UCF Center for Research in Computer Vision

Visual-Language Models Introduction
Part-V: Video LLaVA, Shikara
Lecture-8
CAP6412 Spring 2025

Mubarak Shah

shah@crcv.ucf.edu

Contents

- Lecture-4
  - CLIP
  - CoCA
  - PALI
- Lecture-5
  - FLAMINGO
  - FLAVA
  - Painter
  - BLIP-2
- Lecture-6
  - Image-Bind
  - Language-Bind
  - LLaVA
- Lecture-7
  - Video ChatGPT
  - PG-Video LLaVA
- Lecture-8
  - Video LLaVA
  - Shikra

Video-LLaVA: Learning United Visual Representation by Alignment Before Projection

Bin Lin[1], Yang Ye[1], Bin Zhu[1], Jiaxi Cui[4], Munang Ning[1,2,3], Peng Jin[1,2,3], Li Yuan[1,2,3]

[1]Peking University Shenzhen Graduate School, [2]Peng Cheng Laboratory,
[3]AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School,
[4]PandaVilla Tech Limited

Correspondence: yuanli-ece@pku.edu.cn

GitHub: https://github.com/PKU-YuanGroup/Video-LLaVA

Abstract

Large Vision-Language Model (LVLM) has enhanced the performance of various downstream tasks in visual-language understanding. Most existing approaches encode images and videos into separate feature spaces, which are then fed as inputs to large language models. However, due to the lack of unified tokenization for images and videos, namely misalignment before projection, it becomes challenging for a Large Language Model (LLM) to learn multi-modal interactions from several poor projection layers. In this work, we unify visual representation into the language feature space to advance the foundational LLM towards a unified LVLM. As a result, we establish a simple but robust LVLM baseline, **Video-LLaVA**, which learns from a mixed dataset of images and videos, mutually enhancing each other. As a result, Video-LLaVA outperforms Video-ChatGPT by **5.8%, 9.9%, 18.6%, and 10.1%** on MSRVTT, MSVD, TGIF, and ActivityNet, respectively. Additionally, our Video-LLaVA also achieves superior performances on a broad range of 9 image benchmarks. Notably, extensive experiments demonstrate that Video-LLaVA mutually benefits images and videos within a unified visual representation, outperforming models designed specifically for images or videos. We aim for this work to provide modest insights into the multi-modal inputs for the LLM.

1 Introduction

Recently, LLMs have gained rapid popularity in the AI community, such as GPT-3.5, GPT-4 (OpenAI, 2023), PaLM (Bi et al., 2020; Anil et al., 2023), and BLOOM (Scao et al., 2022). They rely on their powerful language comprehension abilities

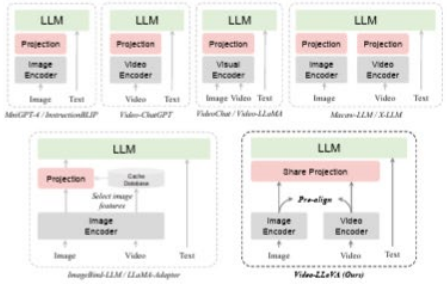


Figure 1: **Comparing Different LVLM Paradigms.** Video-LLaVA aligns images and videos before projection, allowing LLM to learn from a unified visual representation and endowing LLM with the ability to comprehend both images and videos simultaneously.

works (Ye et al., 2023; Zhu et al., 2023b; Alayrac et al., 2022) have mapped images into text-like tokens, enabling LLMs to emerge with the ability to comprehend images. Despite their effectiveness, empowering LLMs to understand videos is more challenging than image-only comprehension tasks. Nevertheless, recent work (Maaz et al., 2023; Li et al., 2023c; Zhang et al., 2023a) has made initial strides in enabling interactions between video and language.

However, most current LVLMs (Li et al., 2023b; Dai et al., 2023; Luo et al., 2023; Li et al., 2023a; Yin et al., 2023; Fu et al., 2023) can primarily handle a single visual modality, either image-language or video-language. We compare different LVLM paradigms as shown in Figure 1, where VideoChat (Li et al., 2023c) and Video-LLaMA (Zhang et al., 2023a) utilize a share visual

Video-LLaVA: Learning United Visual Representation by Alignment Before Projection

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munang Ning, Peng Jin, Li Yuan

Arxiv Preprint 2023

12 citations: 2024

423 citations: 2025

https://www.y outube.com/watch?v=x0dTQcpCA9E

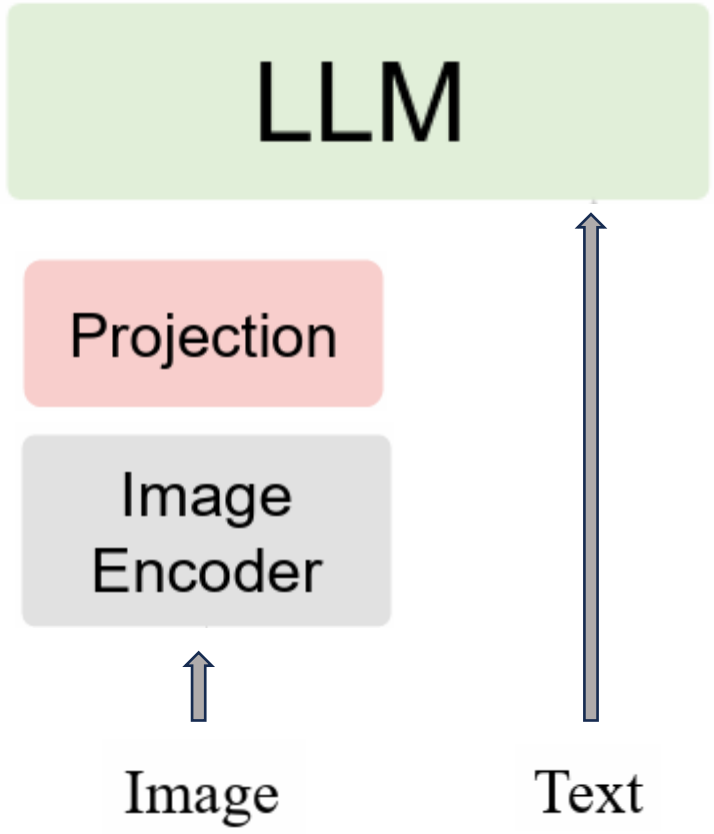
Presented By:

David Shatwell Pittaluga, Anthony Bilic, Kevin Zhai, Zain Ulabedeen Farhat, Kunyang Li

Background/Motivation

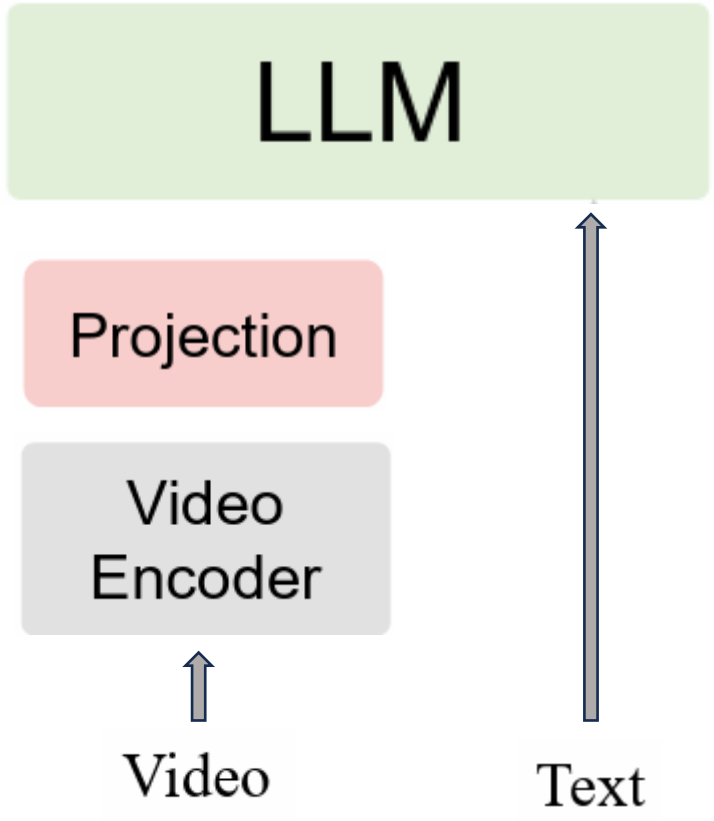
LVLM Paradigms

- MiniGPT-4, InstructionBLIP, LLaVA (Image Only)



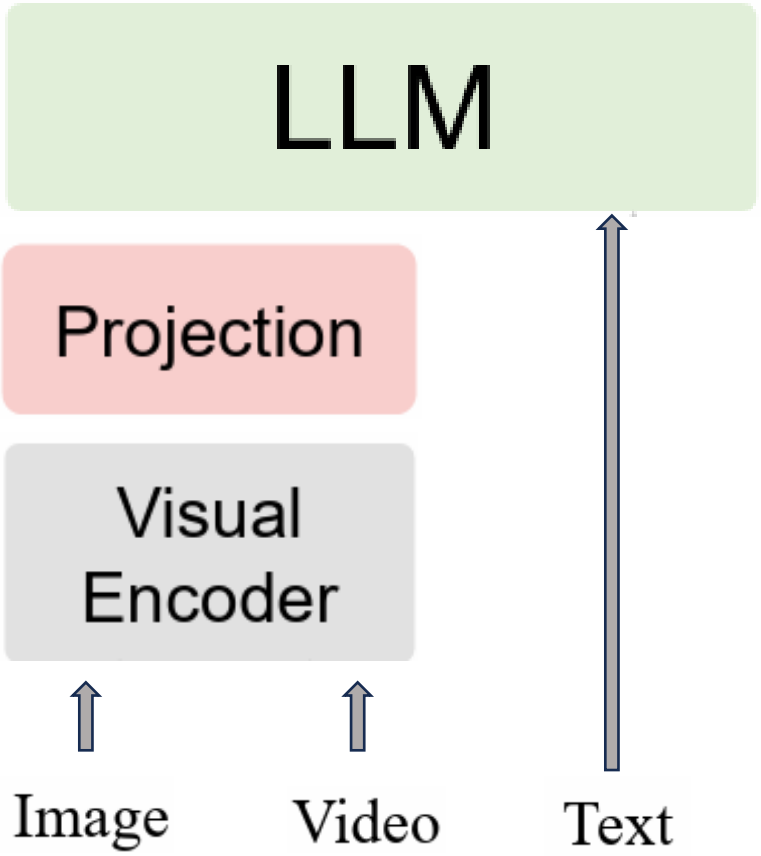
LVLM Paradigms

- Video ChatGPT (Video Only)



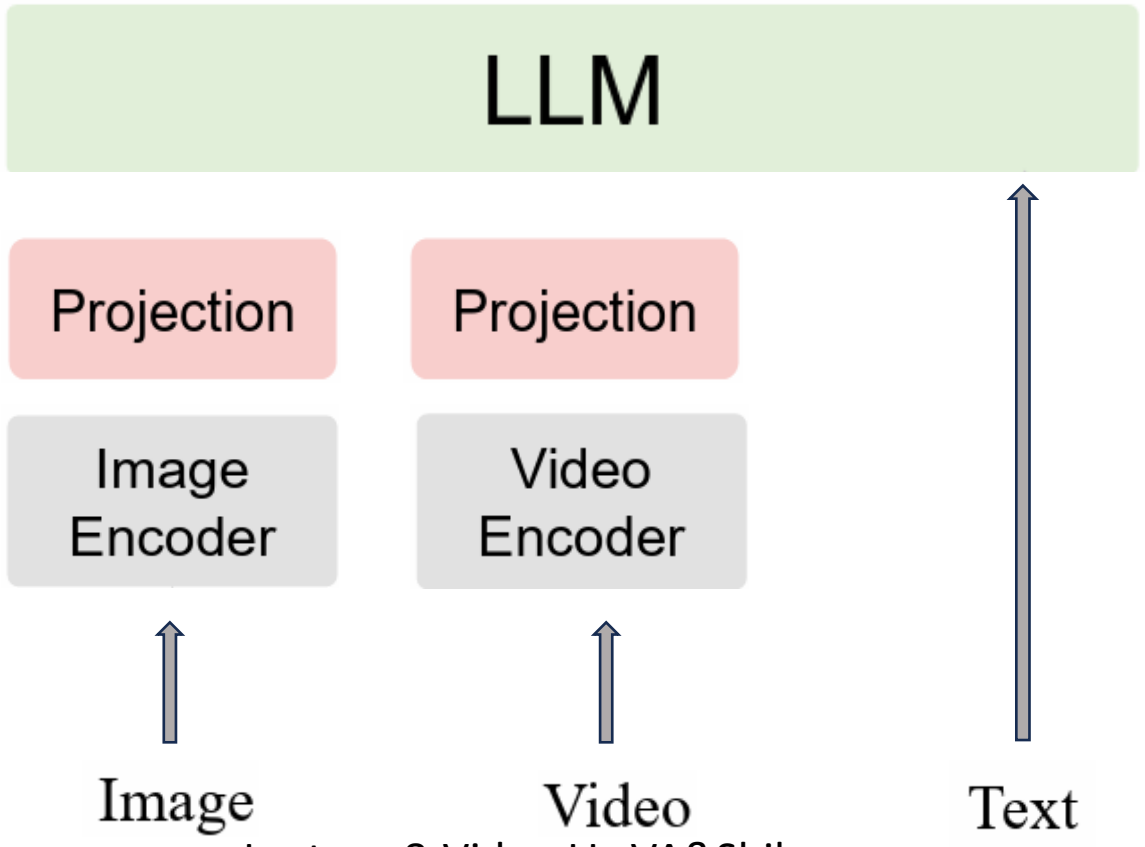
LVLM Paradigms

- VideoChat, Video-LLaMA (Image + Video - Shared)

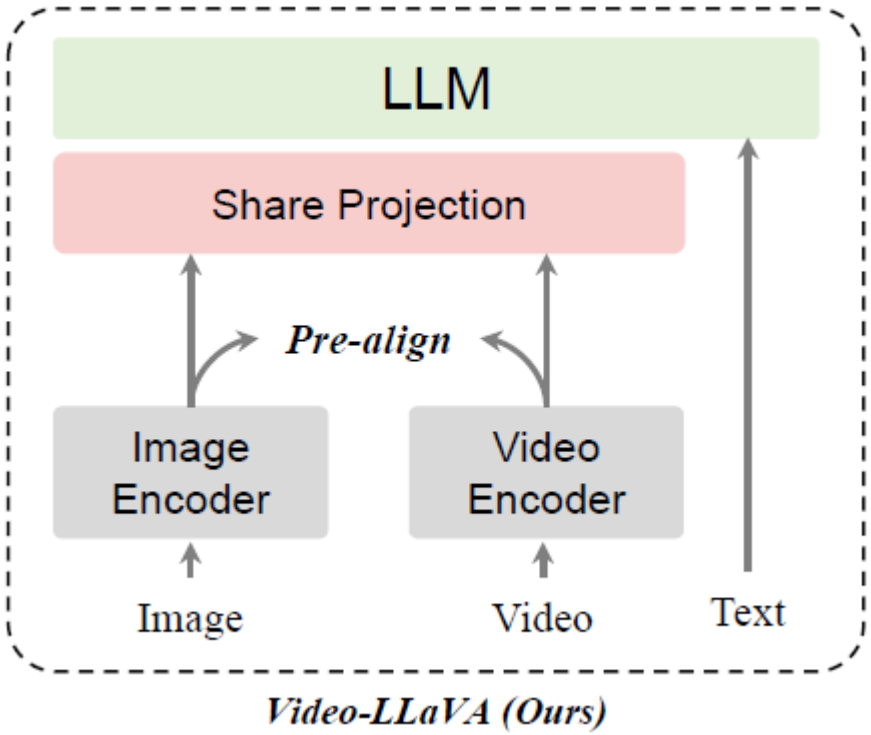


LVLM Paradigms

- Macaw-LLM, X-LLM (Image + Video - Separate)



LVLM Paradigms



Video-LLaVA (Ours)

Model Architecture

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.*



Large Language Model f_L
Vicuna v1.5

V V V V V V V T T T T T

Share Projection f_P

Word Embedding Layer f_W

Encoder Zoo
LanguageBind f_V

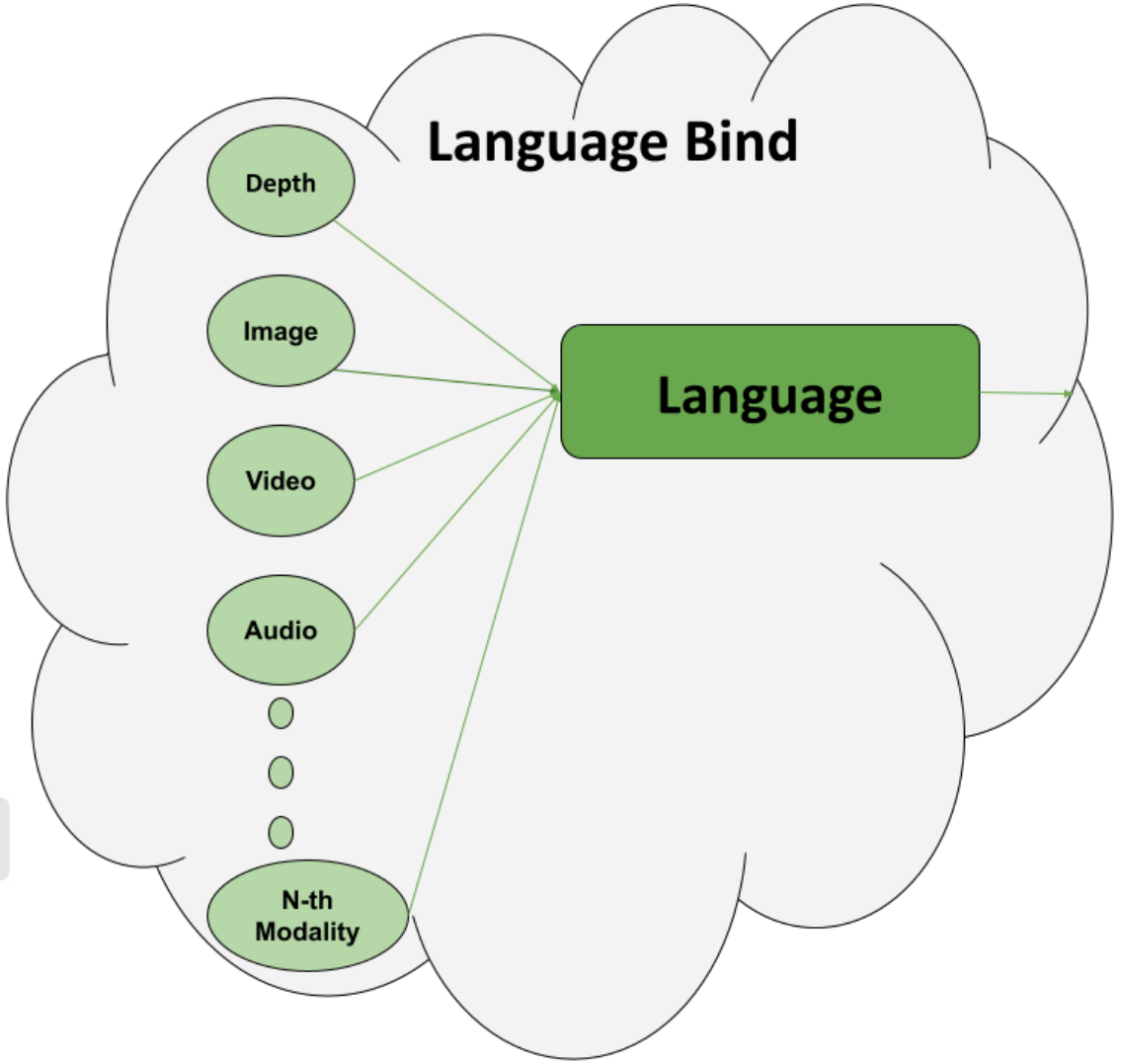


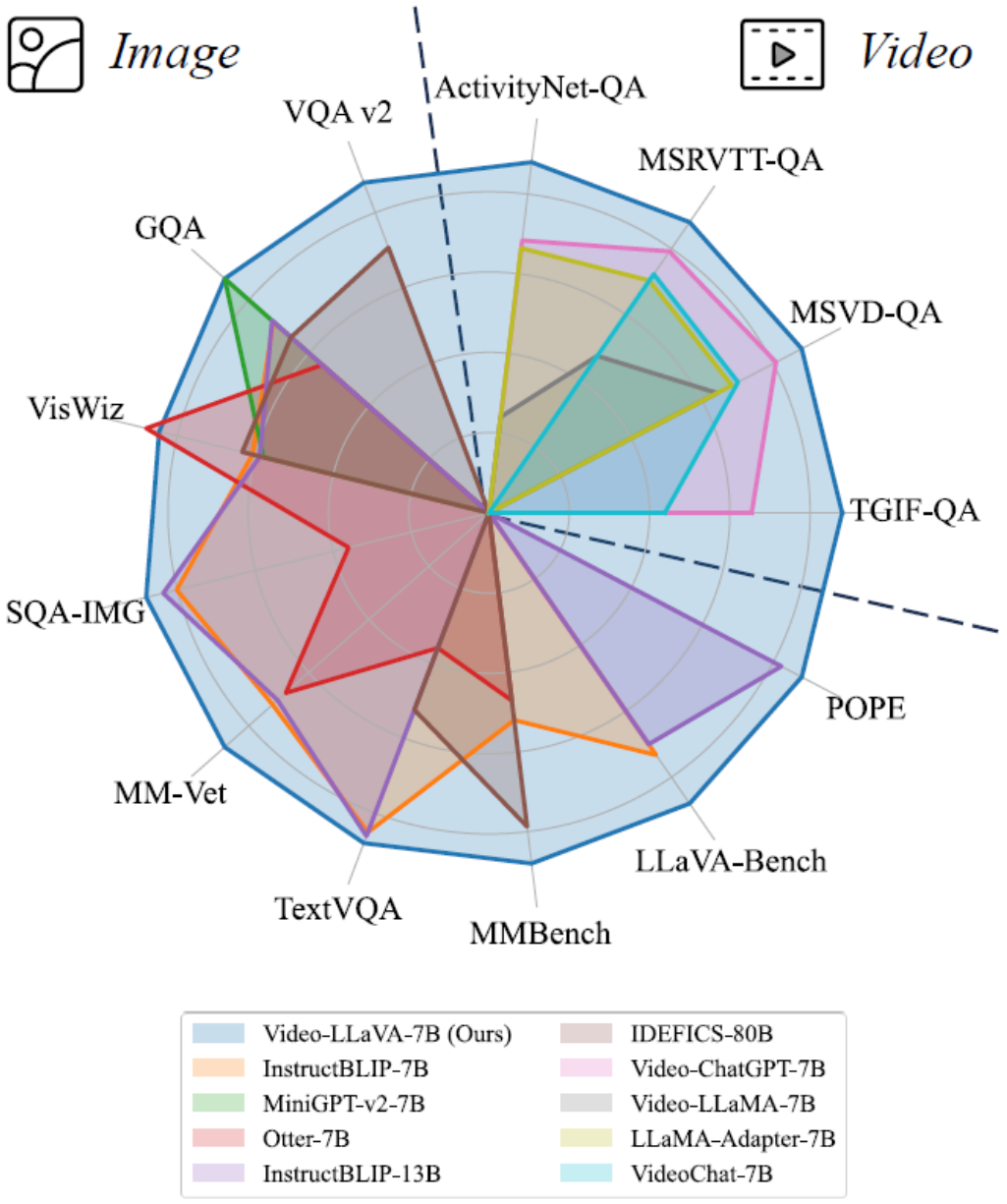
Image



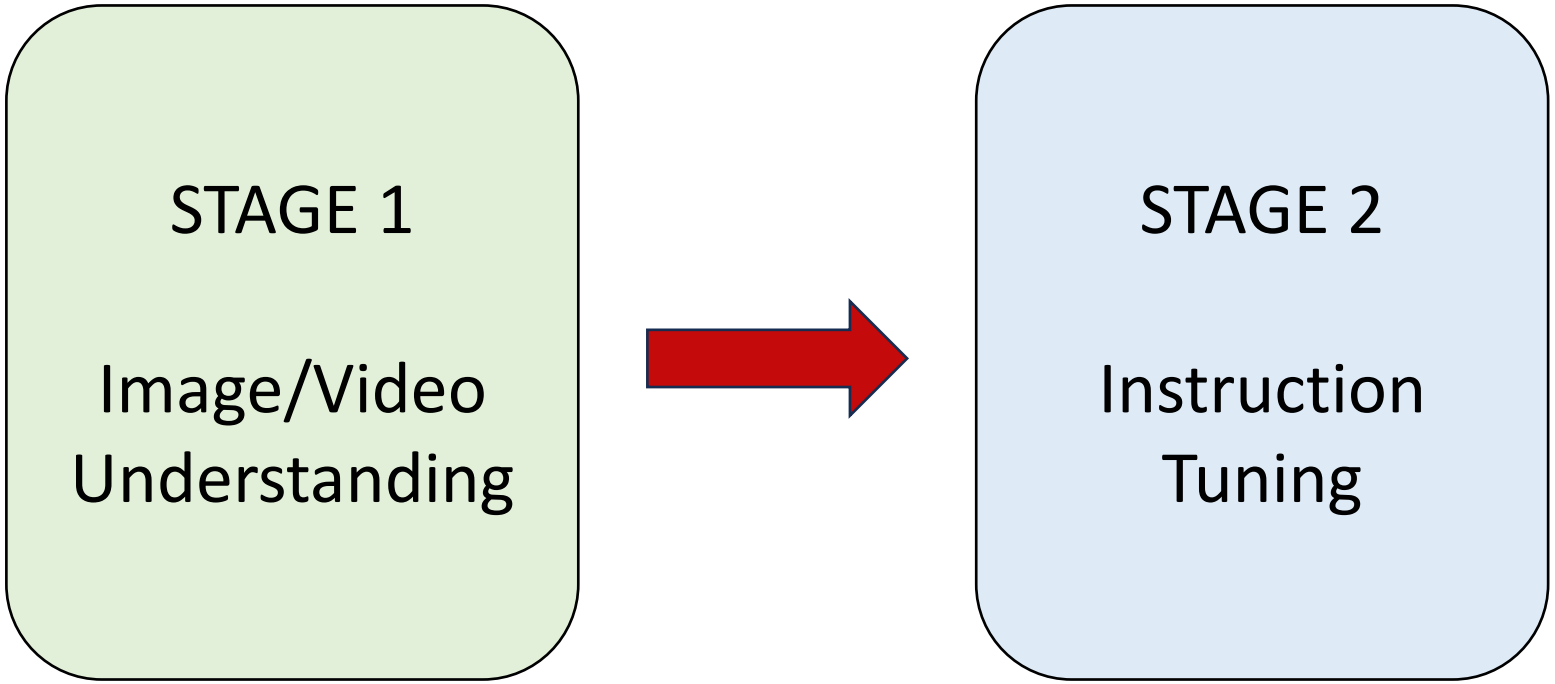
Video

Are ***the image and the video*** *depicting the same place?*



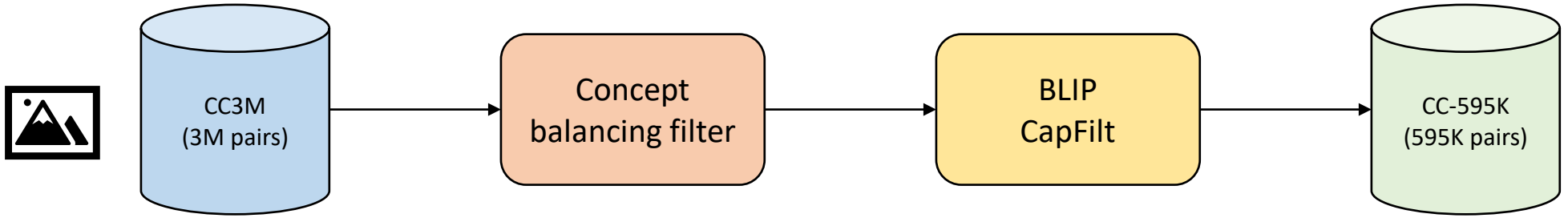

Image
Video
VQA v2
ActivityNet-QA
MSRVTT-QA
GQA
MSVD-QA
VisWiz
TGIF-QA
SQA-IMG
POPE
MM-Vet
LLaVA-Bench
TextVQA
MMBench
Video-LLaVA-7B (Ours)
InstructBLIP-7B
MiniGPT-v2-7B
Otter-7B
InstructBLIP-13B
IDEFICS-80B
Video-ChatGPT-7B
Video-LLaMA-7B
LLaMA-Adapter-7B
VideoChat-7B

Training Pipeline: Overview

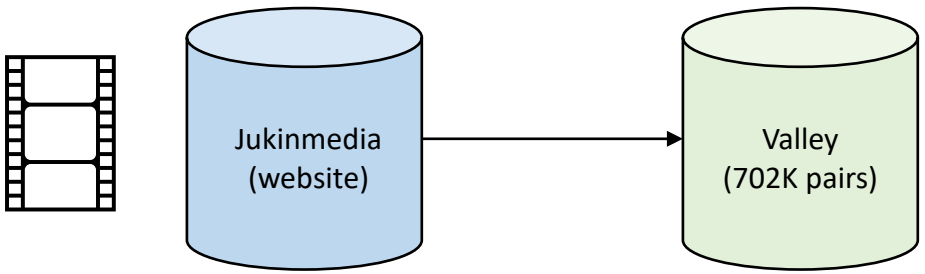


Stage 1: Image/Video Understanding - Dataset

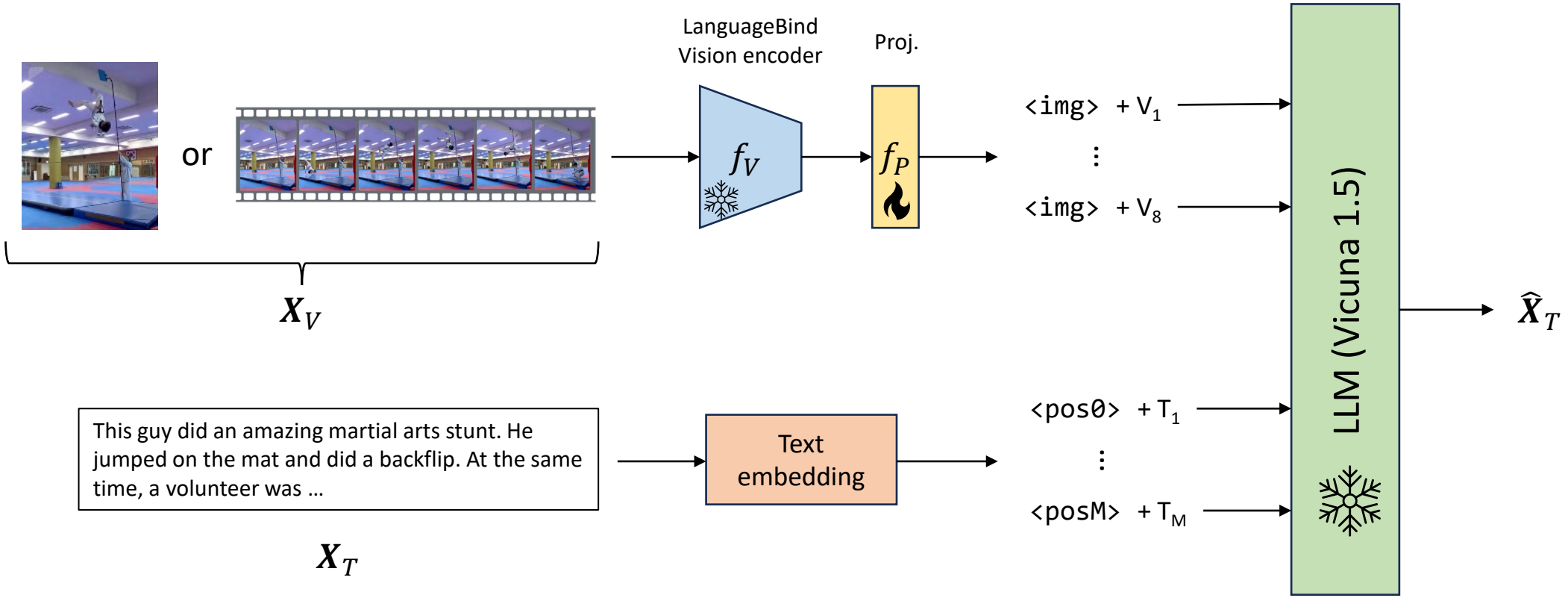
Image dataset: CC-595K (LLaVA)



Video dataset: Valley 702K (Valley)

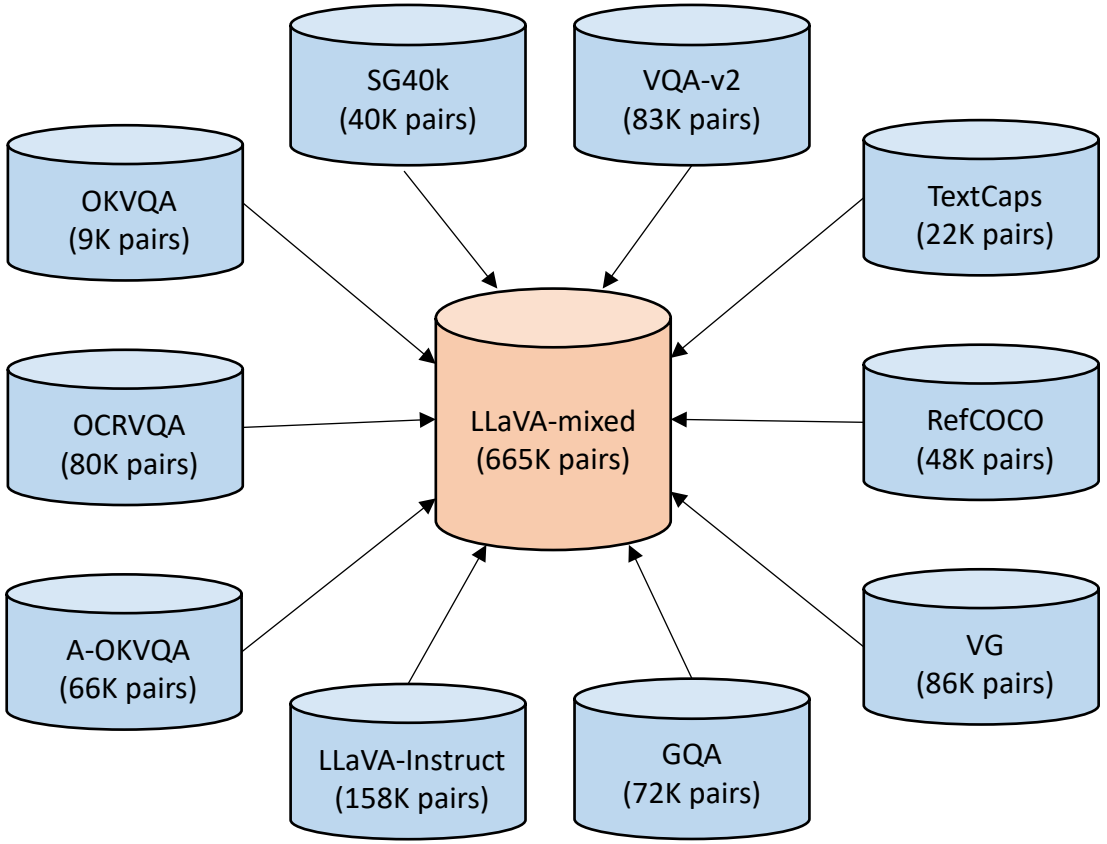


Stage 1: Video/Image Understanding - Training

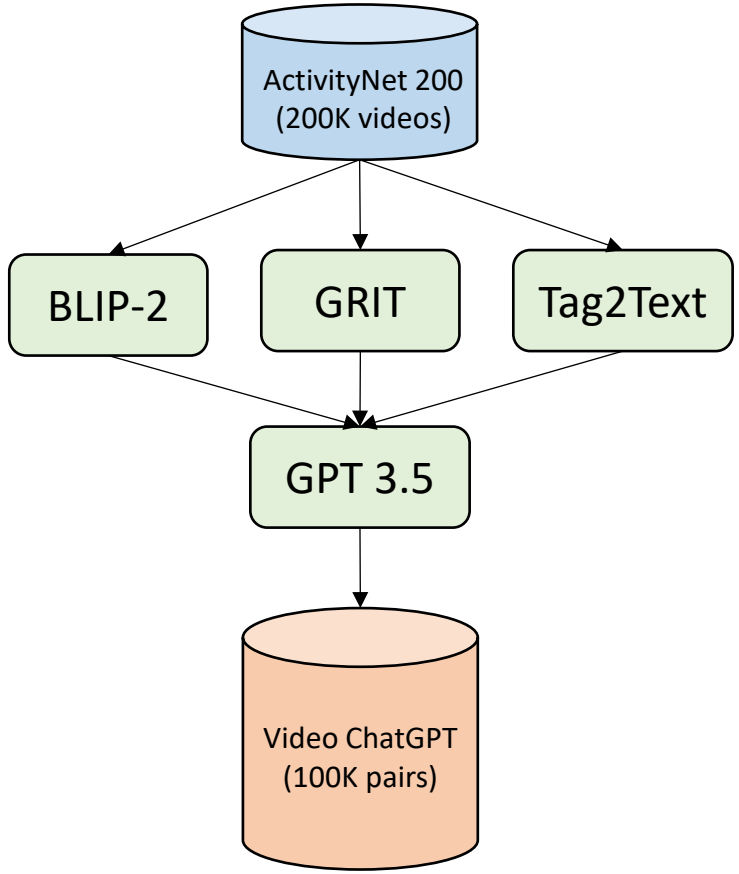


Stage 2: Instruction Tuning - Dataset

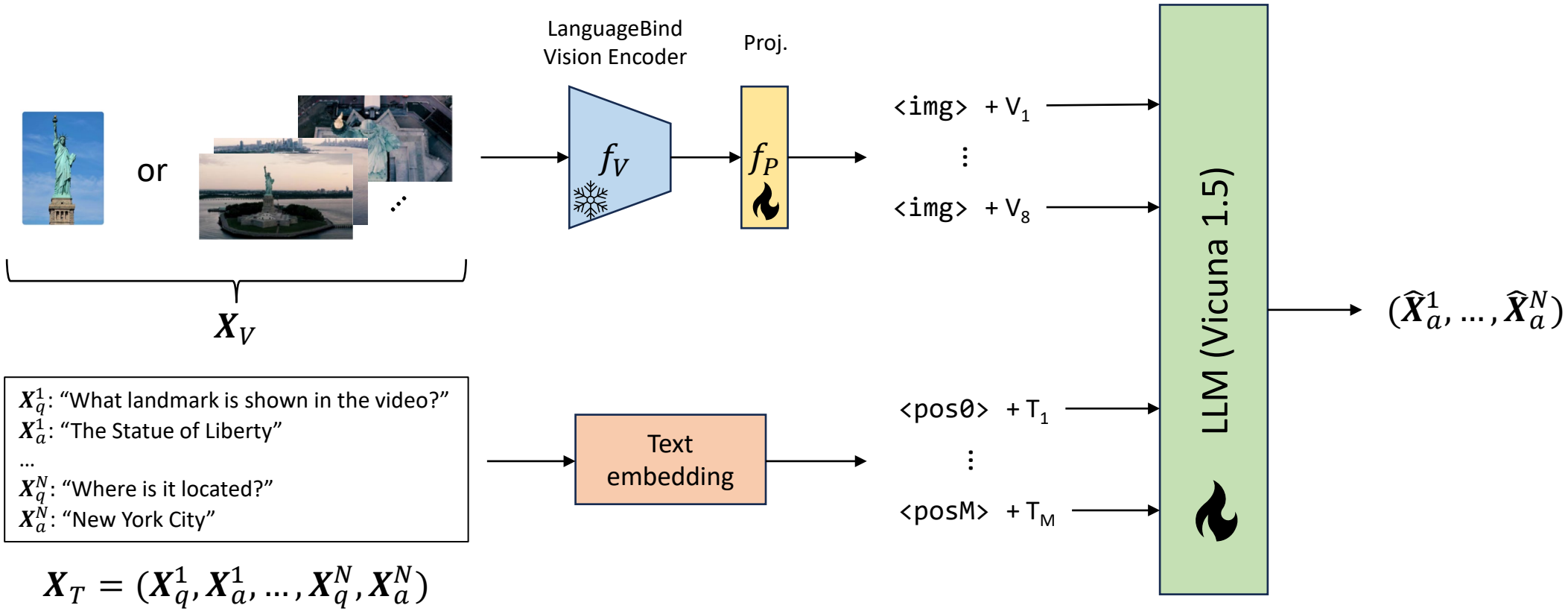
Image dataset: LLaVA-Mixed (LLaVA 1.5)



Video dataset: Video-ChatGPT



Stage 2: Instruction Tuning - Training



Results

Zero-Shot Image Question-Answering

Methods	LLM	Res.	Image Question Answering				
			VQA^{v2}	GQA	VisWiz	SQA^{I}	VQA^{T}
MiniGPT-4	LLaMA-7B	224	-	30.8	47.5	25.4	19.4
IDEFICS-9B	LLaMA-7B	224	50.9	38.4	35.5	-	25.9
mPLUG-Owl	LLaMA-7B	224	-	14.0	39.0	2.8	38.8
Otter	LLaMA-7B	224	-	38.1	**50.0**	27.2	21.2
InstructBLIP	Vicuna-7B	224	-	49.2	34.5	60.5	50.1
Video-LLaVA	Vicuna-7B	224	**74.7***	**60.3***	48.1	**66.4**	**51.8**

∗ denotes that there is some overlap in the training data.

- **Claim**: *Demonstrates strong understanding ability in natural visual environments.*

Image Benchmark Toolkits

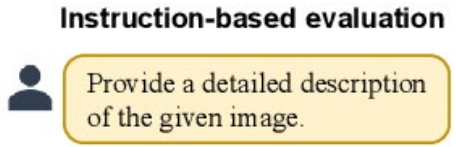
Methods	LLM	Res.	Benchmark Toolkit			
			POPE	MMB	LLaVA^{W}	MM-Vet
MiniGPT-4	LLaMA-7B	224	-	23.0	-	22.1
IDEFICS-9B	LLaMA-7B	224	-	48.2	-	-
mPLUG-Owl	LLaMA-7B	224	-	46.6	-	-
Otter	LLaMA-7B	224	-	32.6	-	24.6
InstructBLIP	Vicuna-7B	224	-	36.0	60.9	26.2
Video-LLaVA	Vicuna-7B	224	**84.4**	**60.9**	**73.1**	**32.0**

- *POPE: polling based, for better evaluation of object hallucination.*
- *MMB: convert free-form predictions into pre-defined choices.*
- *LLaVA^W: challenging tasks and generalizability to novel domains.*
- *MM-Vet: complicated multimodal tasks.*

Object hallucination and POPE method

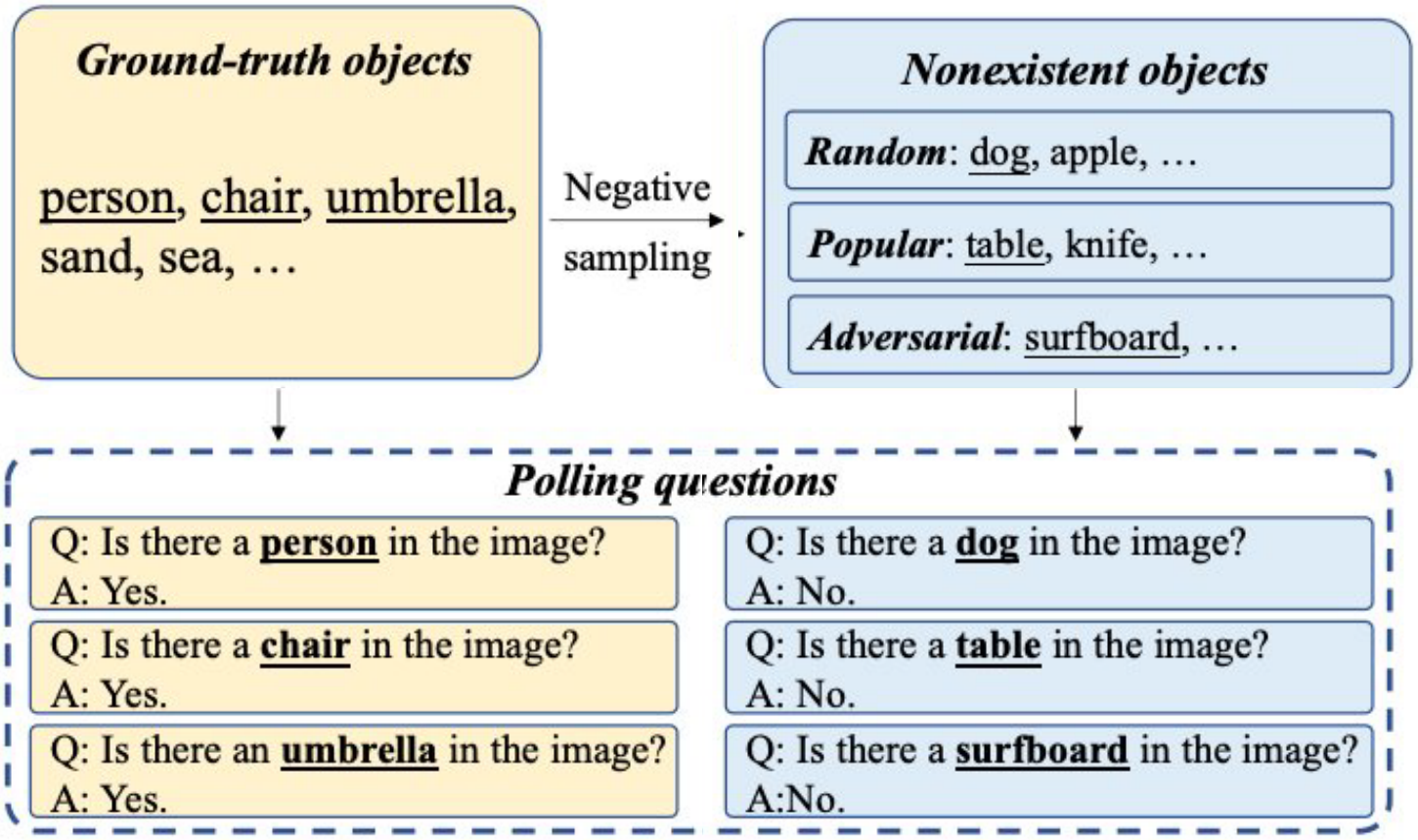
- Random: any absent objects
- Popular: k most popular absent objects
- Adversarial: k absent objects most associated with present





Li et. al., Evaluating Object Hallucination in Large Vision-Language Models, EMNLP 2023

POPE



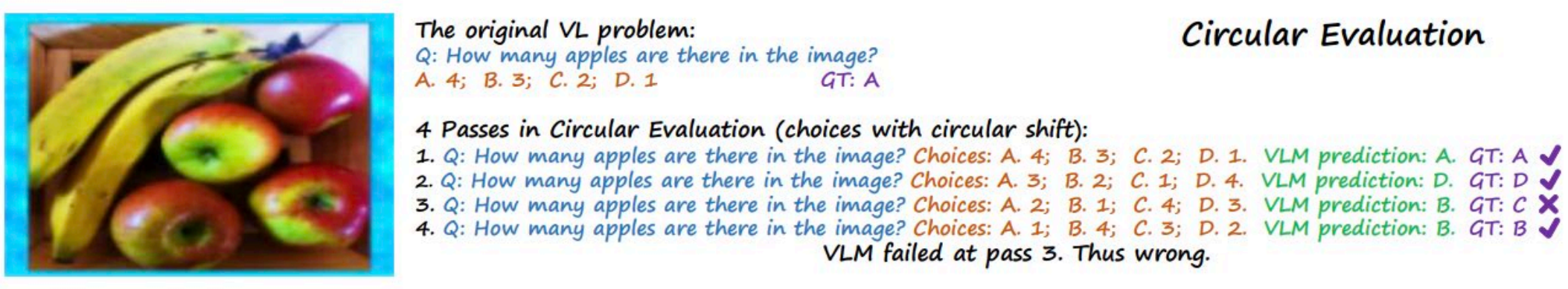


The smaller the value, the better the performance of the model.

MMBench

- *A dataset, more in terms of the number and variety of evaluation questions and abilities.*

- *A novel strategy, that is designed to convert free-form predictions into pre-defined choices.*



The bigger the value, the better the performance of the model.

LLaVA-Bench (In-the-Wild)

- Collect a diverse set of 24 images with 60 questions in total.
- Provide extremely-detailed annotation for each image for an accurate evaluation.



What is the brand of the blueberry-flavored yogurt?

- Require the model to extract details from high resolution image and to have a broad knowledge coverage

The bigger the value, the better the performance of the model.

MM-Vet



Q: What will the girl on the right write on the board?
GT: 14

Required capabilities:
Recognition
Spatial awareness
OCR
Math

VQA v2



Q: Is the boy happy?
GT: Yes

Required capability:
Recognition

The bigger the value, the better the performance of the model.

Image Benchmark Toolkits

Methods	LLM	Res.	Benchmark Toolkit			
			POPE	MMB	LLaVA^{W}	MM-Vet
MiniGPT-4	LLaMA-7B	224	-	23.0	-	22.1
IDEFICS-9B	LLaMA-7B	224	-	48.2	-	-
mPLUG-Owl	LLaMA-7B	224	-	46.6	-	-
Otter	LLaMA-7B	224	-	32.6	-	24.6
InstructBLIP	Vicuna-7B	224	-	36.0	60.9	26.2
Video-LLaVA	Vicuna-7B	224	**84.4**	**60.9**	**73.1**	**32.0**

- **Claim**: *Demonstrates a strong understanding of semantic aspects of scenes.*
- **Claim**: *Is able to answer open-ended and free-form natural language questions about images.*

Image Object Hallucination Evaluation

Evaluation results are reported for the POPE evaluation setting

Methods	LLM	Adersarial			Popular			Random		
		Accuracy	F1-Score	Yes	Accuracy	F1-Score	Yes	Accuracy	F1-Score	Yes
MiniGPT-4	Vicuna-13B	66.6	71.4	66.7	68.3	72.2	64.1	77.8	78.9	54.8
InstructBLIP	Vicuna-13B	74.4	78.5	69.0	81.4	83.5	62.6	**88.7**	**89.3**	55.2
MM-GPT	LLaMA-7B	50.0	66.7	100.0	50.0	66.7	100.0	50.0	66.7	100.0
Video-LLaVA	Vicuna-7B	**81.6**	**80.8**	45.8	**85.3**	**84.0**	42.1	86.2	85.2	42.0

- **Claim**: *Validates the consistency between unified visual representations and the generation of textual descriptions.*

Image Understanding Example



[Input Instruction]
What happens when the glove drops?

Video-LLaVA

When the glove drops, it will fall to the floor, and the ball will roll off the bench. This could potentially cause the ball to roll under the bench or into a hard-to-reach area, making it difficult to retrieve. **The glove, being red**, is hanging from a chain, which suggests that it is being used as a training aid for boxing or other similar sports. The ball, **on the other hand, is blue and is placed on the bench**, possibly for a different purpose or as a prop for the training session. The presence of the glove and the ball on the bench indicates that they are being used for a specific activity or exercise, and the dropping of the glove could disrupt the training or practice session.

GPT-4

It will hit the wood plank and the ball will fly up.

- **Claim**: More comprehensive, intuitive, and logical response.

Zero-shot Video Understanding

Methods	LLM size	MSVD-QA		MSRVTT-QA		TGIF-QA		ActivityNet-QA	
		Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM	1B	32.2	-	16.8	-	41.0	-	24.7	-
VideoChat	7B	56.3	2.8	45.0	2.5	34.4	2.3	-	2.2
LLaMA-Adapter	7B	54.9	3.1	43.8	2.7	-	-	34.2	2.7
Video-LLaMA	7B	51.6	2.5	29.6	1.8	-	-	12.4	1.1
Video-ChatGPT	7B	64.9	3.3	49.3	2.8	51.4	3.0	35.2	2.7
Video-LLaVA	7B	**70.7** +5.8	**3.9** +0.6	**59.2** +9.9	**3.5** +0.7	**70.0** +18.6	**4.0** +1.0	**45.3** +5.1	**3.3** +0.6

- Video-LLaVA outperforms Video-ChatGPT
- **Claim**: *Can comprehend videos and provide contextually appropriate responses based on instructions.*

Video Understanding Example



[Input Instruction] Why is this video funny?
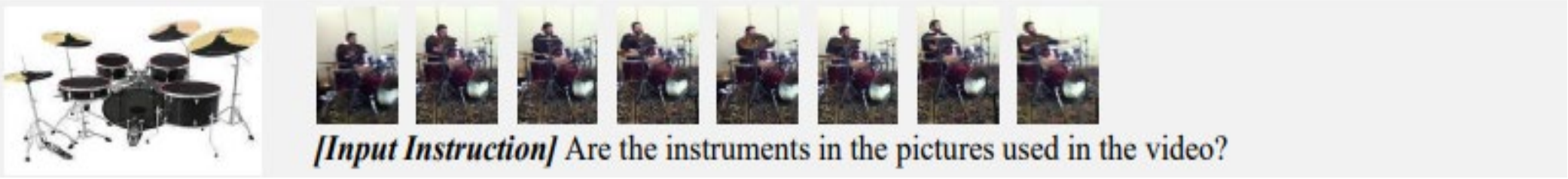
Video-LLaVA

The video is funny because the baby is trying to **read a book while wearing glasses, which is a humorous sight.** The baby's attempts to read the book are also amusing, as they seem to be struggling with the task. Additionally, the baby's facial expressions and movements add to the overall humor of the video.

Video-ChatGPT

The video is funny because it shows a baby girl who is learning to read. She is sitting on a bed and looking at a book, and she is trying to read it. The baby girl is wearing glasses, and she is holding the book with both hands. The baby girl is trying to read the book, and her actions are adorable and endearing. The video captures the innocence and curiosity of a young child as she learns to read, which is a joyful and amusing sight for the viewer.

Image-Video Joint Understanding Example



Video-LLaVA

Yes, the instruments in the images are used in the video. The man is playing a drum set, and the other instruments are also shown in the video.

- **Claim**: Demonstrate the ability to understand inputs composed of both images and videos simultaneously.

Ablations

Ablations: Joint Training

Video Understanding

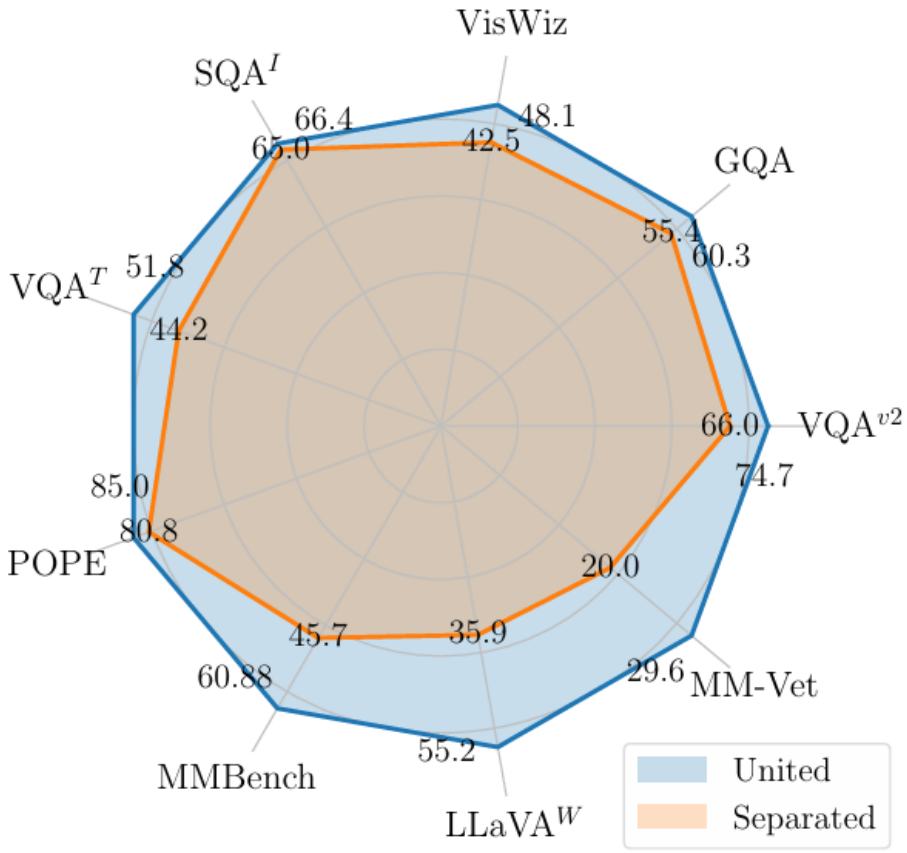
- Joint Training = Images + Videos

Methods	MSVD	MSRVTT	TGIF	ActivityNet
Video-LLaVA*	64.8	58.3	67.8	40.7
Joint with Image	70.7	59.2	70.0	45.3
Δ *Acc.*	***+ 5.9%***	***+ 0.9%***	***+ 2.2%***	***+ 4.6%***

* denotes using only video data in first and second stages

Image-Video Alignment

- Figure compares image understanding tasks
- Unified vs. Separated Visual Representation
  - Unified = LanguageBind
  - Separated = MAE for Images + LanguageBind for Videos



Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic

Keqin Chen[12]*, **Zhao Zhang**[1]†, **Weili Zeng**[3], **Richong Zhang**[2], **Feng Zhu**[1], **Rui Zhao**[14]
[1]SenseTime Research; [2]SKLSDE, Beihang University
{[3]SEIEE, [4]Qing Yuan Research Institute}, Shanghai Jiao Tong University
chenkq@act.buaa.edu.cn; zzhang@mail.nankai.edu.cn

arXiv:2306.15195v2 [cs.CV] 3 Jul 2023

Abstract

In human conversations, individuals can indicate relevant regions within a scene while addressing others. In turn, the other person can then respond by referring to specific regions if necessary. This natural referential ability in dialogue remains absent in current Multimodal Large Language Models (MLLMs). To fill this gap, this paper proposes an MLLM called Shikra, which can handle spatial coordinate inputs and outputs in natural language. Its architecture consists of a vision encoder, an alignment layer, and a LLM. It is designed to be straightforward and simple, ***without*** the need for extra vocabularies, position encoder, pre-/post-detection modules, or external plug-in models. All inputs and outputs are in natural language form. Referential dialogue is a superset of various vision-language (VL) tasks. Shikra can naturally handle location-related tasks like REC and PointQA, as well as conventional VL tasks such as Image Captioning and VQA. Experimental results showcase Shikra's promising performance. Furthermore, it enables numerous exciting applications, like providing mentioned objects' coordinates in chains of thoughts and comparing user-pointed regions similarities. Our code and model are accessed at https://github.com/shikras/shikra.

1 Introduction

In recent months, Multimodal Large Language Models (MLLMs) have witness remarkable progress (Alayrac et al., 2022; Huang et al., 2023; Liu et al., 2023a; Zhu et al., 2023; Li et al., 2023a; Gao et al., 2023; Dai et al., 2023). They brings eyes

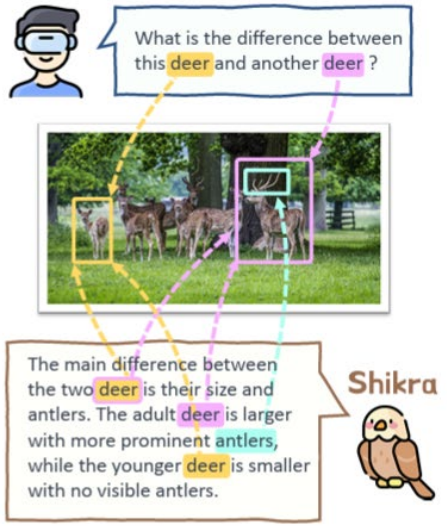


Figure 1: **Demo of Referential Dialogue (RD)**. Users can point to specific areas and ask questions. In turn, Shikra will indicate the specific regions when replying, if necessary. More interesting dialogues can be found in Figure 2 and Appendix C.

precise positions of the content. Users cannot indicate areas of interest in the image, and the models cannot provide the exact locations of the described content. Differently, as shown in Figure 1, in human daily communication, different regions or objects in the scene are often attended to, and people can speak and point to these regions for efficient information exchange. We refer to this interaction mode as Referential Dialogue (RD). If an MLLM

Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, Rui Zhao
SenseTime Research; SKLSDE, Beihang University { SEIEE, Qing Yuan Research Institute}, Shanghai Jiao Tong University

2024 Citations: 119
2025 Citations: 525
https://www.youtube.com/watch?v=5EukpFo7om4

Presented by Group 2: Mahad Ali, Anthony Jackson, Charlee Mione, Rafeeq Shodeinde, Isaac Tuckey

Motivation

- Multimodal LLMs have become popular over the years
- Despite their power, many struggle with grasping positional information
- MLLMs proficient in Referential Dialogue can bridge this gap

Representing position in natural language using numerical values

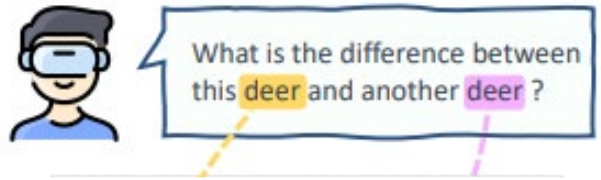
- Bounding box & Center point represented as [x_{min}, y_{min}, x_{max}, y_{max}] & [x_{center}, y_{center}]
- The square brackets ([and]) are treated as tokens just like normal text, and appear normally in text.





What is this man [0.171,0.330,0.650,0.779] scared of?

Referential Dialogue Introduction



Applications

- Mixed Reality headsets
  - Conversing with AI assistants
    - Apple Vision Pro and Siri



Source: https://medium.com/@davidpogue/my-first-immersion-in-apple-vision-pro-heavy-man-d99f0a940c7c

Applications

- Robotics
  - Human Robot interaction
    - Comprehending reference points



Source: https://www.nbcnews.com/mach/science/robots-show-they-can-foresee-their-own-future-ncna833976

Applications

- Consumers
  - Online shopping
    - Specify a product in an image



Source: https://www.etsy.com/listing/1105408268/luxury-window-half-body-female-dress

Referential Dialogue Tasks

Referring Expression Comprehension (REC)

- Input: image + prompt containing object description
- Output: Bounding box for object



Referring Expression Compression

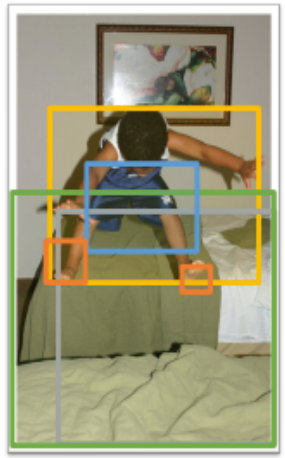
Detect the location of the vanilla dessert with a cashew nose in image and share the coordinates with me, please.

Spotting Captioning

- Spotting Captioning
  - Input: Image + prompt
  - Output: Description of scene + bboxes

Spotting Captioning

Provide a description of the image and include the boxes of the items you reference.



Referring Expression Generation (REG)

- Input: Image + bounding box
- Output: Discriminatory description of region



Referring Expression Generation

Please generate a distinguishing description for the region [0.336, 0.371, 0.515, 0.637] in the image.

PointQA

- PointQA-Point:
  - Input: Image + point + question about point
  - Output: answer

PointQA - Point



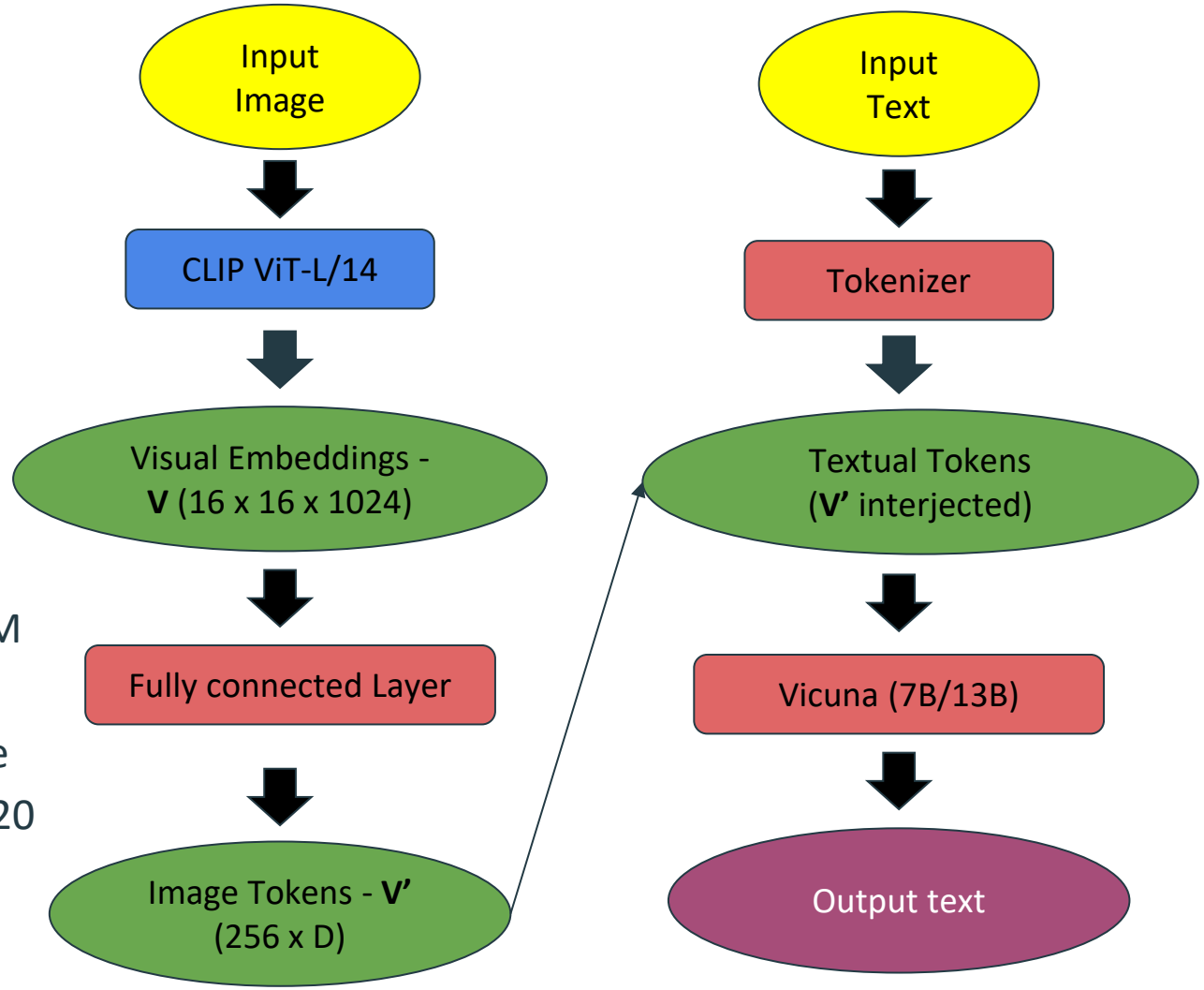
- PointQA-Box:
  - Input: Image + bbox + question about region
  - Output: answer

PointQA-Box



Shikra - Architecture

- Two main components:
  ViT-L/14 from CLIP & Vicuna LLM
- ViT-L/14 is frozen
- Use either Vicuna7B or Vicuna 13B
- Image embeddings used as input tokens to LLM
- FC layer maps image embeddings to match the input dimension of the LLM (4096 for 7B & 5120 for 13B)



Datasets

- Two types of data were used to train Shikra:
  - Reorganized public datasets
  - Shikra-RD: A Referential Dialogue (RD) dataset built using Flickr30K entities
- Public datasets: VQAv2, Captioning (LLaVA-Pretraining), REG/REC (RefCOCO), PointQA (Visual-7W)
- Shikra-RD generated using Flickr30K entities dataset by having GPT4 generate Q&A pairs.

Model Training

- The model is trained in two stages.
- First stage: train for 100,000 steps over public dataset
- Second stage: increase sampling ratio of Shikra-RD & LLaVA-Instruct-150K to 50%
- In both stages the visual encoder is frozen and the LLM is completely unfrozen.
- Stage 1 learns conventional VL tasks, stage 2 fine tunes for RD on RD-specific data

Tasks & Datasets

Task	Dataset
Captioning	LLaVA-Pretraining
Soptting Cap.	Flickr30K Entities
Grounding Cap.	Visual Genome
REG	RefCOCO, RefCOCO+, RefCOCOg
REC	RefCOCO, RefCOCO+, RefCOCOg, Visual Genome
VQA	VQAv2
PointQA	PointQA-Local/Twice, Visual-7W ('which box' subset)
Dialogue	LLaVA-Instruct-150K*
RD	VCR, Shikra-RD (Generated data from Flickr30K Entities)*

Text prompt templates (overview)

- Training was performed on 9 different tasks, with multiple templates used for each task.
- Templates only used for training, natural language can be used for inference.
- GPT4 used to generate 100s of versions of a template for each task to ensure generalizability

Text prompt templates

Task	Prompt
Captioning	Describe this <image> as simply as possible.

Text prompt templates

Task	Prompt
Captioning	Describe this <image> as simply as possible.
Spotting Cap.	Can you provide a description of the <image> and include the coordinates [x0,y0,x1,y1] for each mentioned object?

Text prompt templates

Task	Prompt
Captioning	Describe this <image> as simply as possible.
Spotting Cap.	Can you provide a description of the <image> and include the coordinates [x0,y0,x1,y1] for each mentioned object?
Grounding Cap.	Can you give me a description of the region <objs> in <image>?

Text prompt templates

Task	Prompt
Captioning	Describe this <image> as simply as possible.
Spotting Cap.	Can you provide a description of the <image> and include the coordinates [x0,y0,x1,y1] for each mentioned object?
Grounding Cap.	Can you give me a description of the region <objs> in <image>?
REG	For the given <image>, can you provide a unique description of the area <objs>?

Text prompt templates

Task	Prompt
Captioning	Describe this <image> as simply as possible.
Spotting Cap.	Can you provide a description of the <image> and include the coordinates [x0,y0,x1,y1] for each mentioned object?
Grounding Cap.	Can you give me a description of the region <objs> in <image>?
REG	For the given <image>, can you provide a unique description of the area <objs>?
Q → A	I want to know the answer to '<question>' Refer to the <image> and give a clear response

Text prompt templates

Task	Prompt
Captioning	Describe this <image> as simply as possible.
Spotting Cap.	Can you provide a description of the <image> and include the coordinates [x0,y0,x1,y1] for each mentioned object?
Grounding Cap.	Can you give me a description of the region <objs> in <image>?
REG	For the given <image>, can you provide a unique description of the area <objs>?
Q → A	I want to know the answer to '<question>' Refer to the <image> and give a clear response
REC	In the given <image>, could you find and tell me the coordinates of <expr>?

Text prompt templates

Task	Prompt
Q→CA	Having a look at <image>, can you tell me the answer to my '<question>' and the logic leading to it?
$Q \rightarrow C^{Point}A$	Analyze the <image> and answer '<question>' Include your reasoning process and mark center points of related objects as [cx, cy].
Q→C BoxA	Please explain your reasoning and provide bounding boxes, denoted by square brackets, for the objects mentioned in the picture <image>. <question>

Evaluation

Referring Expression Comprehension Task

- Measures prediction accuracy of bounding boxes
  - IoU > .5 = true positive

Model type	Model	RefCOCO			RefCOCO+			RefCOCOg		GRIT
		val	test-A	test-B	val	test-A	test-B	val-u	test-u	refexp
Generalist VL SOTAs (w/o finetuning)	GPV-2	-	-	-	-	-	-	-	-	51.50
	OFA-L*	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58	61.70
	Unified-IO	-	-	-	-	-	-	-	-	78.60
	OFASys	-	80.10	-	-	-	-	-	-	-
	VisionLLM-H	-	86.70	-	-	-	-	-	-	-
	Shikra-7B	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	69.34
	Shikra-13B	87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16	69.03

Image Captioning and VQA

- Shikra-13B performance on traditional VL tasks

Datasets		Shikra	FM-80B	FM-9B	Kosmos-1	BLIP-2	Unified-IO	VPGTrans	VisionLLM
VQA	VQAv2^{val}								
	VQAv2^{dev}								
	VQAv2^{std}								
	OK-VQA								
Caption	Flickr30k								
	COCO								

Object Hallucination using POPE pipeline

Datasets	Metrics	Shikra	InstructBLIP	MiniGPT-4	LLaVA	MM-GPT	mPLUG-Owl
Random	Accuracy (↑)	86.90	88.57	79.67	50.37	50.10	53.97
	Precision (↑)	94.40	84.09	78.24	50.19	50.05	52.07
	Recall (↑)	79.27	95.13	82.20	99.13	100.00	99.60
	F1-Score (↑)	86.19	89.27	80.17	66.64	66.71	68.39
	Yes	43.26	56.57	52.53	98.77	99.90	95.63
Popular	Accuracy (↑)	83.97	82.77	69.73	49.87	50.00	50.90
	Precision (↑)	87.55	76.27	65.86	49.93	50.00	50.46
	Recall (↑)	79.20	95.13	81.93	99.27	100.00	99.40
	F1-Score (↑)	83.16	84.66	73.02	66.44	66.67	66.94
	Yes	45.23	62.37	62.20	99.40	100.00	98.57
Adversarial	Accuracy (↑)	83.10	72.10	65.17	49.70	50.00	50.67
	Precision (↑)	85.60	65.13	61.19	49.85	50.00	50.34
	Recall (↑)	79.60	95.13	82.93	99.07	100.00	99.33
	F1-Score (↑)	82.49	77.32	70.42	66.32	66.67	66.82
	Yes	46.50	73.03	67.77	99.37	100.00	98.67

Conclusion

- Shikra excels at interpreting positional information
- Obtains competitive performance with respect to specialized models
- Visual information valuable for reducing hallucinations