PG-Video-LLaVA: Pixel Grounding Large Video-Language Models

ArXiv Preprint, December 2023 (2 Citations)

Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, Fahad Khan

Presented by Group 3:
- Tyler VanderMate
- Ashton Frias
- Nicholas Gray
- Wen-Kai Chen
- Abhinav Kotta

What does the child in yellow hold outdoors?
The child in yellow holds a tennis racket.

Who walks to the door and opens it?
A man walks to the door and opens it.
Outline

• Introduction
• Model Architecture
• Training and Results
• Conclusion
• Limitations and Potential Ideas
Introduction

• Current LLM/LMMs to images
  • LLaVa
  • InstructBLIP
  • MiniGPT-4

• Extending to Video is still an open question

• Challenges
  o Massive Data Volume
  o Temporal Complexity
  o Audio Complexity
Recent Approaches

• Recent works that align LLMs to videos
  o VideoChat
  o Video-ChatGPT
  o Video-LLaMA

• Demonstrate video and dialogue comprehension

• Lack visual grounding feature
  o Linking LLM responses to specific objects in the video

• Lack the use of audio signals
Key Contributions

1. The first video-based LMM with pixel-level grounding capabilities
2. Incorporates audio context that enhances video understanding.
3. Improved quantitative benchmarks for video-based conversation models
Frozen LLM (Vicuna-13b-v1.5)

- Ensures reproducibility of results over GPT 3.5-Turbo
- 4 Section Prompt
  - Text System Command
  - Text User Query
  - Video Tokens
  - Text Audio Transcript
PG-Video-LLaVA

Response: The person in the video is a man who is sitting on the ground and petting a lion.

Grounding Module

System Command

Spatio-temporal Video Representation

Audio Transcript

Large Language Model

System Command

User Query

Spatio-temporal Video Representation

Audio Transcript
System Commands & User Query

- "You are PG-Video-LLaVA, a large vision-language assistant."
- "You are able to understand the video content that the user provides, and assist the user with a variety of tasks using natural language."
- "Follow the instructions carefully and explain your answers in detail based on the provided video."

**User Query**
  - What is the person doing in the video?
PG-Video-LLaVA

Grounding Module

Response

The person in the video is a man who is sitting on the ground and petting a lion.

Large Language Model

System Command

You are PG-Video-LLaVA, a large vision language model trained with video instruction data.

User Query

What is the person in the video doing?

Spatio-temporal Video Representation

Audio Transcript

Audio Transcript

Spatial Temporal Video Representation
Audio Modality Integration

1. Voice Activity Detection
Voice Activity Detection & Cut/Merging

• Identifies speech segments
• Filter out noise
  o Lyrics of background music
  o Garbage noise
• Cuts and merges audio into batches
  o Aligns with context window (Cuts up to 30 seconds)
  o Parallelizable
Audio Modality Integration

1. Voice Activity Detection
2. Whisper
3. Audio Tagging
4. Phoneme Model
Whisper

- Whisper Model
  - Speech to transcript generation
Audio Tagging (Whisper-AT)

• Produce audio-tagging output from the sliced original audio
• Consider the top 3 audio classes predicted
  o If “speech” is not among the top 3 predictions, the segment is ignored.
  o If $P[\text{music}] > P[\text{speech}]$ and $P[\text{music}] - P[\text{speech}] > \text{threshold}$, the segment is ignored
Phoneme Model & Forced Alignment

1. Phoneme Model
   - Classifies the smallest unit of speech distinguishing one from another

2. Forced Alignment (Dynamic Time Warping)
   - Temporally aligns transcript with matching phonemes
Audio Transcription & Tagging Result

Transcript Obtained from Whisper

Hi, I'm Stephanie Ragnodden and I'm going to show you step 3 of a 4-step process to groom your horse to shine naturally. This time we're going to use a finishing brush with a cocoa fiber horse hair mix because it just gives us a little bit more of an edge. ... We're struggling with the oil so we'll put the coat and always cleaning our brush on the curry and we're going to do this from the head all the way to the tail. Thanks for watching.

Transcript Obtained from Whisper

Yeah You know that? I didn't think that I had a debt to pay Till the king had take what I left away It was all my fault, you beat it to destiny But I remember you saying that yesterday There was a time when my heart wasn't on the show ... YOUR BLESS BUT YOU are a forever Oh to be the best WOO Kennedy I'm not saying it right here YOUR BLESS I don't know what I thought I might say Seems like we never were talking right away Every other minute I'm fuffing my place

1.5-1.1 2.5-1.1 2.5-1.1 2.5-1.1 2.5-1.1 2.5-1.1 2.5-1.1 2.5-1.1 1.5-1.1

Hallucination

Music

Garbage
PG-Video-LLaVA

The person in the video is a man who is sitting on the ground and petting a lion.
Spatial Temporal Video Representation (Video-ChatGPT)

MLP

\[ z_i \in \mathbb{R}^{T \times D} \]

\[ t_i \in \mathbb{R}^{N \times D} \]

\[ x_i \in \mathbb{R}^{T \times h \times w \times D} \]

\[ N = h \times w \]

Visual Encoder (CLIP ViT-L/14)

\[ V_i \in \mathbb{R}^{T \times H \times W \times 3} \]
Spatial Temporal Video Representation (Video-ChatGPT)

\[ Q_v = g(v_i) \in \mathbb{R}^{(T+N)\times D} \]

MLP

\[ v_i = [t_i \ z_i] \in \mathbb{R}^{(T+N)\times D} \]

Visual Encoder (CLIP ViT-L/14)

\[ V_i \in \mathbb{R}^{T \times H \times W \times 3} \]
PG-Video-LLaVA

The person in the video is a man who is sitting on the ground and petting a lion.
Scene Detection (PySceneDetect)
Tagging Module (RAM SWIN Large Variant)
Grounding Module (GroundingDINO-T)

“lion, man, mountain”

lion | man | mountain
Proposal Refinement (SAM)
Class Agnostic Object Tracker (DEVA Tracker)

- Decouples segmentation and tracking tasks to be class agnostic
- Assigns agnostic IDs to individual objects through the video
The person in the video is a man who is sitting on the ground and petting a lion.
PG-Video-LLaVA

Entity Matching Module

Response

The person in the video is a man who is sitting on the ground and petting a lion.

Large Language Model

System Command

You are PG-Video-LLaVA, a large vision language model trained with video instruction data.

User Query

What is the person in the video doing?

Visual Encoder (CLIP)

Scene Detection

Tagging Module

lion | man | mountain

Phoneme Model

Voice Activity Detection

Spatio-temporal Video Representation

Audio Transcript

Audio Tagging

Class Agnostic Object Tracker

Proposal Refinement

Grounding Module

MLP

Temporal Features

Spatial Pooling

Spatial Features

Temporal Pooling
Training Details

• Training dataset: VideoInstruct100K
  • 100K video instructions derived from ActivityNet-200

• Tune MLP layer for Spatial Temporal Video Representation
  • 3 epochs
  • Learning rate 2e-5
  • AdamW optimizer
  • Batch size 32
  • Causal LM objective
Conversational Benchmark Implementation

• Uses same framework as Video-ChatGPT
  o Queries LLM for 0-5 score of the predicted answer against ground truth
• Replaces GPT-3.5 with Vicuna-1.5 as evaluation LLM
• Measures LLM answer across 5 axes:
  o Correctness of Information
  o Detail Orientation
  o Contextual Understanding
  o Temporal Understanding
  o Consistency
Conversational Benchmark Implementation

• Correctness of Information
  o Verify the accuracy of the generated text, ensuring it aligns with the video content and doesn’t misinterpret or misinform.

• Detail Orientation
  o Evaluate the depth of the model’s responses, looking for both completeness, meaning the model’s response covers all major points from the video, and the inclusion of specific details rather than just generic points in the model’s response.

• Contextual Understanding
  o Assess the model’s understanding of the video’s context, checking if its responses aligns with the overall context of the video content.

• Temporal Understanding
  o Examine the model's grasp of the temporal sequence of events in the video when answering questions.

• Consistency
  o Evaluate the model's consistency across different but similar questions or different sections of the video.
Example Query: Correctness of Information

```
messages=[
  
  
  "role": "system",
  "content":
  "You are an intelligent chatbot designed for evaluating the factual accuracy of generative outputs for video-based question-answer pairs."
  "Your task is to compare the predicted answer with the correct answer and determine if they are factually consistent. Here's how you can accomplish the task:"
  "###INSTRUCTIONS:"
  "Focus on the factual consistency between the predicted answer and the correct answer. The predicted answer should not contain any misinterpretations or misinformation."
  "The predicted answer must be factually accurate and align with the video content."
  "Consider synonyms or paraphrases as valid matches."
  "Evaluate the factual accuracy of the prediction compared to the answer."
  
  }

  "role": "user",
  "content":
  "Please evaluate the following video-based question-answer pair:"
  "Question: {question}"
  "Correct Answer: {answer}"
  "Predicted Answer: {pred}"
  "Provide your evaluation only as a factual accuracy score where the factual accuracy score is an integer value between 0 and 5, with 5 indicating the highest level of factual consistency."
  "Please generate the response in the form of a Python dictionary string with keys 'score', where its value is the factual accuracy score in INTEGER, not STRING."
  "DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string."
  "For example, your response should look like this: {'score': 4.8}."
```
Conversational Quantitative Benchmarks

- Distinct video segments
- Human enriched ground truth annotations

<table>
<thead>
<tr>
<th>Model</th>
<th>Correctness</th>
<th>Detail Orientation</th>
<th>Contextual Understanding</th>
<th>Temporal Understanding</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaMA Adapter [9]</td>
<td>2.34</td>
<td>2.43</td>
<td>2.65</td>
<td>2.20</td>
<td>3.02</td>
</tr>
<tr>
<td>Video Chat [15]</td>
<td>2.49</td>
<td>2.82</td>
<td>2.92</td>
<td>2.27</td>
<td>3.11</td>
</tr>
<tr>
<td>Video-LLaMA [45]</td>
<td>2.29</td>
<td>2.61</td>
<td>2.68</td>
<td>2.17</td>
<td>2.87</td>
</tr>
<tr>
<td>Video-ChatGPT [22]</td>
<td>2.51</td>
<td>2.53</td>
<td>2.85</td>
<td>2.32</td>
<td>3.10</td>
</tr>
<tr>
<td>PG-Video-LLaVA (7B)</td>
<td>2.73</td>
<td>2.89</td>
<td>3.13</td>
<td>2.44</td>
<td>3.40</td>
</tr>
<tr>
<td>PG-Video-LLaVA (13B)</td>
<td><strong>2.86</strong></td>
<td><strong>2.95</strong></td>
<td><strong>3.23</strong></td>
<td><strong>2.53</strong></td>
<td><strong>3.49</strong></td>
</tr>
</tbody>
</table>
Spatial Grounding Evaluation Datasets

VidSTG

VidOR
• Raw videos
• Dense bounding box annotations

VidSTG
• Raw videos
• Dense bounding box annotations
• Caption descriptions
• Given questions for caption

HC-SVTG

HC-SVTG
• Raw videos
• Dense bounding box annotations
• Caption descriptions

Vicuna 1.5

Generated questions for captions
Spatial Grounding Evaluation

Q: Who is sitting next to the lions?
A: The man is sitting next to the lions.

Vicuna 1.5 (Referring Expression Extraction) → “man” → Grounding Module

Ground Truth

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

$$\text{mean IoU} = \frac{1}{|v|} \left( \sum_{v \in \text{object is present in GT}} \text{IoU}(v) \right)$$
### Spatial Grounding Quantitative Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>VidSTG [48]</th>
<th>HC-STVG [34]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grounding DINO [20]</td>
<td>25.3</td>
<td>19.5</td>
</tr>
<tr>
<td>Video-LLaMA [45]</td>
<td>28.6</td>
<td>26.1</td>
</tr>
<tr>
<td>Video-ChatGPT [22]</td>
<td>32.8</td>
<td>20.8</td>
</tr>
<tr>
<td>PG-Video-LLaVA (7B)</td>
<td>34.2</td>
<td><strong>28.3</strong></td>
</tr>
<tr>
<td>PG-Video-LLaVA (13B)</td>
<td><strong>35.1</strong></td>
<td>27.3</td>
</tr>
</tbody>
</table>
Zero Shot Video VQA Datasets

- **MSVD-QA** (520 videos)
  - Based on MSVD dataset
  - QA pairs generated from video descriptions

- **MSRVTT-QA** (2,990 videos)
  - Based on MSR-VTT dataset
  - QA pairs generated from video descriptions

- **TGIF-QA** (8,944 videos)
  - Based on TGIF dataset
  - GIFs, no audio present
  - Crowdsourced QA pairs

- **Activity Net-QA** (500 videos)
  - Based on ActivityNet-200
  - Human annotated QA pairs
# Zero Shot Video VQA Benchmarks

## Question

Vicuna 1.5

## Actual Answer

Correct? (Yes/No)

How Correct? (Score 0-5)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FrozenBiLM [41]</td>
<td>32.2</td>
<td>–</td>
<td>41.0</td>
<td>24.7</td>
</tr>
<tr>
<td>LLaMA Adapter [9]</td>
<td>53.7</td>
<td>3.3</td>
<td>54.3</td>
<td>37.3</td>
</tr>
<tr>
<td>Video LLaMA [45]</td>
<td>48.6</td>
<td>3.2</td>
<td>51.4</td>
<td>27.1</td>
</tr>
<tr>
<td>Video-ChatGPT [22]</td>
<td>62.6</td>
<td>3.6</td>
<td>66.5</td>
<td><strong>40.8</strong></td>
</tr>
<tr>
<td>PG-Video-LLaVA</td>
<td><strong>64.1</strong></td>
<td><strong>3.7</strong></td>
<td>66.8</td>
<td>39.9</td>
</tr>
</tbody>
</table>

The model PG-Video-LLaVA achieves the highest accuracy and score across all benchmarks.
Conversational Qualitative Performance

Video-ChatGPT
The video shows a giraffe performing a trick by standing on a diving board and then jumping into the water.

PG-Video-LLaVA (7B)
The video shows a giraffe standing on a diving board and jumping into a pool. The giraffe is seen standing on the diving board, looking down, and then jumping into the pool. The scene takes place outdoors, with a clear blue sky visible in the background. The giraffe appears to be enjoying the water and the experience of jumping into the pool.

PG-Video-LLaVA (13B)
The video features a giraffe standing on a diving board and jumping into a pool. The giraffe is seen swimming in the pool and then jumping out of it. The video is set in a pool area with a blue sky in the background. The giraffe’s movements are graceful and fluid, and it appears to be enjoying itself. The video captures the unique and unexpected sight of a giraffe engaging in a human activity, showcasing its incredible agility and ability to adapt to new situations. Overall, the video is a delightful and entertaining spectacle that captures the beauty and grace of the giraffe in an unconventional setting.
Spatial Grounding Qualitative Performance

What do you see in this video?

In the video, we see a giraffe standing on a diving board, and then it jumps into the water. The giraffe is the main focus of the video, and we see it in various positions throughout the video.

What do you see in this video?

In the video, I see a large statue of a woman holding a torch in her right hand. The statue is located on top of a hill, and it is surrounded by a large body of water. The statue is lit up at night, and it is visible from a distance. The video also shows a boat passing by the statue.

What do you see in this video?

In the video, a young girl is performing a gymnastics routine on a balance beam. She starts by jumping onto the beam and then proceeds to perform various flips and turns. She finishes her routine by jumping off the beam and landing on a mat.
Conversational + Audio Transcript Performance

Can you suggest a title for this video clip based on what you see and hear?

Without Audio

Based on the video, I would suggest "The Future of Technology and Communication" as a title for the video clip.

What is Ross complaining about?

Without Audio

Ross is complaining about the taste of the beer he is drinking. He is seen making a face and spitting out the beer, while the other people in the scene watch him.
PG-Video-LLaVA Failure Cases

Who does the woman in the yellow dress hug?

The woman in the yellow dress hugs the man in the blue suit.

Who walks to the wooden frame?

A woman walks to the wooden frame in the video.

Who walks to the door and then turns?

A man walks to the door and then turns around to face the camera.

Who pushes the woman to the table?

A man in a white lab coat pushes the woman to the table.
Conclusion

• PG-Video-LLaVA is a novel video-based conversational model with pixel-level grounding capabilities

• Novel addition of filtered audio transcripts to enrich visual understanding

• Grounding module to track and generate pixel-level object grounding in videos

• Improved and reproducible quantitative benchmarks by switching from GPT-3.5 to Vicuna LLM
Limitations and Potential Ideas

• Entity Matching Module uses only text for reasoning
  - Use image encoder or spatial temporal feature extraction on bounding boxes alongside text prompt.

• Audio transcript simplifies complexities of speech
  - Contrastive learning between an audio encoder and audio transcript.
  - Extend audio tagging module to output tone and intensity.

• Multiple potential points of failure (LLM output, grounding module, object tracking, audio transcription, etc.)
  - Use evaluation LLM to check results of individual steps in grounding module.
  - Check between grounding module output and LLM output.