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

Presented By:
David Shatwell Pittaluga, Anthony Bilic, Kevin Zhai, Zain Ulabeledeen 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)
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

- ImageBind-LLM, LLaMA-Adapter (Alignment Before Projection)
Method
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.
Training Pipeline: Overview

STAGE 1
Image/Video Understanding

STAGE 2
Instruction Tuning
Stage 1: Image/Video Understanding - Dataset

Image dataset: CC-595K (LLaVA)

- CC3M (3M pairs) → Concept balancing filter → BLIP CapFilt → CC-595K (595K pairs)

Video dataset: Valley 702K (Valley)

- Jukinmedia (website) → Valley (702K pairs)
Stage 1: Video/Image Understanding - Training

This guy did an amazing martial arts stunt. He jumped on the mat and did a backflip. At the same time, a volunteer was ...
Stage 2: Instruction Tuning - Dataset

**Image dataset:** LLaVA-Mixed (LLaVA 1.5)

- OKVQA (9K pairs)
- OCRVQA (80K pairs)
- A-OKVQA (66K pairs)
- LLaVA-Instruct (158K pairs)
- GQA (72K pairs)
- VQA-v2 (83K pairs)
- SG40k (40K pairs)
- TextCaps (22K pairs)
- RefCOCO (48K pairs)
- GQA (86K pairs)

**Video dataset:** Video-ChatGPT

- ActivityNet 200 (200K videos)
- BLIP-2
- GPT 3.5
- GRIT
- Tag2Text
- Video ChatGPT (100K pairs)
Stage 2: Instruction Tuning - Training

\[ X_V = (X^1_q, X^1_a, ..., X^N_q, X^N_a) \]

\[ X_T = (X^1_q, X^1_a, ..., X^N_q, X^N_a) \]

- \( X^1_q \): “What landmark is shown in the video?”
- \( X^1_a \): “The Statue of Liberty”
- ... 
- \( X^N_q \): “Where is it located?”
- \( X^N_a \): “New York City”

The text embedding is calculated as:

\[ <\text{img}> + V_1 \]
\[ \vdots \]
\[ <\text{img}> + V_8 \]
\[ <\pos0> + T_1 \]
\[ \vdots \]
\[ <\posM> + T_M \]

The resulting input to the LLM (Vicuna 1.5) is:

\[ (\tilde{X}^1_a, ..., \tilde{X}^N_a) \]
Results
Zero-Shot Image Question-Answering

<table>
<thead>
<tr>
<th>Methods</th>
<th>LLM</th>
<th>Res.</th>
<th>Image Question Answering</th>
<th>VQA$^2$</th>
<th>GQA</th>
<th>VisWiz</th>
<th>SQA$^1$</th>
<th>VQA$^T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MiniGPT-4</td>
<td>LLaMA-7B</td>
<td>224</td>
<td></td>
<td>-</td>
<td>30.8</td>
<td>47.5</td>
<td>25.4</td>
<td>19.4</td>
</tr>
<tr>
<td>IDEFICS-9B</td>
<td>LLaMA-7B</td>
<td>224</td>
<td></td>
<td>50.9</td>
<td>38.4</td>
<td>35.5</td>
<td>-</td>
<td>25.9</td>
</tr>
<tr>
<td>mPLUG-Owl</td>
<td>LLaMA-7B</td>
<td>224</td>
<td></td>
<td>14.0</td>
<td>39.0</td>
<td>2.8</td>
<td>38.8</td>
<td></td>
</tr>
<tr>
<td>Otter</td>
<td>LLaMA-7B</td>
<td>224</td>
<td></td>
<td>38.1</td>
<td></td>
<td></td>
<td>50.0</td>
<td>27.2</td>
</tr>
<tr>
<td>InstructBLIP</td>
<td>Vicuna-7B</td>
<td>224</td>
<td></td>
<td></td>
<td>49.2</td>
<td>34.5</td>
<td>60.5</td>
<td>50.1</td>
</tr>
<tr>
<td>Video-LLaVA</td>
<td>Vicuna-7B</td>
<td>224</td>
<td></td>
<td>74.7$^*$</td>
<td>60.3$^*$</td>
<td>48.1</td>
<td>66.4</td>
<td>51.8</td>
</tr>
</tbody>
</table>

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

**Claim:** Demonstrates strong understanding ability in natural visual environments.
Image Benchmark Toolkits

- **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.
POPE

- **Sampling Schemes**
  - Adversarial: Similarly.
  - Popular: Most frequently.
  - Random: Randomly.

```
Ground-truth objects
person, chair, umbrella, sand, sea, ...

Nonexistent objects
Random: dog, apple, ...
Popular: table, knife, ...
Adversarial: surfboard, ...
```

Polling questions

Q: Is there a **person** in the image?  
A: Yes.

Q: Is there a **chair** in the image?  
A: Yes.

Q: Is there an **umbrella** in the image?  
A: Yes.

Q: Is there a **dog** in the image?  
A: No.

Q: Is there a **table** in the image?  
A: No.

Q: Is there a **surfboard** in the image?  
A: No.

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

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>LLM</th>
<th>Res.</th>
<th>Benchmark Toolkit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>POPE</td>
<td>MMB</td>
</tr>
<tr>
<td>MiniGPT-4</td>
<td>LLaMA-7B</td>
<td>224</td>
<td>23.0</td>
</tr>
<tr>
<td>IDEFICS-9B</td>
<td>LLaMA-7B</td>
<td>224</td>
<td>48.2</td>
</tr>
<tr>
<td>mPLUG-Owl</td>
<td>LLaMA-7B</td>
<td>224</td>
<td>46.6</td>
</tr>
<tr>
<td>Otter</td>
<td>LLaMA-7B</td>
<td>224</td>
<td>32.6</td>
</tr>
<tr>
<td>InstructBLIP</td>
<td>Vicuna-7B</td>
<td>224</td>
<td>36.0</td>
</tr>
<tr>
<td><strong>Video-LLaVA</strong></td>
<td><strong>Vicuna-7B</strong></td>
<td><strong>224</strong></td>
<td><strong>84.4</strong></td>
</tr>
</tbody>
</table>
**Image Object Hallucination Evaluation**

Evaluation results are reported for the POPE evaluation setting

<table>
<thead>
<tr>
<th>Methods</th>
<th>LLM</th>
<th>Adversarial</th>
<th></th>
<th>Popular</th>
<th></th>
<th>Random</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
<td>F1-Score</td>
<td>Accuracy</td>
<td>F1-Score</td>
<td>Accuracy</td>
<td>F1-Score</td>
</tr>
<tr>
<td>MiniGPT-4</td>
<td>Vicuna-13B</td>
<td>66.6</td>
<td>71.4</td>
<td>66.7</td>
<td>68.3</td>
<td>72.2</td>
<td>64.1</td>
</tr>
<tr>
<td>InstructBLIP</td>
<td>Vicuna-13B</td>
<td>74.4</td>
<td>78.5</td>
<td>69.0</td>
<td>81.4</td>
<td>83.5</td>
<td>62.6</td>
</tr>
<tr>
<td>MM-GPT</td>
<td>LLaMA-7B</td>
<td>50.0</td>
<td>66.7</td>
<td>100.0</td>
<td>50.0</td>
<td>66.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Video-LLaVA</td>
<td>Vicuna-7B</td>
<td>81.6</td>
<td>80.8</td>
<td>45.8</td>
<td>85.3</td>
<td>84.0</td>
<td>42.1</td>
</tr>
</tbody>
</table>

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

**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

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

- 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.
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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>MSVD</th>
<th>MSRVTT</th>
<th>TGIF</th>
<th>ActivityNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video-LLaVA*</td>
<td>64.8</td>
<td>58.3</td>
<td>67.8</td>
<td>40.7</td>
</tr>
<tr>
<td>Joint with Image</td>
<td>70.7</td>
<td>59.2</td>
<td>70.0</td>
<td>45.3</td>
</tr>
<tr>
<td>Δ Acc.</td>
<td>+ 5.9%</td>
<td>+ 0.9%</td>
<td>+ 2.2%</td>
<td>+ 4.6%</td>
</tr>
</tbody>
</table>
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.)