

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

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

Arxiv Preprint 2023

12 citations

Presented By:

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

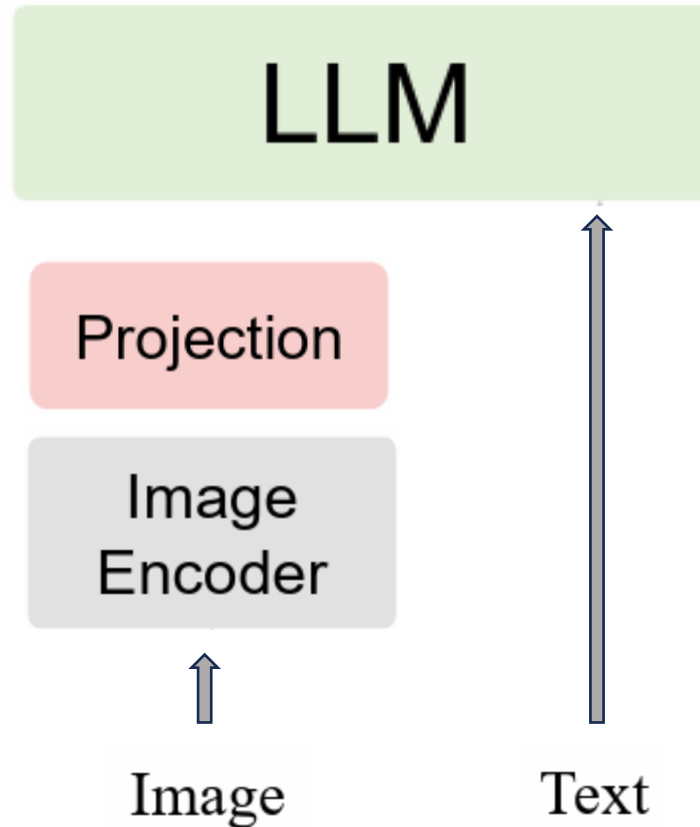
# Outline

1. Background/Motivation
2. Method
3. Results
4. Ablations
5. Conclusion
6. Limitations/Future Considerations

# Background/Motivation

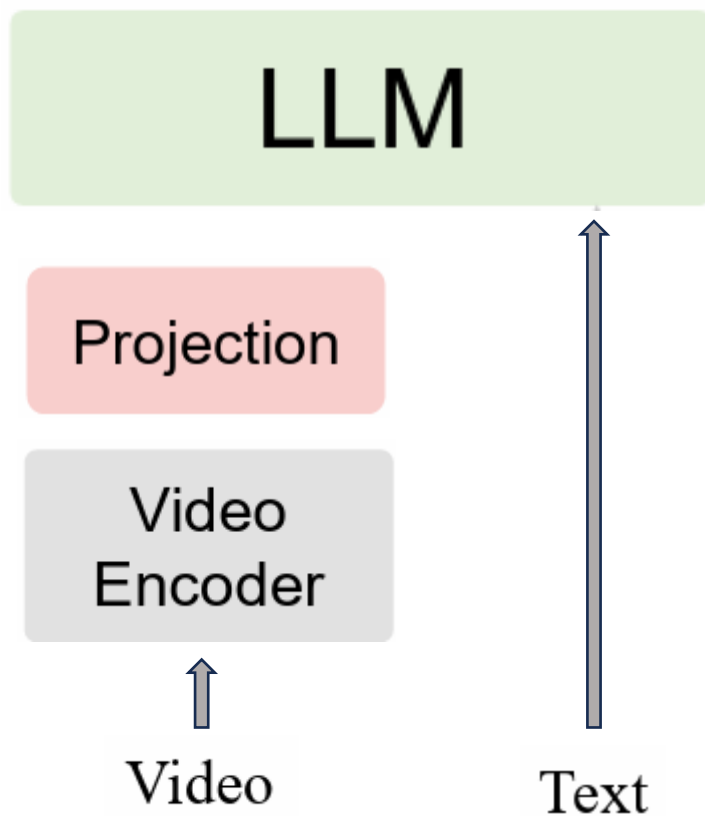
# LVLM Paradigms

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



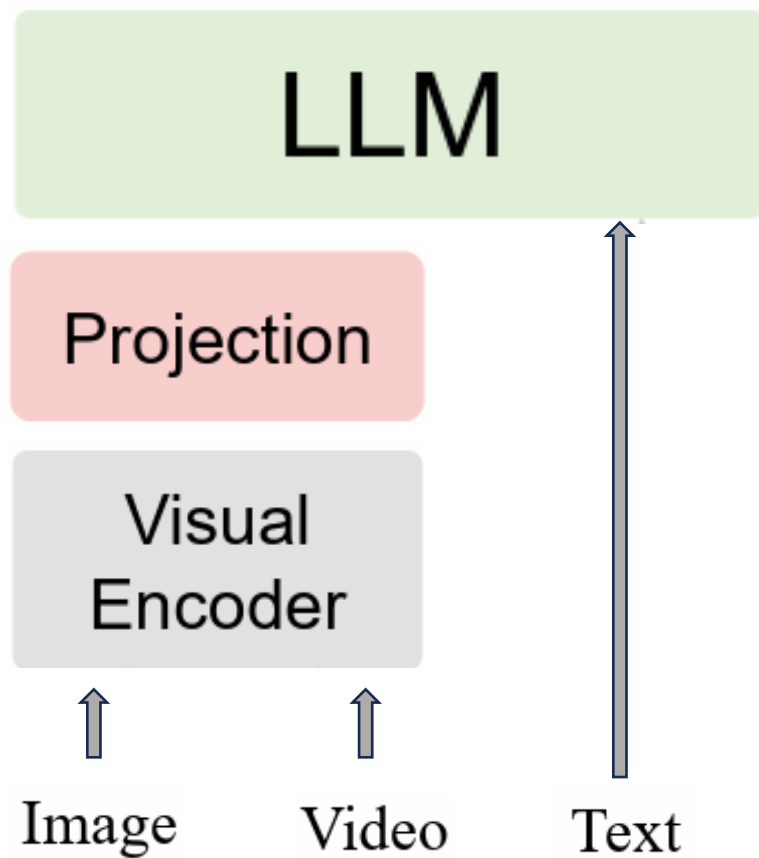
# LVLN Paradigms

- Video ChatGPT (Video Only)



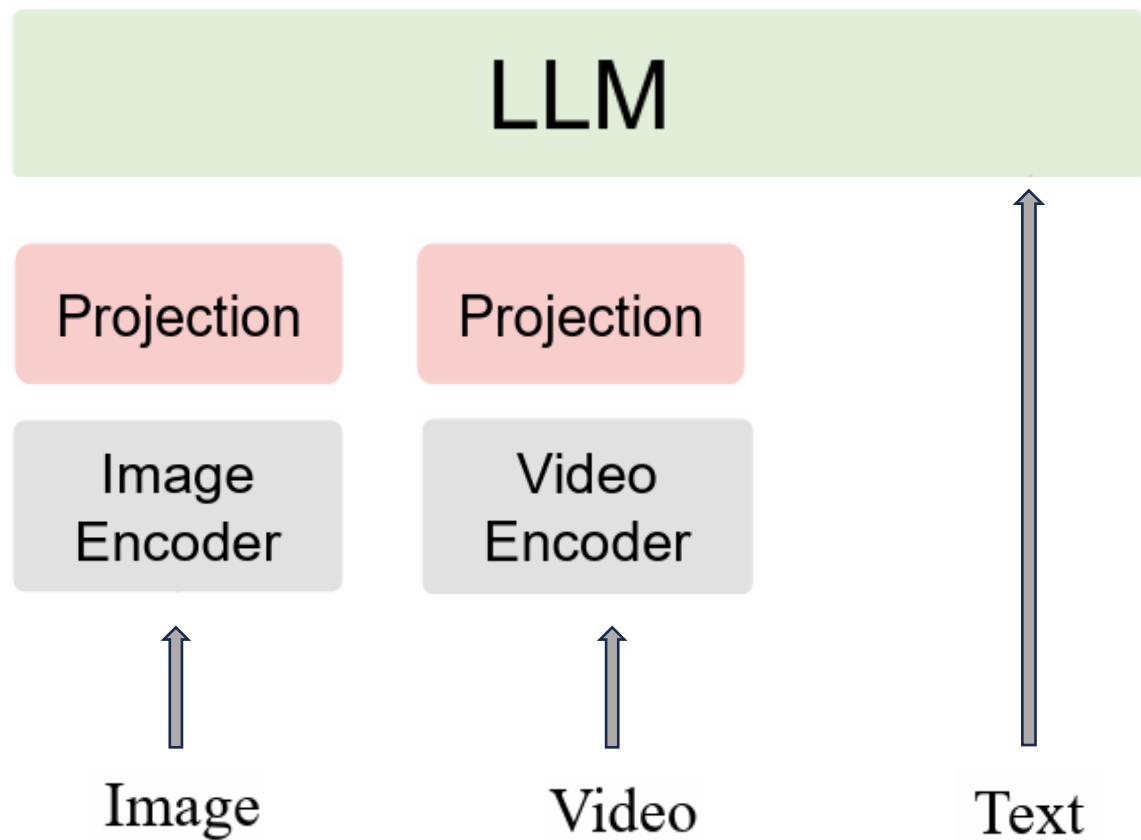
# LVLMM Paradigms

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



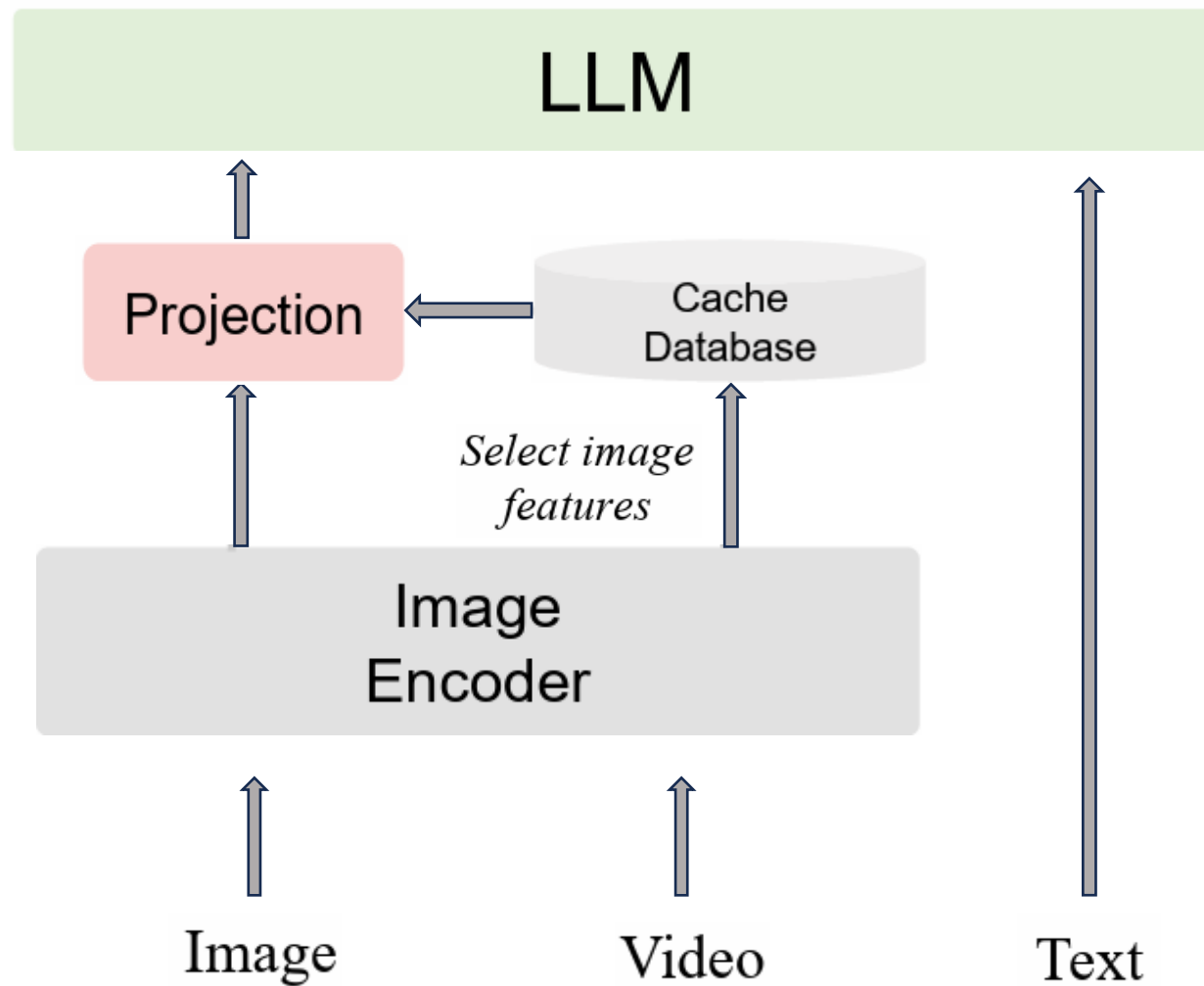
# LVLM Paradigms

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



# LVLM Paradigms

- ImageBind-LLM, LLaMA-Adapter (Alignment Before Projection)



# Method

# 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  $f_P$



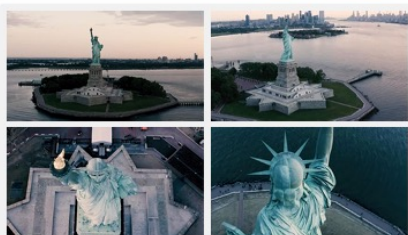
Word Embedding Layer  $f_W$



Encoder Zoo  
LanguageBind  $f_V$

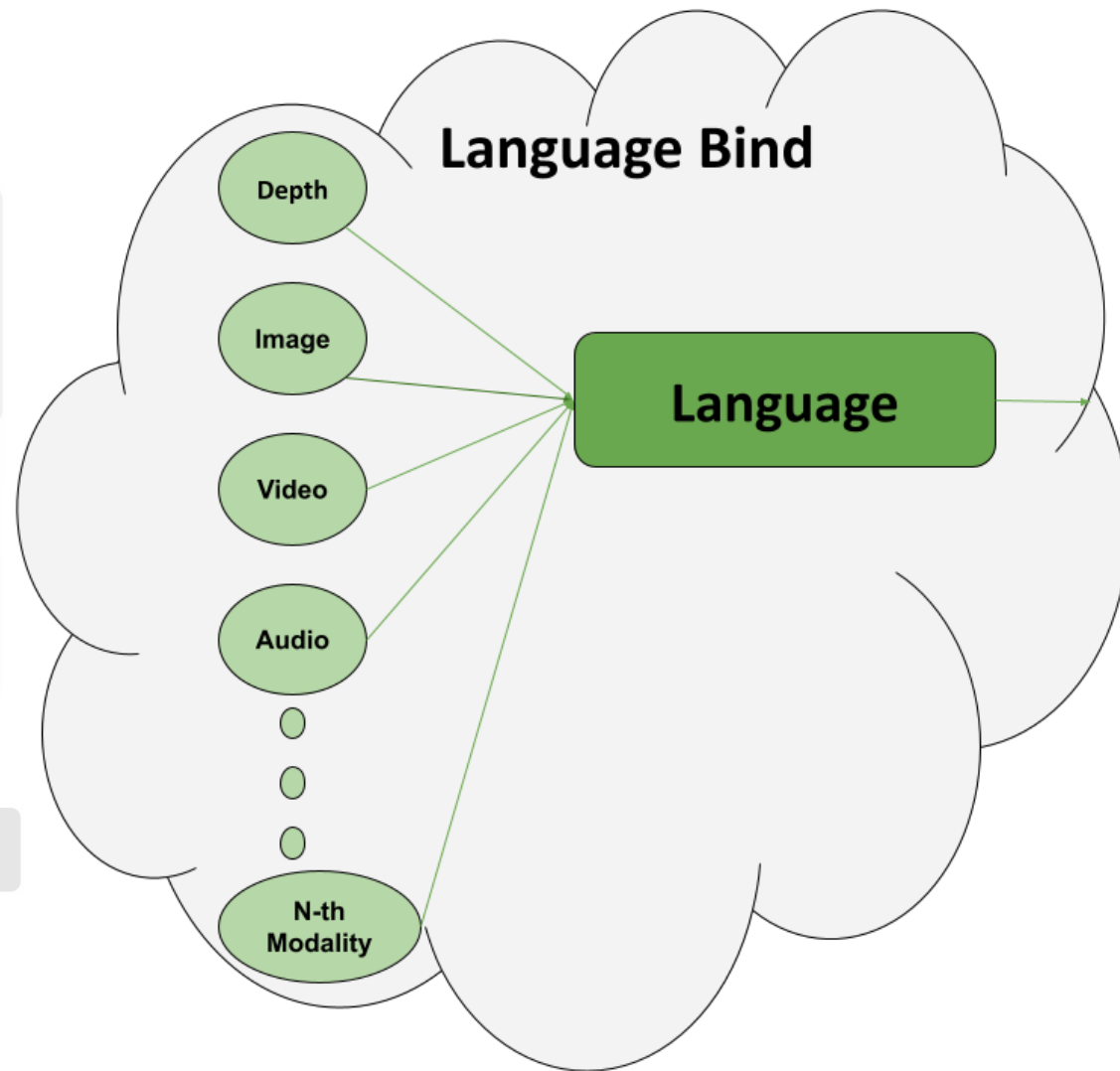


Image

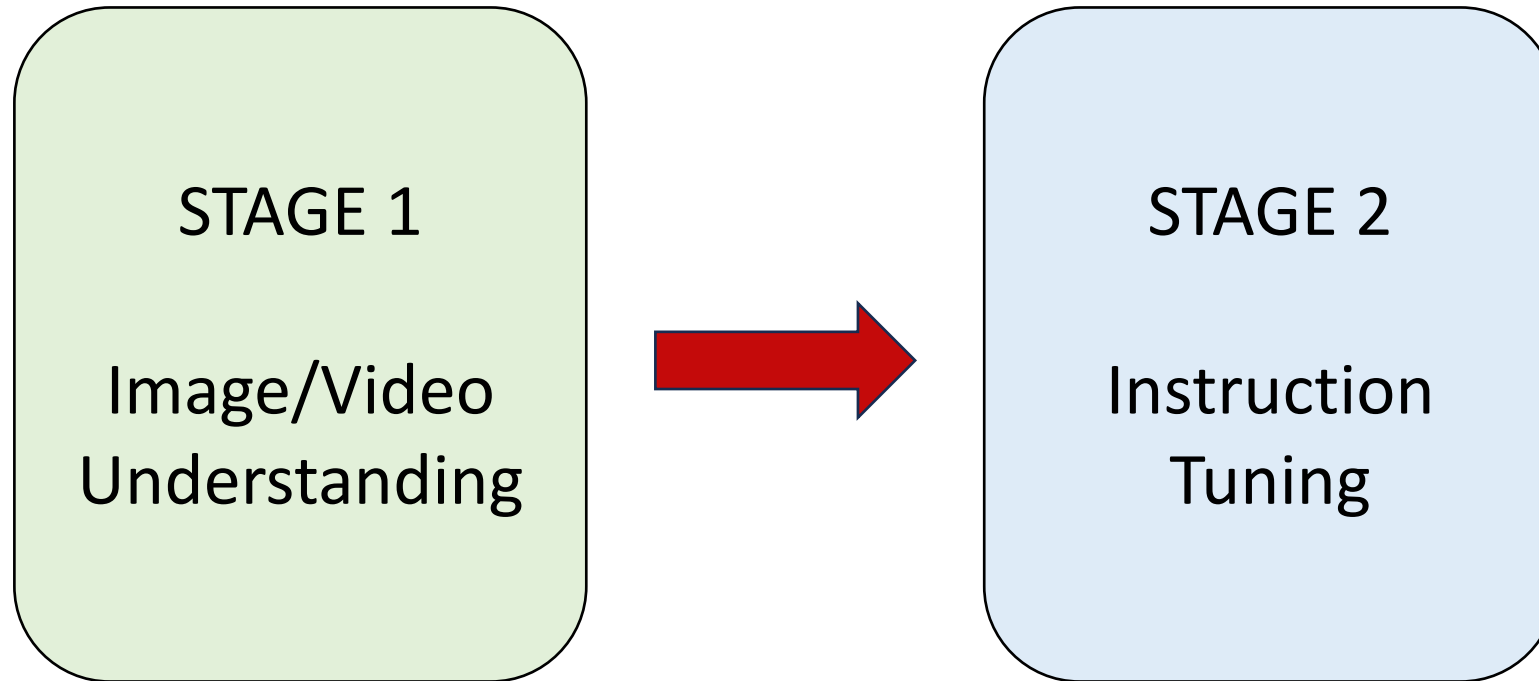


Video

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

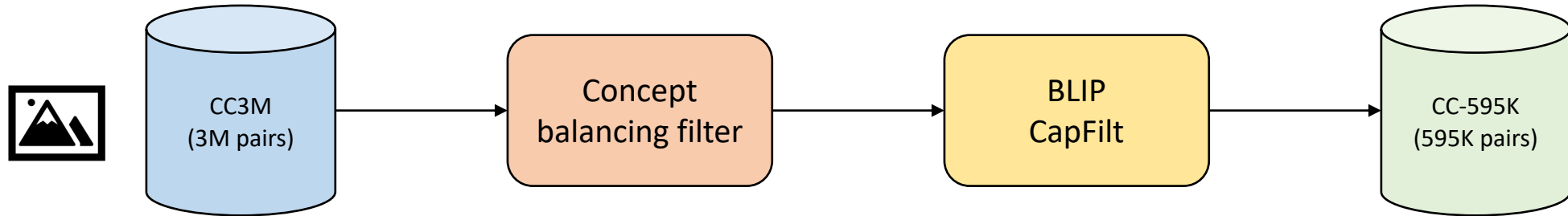


# Training Pipeline: Overview

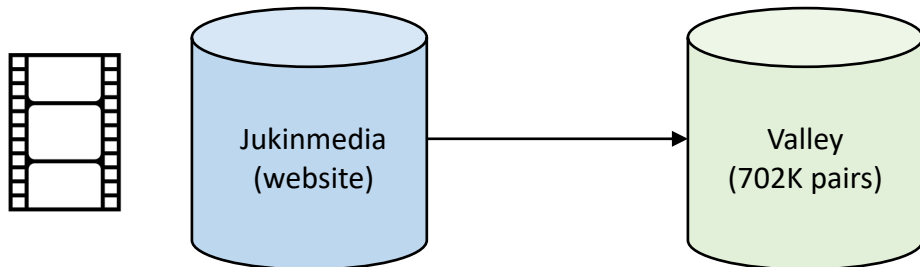


# Stage 1: Image/Video Understanding - Dataset

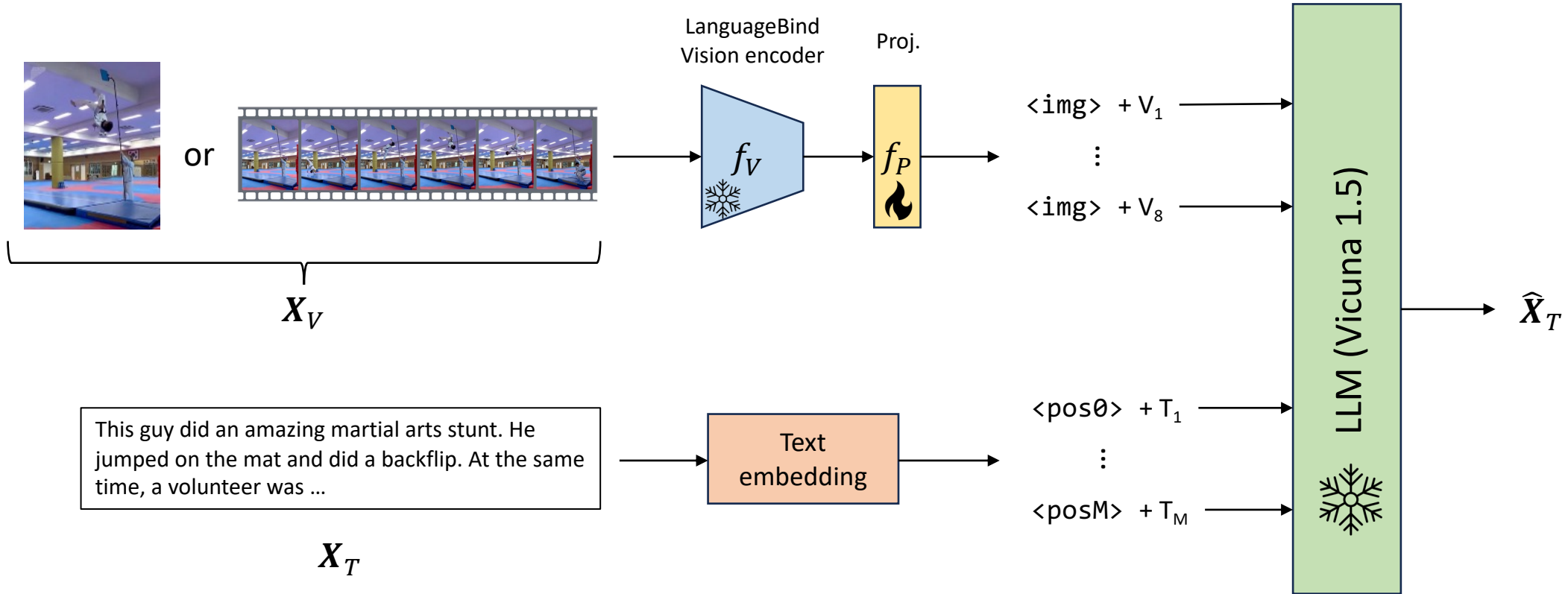
Image dataset: CC-595K (LLaVA)



Video dataset: Valley 702K (Valley)

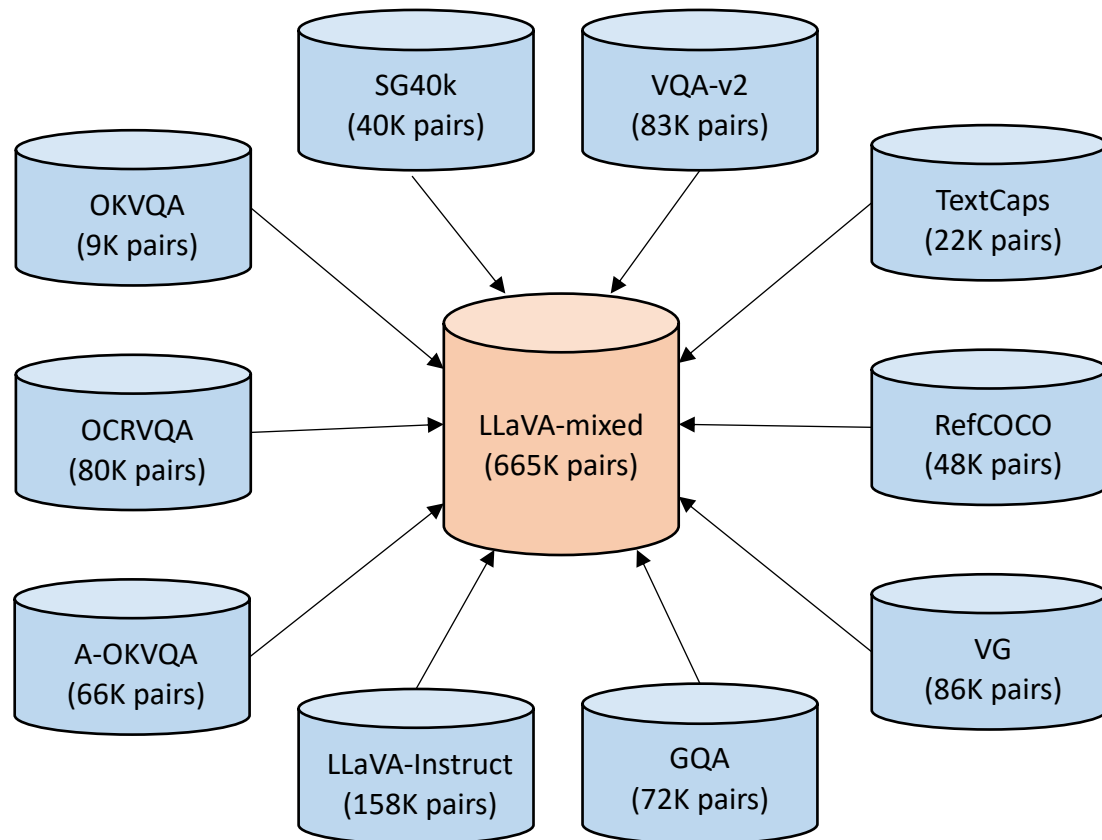


# Stage 1: Video/Image Understanding - Training

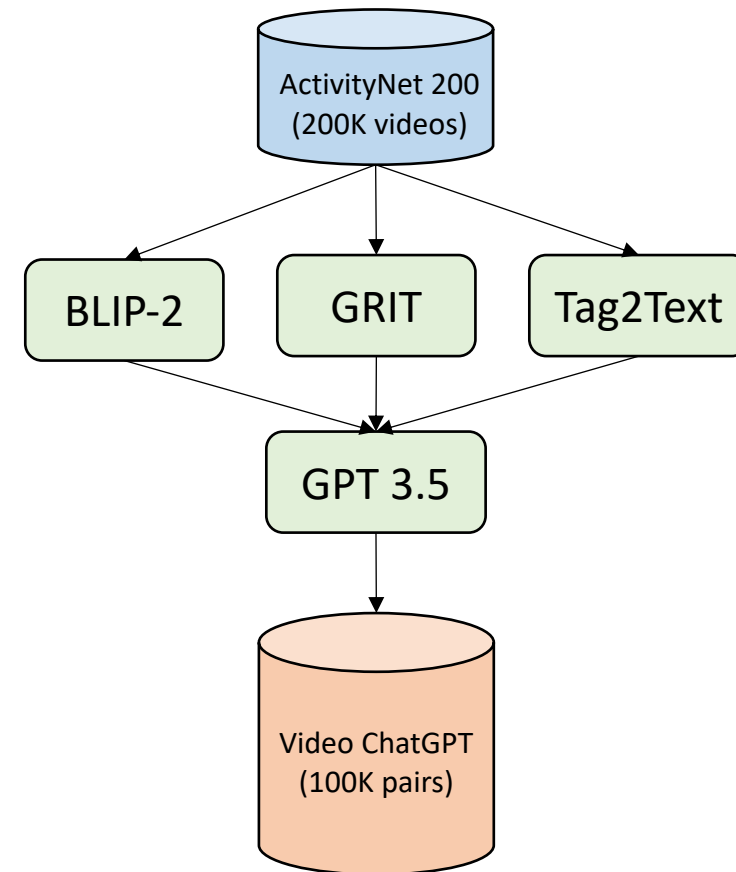


# Stage 2: Instruction Tuning - Dataset

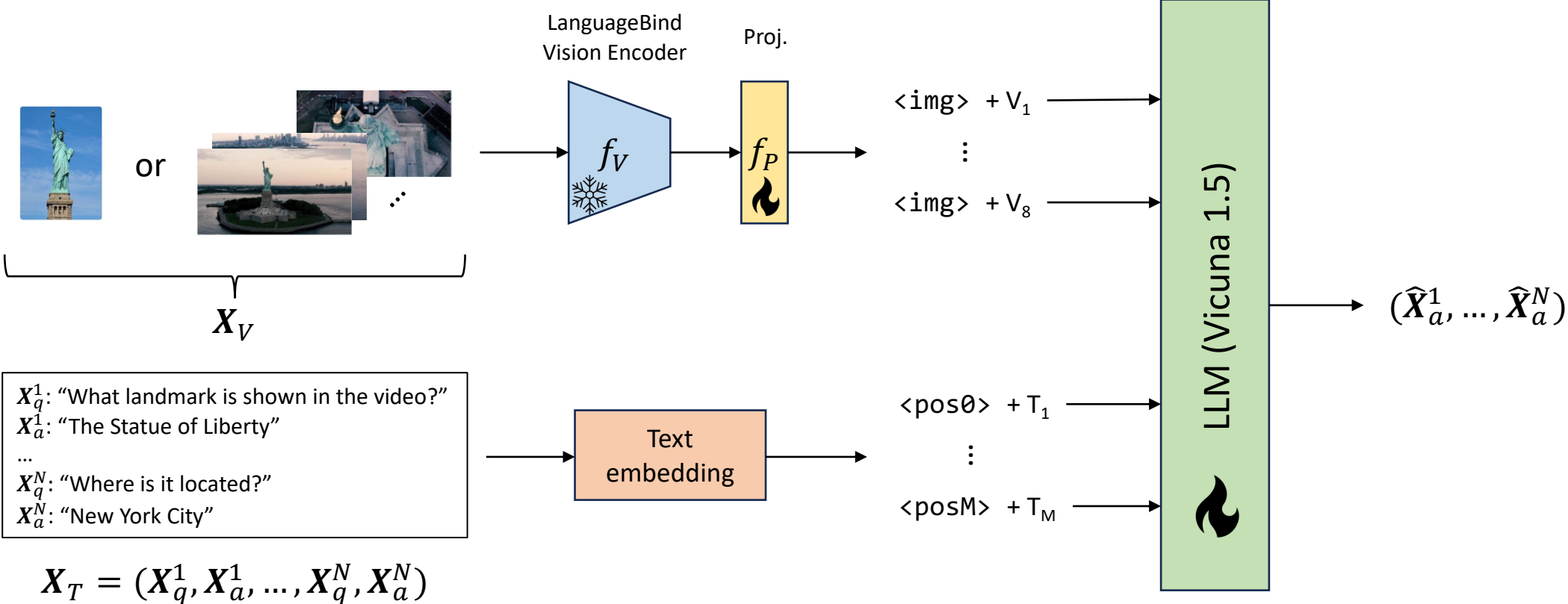
Image dataset: LLaVA-Mixed (LLaVA 1.5)



Video dataset: Video-ChatGPT



# Stage 2: Instruction Tuning - Training



# Results



# Zero-Shot Image Question-Answering

Methods	LLM	Res.	Image Question Answering				
			VQA <sup>v2</sup>	GQA	VisWiz	SQA <sup>1</sup>	VQA <sup>T</sup>
MiniGPT-4	LLaMA-7B	224	-	30.8	47.5	25.4	19.4
IDEFICS-9B	LLaMA-7B	224	<u>50.9</u>	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	<b>50.0</b>	27.2	21.2
InstructBLIP	Vicuna-7B	224	-	<u>49.2</u>	34.5	<u>60.5</u>	<u>50.1</u>
Video-LLaVA	Vicuna-7B	224	<b>74.7*</b>	<b>60.3*</b>	<u>48.1</u>	<b>66.4</b>	<b>51.8</b>

\* denotes that there is some overlap in the training data.

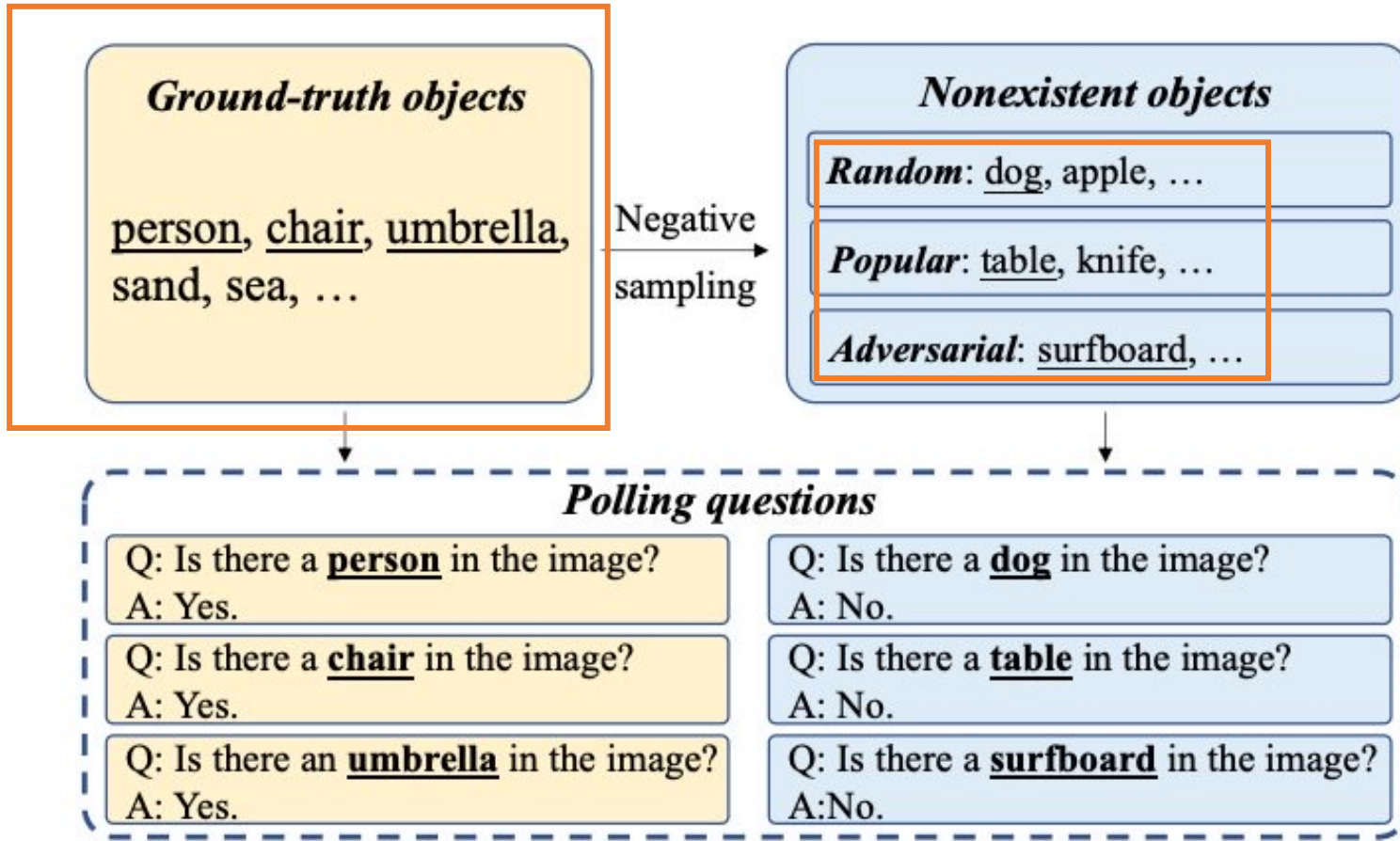
- **Claim:** *Demonstrates strong understanding ability in natural visual environments.*

# Image Benchmark Toolkits

Methods	LLM	Res.	Benchmark Toolkit			
			POPE	MMB	LLaVA <sup>W</sup>	MM-Vet
MiniGPT-4	LLaMA-7B	224	-	23.0	-	22.1
IDEFICS-9B	LLaMA-7B	224	-	<u>48.2</u>	-	-
mPLUG-Owl	LLaMA-7B	224	-	46.6	-	-
Otter	LLaMA-7B	224	-	32.6	-	24.6
InstructBLIP	Vicuna-7B	224	-	36.0	<u>60.9</u>	<u>26.2</u>
Video-LLaVA	Vicuna-7B	224	<b>84.4</b>	<b>60.9</b>	<b>73.1</b>	<b>32.0</b>

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

# POPE



- **Sampling Schemes**

- Adversarial : Similarly.
- Popular : Most frequently.
- Random: Randomly.

**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 original VL problem:

Q: How many apples are there in the image?

A. 4; B. 3; C. 2; D. 1

GT: A

Circular Evaluation

4 Passes in Circular Evaluation (choices with circular shift):

1. Q: How many apples are there in the image? Choices: A. 4; B. 3; C. 2; D. 1. VLM prediction: A. GT: A ✓

2. Q: How many apples are there in the image? Choices: A. 3; B. 2; C. 1; D. 4. VLM prediction: D. GT: D ✓

3. Q: How many apples are there in the image? Choices: A. 2; B. 1; C. 4; D. 3. VLM prediction: B. GT: C ✗

4. Q: How many apples are there in the image? Choices: A. 1; B. 4; C. 3; D. 2. VLM prediction: B. GT: B ✓

VLM failed at pass 3. Thus wrong.

**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 <sup>W</sup>	MM-Vet
MiniGPT-4	LLaMA-7B	224	-	23.0	-	22.1
IDEFICS-9B	LLaMA-7B	224	-	<u>48.2</u>	-	-
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	<b>84.4</b>	<b>60.9</b>	<b>73.1</b>	<b>32.0</b>

- **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	<u>74.4</u>	<u>78.5</u>	69.0	<u>81.4</u>	<u>83.5</u>	62.6	<b>88.7</b>	<b>89.3</b>	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	<b>81.6</b>	<b>80.8</b>	45.8	<b>85.3</b>	<b>84.0</b>	42.1	<u>86.2</u>	<u>85.2</u>	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	<b>70.7 +5.8</b>	<b>3.9 +0.6</b>	<b>59.2 +9.9</b>	<b>3.5 +0.7</b>	<b>70.0 +18.6</b>	<b>4.0 +1.0</b>	<b>45.3 +5.1</b>	<b>3.3 +0.6</b>

- 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**



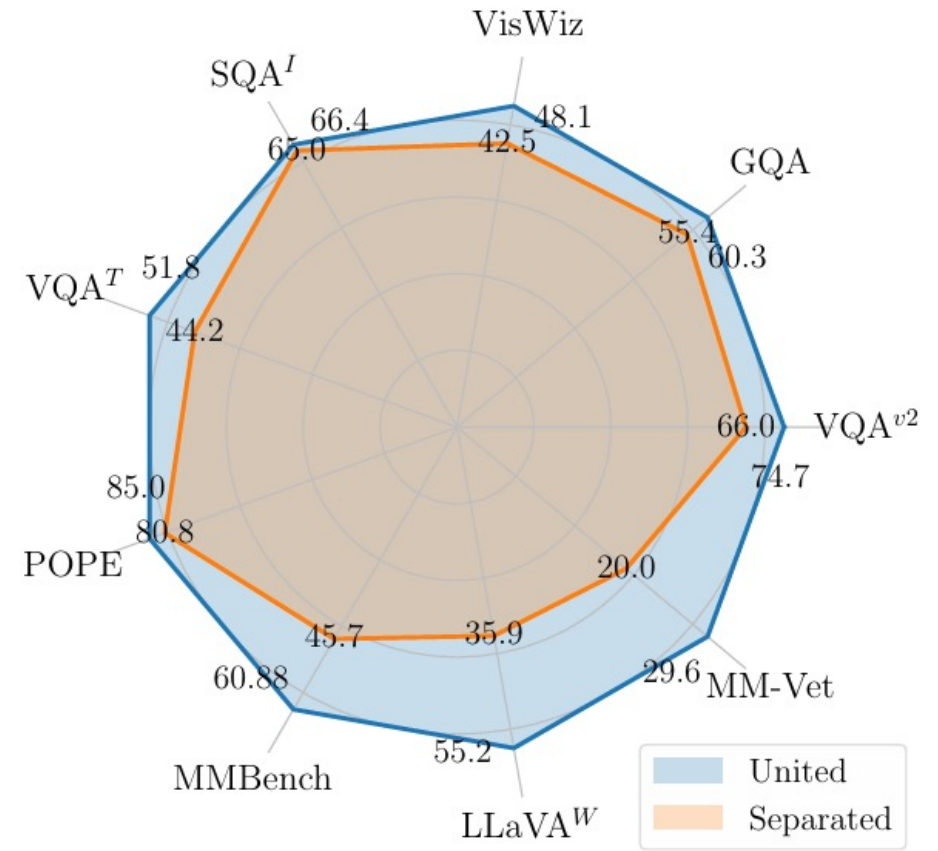
Yes, the instruments in the images are used in the video. The man is playing a drum set, and the other instruments are also shown in the video.

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

# Ablations

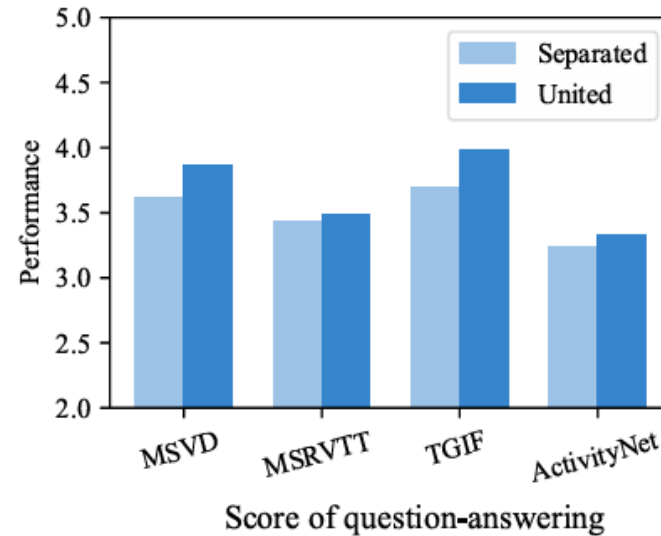
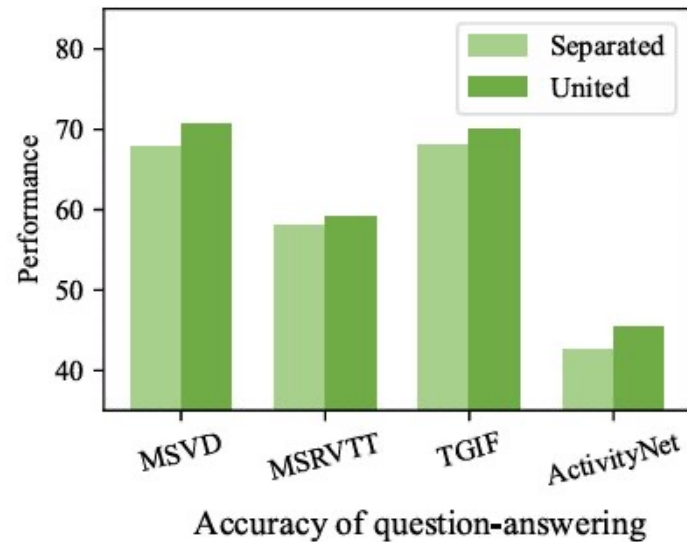
# Image-Video Alignment

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



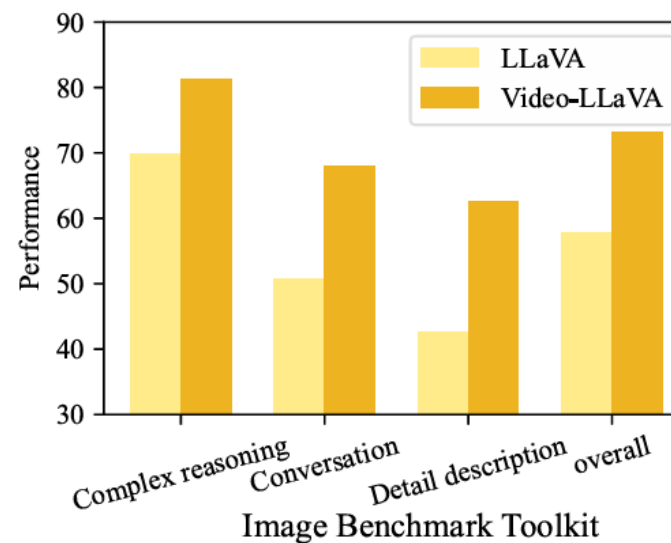
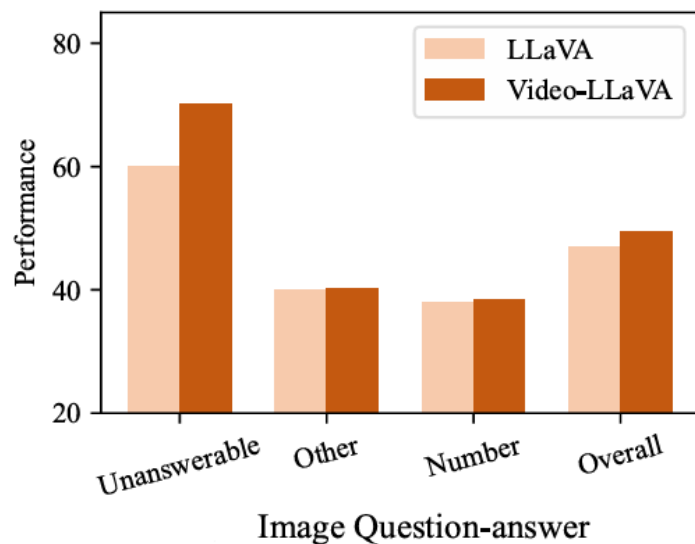
# Image-Video Alignment

- Figure compares image understanding tasks
- Unified vs. Separated Visual Representation



# Joint Image-Video Training

- Figure compares image understanding tasks on the VisWiz dataset
- Unanswerable question performance is notably increased



# Joint Image-Video Training

- Table compares video understanding tasks
- **Claim:** *Enhances the LLM's comprehension of visual representations.*

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
$\Delta$ Acc.	+ 5.9%	+ 0.9%	+ 2.2%	+ 4.6%

Conclusion

# Summary

- Video-LLaVA: An Extension of LLaVA but with Videos
- Alignment Before Projection => Language Bind
- Joint Training => Images + Videos

# Limitations/Future Considerations

# Potential Improvements

- Struggles with Spatio-Temporal Localization
- Struggles with Long-Range Video Understanding
- Improve with Timestamp Embeddings
- Extend to More Visual-Related Modalities (Depth, Infrared, etc.)