

# Visual-Language Models Introduction

## Part-IV: Video ChatGPT, PG-Video LLaVA

### Lecture-7

## CAP6412 Spring 2024

Mubarak Shah

[shah@crcv.ucf.edu](mailto:shah@crcv.ucf.edu)

# Contents

- CoCa
- PALI
- FLAMINGO
- FLAVA
- Painter
- BLIP-2
- Image-Bind
- Language-Bind
- LLaVA
- Video ChatGPT

# **Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models.**

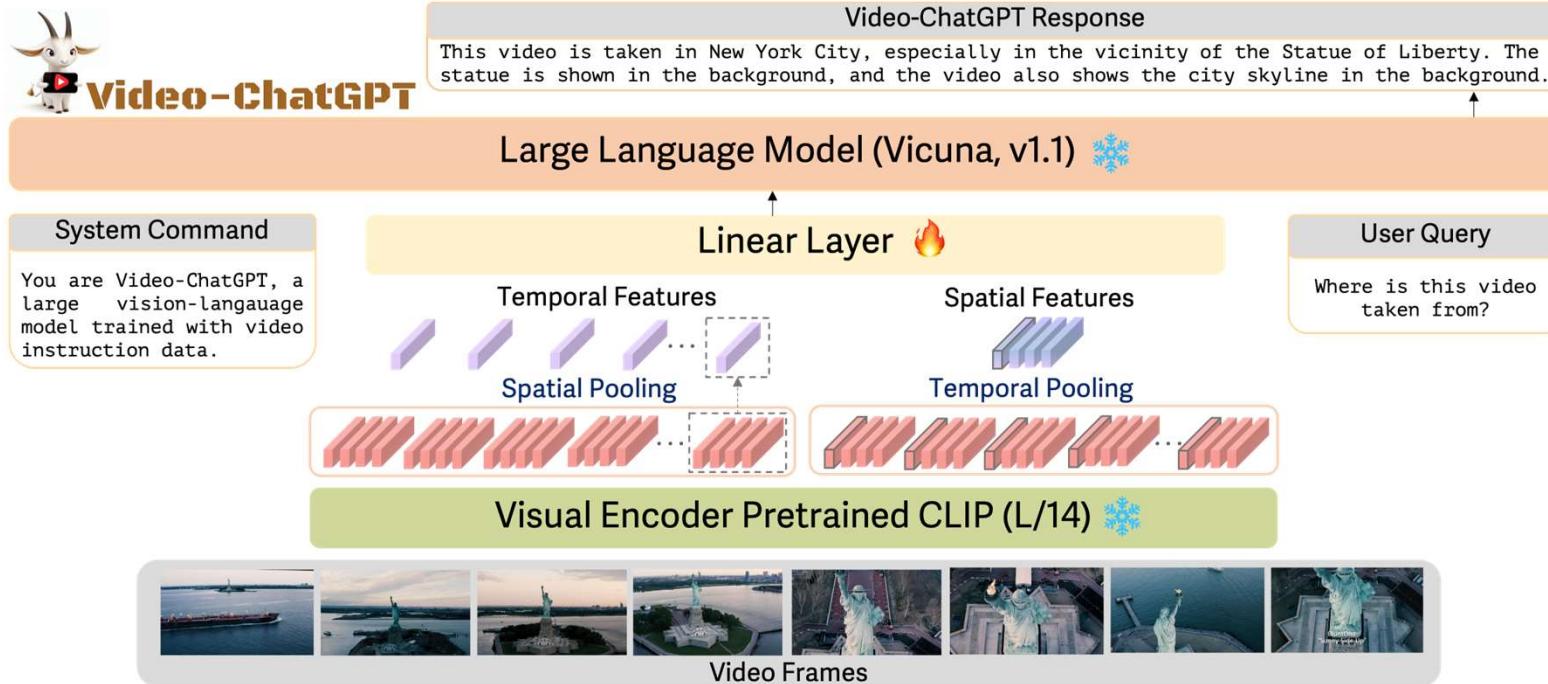
**Muhammad Maaz\*, Hanoona Rasheed\*, Salman Khan, Fahad Shahbaz Khan**  
**Mohamed bin Zayed University of AI**

Maaz et al., *Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models*. Arxiv 2023.

# Video ChatGPT

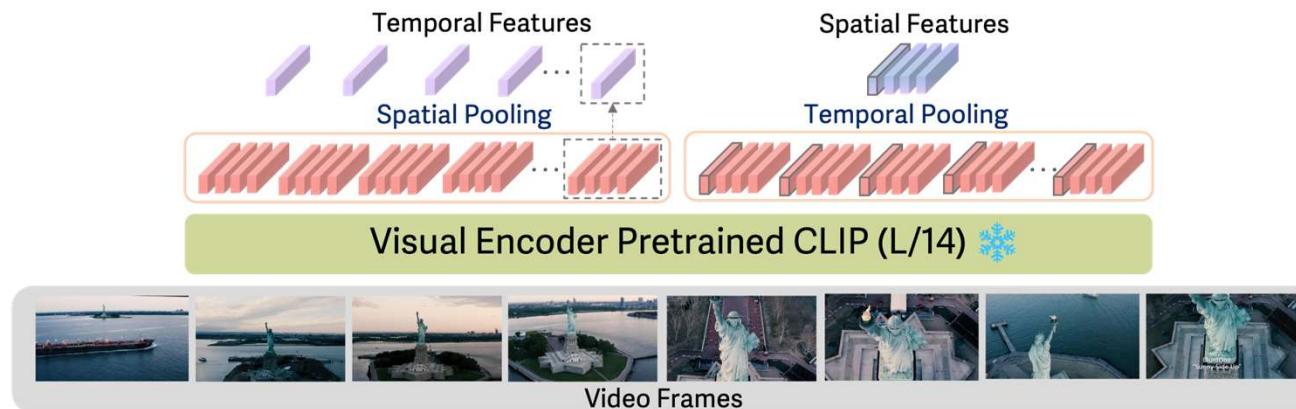
- Simple Video Conversational Model
- Semi-automated annotation framework
- 100K high quality instructions data for videos
- First quantitative video conversation evaluation

# Video-ChatGPT: Architecture



# Video-ChatGPT: Architecture

- Spatiotemporal features extraction
- Spatial pooling for temporal feature extraction
- Temporal pooling for spatial feature extraction



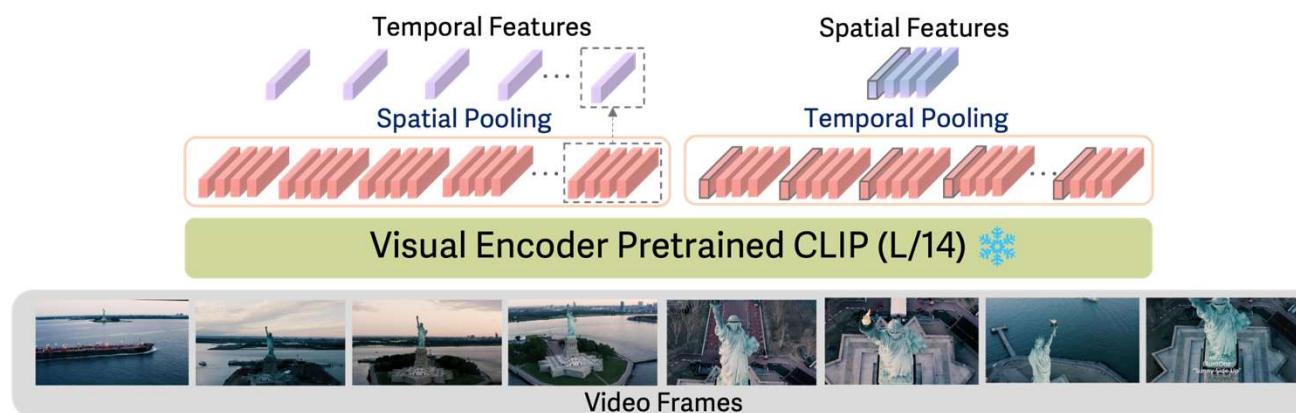
# Video-ChatGPT: Architecture

$$x_i \in \mathbb{R}^{T \times h \times w \times D}, \quad N = h \times w$$

average pooling across the temporal dimension  $t_i \in \mathbb{R}^{N \times D}$

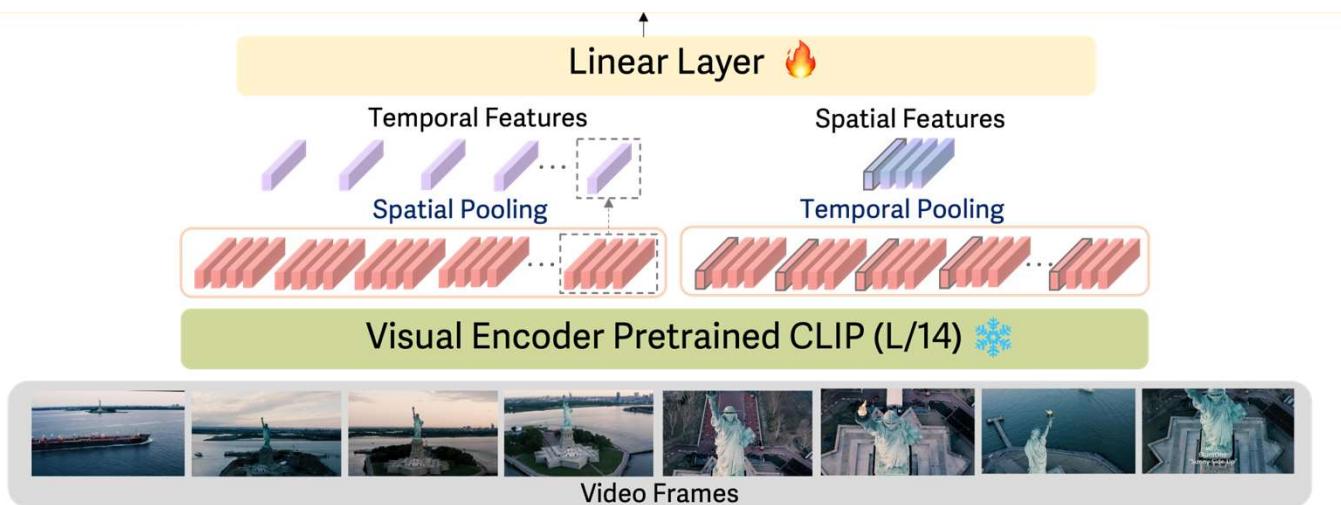
average pooling along the spatial dimension.  $z_i \in \mathbb{R}^{T \times D}$

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



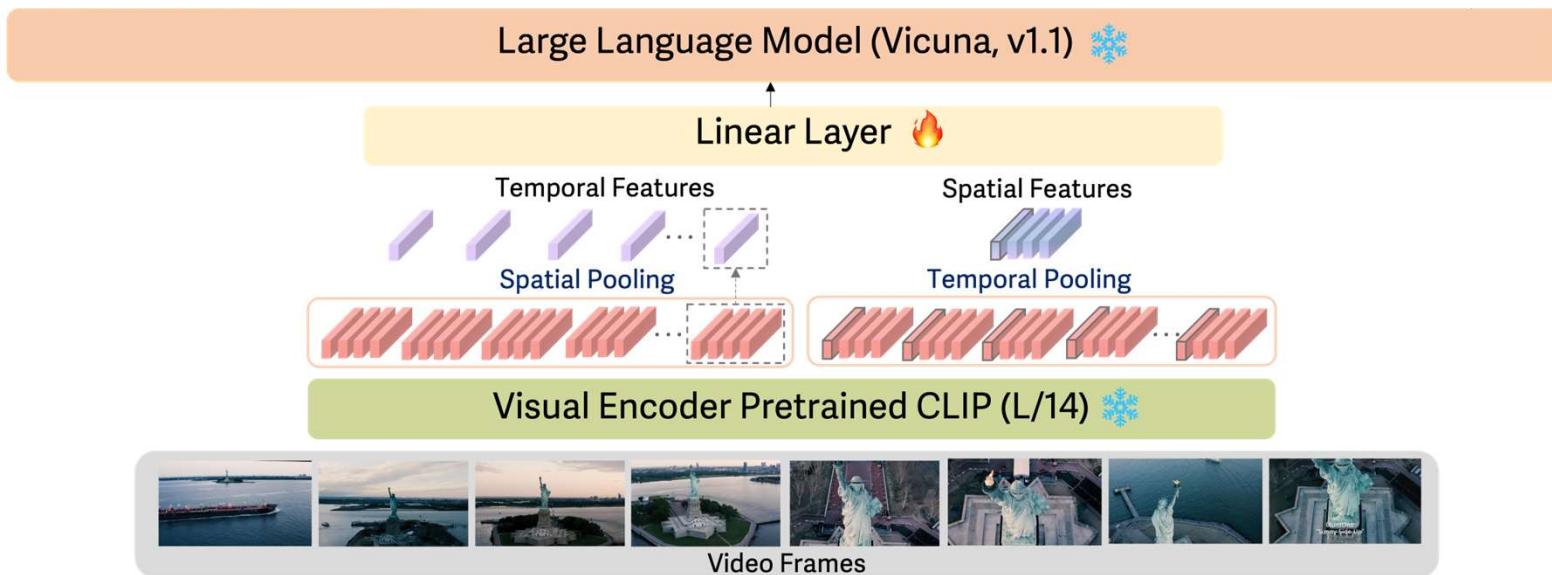
# Video-ChatGPT: Architecture

- Linear layer projects spatiotemporal video features to LLM input space

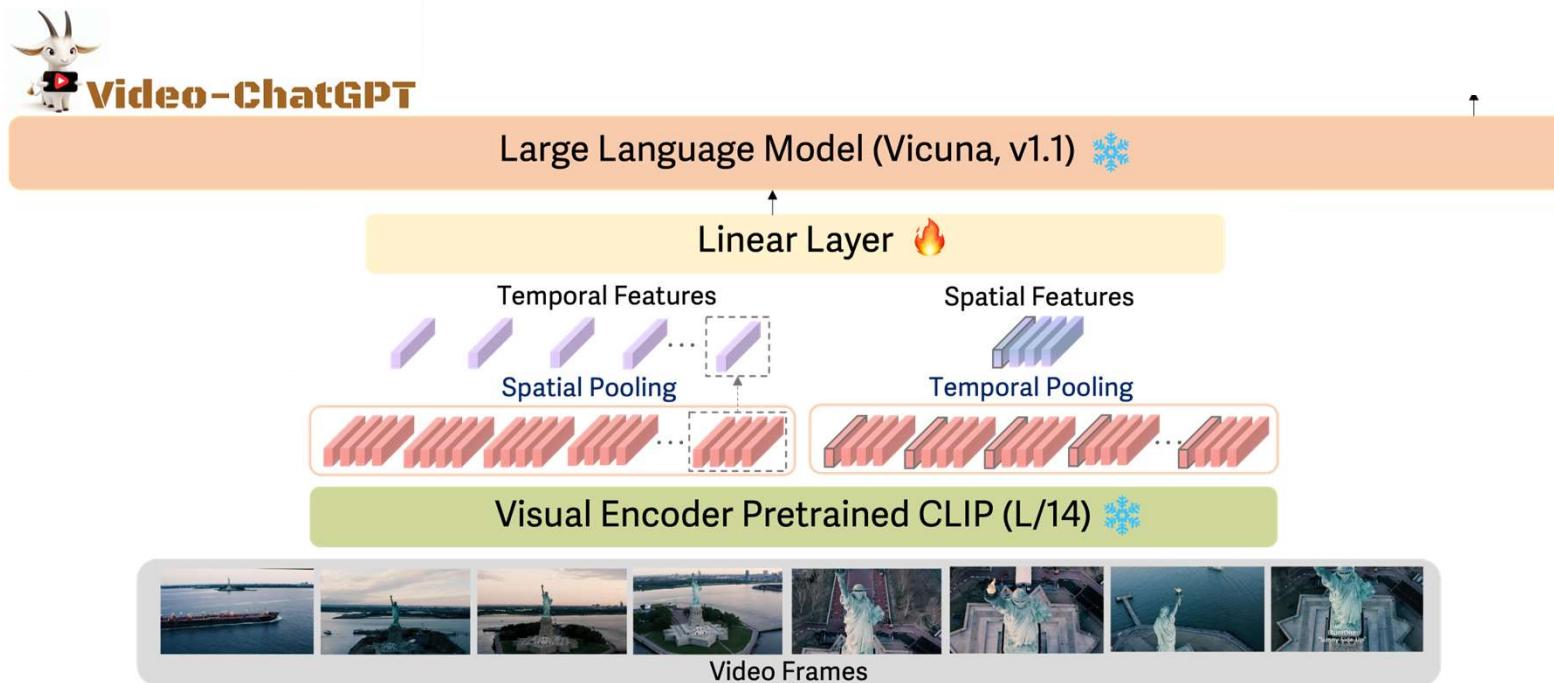


# Video-ChatGPT: Architecture

- Linear layer projects spatiotemporal video features to LLM input space



# Video-ChatGPT: Architecture



## Video-ChatGPT: Video Instruction Set Creation

- Human Assisted Annotations
- Videos and captions from ActivityNet-200
- Annotators enrich original captions by adding information about
  - Physical appearances
  - Spatial and temporal localization

# Video-ChatGPT: Video Instruction Set Creation



## Original Dataset Annotation:

An intro leads into various pictures of horse tools and a woman speaking to the camera holding a brush. She then brushes the horse's mane thoroughly while still looking to the camera and speaking.

## Human Annotated:

The video begins with an intro slide, where there is a picture of the woman standing next to a brown and white horse. The woman is wearing a green jacket over a white shirt. Then an intro leads into various pictures of horse tools and a woman speaking to the camera holding a brush. The video is about "how to groom your horse to shine in four easy steps with a finishing brush. She then brushes the horse's mane thoroughly while still looking at the camera and speaking. The horse is black and the video is shot inside a stable. She wears a green jacket over a long-sleeved purple t-shirt. She is explaining how to use a finishing brush. She shows that the brush has a raised edge, which will help remove more dirt. She uses long strokes to brush the hair of the horse in the direction of hair growth. She cleans from the top to the legs. She then explains that the brush needs to be cleaned with a tool named curry. Finally, the video ends with a picture of the black horse all groomed up and credits to the video.



## Original Dataset Annotation:

A close up of a Christmas tree is shown followed by close ups of ornaments. Two people are then seen moving around the tree decorating as well as turning the lights off. They finish decorating the tree and playing with one another and laughing. In the end close ups of the trees are shown as well as a bear.

## Human Annotated:

In the video, we see a beautifully decorated Christmas tree with lush green branches adorned with bright and colorful ornaments. As the camera pans over the ornaments, they glisten in the light, reflecting the colors of the rainbow. Two people are then shown moving around the tree, hanging ornaments and stringing lights, carefully placing each ornament in its designated spot. As they work, they chat and joke around, enjoying each other's company and the festive spirit. After they finish hanging the ornaments, they step back and admire their work, giggling and hugging each other. The camera captures close-ups of the finished tree, showing off all of its beauty and sparkle. The lights twinkle and glow, casting a warm and cozy ambiance around the room. The scene is one of holiday cheer and joy, a perfect representation of the Christmas spirit. Finally, the camera zooms out and we see a cute little bear toy dancing under the tree, surrounded by beautifully wrapped presents. Overall, the video showcases the joy and warmth of the festive season, with a beautifully decorated tree, love, and the spirit of giving and sharing.

# Video-ChatGPT: Video Instruction Set Creation



## Original Dataset Annotation:

A close up of a christmas tree is shown followed by close ups of ornaments. Two people are then seen moving around the tree decorating as well as turning the lights off. They finish decorating the tree and playing with one another and laughing. In the end close ups of the trees are shown as well as a bear.

## Human Annotated:

In the video, we see a beautifully decorated Christmas tree with lush green branches adorned with bright and colorful ornaments. As the camera pans over the ornaments, they glisten in the light, reflecting the colors of the rainbow. Two people are then shown moving around the tree, hanging ornaments and stringing lights, carefully placing each ornament in its designated spot. As they work, they chat and joke around, enjoying each other's company and the festive spirit. After they finish hanging the ornaments, they step back and admire their work, giggling and hugging each other. The camera captures close-ups of the finished tree, showing off all of its beauty and sparkle. The lights twinkle and glow, casting a warm and cozy ambiance around the room. The scene is one of holiday cheer and joy, a perfect representation of the Christmas spirit. Finally, the camera zooms out and we see a cute little bear toy dancing under the tree, surrounded by beautifully wrapped presents. Overall, the video showcases the joy and warmth of the festive season, with a beautifully decorated tree, love, and the spirit of giving and sharing.

# Video Instruction Set Creation

- **Semi-automatic Annotation Framework:**
- Enrich the ground truth captions using the off-the-shelf Models
  - BLIP-2 – Frame level image captioning
  - GRiT – Detailed captions for scene objects
  - Tag2Text – Tags for each key frame

# **Video-ChatGPT: Video Instruction Set Creation**

## **Semi-automatic Annotation Framework:**

- To ensure high-quality data and mitigate noise
  - High prediction threshold for all the off-the-shelf models
  - Remove BLIP-2 or GRiT caption not matching with Tag2Text frame-level tags.
  - Finally, merge frame-level captions and use GPT-3.5 to generate coherent caption.

# Video-ChatGPT: Video Instruction Set Creation

## Semi-automatic Annotation Framework:



### Original Dataset Annotation:

A countdown is shown followed by a woman catching balls thrown at her. The number counts down more and shows more clips of women playing dodgeball. Several more numbers and clips are shown ending with people performing impressive throws.

### Semi-Automatically Annotated:

The video starts with a countdown, and as the numbers decrease, we see a woman standing in an arena. Suddenly, multiple balls are thrown at her from different directions, and we see her successfully catch them one by one with great agility and precision. As the countdown continues, we see more clips of women playing dodgeball in the same arena, demonstrating their skills in avoiding being hit by the balls. The countdown continues to show more numbers and clips of people playing dodgeball. We see players on different teams with colorful jerseys, throwing balls at their opponents and trying to avoid getting hit themselves. In some of the clips, we see players making impressive throws, spinning and turning before they release the ball, which adds to the excitement of the game. Towards the end of the video, we see people performing some incredible throws, bouncing the ball off the floor or walls before hitting their targets. The video ends with a final impressive move, where someone jumps from a trampoline and throws the ball with great power and accuracy. Overall, the video showcases the excitement, skills, and competitiveness of the game of dodgeball, emphasizing the agility and precision required to play it successfully.



### Original Dataset Annotation:

Several individuals play against each other in a table football match. Two individuals walk by in the background. Another two individuals walk by in the background.

### Semi-Automatically Annotated:

This video features a table football match between several individuals. In the background, we can see two individuals walking by, followed by another two individuals. The main focus of the video is the intense match taking place on the table football. The players seem to be engaged in a competitive game, with each team trying to score as many goals as possible. The players are constantly moving their rods back and forth, trying to defend their goal and score against their opponents. As we observe the video, we can see that it takes place indoors, presumably in a recreational center or someone's home. The noise level is moderate, with the sounds of the table football game and quiet conversations in the background. The lighting is adequate, with a soft yellowish glow coming from ceiling lights. The table football itself is a standard one, with green color cloth and black and white football players. In conclusion, this video is a portrayal of an enthusiastic and competitive table football game, with several individuals trying their best to defeat their opponents. The peripheral characters in the background make it evident that the video takes place indoors, most probably in a recreational setting.

# Video-ChatGPT: Video Instruction Set Creation

## Semi-automatic Annotation Framework:



### Original Dataset Annotation:

A countdown is shown followed by a woman catching balls thrown at her. The number counts down more and shows more clips of women playing dodgeball. Several more numbers and clips are shown ending with people performing impressive throws.

### Semi-Automatically Annotated:

The video starts with a countdown, and as the numbers decrease, we see a woman standing in an arena. Suddenly, multiple balls are thrown at her from different directions, and we see her successfully catch them one by one with great agility and precision. As the countdown continues, we see more clips of women playing dodgeball in the same arena, demonstrating their skills in avoiding being hit by the balls. The countdown continues to show more numbers and clips of people playing dodgeball. We see players on different teams with colorful jerseys, throwing balls at their opponents and trying to avoid getting hit themselves. In some of the clips, we see players making impressive throws, spinning and turning before they release the ball, which adds to the excitement of the game. Towards the end of the video, we see people performing some incredible throws, bouncing the ball off the floor or walls before hitting their targets. The video ends with a final impressive move, where someone jumps from a trampoline and throws the ball with great power and accuracy. Overall, the video showcases the excitement, skills, and competitiveness of the game of dodgeball, emphasizing the agility and precision required to play it successfully.

# Video-ChatGPT: Video Instruction Set Creation

Semi-automatic Annotation Framework:



Original Dataset Annotation:

"A man is playing the bagpipes in front of people., "The people on the couch in front of him start laughing."

# Video-ChatGPT: Video Instruction Set Creation

Semi-automatic Annotation Framework:

Generating Off-the-Shelf Dense Predictions:

BLIP2 Caption: ["**three women** are sitting on a couch"]

GRIT Descriptions: "**a women in a pink outfit**: [45, 23, 338, 288]; "

Tags: ["table | couch | pajama | woman | living room | chair | girl | sit | wear | sit on | pink | red"]

BLIP2 Caption: ["**a man and woman** sitting on a couch"]

GRIT Descriptions: "**man sitting on couch**: [172, 13, 379, 287];

a woman sitting on a couch: [31, 39, 181, 285]; couch with flower pattern: [1, 98, 383, 286];

a woman holding a remote: [268, 17, 310, 108]; "

Tags: ["couch | woman | chair | people | show | person | man | sit | sit in | sit on"], [320, 240]]

BLIP2 Caption: ["two women sitting on a couch laughing"]

GRIT Descriptions: "**woman wearing a blue shirt**: [31, 36, 246, 287];

woman in pink shirt on phone: [177, 22, 331, 252]; "

Tags: ["couch | woman | blue | shirt | girl | sit | sit on | talk"], [320, 240]]

BLIP2 Caption: ["a man holding a bagpipes **in front of a television**"]

GRIT Descriptions: "**a man in a black shirt**: [116, 6, 315, 287];

a green plant next to a couch: [0, 201, 97, 287]; "

Tags: ["microphone | person | man | black | shirt| hold | stand"], [320, 240]]

BLIP2 Caption: ["a man playing the bagpipes **in a living room**"]

GRIT Descriptions: "**a man in a black shirt**: [109, 18, 290, 287];

a green plant next to the man: [10, 195, 131, 287];

Tags: ["pipe | black | shirt | tv | person | man | play | stand"], [320, 240]]

# Video-ChatGPT: Video Instruction Set Creation

## Semi-automatic Annotation Framework:

### Filtering Off-the-Shelf Dense Predictions: **Removing noisy information**

BLIP2 Caption: ["three women are sitting on a couch"]

GRIT Descriptions: "a women in a pink outfit: [45, 23, 338, 288]; "

Tags: ["table | couch | pajama | woman | living room | chair | girl | sit | wear | sit on | pink | red"]

BLIP2 Caption: ["a man and woman sitting on a couch"]

GRIT Descriptions: "man sitting on couch: [172, 13, 379, 287];

a woman sitting on a couch: [31, 39, 181, 285]; couch with flower pattern: [1, 98, 383, 286];

a woman holding a remote: [268, 17, 310, 108]; "

Tags: ["couch | woman | chair | people | show | person | man | sit | sit in | sit on"], [320, 240]]

BLIP2 Caption: ["two women sitting on a couch laughing"]

GRIT Descriptions: "woman wearing a blue sweater: [31, 36, 246, 287];

woman in pink shirt on phone: [177, 22, 331, 252]; "

Tags: ["couch | woman | blue | sweater | girl | sit | sit on | talk"], [320, 240]]

BLIP2 Caption: ["a man holding a bagpipes in front of a television"]

GRIT Descriptions: "a man in a black shirt: [116, 6, 315, 287];

a green plant next to a couch: [0, 201, 97, 287]; "

Tags: ["microphone | person | man | black | shirt| hold | stand"], [320, 240]]

BLIP2 Caption: ["a man playing the bagpipes in a living room"]

GRIT Descriptions: "a man in a black shirt: [109, 18, 290, 287];

a green plant next to the man: [10, 195, 131, 287];

Tags: ["pipe | black | shirt | tv | person | man | play | stand"], [320, 240]]

# Question & Answer Pairs from GPT 3.5

- These aspects include
  - Detailed descriptions,
  - Summarizations,
  - Question-answer pairs,
  - Tasks that stimulate creativity or the generation of new ideas,
  - Conversational tasks

[{"q": "What are the main activities that take place in the video?", "a": "The main activities that take place in the video are the preparation of camera equipment by a man, a group of men riding a helicopter, and a man sailing a boat through the water.", "video\_id": "v\_k\_ZXmr8pmrs"}, {"q": "Is the mime the only person who does something other than singing?", "a": "Yes, the mime walks in with fireworks while the group still sings and many around are seen laughing and smiling.", "video\_id": "v\_GqWH2IYPw6U"}, {"q": "Could you briefly describe the video content?", "a": "Certainly! The video shows a woman in a blue/white outfit lifting a heavy yellow weight after standing still for a while. While lifting the weight, someone watches the woman and yells in a peremptory manner. Shortly after, the woman drops the weight back down to the ground.", "video\_id": "v\_juLxWt\_3omw"}, {"q": "Can you give me an overview of the video content?", "a": "The video showcases a group of men washing cars in a parking lot using various tools and techniques, such as pressure washing hoses, hand towels, and brushes. There are also text displays of special offers and business details throughout the video. The men work on different cars, including a white car and a black SUV.", "video\_id": "v\_7bUu05RIksU"}, {"q": "Can you describe what the woman is doing on the lake?", "a": "The woman is pulled forward in the lake and then begins to glide on the water using a surfboard. She sways left and right while sliding, and then slowly squats down, leaving the surfboard and falling into the water and floating on the surface.", "video\_id": "v\_TpB\_zMG3XBA"}, {"q": "Can you summarize what the girls are doing in the video?", "a": "Sure. The video shows a girl brushing the hair of another girl who keeps moving her face around while the first girl keeps brushing the hair.", "video\_id": "v\_AmhfmeKk6Bg"}, {"q": "What is the main activity happening in the video?", "a": "The main activity in the video is a boy attempting to pole vault over a bar in an indoor field while being watched by several people.", "video\_id": "v\_rBMQFpHspmo"}, {"q": "Can you describe what happens in the video of \"tree mulching made easy\"?", "a": "Sure! The video starts with an introduction of a tree in a grassy field, with a house visible in the background. The title \"tree mulching made easy\" appears on the screen. Then, the scene shows a man preparing the area around the tree, with a pair of gloves, a rake, and 4"}]



# Video-ChatGPT: Video Instruction Set Creation

Prompt GPT-3.5 model to create question-answer pairs from the enriched and detailed captions

H: Can you summarize the events in the video?

AI: Yes, the video shows a man playing the bagpipes in front of a group of people sitting on a couch. They start laughing at some point, creating a friendly atmosphere.

H: Can you describe the appearance of the bagpipes player in the video?

AI: Yes, the bagpipes player is dressed in a black shirt and holding the bagpipes in front of a television.

H: Who are the people sitting on the couch in front of the bagpipes player?

AI: There are three women and a man sitting on the couch.

H: Can you describe the setting and atmosphere in the video?

AI: The video captures a relaxed and friendly atmosphere in a domestic setting. Specifically, it shows a group of people sitting on a couch chatting and laughing while a man in a black shirt plays the bagpipes in front of them. The people are dressed casually, and in some scenes, a woman in a pink outfit, another dressed in blue sweater sits next to a man on the couch and continue to chat.

## **Video-ChatGPT: Quantitative Evaluation**

- Introduce quantitative benchmarks for video conversation task
- Two types of evaluations
  - Video-based Generative Performance Benchmarking
  - Zero-shot question-answer evaluation

# Video-ChatGPT: Quantitative Evaluation

- Curate a dataset using ActivityNet-200 test set
- Human annotated detailed descriptions
- GPT-3.5 powered evaluation pipeline
  - Assesses various capabilities of the model and
  - Assigns a relative score to the generated predictions on a scale of 1-5.

Correctness of information

Detail orientation

Contextual understanding

Temporal understanding

Consistency

# Video-ChatGPT: Quantitative Evaluation

- **Correctness of Information:**
  - Verify the accuracy of the generated text, ensuring it aligns with the video content and doesn't misinterpret or misinform.
- **Detail Orientation:**
  - 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:**
  - 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:**
  - Examine the model's grasp of the temporal sequence of events in the video when answering questions.
- **Consistency:**
  - Evaluate the model's consistency across different but similar questions or different sections of the video.

# Instructions to GPT for correctness evaluation

```
• 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.\n"  
•             "- The predicted answer must be factually accurate and align with the video content.\n"  
•             "- Consider synonyms or paraphrases as valid matches.\n"  
•             "- Evaluate the factual accuracy of the prediction compared to the answer."  
•     }  
•     {  
•         "role": "user",  
•         "content":  
•             "Please evaluate the following video-based question-answer pair:\n\n"  
•             f"Question: {question}\n"  
•             f"Correct Answer: {answer}\n"  
•             f"Predicted Answer: {pred}\n\n"  
•             "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}."  
•     }  
• ] }
```

## Video-ChatGPT: Quantitative Evaluation

Evaluation Aspect	Video Chat	LLaMA Adapter	Video-LLaMA	Video-ChatGPT
Correctness of Information	2.23	2.03	1.96	<b>2.40</b>
Detail Orientation	2.50	2.32	2.18	<b>2.52</b>
Contextual Understanding	2.53	2.30	2.16	<b>2.62</b>
Temporal Understanding	1.94	<b>1.98</b>	1.82	<b>1.98</b>
Consistency	2.24	2.15	1.79	<b>2.37</b>

# Video-ChatGPT: Quantitative Evaluation

## Zero-shot Evaluation:

- Conducted evaluation on four commonly used open-ended question-answer datasets.
  - MSRVTT-QA
  - MSVD-QA
  - TGIF-QA FrameQA
  - ActivityNet-QA
- Use GPT-3.5 to measure accuracy of the model and assigns score of 1-5

# Video-ChatGPT: Quantitative Evaluation

Zero-shot Evaluation:

Model	MSVD-QA		MSRVTT-QA		TGIF-QA		Activity Net-QA	
	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM	32.2	–	16.8	–	41.0	–	24.7	–
Video Chat	56.3	2.8	45.0	2.5	34.4	2.3	26.5	2.2
LLaMA Adapter	54.9	3.1	43.8	2.7	-	-	34.2	2.7
Video LLaMA	51.6	2.5	29.6	1.8	-	-	12.4	1.1
Video-ChatGPT	<b>64.9</b>	<b>3.3</b>	<b>49.3</b>	<b>2.8</b>	<b>51.4</b>	<b>3.0</b>	<b>35.2</b>	<b>2.8</b>

# Video-ChatGPT: Quantitative Evaluation

## Zero-shot Evaluation:

- Compare performance with other significant models such as FrozenBiLM and generative video models including Video Chat, LLaMA Adapter and Video LLaMA.

Model	MSVD-QA		MSRVTT-QA		TGIF-QA		Activity Net-QA	
	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM	32.2	–	16.8	–	41.0	–	24.7	–
Video Chat	56.3	2.8	45.0	2.5	34.4	2.3	26.5	2.2
LLaMA Adapter	54.9	3.1	43.8	2.7	-	-	34.2	2.7
Video LLaMA	51.6	2.5	29.6	1.8	-	-	12.4	1.1
Video-ChatGPT	<b>64.9</b>	<b>3.3</b>	<b>49.3</b>	<b>2.8</b>	<b>51.4</b>	<b>3.0</b>	<b>35.2</b>	<b>2.8</b>

# Video-ChatGPT: Quantitative Evaluation

## Zero-shot Evaluation:

- FunQA: Towards Surprising Video Comprehension paper on the FunQA benchmarking

	HumorQA				CreativeQA					MagicQA		
	H1	H2	H3	H4	C1	C2	C3	C4	C5	M1	M2	M3
<b>- Caption-based Model</b>												
mPLUG [47]	-	<b>1.5 / 16.4 / 1.0</b>	<b>1.1 / 12.5 / 0.4</b>	<b>0.6 / 7.5 / 0.1</b>	-	0.4 / 13.4 / 0.0	0.7 / 12.6 / 0.1	0.3 / 3.2 / 0.0	-	1.2 / 15.8 / 0.5	0.9 / 8.9 / 0.4	-
	-	19.9 / 16.0	25.7 / 18.1	22.1 / 17.3	-	14.9 / <b>24.3</b>	24.2 / 9.0	20.8 / 13.7	-	19.7 / 16.9	21.2 / 8.8	-
GIT (L.M.) [52]	-	0.5 / 12.8 / 0.2	-	<b>1.1 / 7.7 / 0.7</b>	-	0.0 / 6.4 / 0.0	-	0.3 / 1.5 / 0.2	-	0.2 / 11.2 / 0.1	-	-
	-	22.4 / 22.0	-	17.0 / 26.8	-	14.4 / 5.0	-	7.1 / 25.2	-	19.4 / 12.7	-	-
GIT (L.V.) [52]	-	1.2 / 16.9 / 0.6	-	<b>1.0 / 8.8 / 0.7</b>	-	0.1 / 8.3 / 0.0	-	<b>0.5 / 2.8 / 0.4</b>	-	0.6 / 13.7 / 0.1	-	-
	-	33.3 / 31.5	-	<b>25.9 / 33.2</b>	-	20.5 / 5.0	-	10.5 / 23.3	-	29.8 / <b>21.4</b>	-	-
<b>- Instruction-based Model</b>												
VideoChat [35]	-	0.5 / 13.7 / 0.0	0.5 / 13.5 / 0.0	0.8 / 5.1 / 0.5	-	0.3 / 7.5 / 0.0	0.3 / 7.7 / 0.0	0.2 / 1.2 / 0.2	67.5	0.6 / 15.5 / 0.0	0.3 / 9.2 / 0.0	-
	-	<b>44.0 / 37.9</b>	<b>45.4 / 31.9</b>	<b>20.2 / 61.7</b>	-	21.7 / 10.9	22.8 / 27.7	7.3 / <b>51.1</b>	-	47.4 / 14.2	43.1 / 24.6	-
Video-ChatGPT [36]	-	0.5 / 14.0 / 0.1	0.7 / 12.4 / 0.1	0.4 / 3.2 / 0.2	-	<b>1.1 / 19.8 / 0.2</b>	0.8 / 17.3 / 0.1	0.2 / 1.9 / 0.2	<b>85.4</b>	0.7 / 20.8 / 0.0	0.5 / 11.3 / 0.0	-
	-	39.9 / 20.7	40.1 / <b>33.0</b>	18.6 / 47.5	-	<b>45.8 / 19.1</b>	<b>45.2 / 30.1</b>	18.8 / 44.5	-	<b>50.0 / 11.8</b>	43.3 / <b>29.2</b>	-
Otter (D.C.) [34]	-	1.1 / 14.3 / 0.4	1.2 / 14.2 / 0.4	0.5 / 5.4 / 0.1	-	0.5 / 13.8 / 0.1	1.0 / 16.8 / 0.2	0.3 / 2.3 / 0.1	45.0	1.0 / 15.0 / 0.3	1.1 / 12.8 / 0.2	-
	-	30.2 / 9.8	32.3 / 13.9	21.7 / 13.3	-	28.7 / 11.0	32.9 / 10.6	17.7 / 4.2	-	32.5 / 14.4	27.3 / 13.7	-
Otter (FunQA) [34]	-	<b>1.5 / 18.1 / 0.9</b>	<b>1.3 / 15.4 / 0.5</b>	0.8 / 5.9 / 0.5	-	<b>1.5 / 19.6 / 0.5</b>	<b>2.2 / 21.2 / 0.5</b>	0.3 / <b>4.3 / 0.3</b>	69.4	<b>2.6 / 23.8 / 1.6</b>	<b>3.4 / 20.3 / 2.6</b>	-
	-	38.4 / 22.2	42.6 / 31.0	24.5 / 51.3	-	40.0 / 21.9	41.1 / 21.1	<b>21.7 / 43.9</b>	-	44.7 / 18.4	<b>44.5 / 19.8</b>	-

BLEU-4 / ROUGE-L / CIDEr (shown in the first row) and BLEURT / GPT-4 (shown in the second row).

# **Video-ChatGPT: Qualitative Evaluation**

## **Video Understanding Task**

# Video-ChatGPT: Qualitative Evaluation

Video Understanding Task:



# Video-ChatGPT: Qualitative Evaluation

## Video Understanding Task:



Video Chatbot

Start Chatting

NOTE: Please make sure you press the 'Upload Video' button and wait for it to display 'Start Chatting' before submitting question to Video-ChatGPT.

Examples

Why is this video strange?

Can you write a short poem inspired from the video.

Where is this video taken? What place/landmark is shown in the video?

What is the main attraction seen by the people on the boat

Why is this video funny?

Upvote Downvote Flag Regenerate Clear history Submit

A screenshot of the Video-ChatGPT web application. At the top right is a logo featuring a cartoon goat and the text "Video-ChatGPT". Below the logo is a video player showing a child reading a book. A watermark for "clideo.com" is visible in the bottom right corner of the video frame. To the left of the video player is a "Start Chatting" button. Below the video player is a note: "NOTE: Please make sure you press the 'Upload Video' button and wait for it to display 'Start Chatting' before submitting question to Video-ChatGPT.". Underneath this note is a section titled "Examples" containing four questions with small thumbnail images: "Why is this video strange?", "Can you write a short poem inspired from the video.", "Where is this video taken? What place/landmark is shown in the video?", and "What is the main attraction seen by the people on the boat". At the bottom of the interface are buttons for "Upvote", "Downvote", "Flag", "Regenerate", and "Clear history", along with a large "Submit" button.

# Video-ChatGPT: Qualitative Evaluation

## Video Understanding Task:

**Video-LLaMA**



Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding

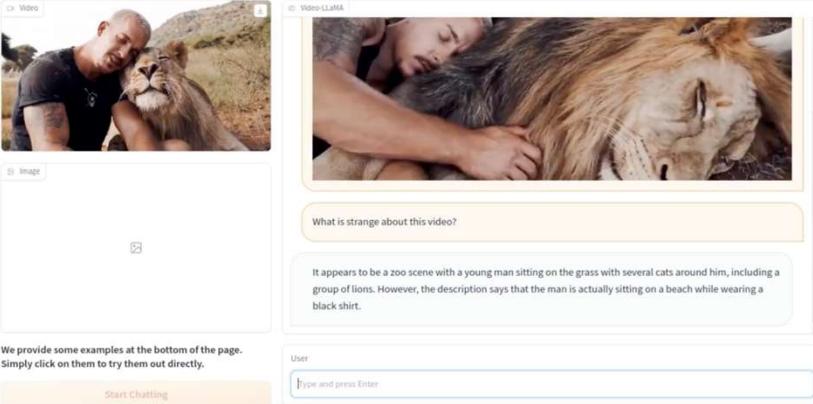
Introduction: Video-LLaMA is a multi-modal large language model that achieves video-grounded conversations between humans and computers by connecting language decoder with off-the-shelf unimodal pre-trained models.

[Github](#) [Code](#) [Hugging Face Spaces](#) [Hugging Face Model](#) [ModelScope Demo](#) [Paper PDF](#)

If you find Video-LLaMA interesting, please give us a star on GitHub.

Current online demo uses the 7B version of Video-LLaMA due to resource limitations. We have released the 13B version on our GitHub repository.

Video



What is strange about this video?

It appears to be a zoo scene with a young man sitting on the grass with several cats around him, including a group of lions. However, the description says that the man is actually sitting on a beach while wearing a black shirt.

We provide some examples at the bottom of the page.  
Simply click on them to try them out directly.

User: Type and press Enter.

Start Chatting

Qualitative comparison with Video LLaMA on **video reasoning task**

# Video-ChatGPT: Qualitative Evaluation

## Video Understanding Task:



Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding

Introduction: Video-LLaMA is a multi-modal large language model that achieves video-grounded conversations between humans and computers by connecting language decoder with off-the-shelf unimodal pre-trained models.

[GitHub](#) [Code](#) [Hugging Face Spaces](#) [Hugging Face Model](#) [ModelScope Demo](#) [Paper PDF](#)

Thank you for using the Video-LLaMA Demo Page! If you have any questions or feedback, feel free to contact us.

If you find Video-LLaMA interesting, please give us a star on GitHub.

Current online demo uses the v8 version of Video-LLaMA due to resource limitations. We have released the 13B version on our GitHub repository.

The screenshot shows the Video-LLaMA demo page. At the top, there's a logo of a llama wearing sunglasses and holding a drink. Below it, the text "Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding". Underneath, there's a brief introduction about the model being a multi-modal LLM that connects language decoder with unimodal pre-trained models. It includes links to GitHub, Code, Hugging Face Spaces, Hugging Face Model, ModelScope Demo, and Paper PDF. A note says "Thank you for using the Video-LLaMA Demo Page! If you have any questions or feedback, feel free to contact us." Another note says "If you find Video-LLaMA interesting, please give us a star on GitHub." A third note says "Current online demo uses the v8 version of Video-LLaMA due to resource limitations. We have released the 13B version on our GitHub repository." The main area has tabs for "Video" and "Image". Under "Video", there's a video player showing a man sitting with lions. Below the video player is a text input field with placeholder "What is strange about this video?". A scrollable box below contains a sample response: "It appears to be a zoo scene with a young man sitting on the grass with several cats around him, including a group of lions. However, the description says that the man is actually sitting on a beach while wearing a black shirt." At the bottom, there's a "Start Chatting" button and a user input field with placeholder "User: Type and press Enter". A note at the bottom says "We provide some examples at the bottom of the page. Simply click on them to try them out directly."



Video-ChatGPT

The screenshot shows the Video-ChatGPT demo page. At the top, there's a logo of a goat with a YouTube play button and the text "Video-ChatGPT". Below it, the text "Video-ChatGPT Chatbot". The main area has tabs for "Video" and "Image". Under "Video", there's a video player showing a man sitting with lions. Below the video player is a text input field with placeholder "Start Chatting". A note says "NOTE: Please make sure you press the 'Upload Video' button and wait for it to display 'Start Chatting' before submitting question to Video-ChatGPT." Below the video player, there's a section titled "Examples" with a sample response: "Why is this video strange?". Another section asks "Can you write a short poem inspired from the video.". A third section asks "Where is this video taken? What place/landmark is shown in the video?". At the bottom, there's a text input field with placeholder "What is strange about this video?" and a "Submit" button.

Qualitative comparison with Video LLaMA on **video reasoning task**

# Video-ChatGPT: Qualitative Evaluation

## Video Understanding Task:



The interface shows a video player window displaying a giraffe's head against a blue sky. The video duration is 0:08. Below the video is a "Start Chatting" button.

**NOTE:** Please make sure you **press the 'Upload Video' button** and wait for it to display 'Start Chatting' before submitting question to Video-ChatGPT.

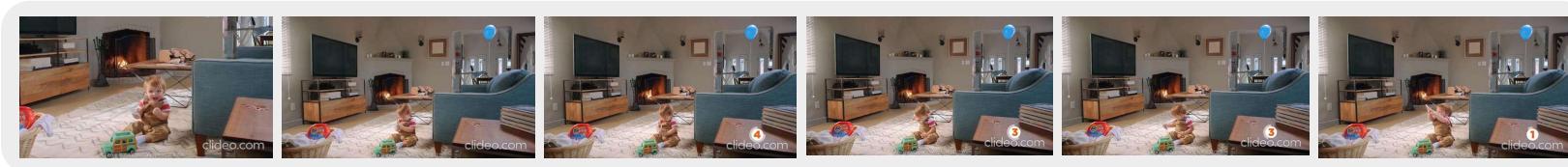
**Examples**

	Why is this video strange?
	Can you write a short poem inspired from the video.
	Where is this video taken? What place/landmark is shown in the video?

What is unusual about this video?  **Submit**

# Video-ChatGPT: Qualitative Evaluation

## Video Understanding Task:



Can you identify any safety hazards in this video?

Yes, there is a safety hazard in the video. A child is playing with a toy car on the floor, and there is a fireplace nearby. The child is sitting on the floor in front of the fireplace, which could potentially be a safety hazard if the child gets too close to the fireplace. It is important to keep children away from fireplaces and other heat sources to prevent burns or other injuries.



# **Video-ChatGPT: Qualitative Evaluation**

Video Understanding Task:



**How To Clean  
Snow off  
Your Car**

# Video-ChatGPT: Qualitative Evaluation

Video Understanding Task:

**How To Clean  
Snow off  
Your Car**



Describe the video in detail.

The video shows a man cleaning snow off his car using a snow brush. He starts by brushing the snow off the hood of his car, then moves on to the windshield and finally the side mirrors. He uses a snow brush to remove the snow from the car's surface.



# **Video-ChatGPT: Qualitative Evaluation**

## **Creative and Generative Tasks**

# Video-ChatGPT: Qualitative Evaluation

Creative and Generative Task:



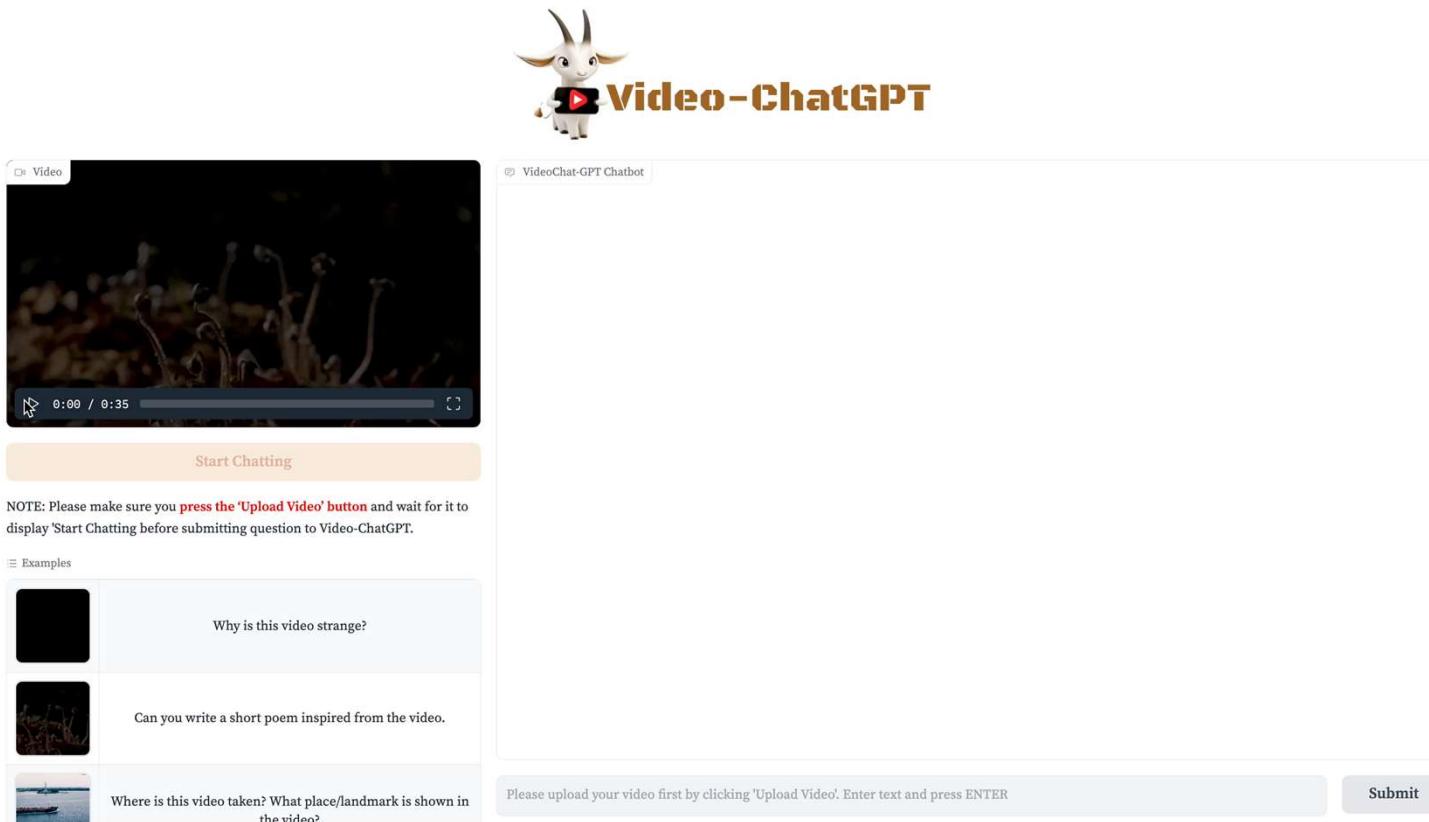
# Video-ChatGPT: Qualitative Evaluation

Creative and Generative Task:

The image shows a composite screenshot of the Video-ChatGPT application. On the left, there is a thumbnail of a video frame showing a woman's eyes and a product bottle labeled "skinauthority Sunscreen Moisturizer". Above this thumbnail is a larger image of the same woman's eyes. In the center, a video player window displays the same scene with a play button and a progress bar showing 0:00 / 0:27. Below the video player is a button labeled "Start Chatting". A note below the video player reads: "NOTE: Please make sure you press the 'Upload Video' button and wait for it to display 'Start Chatting' before submitting question to Video-ChatGPT." To the right of the video player is a "VideoChat-GPT Chatbot" section. It contains three examples of questions generated by the AI: 1) A black square thumbnail with the text "Why is this video strange?". 2) A thumbnail of a person in a red shirt with the text "Can you write a short poem inspired from the video.". 3) A thumbnail of a bridge over water with the text "Where is this video taken? What place/landmark is shown in the video?". At the bottom of the chatbot section is a message box containing the text "Please upload your video first by clicking 'Upload Video'. Enter text and press ENTER" and a "Submit" button.

# Video-ChatGPT: Qualitative Evaluation

Creative and Generative Task:



The screenshot shows the Video-ChatGPT interface. At the top center is a logo featuring a white goat holding a play button icon. To the left is a video player window showing a dark, abstract video thumbnail with a play button and a progress bar at 0:00 / 0:35. Below the video player is a large orange "Start Chatting" button. To the right is a white rectangular area labeled "VideoChat-GPT Chatbot". Inside this area, there is a note: "NOTE: Please make sure you press the 'Upload Video' button and wait for it to display 'Start Chatting' before submitting question to Video-ChatGPT." Below the note is a section titled "Examples" with three rows. Each row contains a small thumbnail image on the left and a question on the right. The first example shows a black square thumbnail and the question "Why is this video strange?". The second example shows a thumbnail of a plant and the question "Can you write a short poem inspired from the video.". The third example shows a thumbnail of a boat on water and the question "Where is this video taken? What place/landmark is shown in the video?". At the bottom of the white area, there is a light gray input field with the placeholder text "Please upload your video first by clicking 'Upload Video'. Enter text and press ENTER" and a "Submit" button on the right.

Video

0:00 / 0:35

Start Chatting

NOTE: Please make sure you press the 'Upload Video' button and wait for it to display 'Start Chatting' before submitting question to Video-ChatGPT.

Examples

Why is this video strange?

Can you write a short poem inspired from the video.

Where is this video taken? What place/landmark is shown in the video?

Please upload your video first by clicking 'Upload Video'. Enter text and press ENTER

Submit

# Video-ChatGPT: Qualitative Evaluation

## Creative and Generative Task:

The screenshot shows the Video-ChatGPT interface. At the top, there is a logo featuring a cartoon goat and the text "Video-ChatGPT". Below the logo, on the left, is a video player window displaying a scene from a movie with a woman in a blue dress. The video player has a play button, a timestamp (0:00 / 0:48), and a volume control. On the right, there is a large white input field labeled "VideoChat-GPT Chatbot". Below the input field, there is a note: "NOTE: Please make sure you press the 'Upload Video' button and wait for it to display 'Start Chatting' before submitting question to Video-ChatGPT." Underneath this note, there is a section titled "Examples" with three rows of questions and their corresponding video thumbnails.

Thumbnail	Question
	Why is this video strange?
	Can you write a short poem inspired from the video.
	Where is this video taken? What place/landmark is shown in the video?

Please upload your video first by clicking 'Upload Video'. Enter text and press ENTER

# **Video-ChatGPT: Qualitative Evaluation**

## **Spatial Understanding Task**

# Video-ChatGPT: Qualitative Evaluation

Spatial Understanding Task:



# Video-ChatGPT: Qualitative Evaluation

## Spatial Understanding Task:



Where is this video taken from?  
What place/landmark is shown in  
the video?

This video is taken in New York City,  
especially in the vicinity of the Statue  
of Liberty. The statue is shown in the  
background, and the video also shows the  
city skyline in the background.



# Video-ChatGPT: Qualitative Evaluation

## Spatial Understanding Task:



Where is this video taken from?  
What place/landmark is shown in  
the video?

This video is taken in New York City,  
especially in the vicinity of the Statue  
of Liberty. The statue is shown in the  
background, and the video also shows the  
city skyline in the background.



# Video-ChatGPT: Qualitative Evaluation

## Spatial Understanding Task:



Describe the video.

The video shows a person cutting up and peeling a red pepper, and then cutting it into slices. They then cut the slices into smaller pieces and place them on the baking sheet.



How many slices are on the baking sheet.

There are four slices of red pepper on the baking sheet.



# **Video-ChatGPT: Qualitative Evaluation**

**Temporal Understanding Task**

# Video-ChatGPT: Qualitative Evaluation

Temporal Understanding Task:



# Video-ChatGPT: Qualitative Evaluation

## Temporal Understanding Task:



Describe the video in detail.



The video shows a man standing on a snowy field, throwing a javelin. He throws the javelin and then runs towards it, picking it up and throwing it again. He then runs towards the javelin and picks it up, throwing it once more. The man then runs towards the javelin and picks it up, throwing it for the fourth time. The video ends with the man running towards the javelin and picking it up, throwing it for the fifth time.

# Video-ChatGPT: Qualitative Evaluation

Temporal Understanding Task:



# Video-ChatGPT: Qualitative Evaluation

## Temporal Understanding Task:



Step-by-step describe the cooking recipe shown in the video.

The video shows a person cooking a dish in a pan. They add ingredients such as noodles, and spices to the pan, and then stir the mixture. The person then pours a sauce into the pan and stirs it until it thickens. Finally, they add some more ingredients and stir the mixture again. The dish is then served on a plate.



# Video-ChatGPT: Qualitative Evaluation

## Temporal Understanding Task:



Briefly describe the video.

The video shows a woman who is grooming a horse with a brush. She is also seen talking to the camera and explaining how to groom a horse.



How many different horses are seen in the video?

There are two different horses seen in the video.



## **Summary: Video-ChatGPT**

- A multimodal model for video understanding and conversations.
- Leverages an adapter on top of pretrained LLM and vision backbones
- Fine-tuned on video-instruction data to capture temporal dynamics and spatial consistency relationships.
- A dataset of 100,000 video-instruction
- Quantitative video conversation evaluation framework for benchmarking.



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

Shehan Munasinghe<sup>1\*</sup>, Rusuru Thushara<sup>1\*</sup>, Muhammad Maaz<sup>1</sup>, Hanoona Abdul Rasheed<sup>1</sup>,  
Salman Khan<sup>1,2</sup>, Mubarak Shah<sup>3</sup>, Fahad Khan<sup>1,3</sup>

<sup>1</sup>Mohamed bin Zayed University of AI, <sup>2</sup>Australian National University  
<sup>3</sup>Linköping University, <sup>4</sup>University of Central Florida

Project: <https://github.com/mbzuai-oryx/Video-LLaVA>

### Abstract

Extending image-based Large Multimodal Models (LMM) to videos is challenging due to the inherent complexity of video data. The recent approaches extending image-based LMM to videos either lack the grounding capabilities (e.g., VideoChat, Video-CharGPT, Video-LLaMA) or do not utilize the audio-signals for better video understanding (e.g., Video-ChatGPT). Addressing these gaps, we propose PG-Video-LLaVA, the first LMM with pixel-level grounding capability, integrating audio cues by transcribing them into text to enrich video-context understanding. Our framework uses an off-the-shelf tracker and a novel grounding module, enabling it to spatially and temporally localize objects in videos following user instructions. We evaluate PG-Video-LLaVA using video-based generative and question-answering benchmarks and introduce new benchmarks specifically designed to measure prompt-based object grounding performance in videos. Further, we propose the use of Vicuna over GPT-3.5, as utilized in Video-ChatGPT, for video-based conversation benchmarking, ensuring reproducibility of results which is a concern with the proprietary nature of GPT-3.5. Our framework builds on SoTA image-based LLaVA model and extends its advantages to the video domain, delivering promising gains on video-based conversation and grounding tasks.

### 1. Introduction

Recent efforts on Large Multimodal Models (LMMs), spearheaded by GPT-4V [25], allow detailed conversations about images but generally do not scale well to videos. The magnitude of video data scales far beyond other modalities due to its massive volume on social and internet media. Furthermore, extending LMMs to videos is challenging due to their complex dynamics with long temporal context that needs to be understood accurately. Although recent

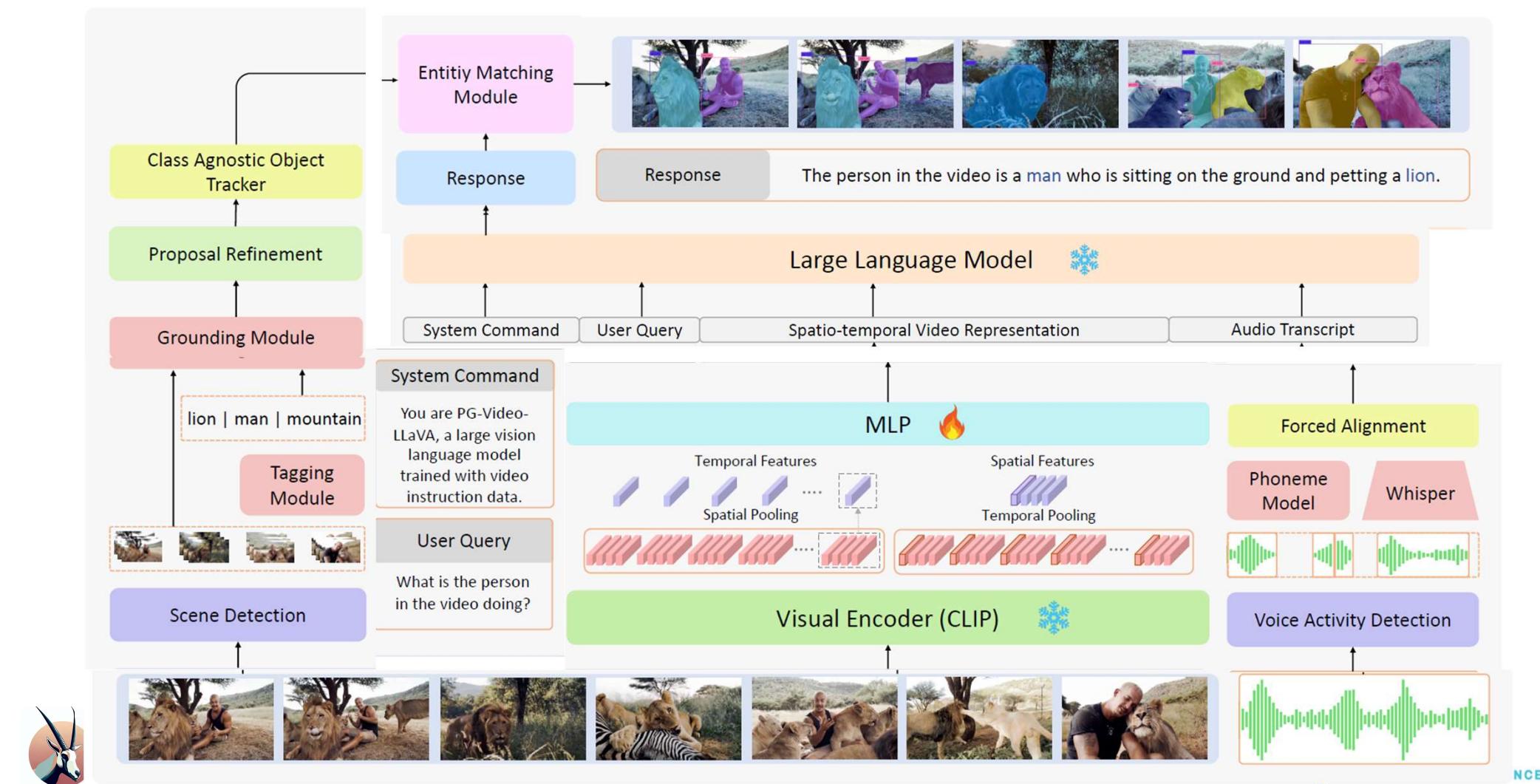
\*Equal Contribution



Figure 1. Spatio-temporal video grounding on example videos from Vid-STVG [48] (above) and HC-STVG [34] (below) datasets. PG-Video-LLaVA can generate textual responses with referred objects grounded in the video content (*tennis racket* and *man* are localized in the top and bottom examples, respectively).

approaches towards video-LMMs such as VideoChat [15], Video-LLaMA [45], and Video-ChatGPT [22] have demonstrated capabilities in video comprehension and dialogue, they lack the crucial feature of visual grounding. Visual grounding in videos aims to associate the LMM responses to specific objects within the video input. Addressing this gap, we introduce PG-Video-LLaVA, the first video-LMM capable of localizing objects appearing in LMM responses. This task leads to enhanced intractability and demonstrates deep understanding of video content.

In PG-Video-LLaVA, we address the unique challenges posed by video data. The model is designed to track objects within shorter video clips that maintain consistent camera views, enabling accurate visual grounding across scenes and motions. This tracking links spatio-temporal segments directly to conversational elements, enhancing the model's



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

Model	MSVD-QA [39]		MSRVT-TT-QA [40]		TGIF-QA [16]		Activity Net-QA [44]	
	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM [41]	32.2	–	16.8	–	41.0	–	24.7	–
LLaMA Adapter [9]	53.7	3.3	45.6	3.2	54.3	3.3	37.3	3.2
Video LLaMA [45]	48.6	3.2	32.8	2.8	51.4	3.4	27.1	2.9
Video-ChatGPT [22]	62.6	3.6	50.0	3.3	66.5	3.7	<b>40.8</b>	<b>3.3</b>
PG-Video-LLaVA	<b>64.1</b>	<b>3.7</b>	<b>51.6</b>	<b>3.3</b>	<b>66.8</b>	<b>3.8</b>	39.9	<b>3.3</b>



ORYX - Video-ChatGPT





Describe this Video.

Video-ChatGPT

The video shows a small child sitting on the floor in front of a fireplace, playing with a toy. The child is holding a toy and appears to be playing with it while sitting on the floor.

PG-Video-LLaVA (7B)

The video shows a young child playing with a toy in a living room. The child is sitting on the floor and playing with a toy car, which is placed on a rug. The room is furnished with a couch, a chair, and a TV. The child appears to be enjoying their playtime, and the scene captures a moment of innocent childhood fun.

PG-Video-LLaVA (13B)

The video shows a young child sitting on the floor in a living room, playing with a toy. The room is furnished with a couch, a chair, and a television. The child is holding a toy in his hand while playing with it. The room appears to be well-lit, with a fireplace in the background. The child seems to be enjoying himself, as he is laughing and smiling while playing with the toy. Overall, the video captures a joyful moment of a child's life, as he engages in playful activities in a comfortable and cozy living room setting.



ORYX - Video-ChatGPT



MOHAMED BIN ZAYED  
UNIVERSITY OF  
ARTIFICIAL INTELLIGENCE



Describe this Video.

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.



ORYX - Video-ChatGPT





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.



ORYX - Video-ChatGPT



MOHAMED BIN ZAYED  
UNIVERSITY OF  
ARTIFICIAL INTELLIGENCE



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.



ORYX - Video-ChatGPT





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.

With Audio

Based on the video, the title could be "The Future of AI: A Discussion on GPT-4 and Chat GPT."

**Audio Transcript:** ... just how amazing Chat GPT turn out to be and then the improvement in GPT for. So GPT 4 has not surprised me at all in terms of reception there. Chat GPT surprised us a little bit, but I still ....  
... You know, they're like, oh, it's better than 3.5, but I thought it was going to be better than 3.5. ...



ORYX - Video-ChatGPT



## Summary

- We introduce Video-ChatGPT, a multimodal model that merges a pretrained visual encoder with a large language model (LLM) to enable video understanding and conversations based on videos.
- Video-ChatGPT leverages an adapter on top of pretrained LLM and vision backbones and is fine-tuned on video-instruction data to capture temporal dynamics and spatial consistency relationships.
- A dataset of 100,000 video-instruction pairs is created to enhance Video-ChatGPT's video-specific understanding and conversation capabilities.
- We introduce quantitative video conversation evaluation framework for benchmarking, evaluating models on a diverse set of capabilities including conventional video question answering as well as open-ended descriptions.

**Thank You!**