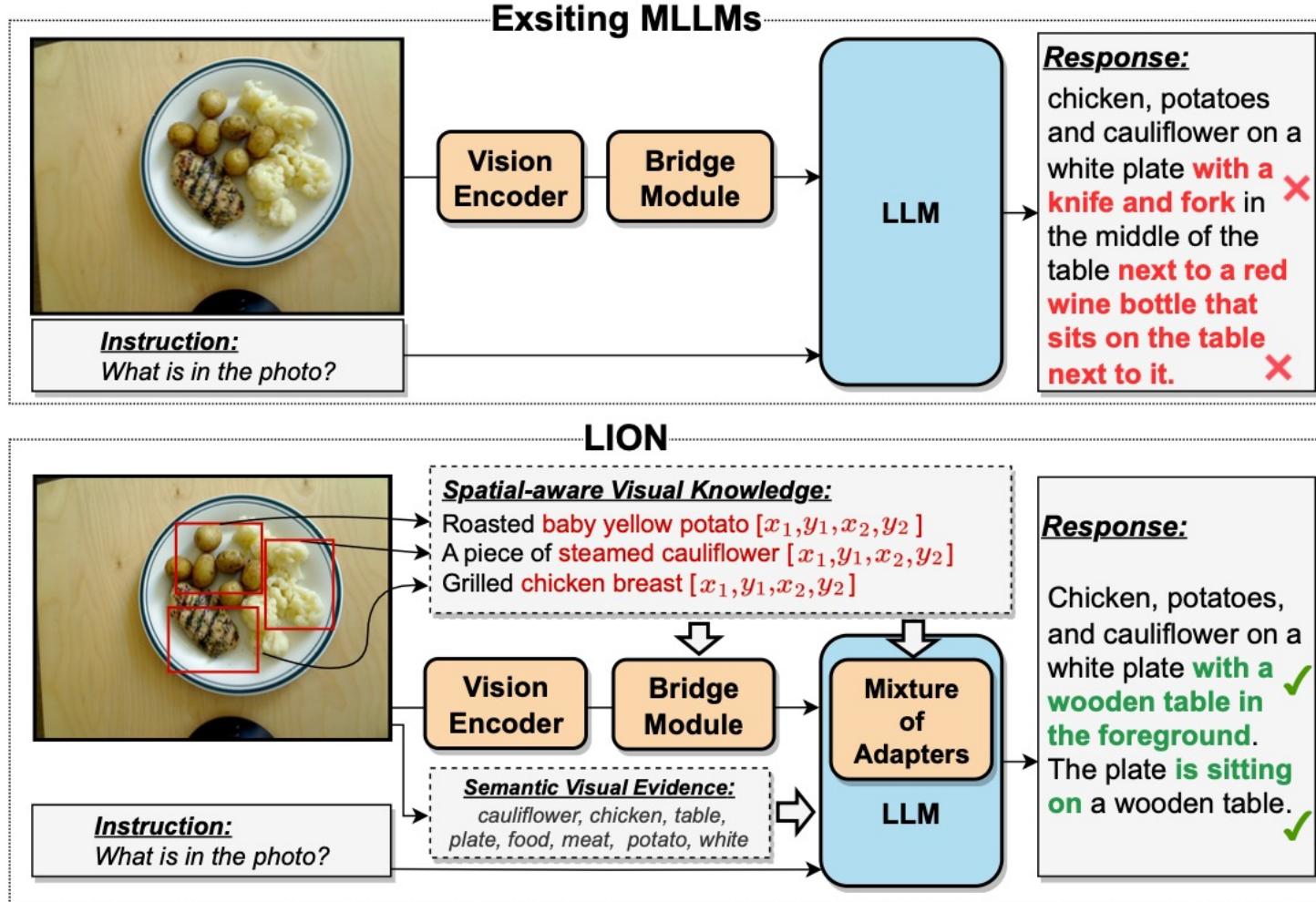
Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

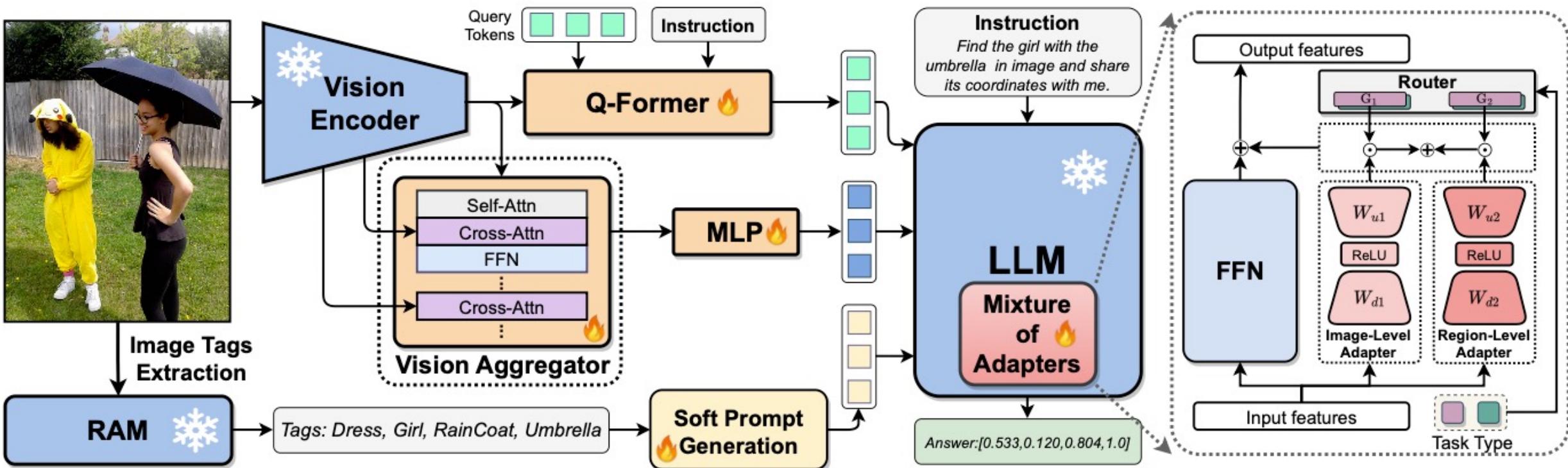
Outline

- Tutorial
- LLaVA
- LION
 - Framework
 - Experiments
- GILL
- GLaMM
- SmartEdit
- Conclusion

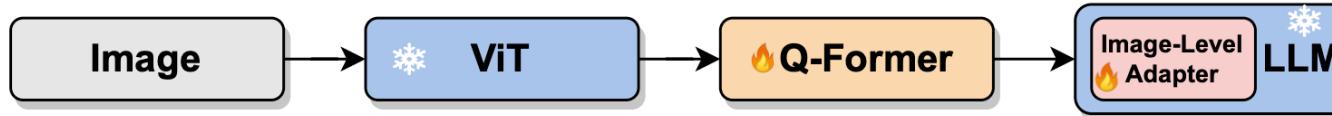
	Modality (w/ Language)			
	Image	Video	Audio	3D
Input-side Perceiving	Flamingo, Kosmos-1, Blip2, mPLUG-Owl, Mini-GPT4, LLaVA, InstructBLIP, VPGTrans, CogVLM, Monkey, Chameleon, Otter, Qwen-VL, GPT-4v, SPHINX, Yi-VL, Fuyu, ... [Pixel-wise] GPT4RoI, LION, MiniGPT-v2, NExT-Chat, Kosmos-2, GLaMM, LISA, DetGPT,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



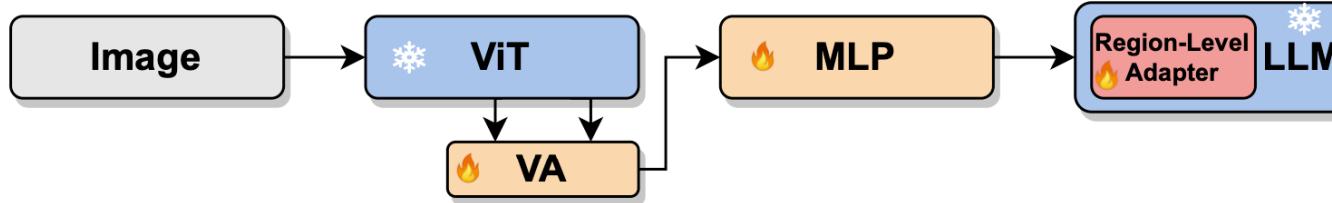
- dual-Level vIsual knOwledge eNhanced Multimodal Large Language Model (**LION**), which enriches the visual information in MLLMs in two levels.
 - 1) Progressive incorporation of fine-grained spatial-aware visual knowledge
 - 2) Soft prompting of high-level semantic visual evidence



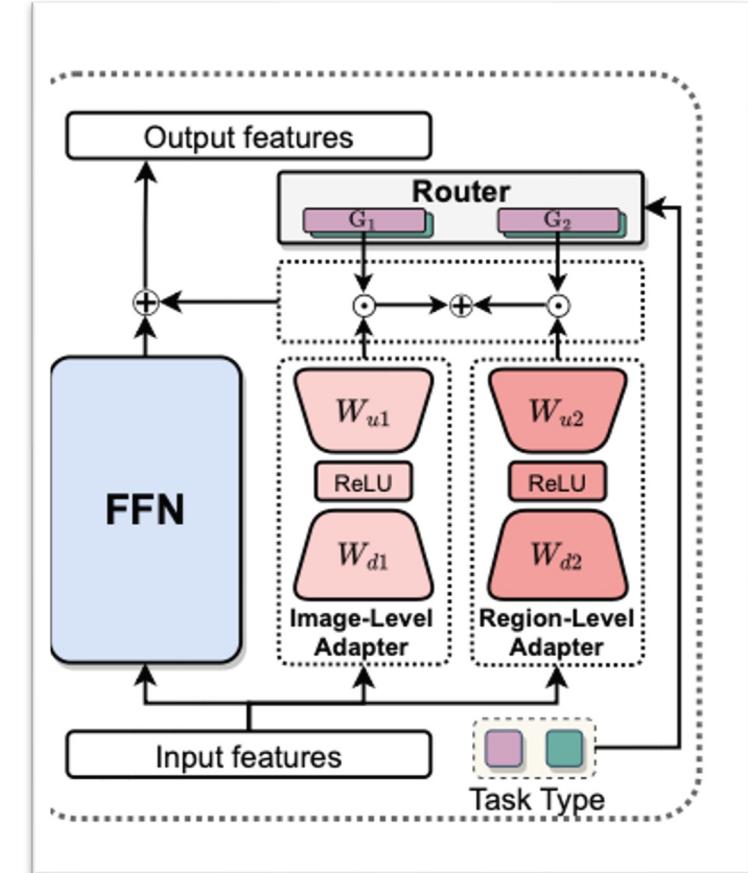
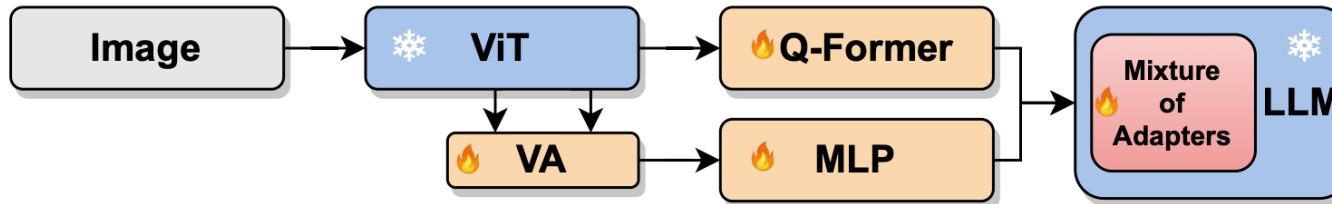
Stage 1: Instruction-Tuning on Image-Level VL Tasks



Stage 2: Instruction-Tuning on Region-Level VL Tasks



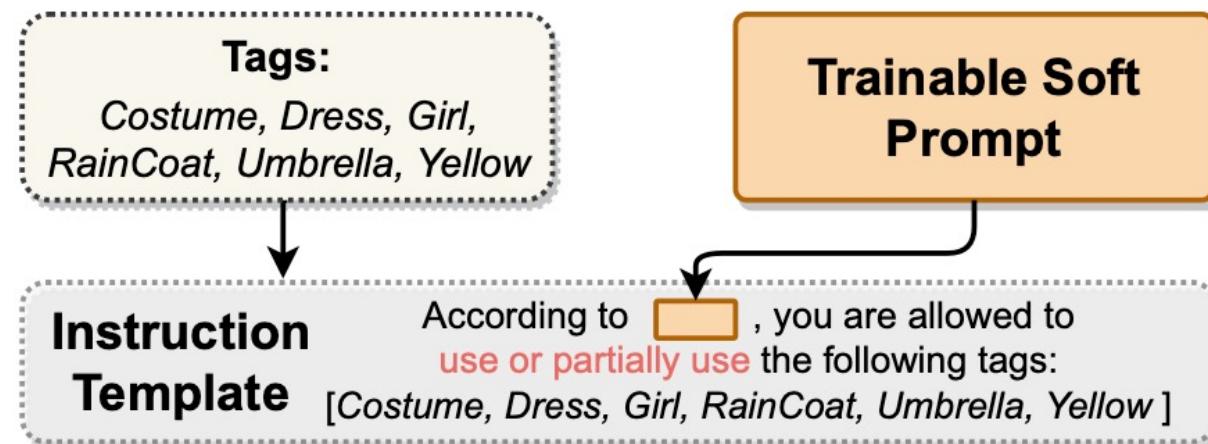
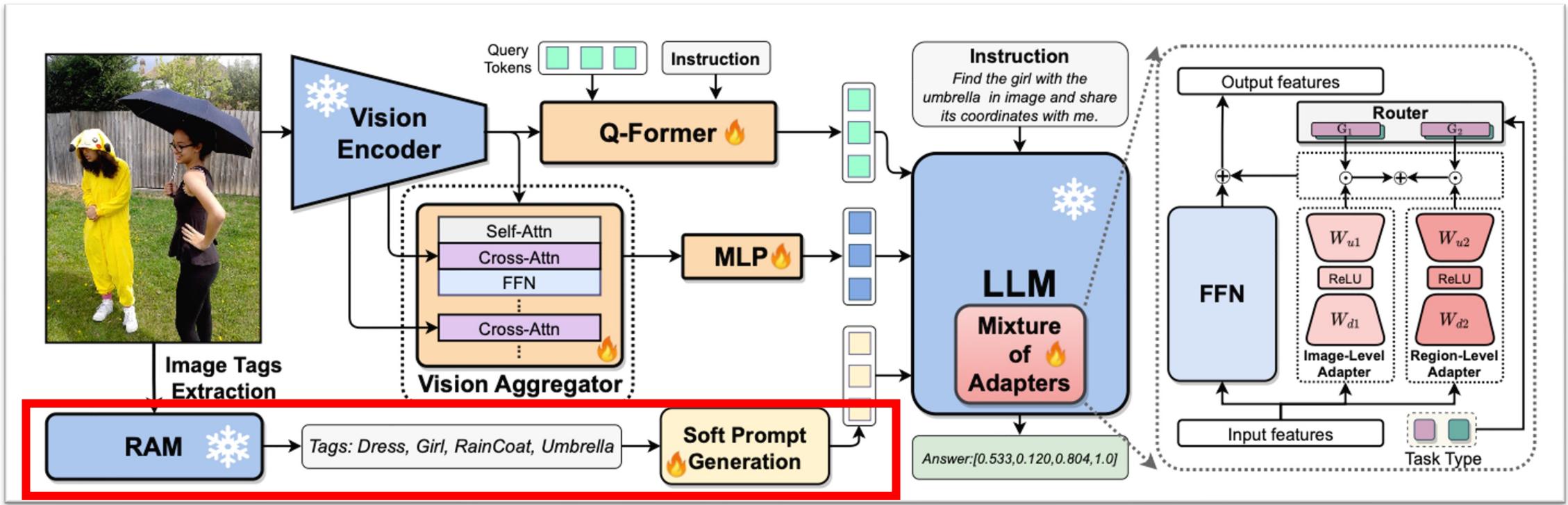
Stage 3: Instruction-Tuning on Image-Level and Region-Level VL Tasks

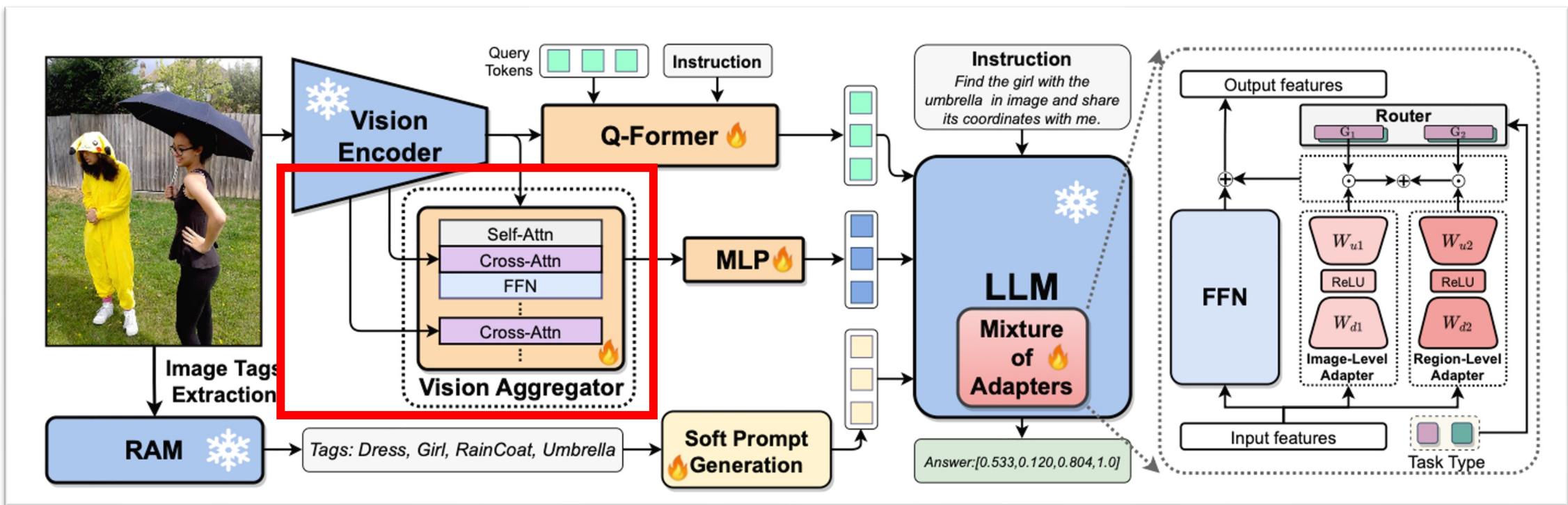


$$O = \mathbf{F}(X) + \mathbf{H}(X), \quad (1)$$

$$\mathbf{H}(X) = W_u(\sigma(W_d X)), \quad (2)$$

$$O^t = \mathbf{F}(X) + \sum_{k=1}^K \mathbf{G}_k^t \odot \mathbf{H}_k(X). \quad (3)$$





$$\bar{V} = \mathcal{B}_2(\mathcal{B}_1(V_i; V_j); V_k), \quad (4)$$

$$\mathcal{B}(X; Y) = \text{FFN}(\text{XAttn}(\text{Attn}(X), Y)). \quad (5)$$

Outline

- Tutorial
- LLaVA
- LION
 - Framework
 - Experiments
- GILL
- GLaMM
- SmartEdit
- Conclusion

VQA	<Image>Given the image, answer the following question with no more than three words. {Question} <Image>Based on the image, respond to this question with a short answer: {Question}. Answer: <Image>Use the provided image to answer the question: {Question} Provide your answer as short as possible: <Image>What is the answer to the following question? "{Question}" <Image>The question "{Question}" can be answered using the image. A short answer is
VQG	<Image>Based on the image, provide a question with the answer: {Answer}. Question: <Image>Given the visual representation, create a question for which the answer is "{Answer}." <Image>From the image provided, craft a question that leads to the reply: {Answer}. Question: <Image>Considering the picture, come up with a question where the answer is: {Answer}. <Image>Taking the image into account, generate a question that has the answer: {Answer}. Question:
Image Captioning	<Image>Can you briefly explain what you see in the image? <Image>Could you use a few words to describe what you perceive in the photo? <Image>Please provide a short depiction of the picture. <Image>Using language, provide a short account of the image. <Image>Use a few words to illustrate what is happening in the picture.
REC	<image>Identify the position of {expr} in image and share its coordinates. <image>I'd like to request the coordinates of {expr} within the photo. <image>How can I locate {expr} in the image? Please provide the coordinates. <image>I am interested in knowing the coordinates of {expr} in the picture. <image>Assist me in locating the position of {expr} in the photograph and its bounding box coordinates. <image>In the image, I need to find {expr} and know its coordinates. Can you please help?
REG	<image>What are the unique characteristics of the rectangular section {BBox} in image? <image>Describe the novel qualities of the selected bounding box {BBox} in image. <image>What sets the chosen region {BBox} in image apart from its surroundings? <image>Provide a one-of-a-kind depiction for the area enclosed by {BBox} in image. <image>How would you portray the unique features of the designated box {BBox} in image? <image>Explain the distinguishing characteristics of the marked bounding box {BBox} in image.

Model	Flickr30K	COCOcap	TextCaps	OKVQA	AOKVQA	GQA	IconQA	VSR	VisDial
Flamingo-3B [1]	60.60	73.00	-	-	-	-	-	-	46.10
Flamingo-9B [1]	61.50	79.40	-	44.70	-	-	-	-	48.00
Kosmos-1 [16]	67.10	84.70	-	-	-	-	-	-	-
Kosmos-2 [35]	80.50	-	-	-	-	-	-	-	-
AdapterV2 [13]	-	122.20	-	-	-	-	-	-	-
Shikra [4]	73.90	117.50	-	47.16	-	52.60	47.80	66.30	-
Pink [45]	-	-	-	59.50	-	30.80	37.60	41.60	16.52*
MiniGPT4 [54]	17.75*	17.04*	24.06*	37.50	34.51*	30.80	37.60	41.60	-
LLaVA [29]	48.03*	73.85*	45.54*	54.40	34.51*	41.30	43.00	51.20	8.65*
MiniGPTV2 [3]	80.75*	129.16*	80.60*	56.90	-	60.30	47.70	60.60	8.47*
BLIVA [15]	87.10	-	-	-	-	44.88	62.20	45.63	-
InstructBLIP† (T5XL) [7]	84.50	138.21	82.55	49.28	57.86	48.40	50.00	64.80	46.60
InstructBLIP† (T5XXL) [7]	83.50	138.28	82.53	48.59	56.16	47.90	51.20	65.60	48.50
InstructBLIP (T5XL) [7]	83.71	135.47	104.17	47.38	56.12	46.34	52.47	69.93	48.75
InstructBLIP (T5XXL) [7]	85.79	138.63	105.44	53.02	59.38	47.74	53.18	68.46	50.41
LION-4B	85.57	138.20	104.87	51.08	59.98	49.50	54.91	72.96	50.02
LION-12B	87.12	139.25	108.76	<u>57.33</u>	<u>60.87</u>	51.56	<u>54.89</u>	<u>73.77</u>	<u>50.42</u>

Model	RefCOCO			RefCOCO+			RefCOCOg		Avg.
	val	test-A	test-B	val	test-A	test-B	val	test	
<i>Zero-shot Setting</i>									
Kosmos-2 [35]	52.32	57.42	47.26	45.48	50.73	42.24	60.57	61.65	52.21
GRILL [18]	-	-	-	-	-	-	-	47.50	-
Pink [45]	54.10	61.20	44.20	43.90	50.70	35.00	59.10	60.10	51.00
LION-4B	57.89	56.07	58.40	46.38	45.29	47.50	64.74	63.56	54.98
LION-12B	58.54	56.41	59.36	45.93	45.73	47.89	66.12	64.69	55.58
<i>Fine-tuning Setting</i>									
OFA-L [42]	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58	72.65
VisionLLM-H [43]	-	86.70	-	-	-	-	-	-	-
Shikra-7B [4]	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	82.93
Shikra-13B [4]	87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16	83.96
Pink [45]	88.30	91.70	84.00	81.40	87.50	73.70	83.70	83.70	84.25
Ferret-7B [47]	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76	83.91
Ferret-13B [47]	89.48	92.41	84.36	82.81	88.14	75.17	85.83	86.34	85.57
MiniGPTv2 [3]	88.69	91.65	85.33	79.97	85.12	74.45	84.44	84.66	84.29
LION-4B	89.73	92.29	84.82	83.60	88.72	77.34	85.69	85.63	85.98
LION-12B	89.80	93.02	85.57	83.95	89.22	78.06	85.52	85.74	86.36

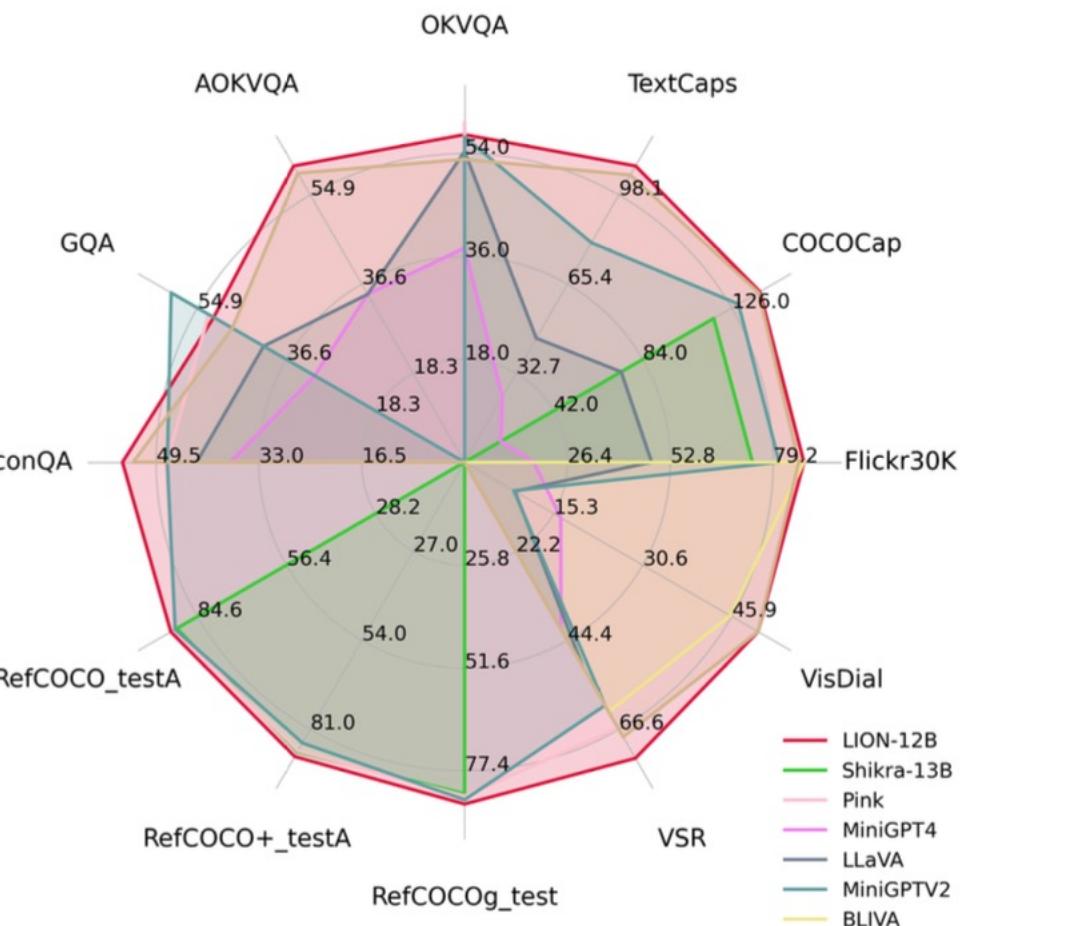


Image Captioning



User

What is in the photo?

InstructBLIP

boy in karate uniform with **white belt and black belts**
watching him do a kick in a martial arts class at a karate club
in california.

LION

A young boy wearing **a white karate uniform** is performing a kick in front of **a yellow wall**, with **the word "Kwon"** written on the wall behind him.

Visual Question Answering



User

On which side of the photo are the **clay pots**?

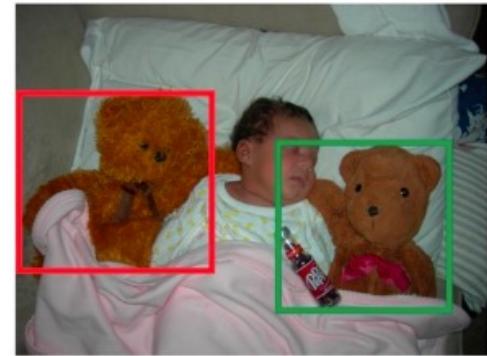
InstructBLIP

The pots are on **the left** of the image.

LION

The pots are on **the right** of the image.

Referring Expression Comprehension



User

Help me locate the coordinates of a brown bear **near a soda bottle** in the image.

Shikra

The brown bear near a soda bottle is located at
[0.002, 0.248, 0.416, 0.760]

LION

The brown bear near a soda bottle is located at
[0.550, 0.387, 0.920, 0.867]

Outline

- Tutorial

- LLaVA

- LION

- GILL

- Framework

- Experiments

- GLaMM

- SmartEdit

- Conclusion

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

Generating Images with Multimodal Language Models

Jing Yu Koh
Carnegie Mellon University
jingyuk@cs.cmu.edu

Daniel Fried
Carnegie Mellon University
dfried@cs.cmu.edu

Ruslan Salakhutdinov
Carnegie Mellon University
rsalakhu@cs.cmu.edu

NeurIPS 2023

Presenter: Hao Wang
Advisor: Prof. Chia-Wen Lin

Introduction



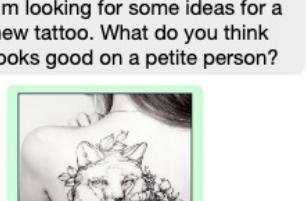
How can I make this more nutritious?

You can add vegetables to your ramen noodles, but you should be careful not to overdo it.



What are some vegetables I can add to it?

Broccoli, carrots, and green beans are all good choices.



I'm looking for some ideas for a new tattoo. What do you think looks good on a petite person?



I want a tattoo of a sunflower, but I'm not sure what style would look best on me.

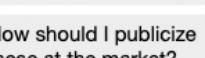


That's beautiful. Show me one that would work well as a complementary piece



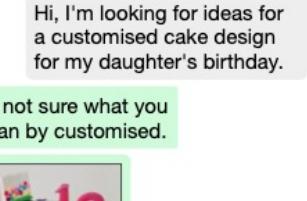


How should I publicize these at the market?



I would suggest a simple sign with the name of the business and a picture of the cupcakes.





Hi, I'm looking for ideas for a customised cake design for my daughter's birthday.

I'm not sure what you mean by customised.



I want a cake with a unicorn theme, with pastel colors and a whimsical feel.

I think



Can you also show me some decorations that would work well for the party?

I can show you some ideas

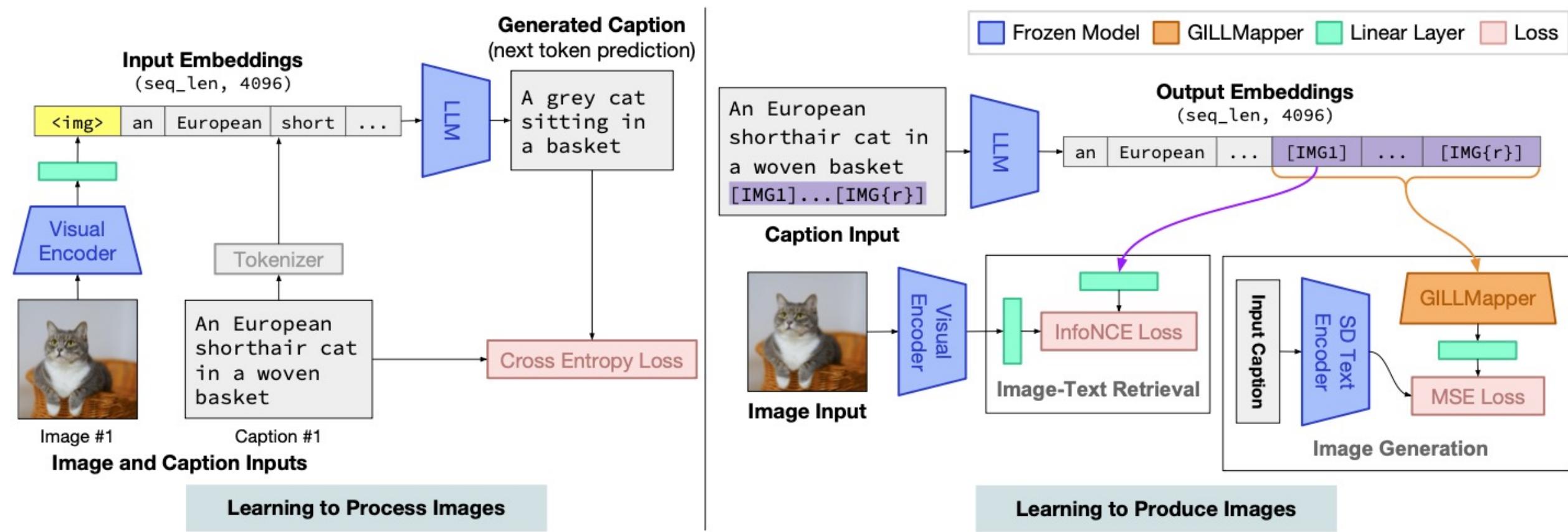


Multimodal Dialogue

Our model can generate multimodal dialogue, weaving together text, retrieved images, and generated images.

■ User prompts ■ Retrieved ■ Generated

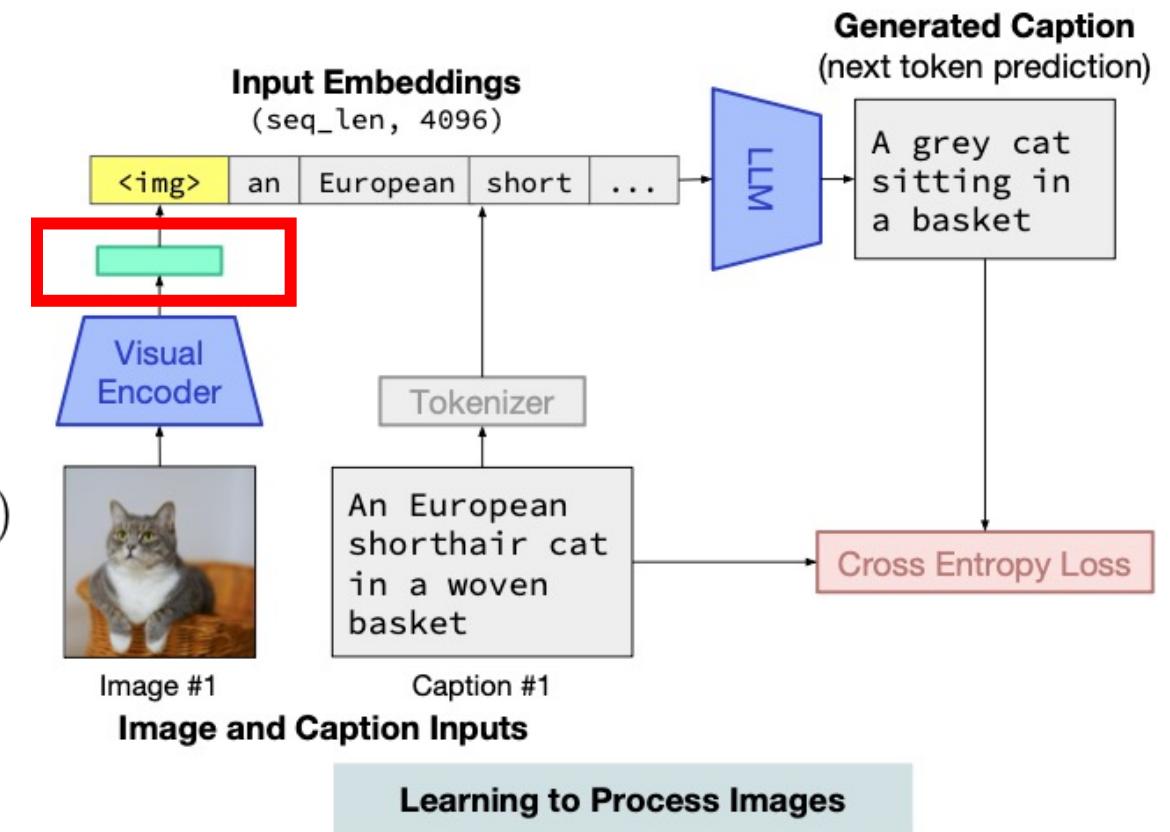
Framework



Method

- Learning to Process Images
 - v_ϕ is CLIP

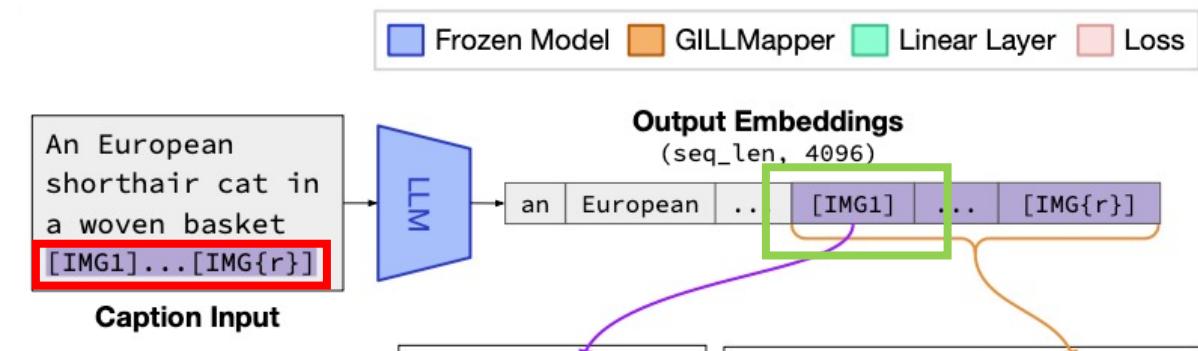
$$l_c(x, y) = - \sum_{t=1}^T \log p_\theta(s_t | v_\phi(x)^T \mathbf{W}_{\text{cap}}, s_1, \dots, s_{t-1})$$



Method

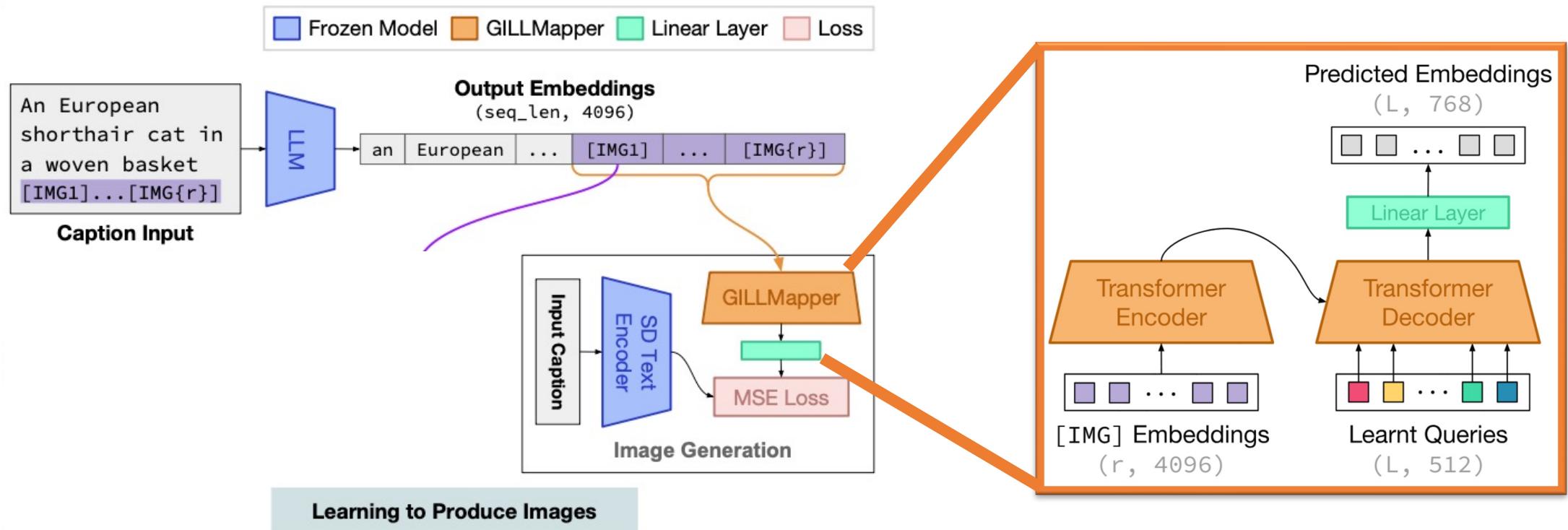
- Learning to Produce Images

$$l_p(y) = -\log p_{\{\theta \cup \mathbf{E}_{\text{img}}\}}([\text{IMG1}] | s_1, \dots, s_t)$$



Method

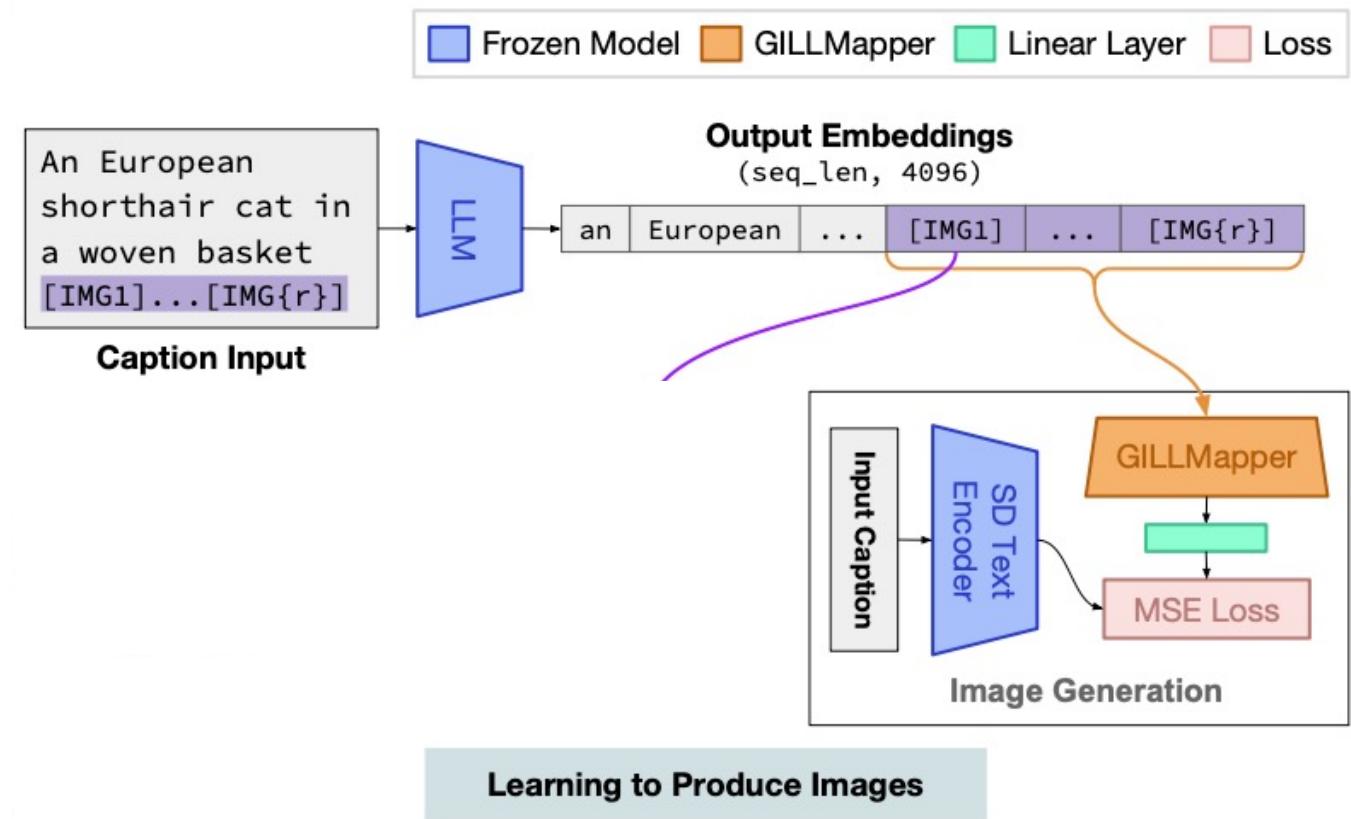
- Learning to Process Images
 - Novel Image Generation



$$l_g(y) = \| f_{\omega}(h_{\{\theta \cup \mathbf{E}_{\text{img}}\}}(y, [\text{IMG}\{1\}]), \dots, h_{\{\theta \cup \mathbf{E}_{\text{img}}\}}(y, [\text{IMG}\{r\}]), q_1, \dots, q_L) - T_{\psi}(y) \|_2^2$$

Method

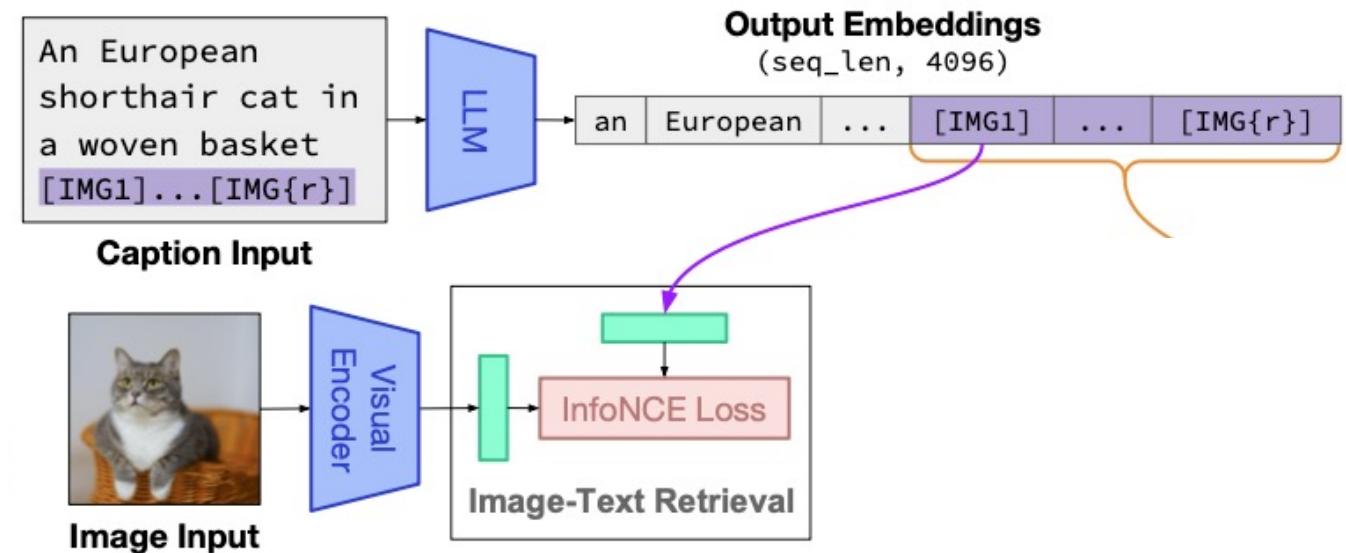
- Learning to Process Images
 - Novel Image Generation
 - G_ψ is Stable Diffusion



$$\text{Generated Image} = G_\psi(f_\omega(h_{\{\theta \cup \mathbf{E}_{\text{img}}\}}(y, [\text{IMG}\{1\}]), \dots, h_{\{\theta \cup \mathbf{E}_{\text{img}}\}}(y, [\text{IMG}\{r\}]), q_1, \dots, q_L))$$

Method

- Image Retrieval



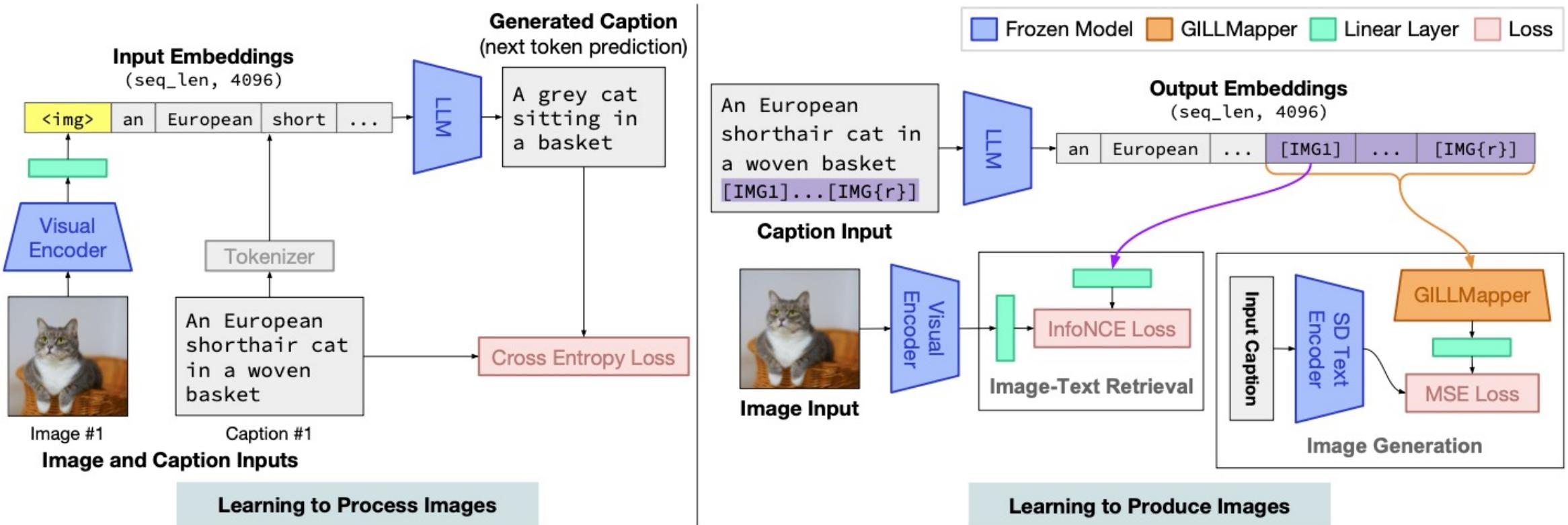
Learning to Produce Images

$$l_r(\mathbf{x}_i, \mathbf{y}_i) = -\log \frac{\exp(\text{sim}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{W}_{t2i})/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{x}_j, \mathbf{y}_i, \mathbf{W}_{t2i})/\tau)} - \log \frac{\exp(\text{sim}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{W}_{i2t})/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{x}_i, \mathbf{y}_j, \mathbf{W}_{i2t})/\tau)}$$

where the similarity is computed as

$$\text{sim}(x, y, \mathbf{W}) = \frac{(\mathbf{W}^T v_\phi(x))^T (\mathbf{W}^T h_{\{\theta \cup \mathbf{E}_{\text{img}}\}}(y, [\text{IMG1}]))}{\|\mathbf{W}^T v_\phi(x)\| \|\mathbf{W}^T h_{\{\theta \cup \mathbf{E}_{\text{img}}\}}(y, [\text{IMG1}])\|}$$

Method

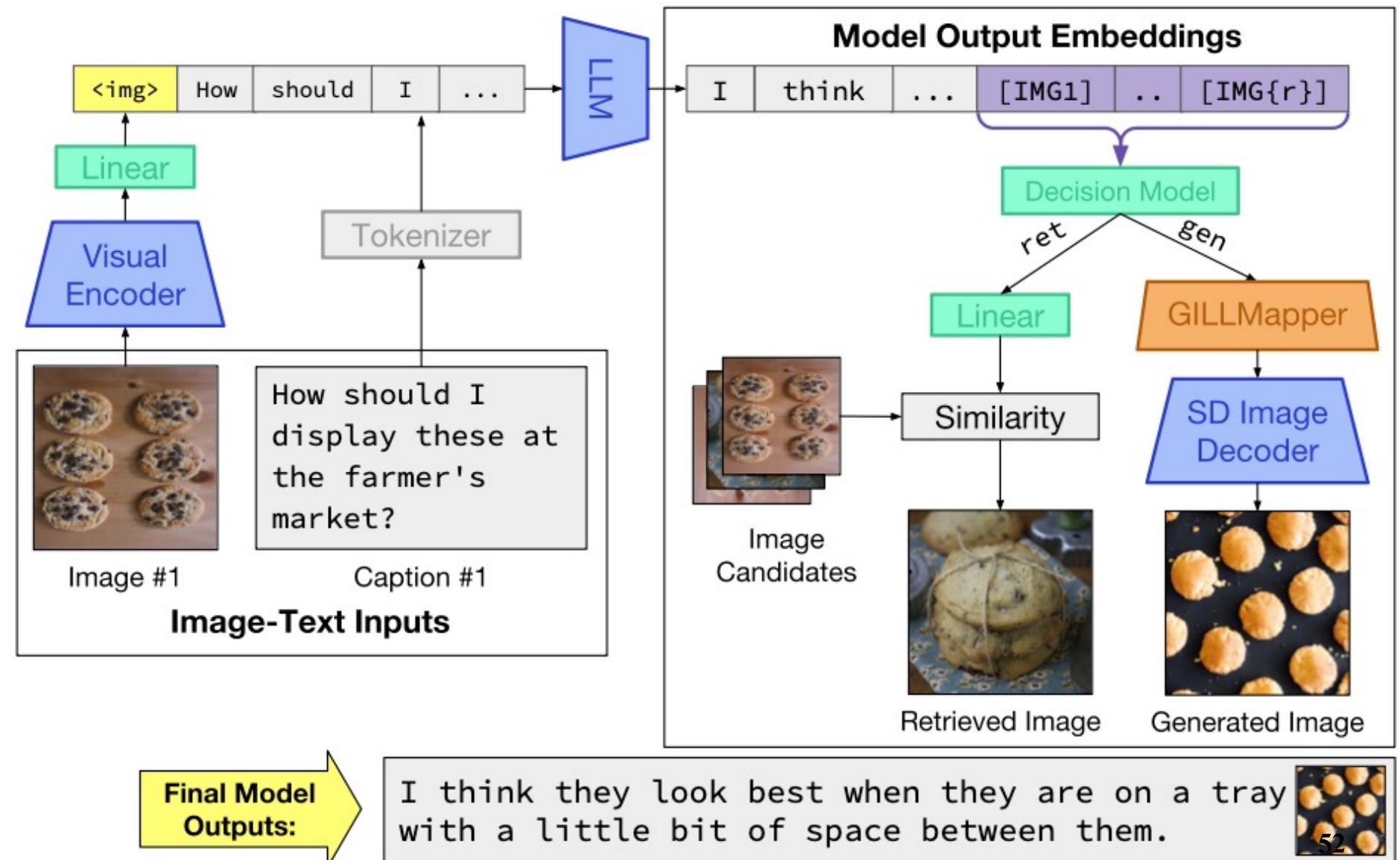


- Total loss

$$\min_{\mathbf{W}_{i2t}, \mathbf{W}_{t2i}, \mathbf{W}_{cap}, \mathbf{E}_{img}, \omega, q_{1:L}} \frac{1}{N} \sum_{i=1}^N (l_c(\mathbf{x}_i, \mathbf{y}_i) + l_p(\mathbf{y}_i) + l_g(\mathbf{y}_i) + l_r(\mathbf{x}_i, \mathbf{y}_i))$$

Method

- Inference



Outline

- Tutorial
- LLaVA
- LION
- GILL
- GLaMM
 - Framework
 - Experiments
- SmartEdit
- Conclusion

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

GLaMM: Pixel Grounding Large Multimodal Model

Hanoona Rasheed^{1*}, Muhammad Maaz^{1*}, Sahal Shaji¹, Abdelrahman Shaker¹, Salman Khan^{1 2}
Hisham Cholakkal¹, Rao M. Anwer^{1 3}, Eric Xing^{1 4}, Ming-Hsuan Yang^{5 7}, Fahad S. Khan^{1 6}

¹Mohamed bin Zayed University of AI, ²Australian National University, ³Aalto University

⁴Carnegie Mellon University, ⁵University of California - Merced, ⁶Linköping University, ⁷Google Research
hanoona.bangalath@mbzuai.ac.ae, muhammad.maaz@mbzuai.ac.ae

<https://github.com/mbzuai-oryx/groundingLMM>, <https://grounding-anything.com>

CVPR 2024

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin



A man and a boy sit on a bench next to an old white car.



A woman in a navy blue jacket and hat with a hair ribbon in her hair.



A soccer player in a red uniform is about to kick the ball while a player in a white uniform is trying to block the shot.



GLaMM : Grounding Large Multimodal Model

Haneem Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan,
Hisham Cholakkal, Rao M. Arwer, Eric Xing, Ming-Hsuan Yang, Fahad S. Khan



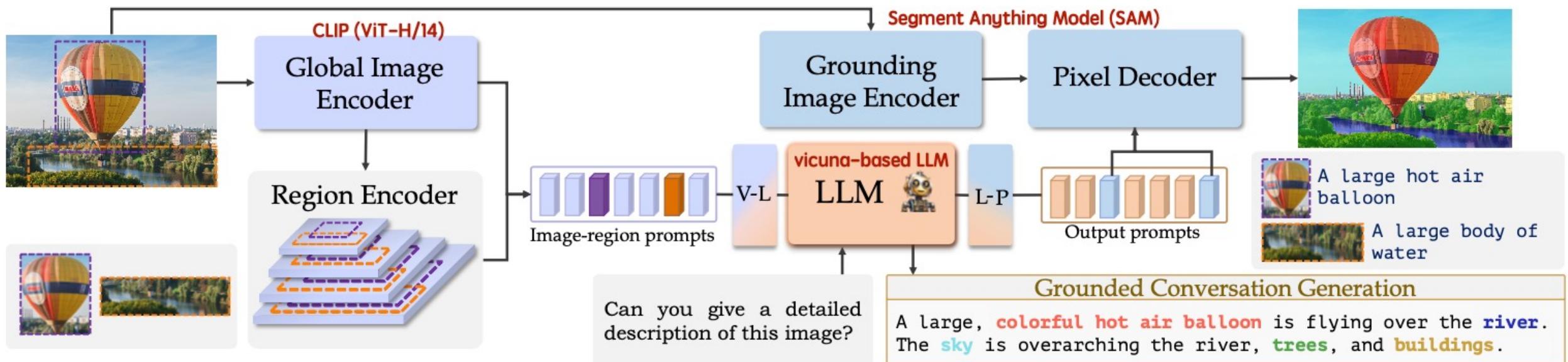
MBZUAI ORYX



MOHAMED BIN ZAYED UNIVERSITY
OF ARTIFICIAL INTELLIGENCE

IVAL
INTELLIGENT VISUAL ANALYTICS LAB

Method	Image	Input / Output		Region Enc. / Dec.	Pixel-Wise Grounding	Multi-turn Conversation	End-End Model
		Region	Multi-Region				
MM-REACT (arXiv-23) [51]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✗
LLaVA (NeurIPS-23) [29]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
miniGPT4 (arXiv-23) [61]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
mPLUG-OWL (arXiv-23) [52]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
LLaMA-Adapter v2 (arXiv-23) [8]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
Otter (arXiv-23) [22]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✗	✓
Instruct-BLIP (arXiv-23) [6]	✓	✗ / ✗	✗ / ✗	✗ / ✗	✗	✓	✓
InternGPT (arXiv-23) [31]	✓	✓ / ✗	✗ / ✗	✗ / ✗	✗	✓	✗
Bubo-GPT (arXiv-23) [59]	✓	✗ / ✓	✗ / ✓	✗ / ✗	✗	✓	✗
Vision-LLM (arXiv-23) [44]	✓	✗ / ✓	✗ / ✓	✗ / ✗	✗	✗	✓
Det-GPT (arXiv-23) [36]	✓	✓ / ✓	✓ / ✓	✗ / ✗	✗	✓	✓
Shikra (arXiv-23) [5]	✓	✓ / ✓	✗ / ✗	✗ / ✗	✗	✗	✓
Kosmos-2 (arXiv-23) [35]	✓	✓ / ✓	✓ / ✓	✗ / ✗	✗	✗	✓
GPT4RoI (arXiv-23) [57]	✓	✓ / ✗	✓ / ✗	✓ / ✗	✗	✓	✓
ASM (arXiv-23) [45]	✓	✓ / ✗	✗ / ✗	✓ / ✗	✗	✗	✓
LISA (arXiv-23) [21]	✓	✗ / ✓	✗ / ✗	✗ / ✓	✓	✗	✓
GLaMM (ours)	✓	✓ / ✓	✓ / ✓	✓ / ✓	✓	✓	✓



Referring Expression Segmentation

Can you segment the balloon in this given picture?

Sure, the segmentation is



Image Level Captioning

Please describe the given image.

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

Region Level Captioning



Could you tell me more about this region?

A yellow, red and blue hot air balloon

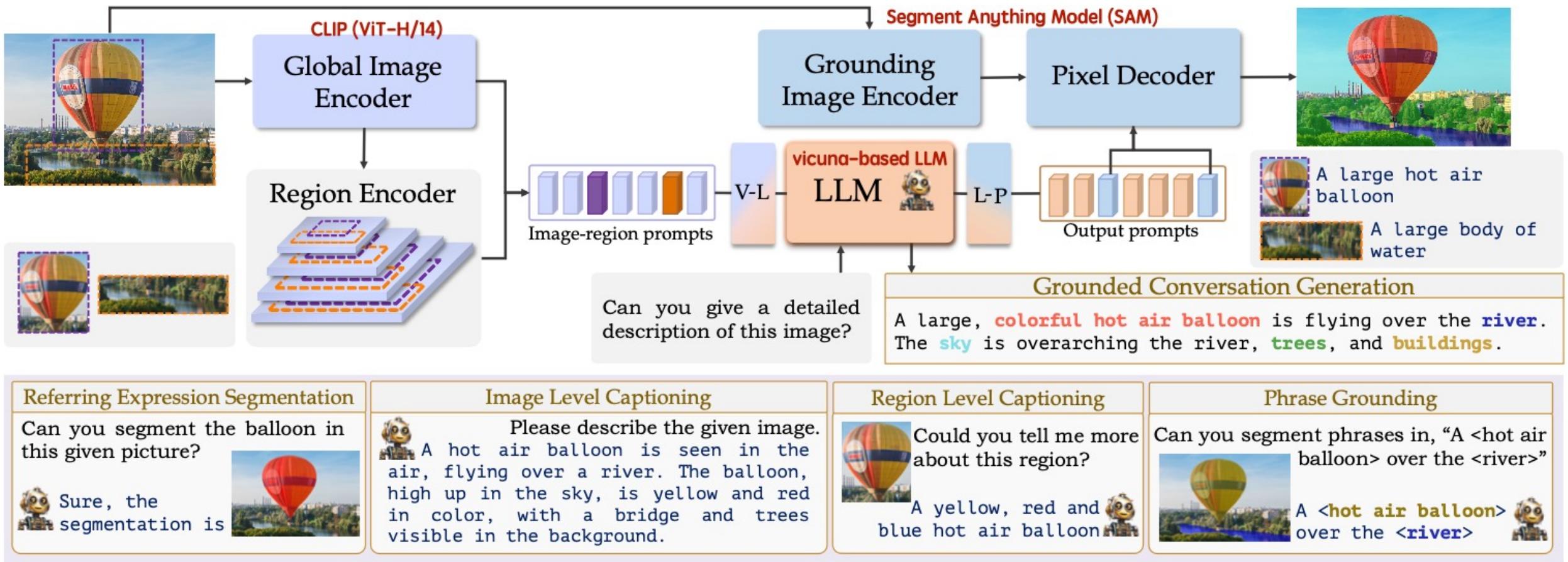
Phrase Grounding

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



A **<hot air balloon>** over the **<river>**





- Scene-Level Understanding

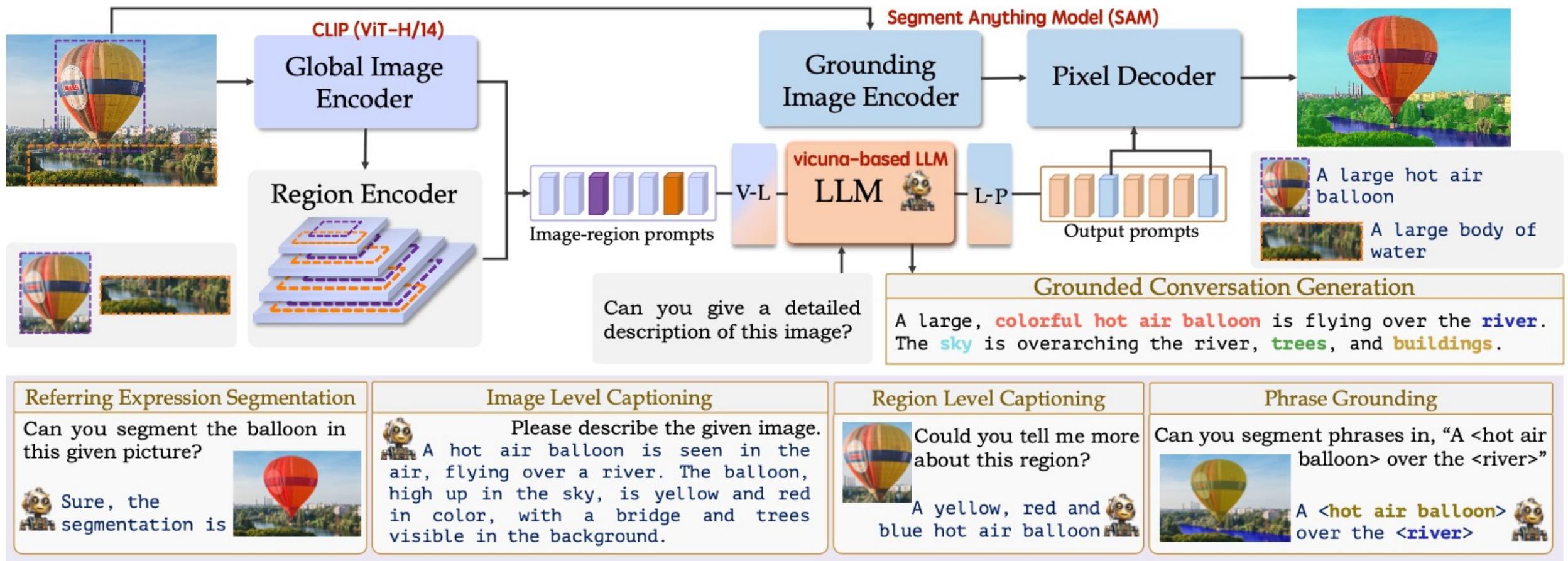
global image encoder (\mathcal{I})

$$I_x = \mathcal{I}(x_{\text{img}}) \in \mathbb{R}^{D_v}$$

$$f(I_x) \in \mathbb{R}^{D_t}$$

$$y_t = \mathcal{L}\left(f(I_x), x_t\right)$$

- “The **image** provides an overview of the image. Could you please give me a detailed description of the image?”



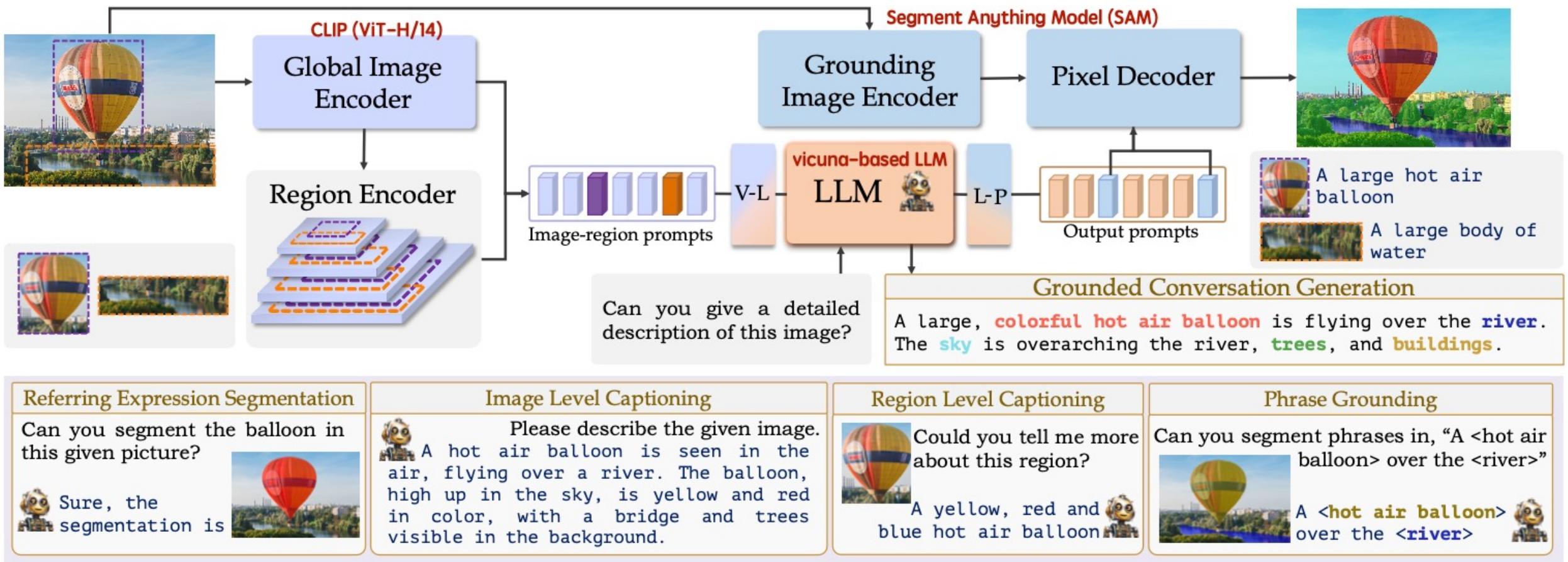
- Region-Level Understanding

$$R_x = \mathcal{R}(I_x, r)$$

$$x'_t = [x_t \leftarrow f(R_x)]$$

- “The **<image>** provides an overview of the image. Can you provide a detailed description of the region **<bbox>**?”

$$y_t = \mathcal{L}\left(f(I_x), x'_t\right)$$



• Pixel-Level Grounding

- grounding image SAM encoder denoted as V and the pixel SAM decoder represented as P

$$M = \mathcal{P} \left(g(l_{seg}), \mathcal{V}(x_{img}) \right), \text{ s.t., } M_i \in \{0, 1\}$$

Outline

- Tutorial
- LLaVA
- LION
- GILL
- GLaMM
 - Framework
 - Experiments
- SmartEdit
- Conclusion

Objects and Attributes

- 1 dog, pub dog, a brown and white dog
- 2 dog collar, black color, chain collar
- 3 bell, cowbell
- 4 steps, stairs, the steps of a building
- 5 sack, a large white bag with black writing

Relationships and Landmarks

A dog sitting on the steps

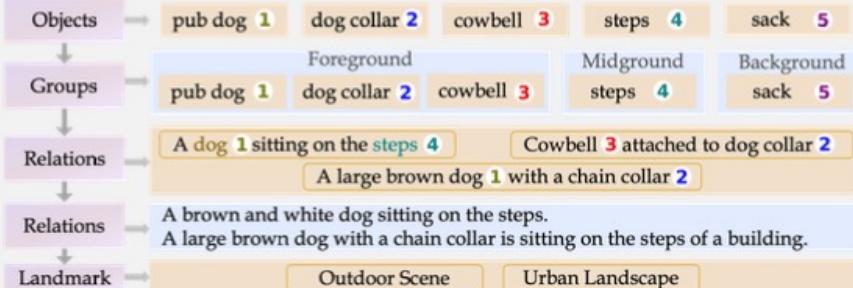
A large brown dog wearing a chain collar

Cowbell attached to dog collar

Landmarks: Outdoor – Urban Landscape



Scene Graph



Dense Grounded Caption

A large brown dog is sitting on the steps of a building. It is wearing a black chain dog collar. The collar has a cowbell attached to it. There is a bag in the background with black writings on it.

Extra Context

Dogs, especially pugs and bulldogs, have been a part of human families for thousands of years, serving as loyal companions. They have been bred for specific traits, making them popular pets. Dogs have been trained for various tasks, including assisting people with disabilities and serving as search and rescue animals. Dog collars, often bearing identification tags, are essential for keeping pets safe and ensuring they can be returned home if lost. Cowbells, once used to signal the arrival of a cow, have been repurposed as dog collars, providing a distinct sound to help locate a dog if it wanders off. In outdoor urban landscape, dogs are often found sitting on steps, as they may choose to rest in spots that offer a good view of their surroundings.

Level-1

Object localization and attributes

- Image Tagging and Object Detection
- Open Vocabulary Detection
- Region Attribute Detection

Level-2

Relationships

- Short Captions and Phrase extraction
- Grounding expression
- Landmarks

Level-3

Scene Graph & Dense Captioning

- Hierarchical Scene Graph
- In-context Learning with LLM
- Verification Pipeline

Level-4

Extra Contextual Insights

- Lanmark Details
- History and Background
- Precautionary Measures

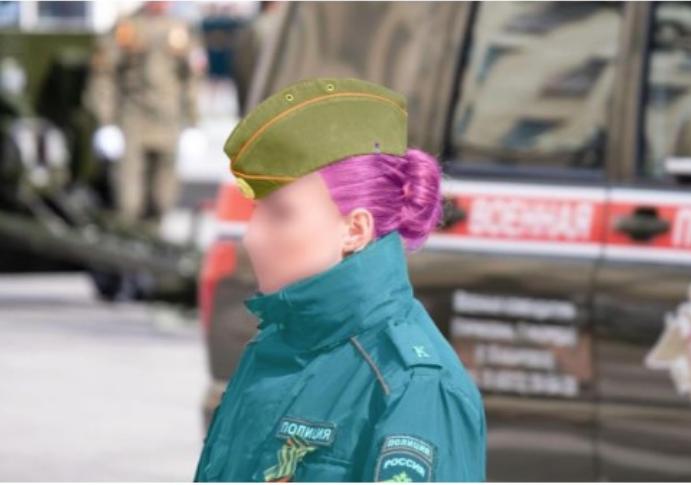
Dataset	Images	Regions	Concepts	Tokens	Captions[†]
COCO [25]	0.1M	0.9M	80	-	-
LVIS [9]	0.1M	1.5M	1,203	-	-
Objects365 [42]	0.6M	10.1M	365	-	-
Open Images [20]	1.5M	14.8M	600	-	-
BigDetection [4]	3.5M	36.0M	600	-	-
V3Det [43]	0.2M	1.5M	13,029	-	-
VG [19]	0.1M	0.3M	18,136	51.2M	-
SA-1B [18]	11M	1.1B	-	-	-
AS-1B [45]	11M	1.2B	3.5M	132.2B	-
GranD (Ours)	11M	810M	7.5M	5.0B	33M

Model	Validation Set					Test Set				
	M	C	AP50	mIoU	Recall	M	C	AP50	mIoU	Recall
BuboGPT [59]	17.2	3.6	19.1	54.0	29.4	17.1	3.5	17.3	54.1	27.0
Kosmos-2 [35]	16.1	27.6	17.1	55.6	28.3	15.8	27.2	17.2	56.8	29.0
LISA* [21]	13.0	33.9	25.2	62.0	36.3	12.9	32.2	24.8	61.7	35.5
GLaMM†	15.2	43.1	28.9	65.8	39.6	14.6	37.9	27.2	64.6	38.0
GLaMM	16.2	47.2	30.8	66.3	41.8	15.8	43.5	29.2	65.6	40.8

Method	refCOCO			refCOCO+			refCOCOg	
	val	testA	testB	val	testA	testB	val(U)	test(U)
CRIS [47]	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4
LAVT [50]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
GRES [26]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
X-Decoder [63]	-	-	-	-	-	-	64.6	-
SEEM [64]	-	-	-	-	-	-	65.7	-
LISA-7B [21]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
GLaMM	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9



A man and a boy sit on a bench next to an old white car.



A woman in a navy blue jacket and hat with a hair ribbon in her hair.



A soccer player in a red uniform is about to kick the ball while a player in a white uniform is trying to block the shot.



A Yellow Volkswagen Beetle

Is this car a new model?

No, it is an old model



Can you please segment the smaller cup now?



Two girls in costume stand in front of a black curtain with a smiling face and skeleton head on it. They are next to a poster of a man in a tuxedo.

Outline

- Tutorial
- LLaVA
- LION
- GILL
- GLaMM
- SmartEdit
 - Framework
 - Experiments
- Conclusion

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

- SmartEdit

SmartEdit: Exploring Complex Instruction-based Image Editing with Multimodal Large Language Models

Yuzhou Huang^{*1,2#} Liangbin Xie^{*2,3,5#} Xintao Wang^{2,4†} Ziyang Yuan^{2,8#} Xiaodong Cun⁴
Yixiao Ge^{2,4} Jiantao Zhou³ Chao Dong^{5,7} Rui Huang⁶ Ruimao Zhang^{1†} Ying Shan^{2,4}

¹School of Data Science, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China

²ARC Lab, Tencent PCG ³University of Macau ⁴Tencent AI Lab

⁵Shenzhen Institute of Advanced Technology ⁶School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China

⁷Shanghai Artificial Intelligence Laboratory ⁸Tsinghua University

CVPR 2024

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Introduction

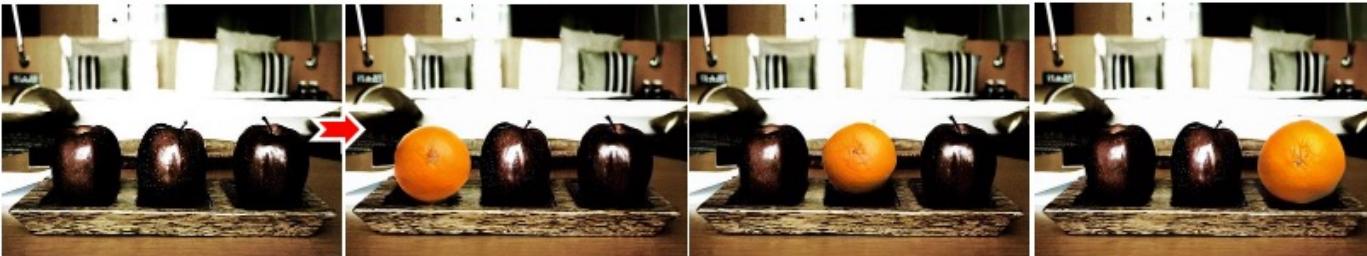
"Change the left/right animal to a white fox"



"Change the bigger/smaller bear to a wolf"



"Change the left/middle/right apple to an orange"



"Change the red/green apple to a peach"



"Change the dog in mirror to a tiger"



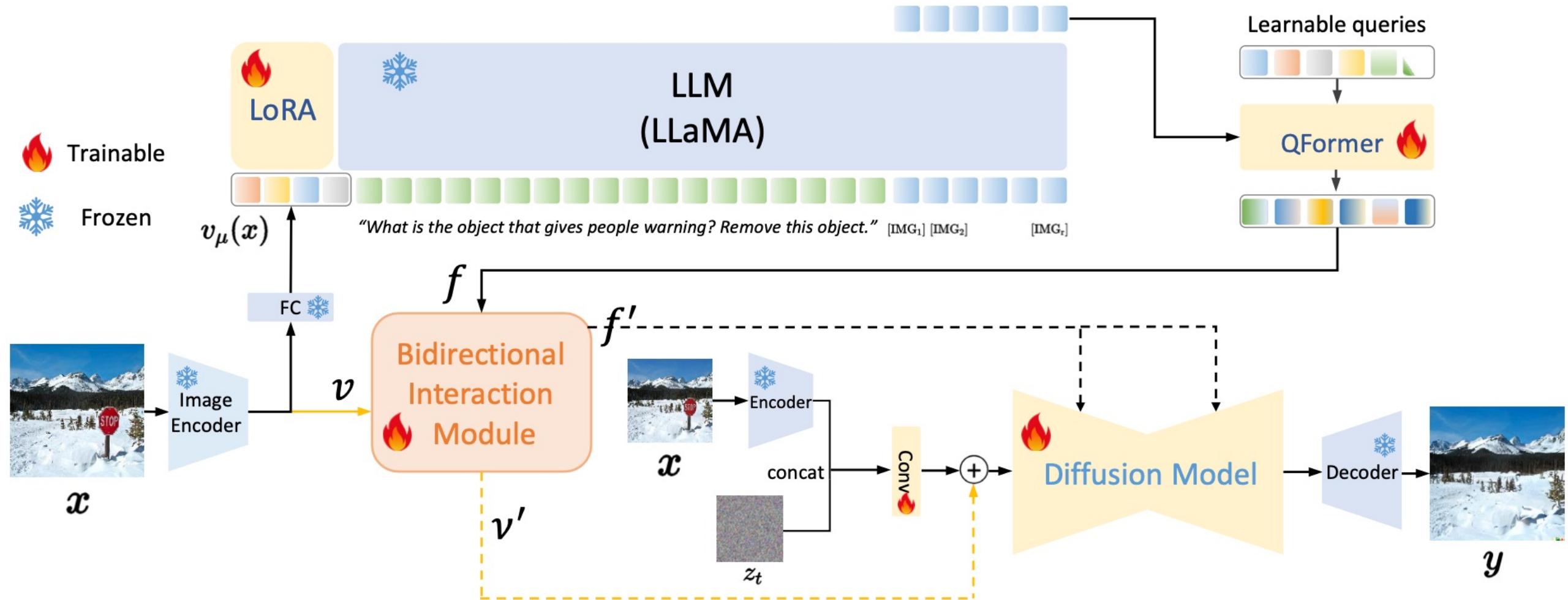
"Please replace the animal that is usually known as friend of human's with a tiger"



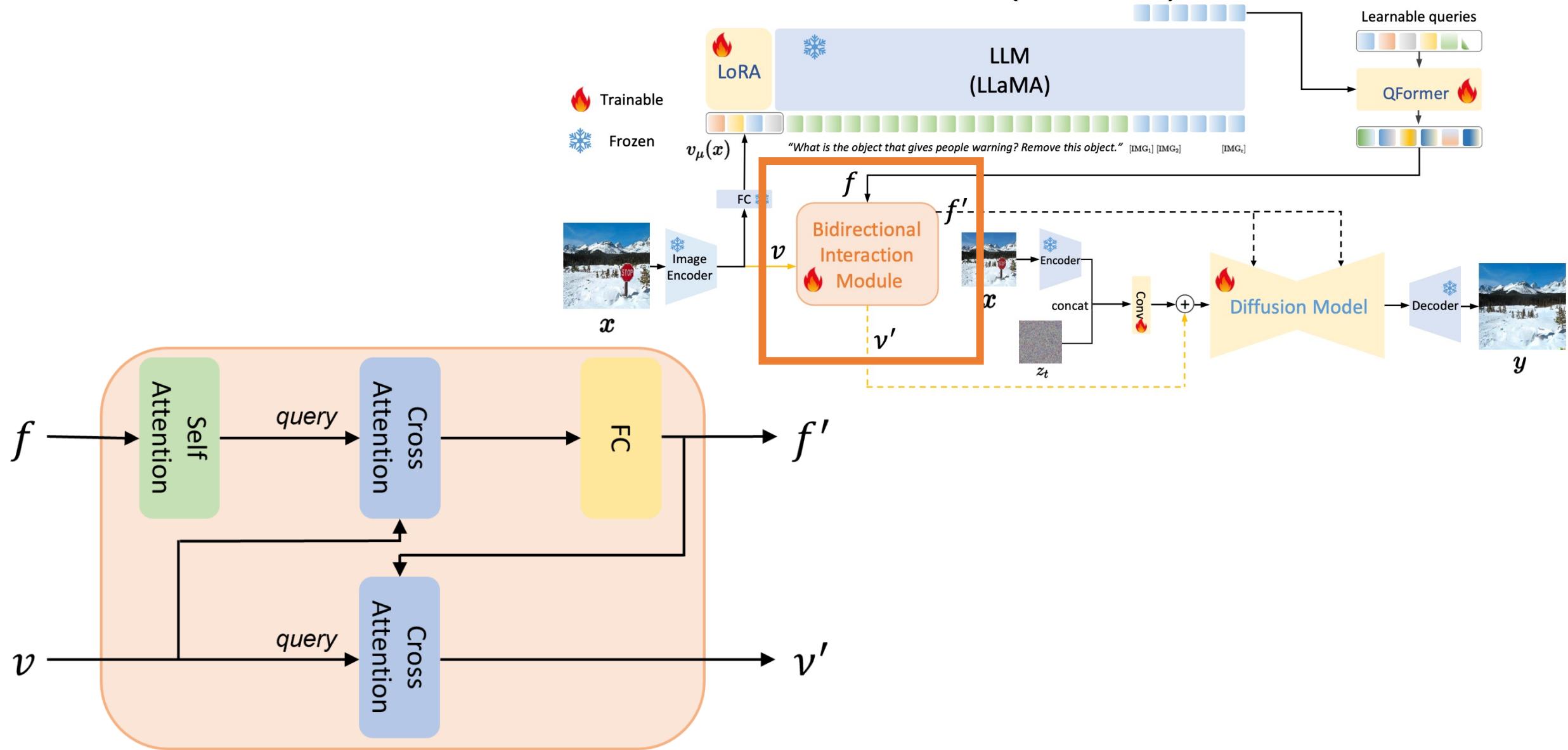
"Please remove the object that can tell the time"



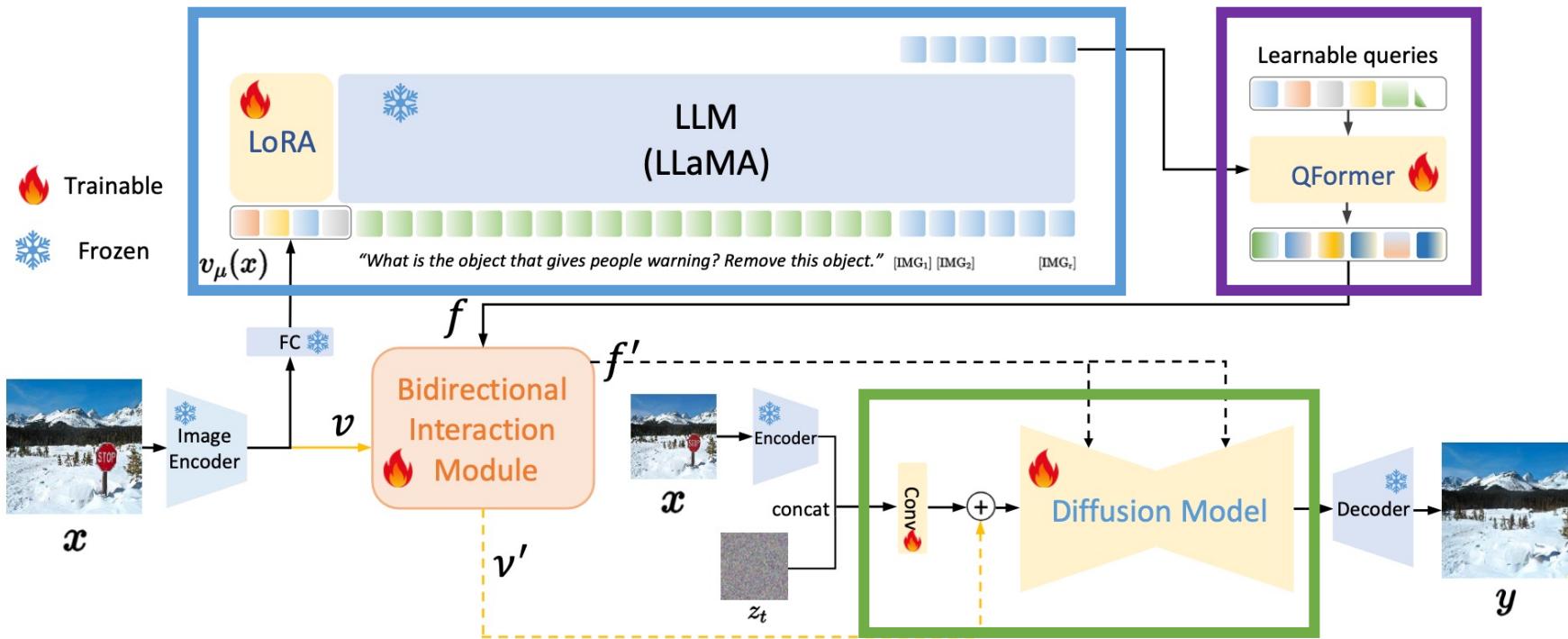
Framework



Bidirectional Interaction Module (BIM)



Implementation Details



- The first stage training objectives are the combination of the **MSE loss** between the output of LLaVA and CLIP text encoder, and the **language model loss**.
- In the second stage, the loss function is composed of two parts: **the language model loss** and the **diffusion loss**.

Outline

- Tutorial
- LLaVA
- LION
- GILL
- GLaMM
- SmartEdit
 - Framework
 - Experiments
- Conclusion

Quantitative comparison

Methods	Understanding Scenarios					Reasoning Scenarios				
	PSNR(dB)↑	SSIM↑	LPIPS↓	CLIP Score↑	Ins-align↑	PSNR(dB)	SSIM	LPIPS	CLIP Score	Ins-align↑
InstructPix2Pix	21.576	0.721	0.089	22.762	0.537	24.234	0.707	0.083	19.413	0.344
MagicBrush	18.120	0.68	0.143	22.620	0.290	22.101	0.694	0.113	19.755	0.283
InstructDiffusion	23.258	0.743	0.067	23.080	0.697	21.453	0.666	0.117	19.523	0.483
SmartEdit-7B	22.049	0.731	0.087	23.611	0.712	25.258	0.742	0.055	20.950	0.789
SmartEdit-13B	23.596	0.751	0.068	23.536	0.771	25.757	0.747	0.051	20.777	0.817

- Adopt three metrics for the **background area**: PSNR, SSIM, and LPIPS.
- For the **foreground area**, we calculate the CLIP Score.
- **Human evaluate** the results on Reason-Edit, which is Instruction-Alignment (Ins-align).

"What is the tool that is used to cut fruits. Remove this tool."



"Please replace food contains most vitamin with an orange."



"Please remove the object that can be used to have meals."



"Change the middle panda to a cat"



"Change the cat in mirror to a tiger."



Input Image

InstructPix2Pix

MagicBrush

InstructDiffusion

SmartEdit-7B

SmartEdit-13B

Conclusion

- Tutorial
- LLaVA

- LION
- GILL

- GLaMM
- SmartEdit