UCF Center for Research in Computer Vision

Visual-Language Models Introduction
Part-IV: Video ChatGPT, PG-Video LLaVA
Lecture-7
CAP6412 Spring 2025

Mubarak Shah

shah@crcv.ucf.edu

Contents

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

UCF Center for Research in Computer Vision

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

Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Khan

Abstract

Conversation agents fueled by Large Language Models (LLMs) are providing a new way to interact with visual data. While there have been initial attempts for image-based conversation models, this work addresses the under-explored field of *video-based conversation* by introducing Video-ChatGPT. It is a multimodal model that merges a video-adapted visual encoder with an LLM. The resulting model is capable of understanding and generating detailed conversations about videos. We introduce a new dataset of 100,000 video-instruction pairs used to train Video-ChatGPT acquired via manual and semi-automated pipeline that is easily scalable and robust to label noise. We also develop a quantitative evaluation framework for video-based dialogue models to objectively analyze the strengths and weaknesses of video-based dialogue models. Code: https://github.com/mbzuai-oryx/Video-ChatGPT.

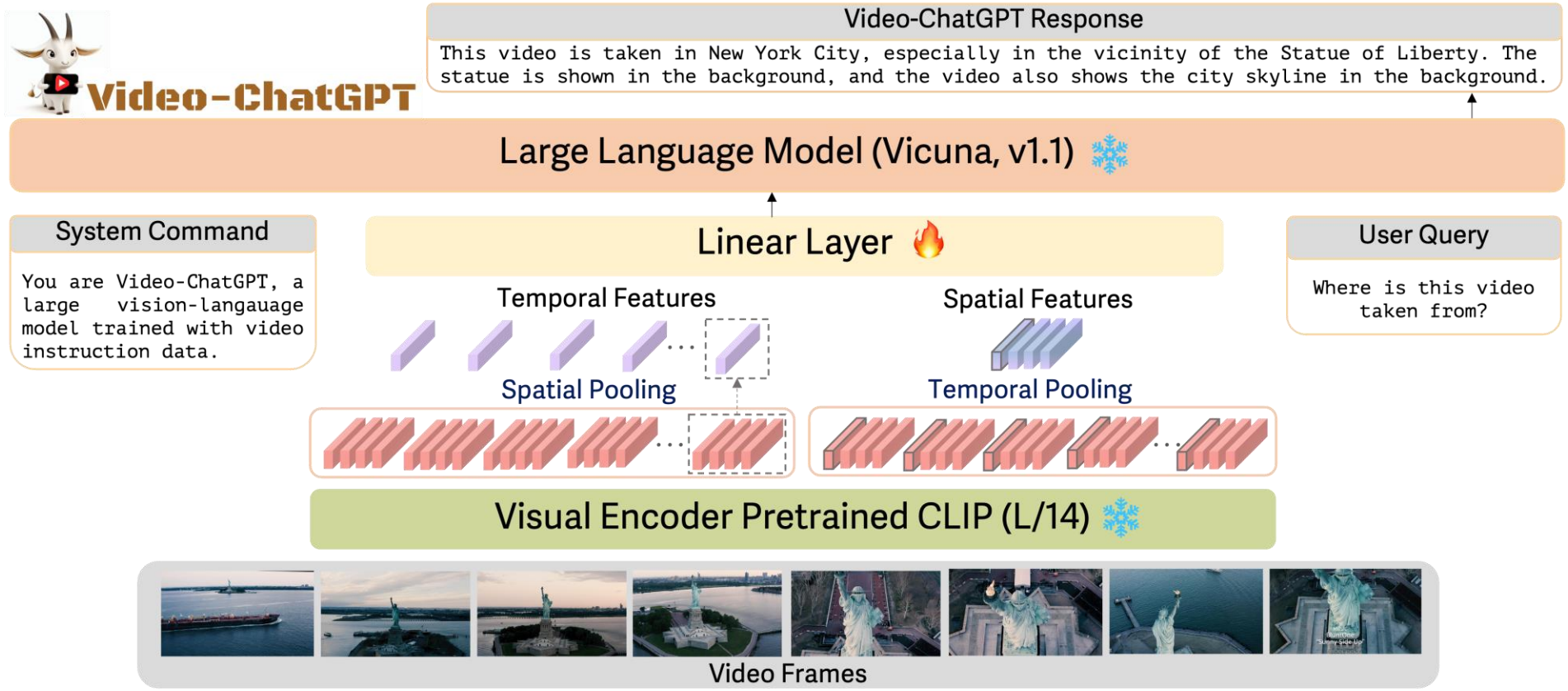
PDF
Cite
Search
Fix data

Anthology ID: 2024.acl-long.679
Volume: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
Month: August
Year: 2024
Address: Bangkok, Thailand

Video ChatGPT

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

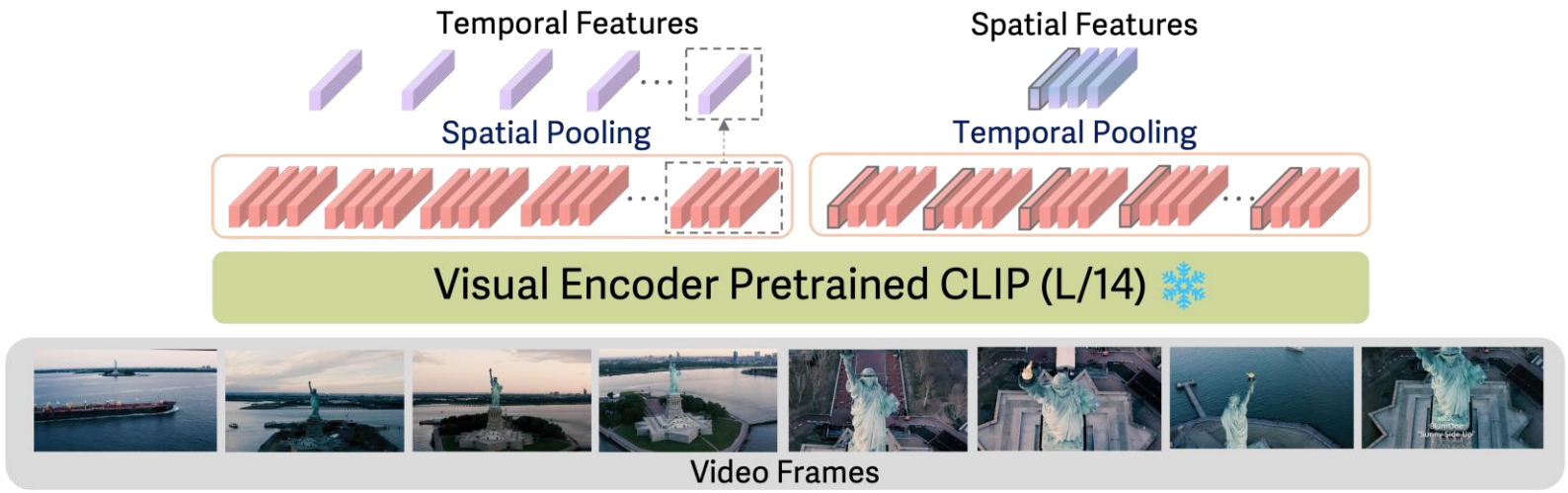
Video-ChatGPT: Architecture



Video ChatGPT, PG-Video LLaVA

Video-ChatGPT: Architecture

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



Video ChatGPT, PG-Video LLaVA

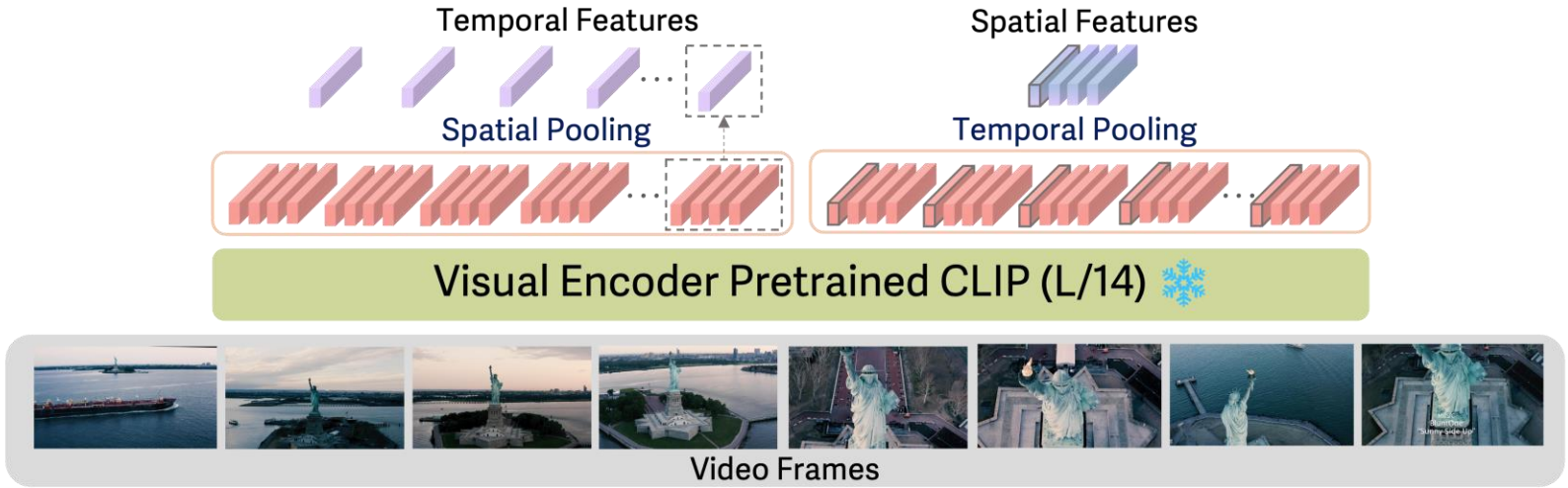
Video-ChatGPT: Architecture

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

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

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

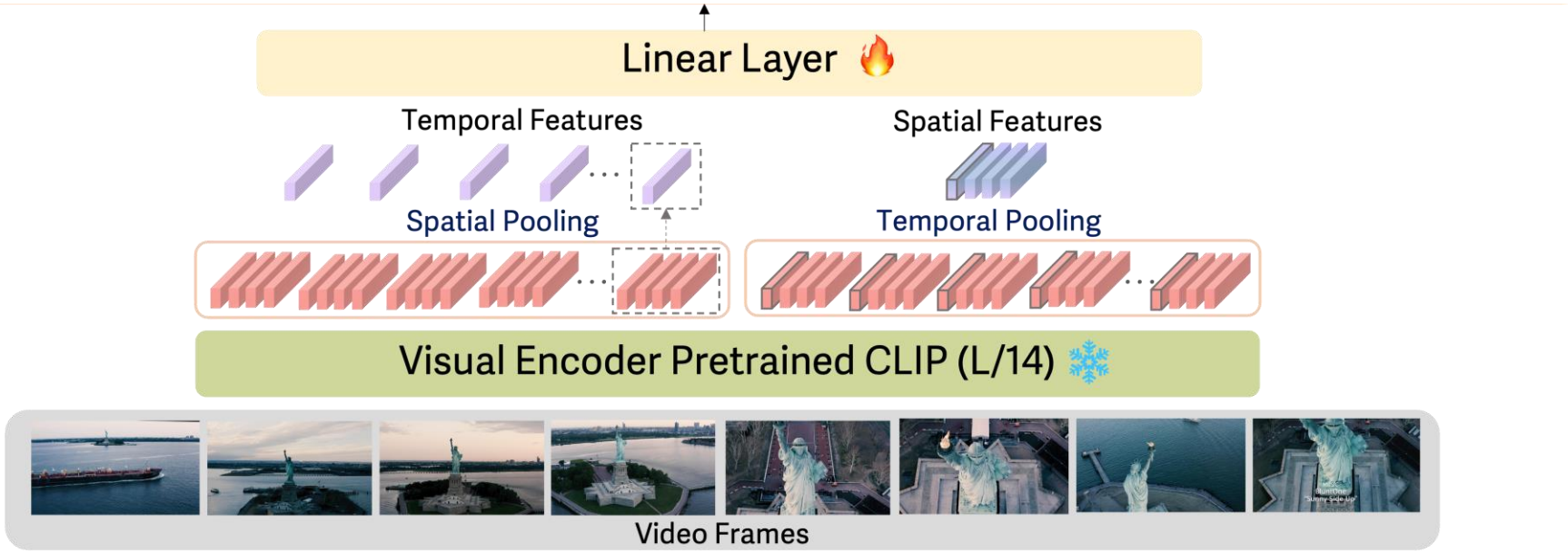
$v_i = [t_i \quad z_i] \in \mathbb{R}^{(T+N) \times D}$



Video ChatGPT, PG-Video LLaVA

Video-ChatGPT: Architecture

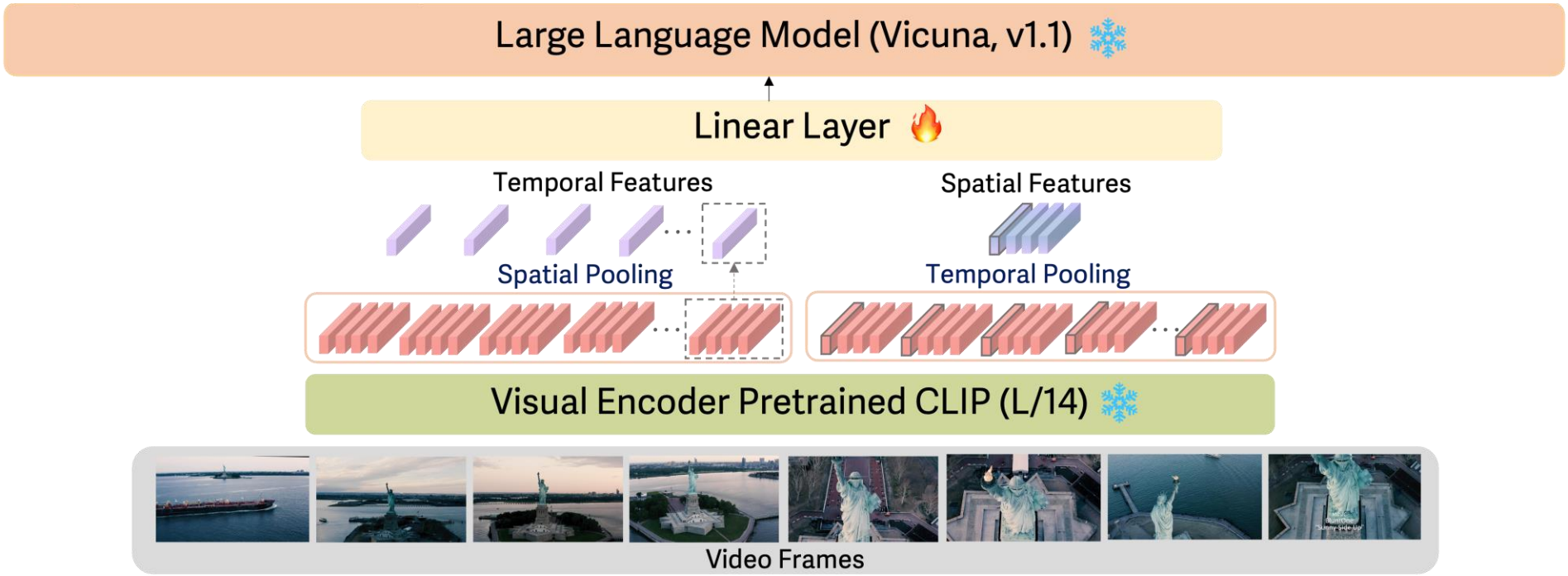
- Linear layer projects spatiotemporal video features to LLM input space



Video ChatGPT, PG-Video LLaVA

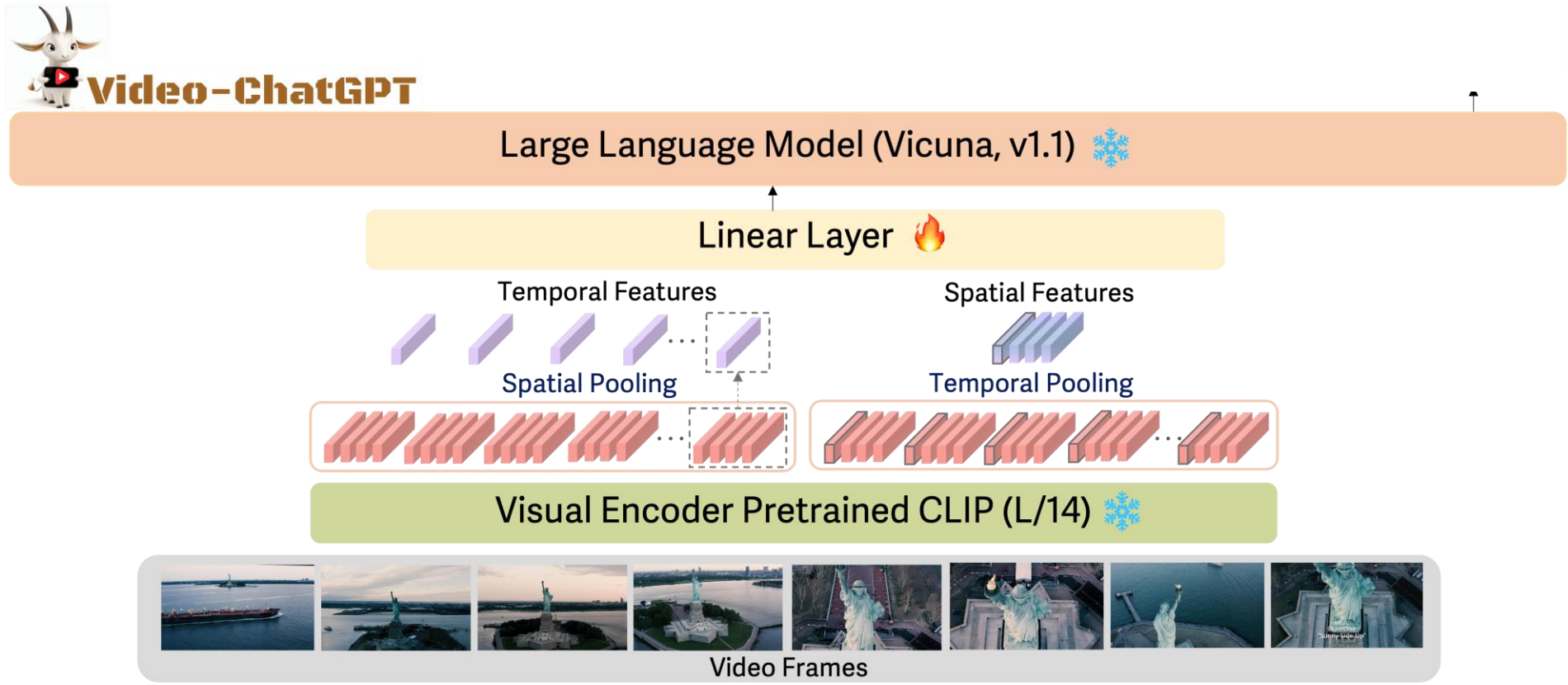
Video-ChatGPT: Architecture

- Linear layer projects spatiotemporal video features to LLM input space



Video ChatGPT, PG-Video LLaVA

Video-ChatGPT: Architecture



Video ChatGPT, PG-Video LLaVA

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:

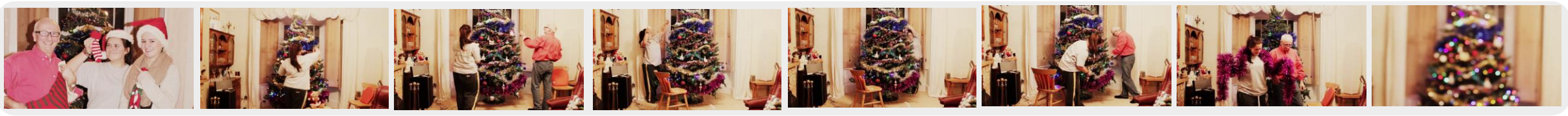
A close up of a christmas tree is shown followed by close ups of ornaments. Two people are then seen moving around the tree decorating as well as turning the lights off. They finish decorating the tree and playing with one another and laughing. In the end close ups of the trees are shown as well as a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

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, PG-Video LLaVA

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, PG-Video LLaVA

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, PG-Video LLaVA

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

Video ChatGPT, PG-Video LLaVA

[{"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, PG-Video LLaVA

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, PG-Video LLaVA

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.

Video ChatGPT, PG-Video LLaVA

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	**2.40**
Detail Orientation	2.50	2.32	2.18	**2.52**
Contextual Understanding	2.53	2.30	2.16	**2.62**
Temporal Understanding	1.94	**1.98**	1.82	**1.98**
Consistency	2.24	2.15	1.79	**2.37**

Video ChatGPT, PG-Video LLaVA

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	**64.9**	**3.3**	**49.3**	**2.8**	**51.4**	**3.0**	**35.2**	**2.8**

Video ChatGPT, PG-Video LLaVA

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

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

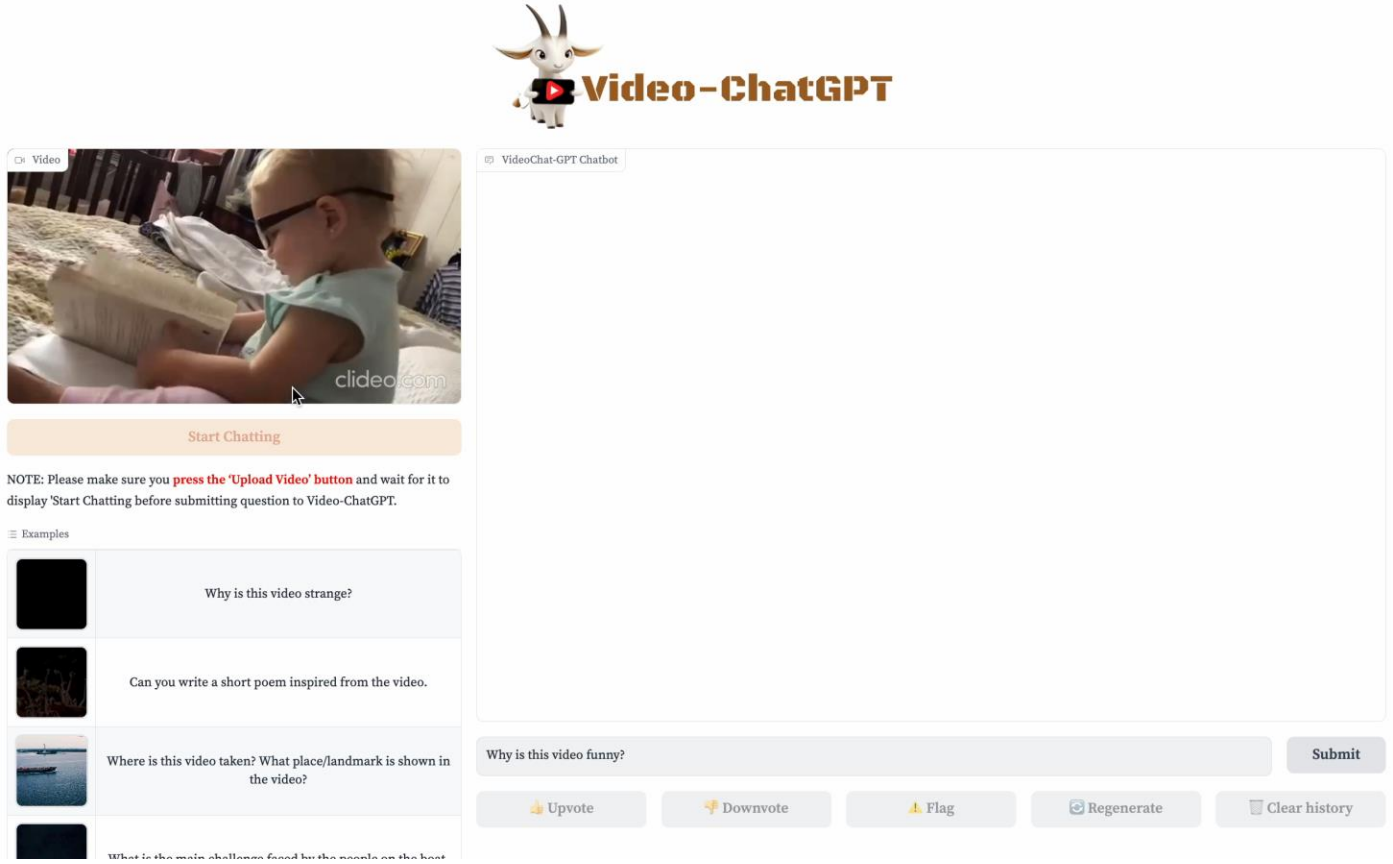
Video Understanding Task:



Video-ChatGPT: Qualitative Evaluation

Video Understanding Task:

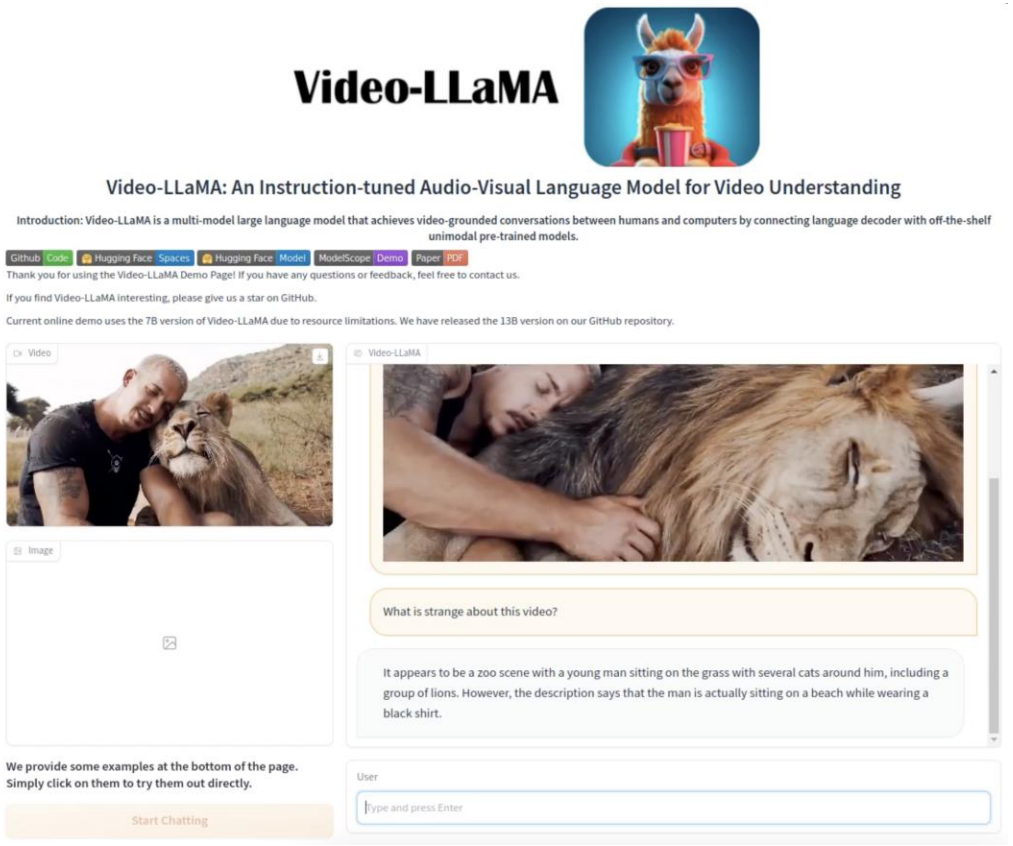




Video ChatGPT, PG-Video LLaVA

Video-ChatGPT: Qualitative Evaluation

Video Understanding Task:

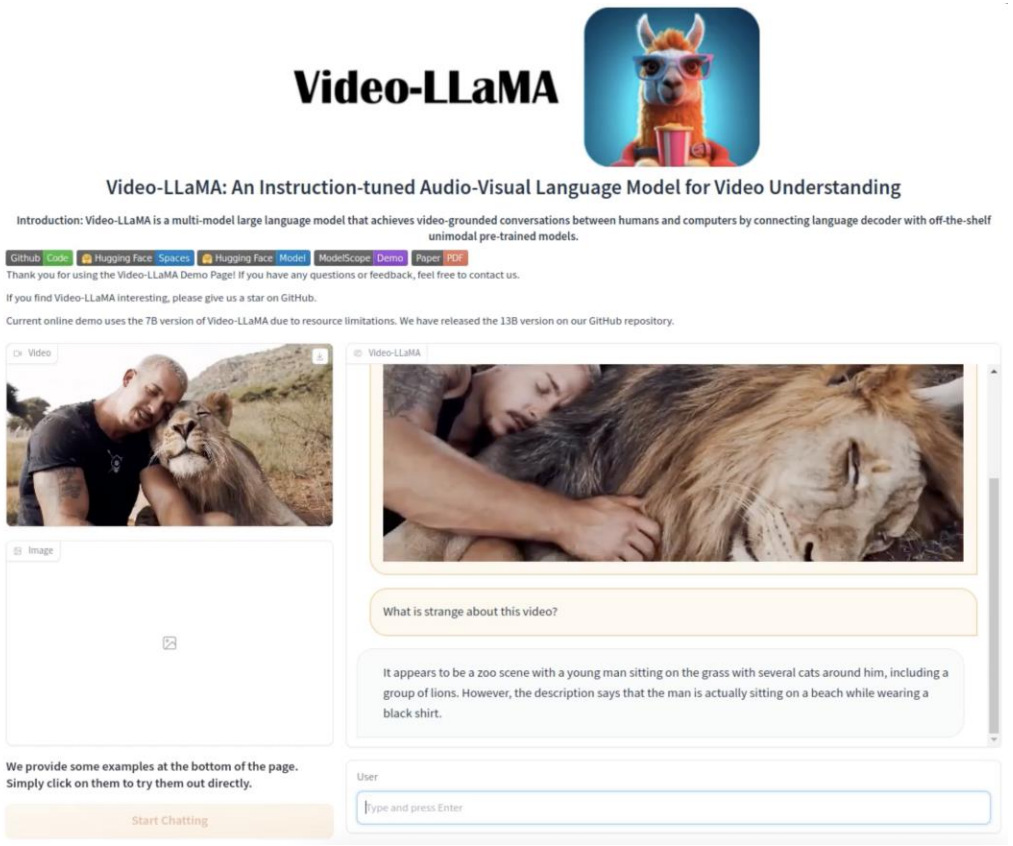


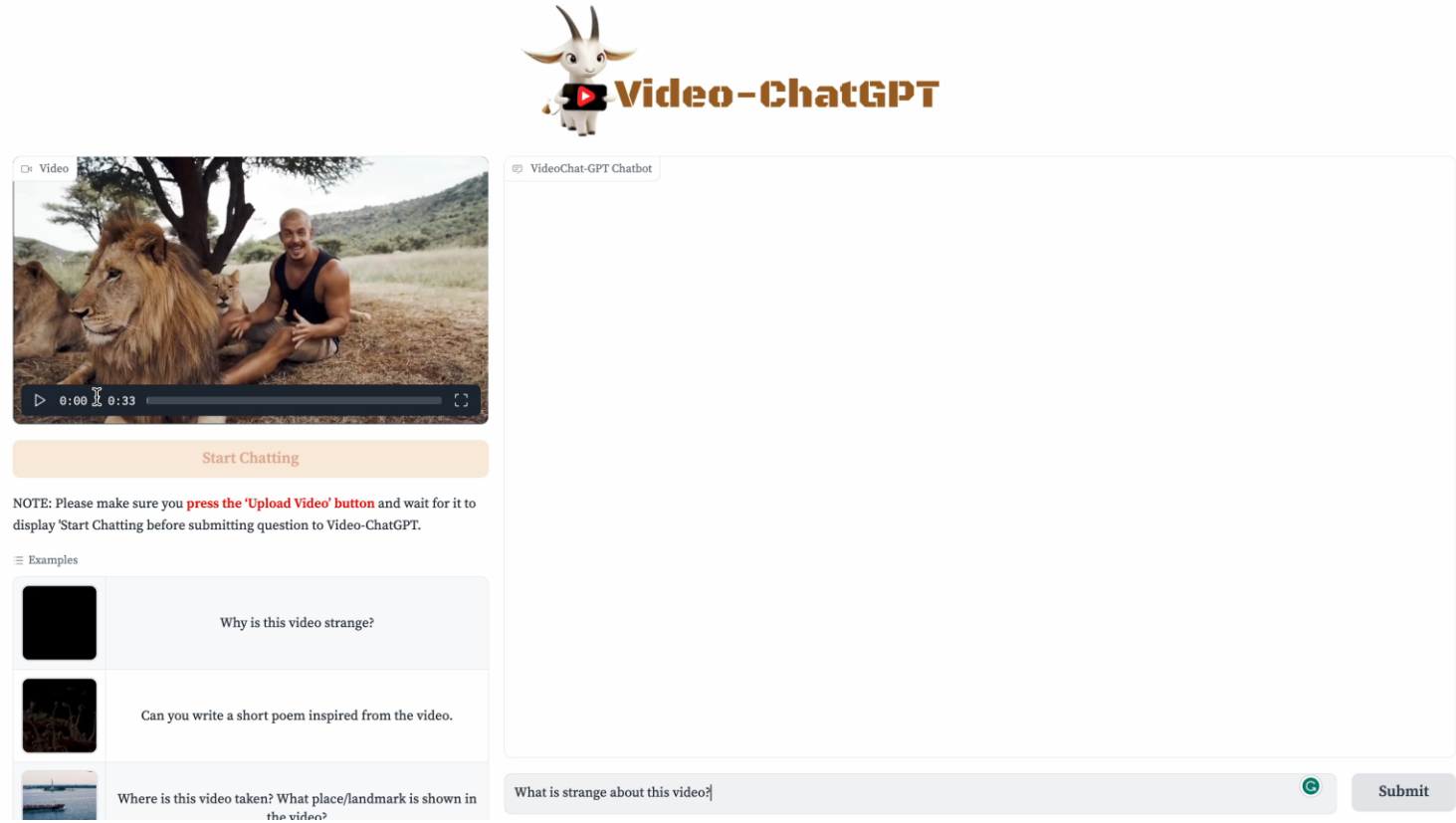
Qualitative comparison with Video LLaMA on video reasoning task

Video ChatGPT, PG-Video LLaVA

Video-ChatGPT: Qualitative Evaluation

Video Understanding Task:



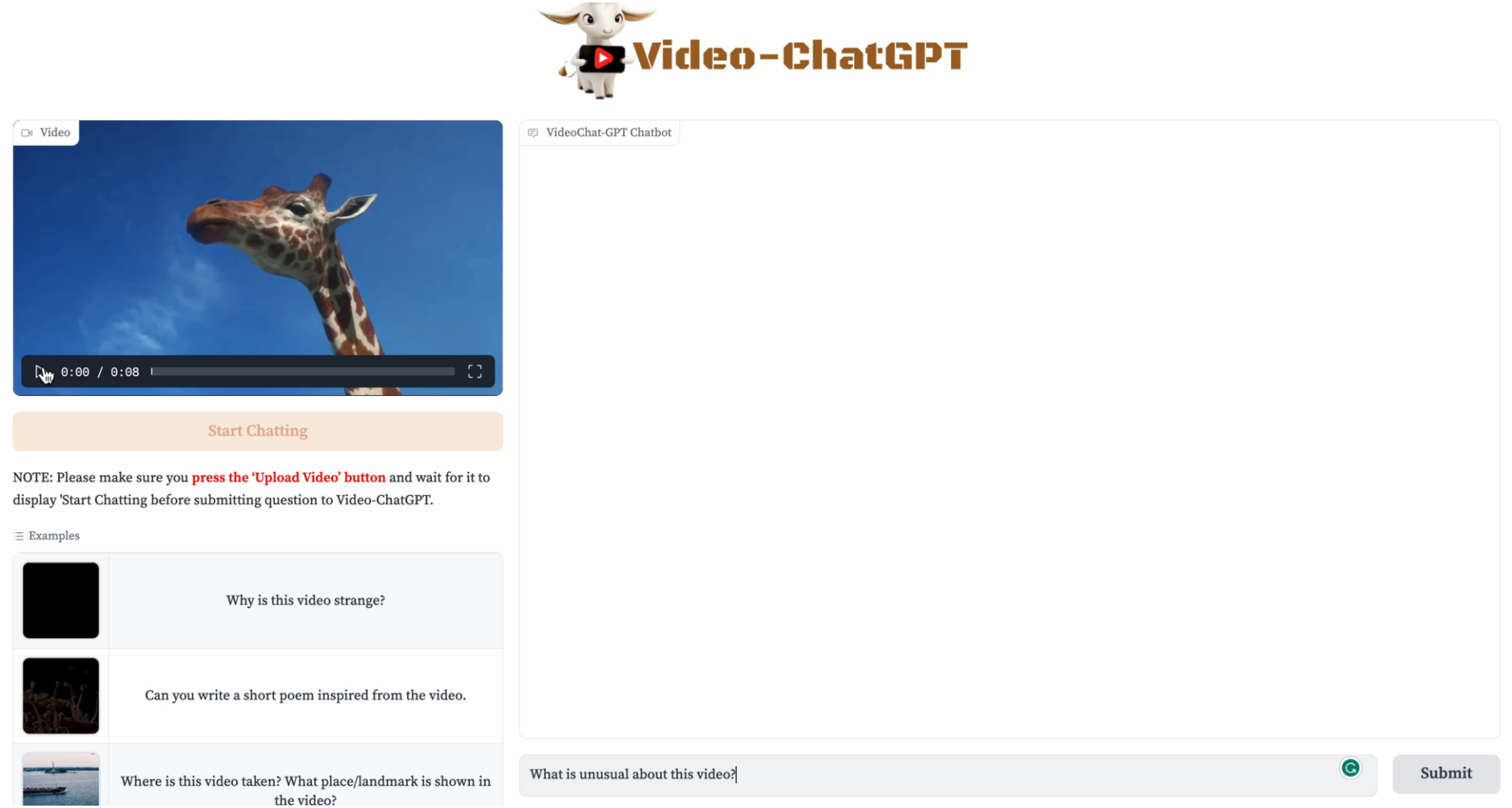


Qualitative comparison with Video LLaMA on video reasoning task

Video ChatGPT, PG-Video LLaVA

Video-ChatGPT: Qualitative Evaluation

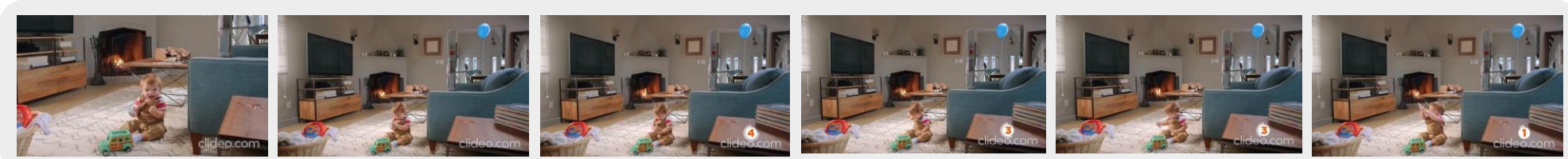
Video Understanding Task:



Video ChatGPT, PG-Video LLaVA

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, PG-Video LLaVA

Video-ChatGPT: Qualitative Evaluation

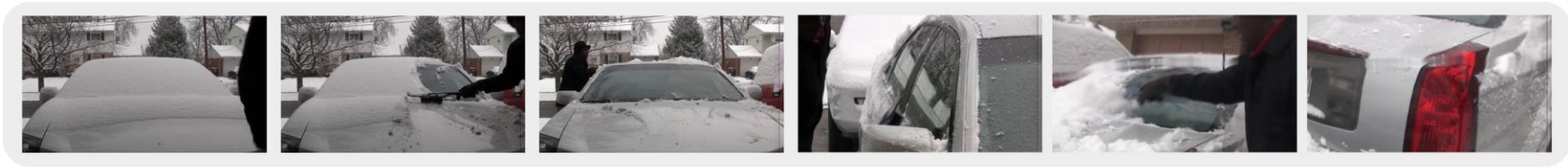
Video Understanding Task:



Video-ChatGPT: Qualitative Evaluation

Video Understanding Task:





Describe the video in detail.

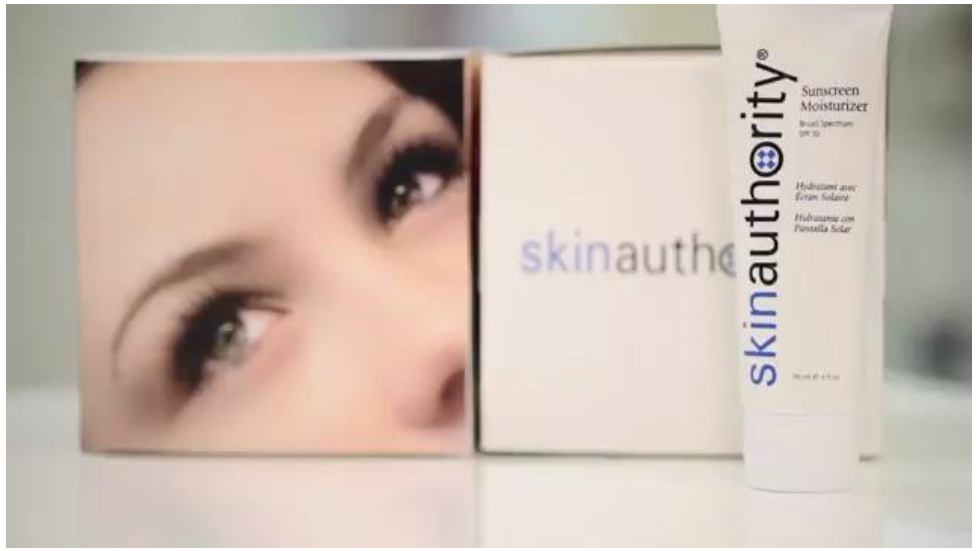
The video shows a man cleaning snow off his car using a snow brush. He starts by brushing the snow off the hood of his car, then moves on to the windshield and finally the side mirrors. He uses a snow brush to remove the snow from the car's surface.

Video-ChatGPT: Qualitative Evaluation

Creative and Generative Tasks

Video-ChatGPT: Qualitative Evaluation

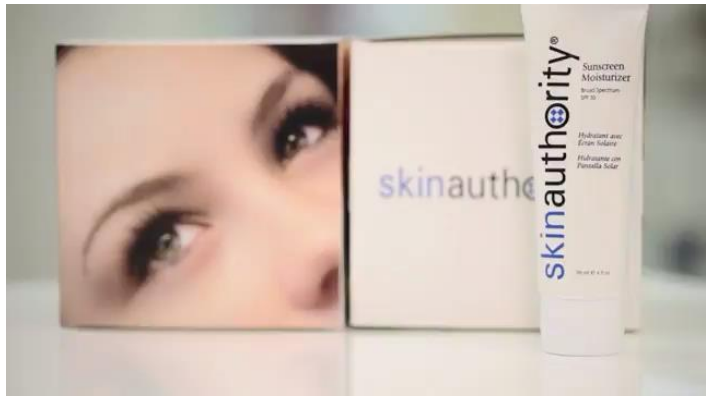
Creative and Generative Task:



Video ChatGPT, PG-Video LLaVA

Video-ChatGPT: Qualitative Evaluation

Creative and Generative Task:

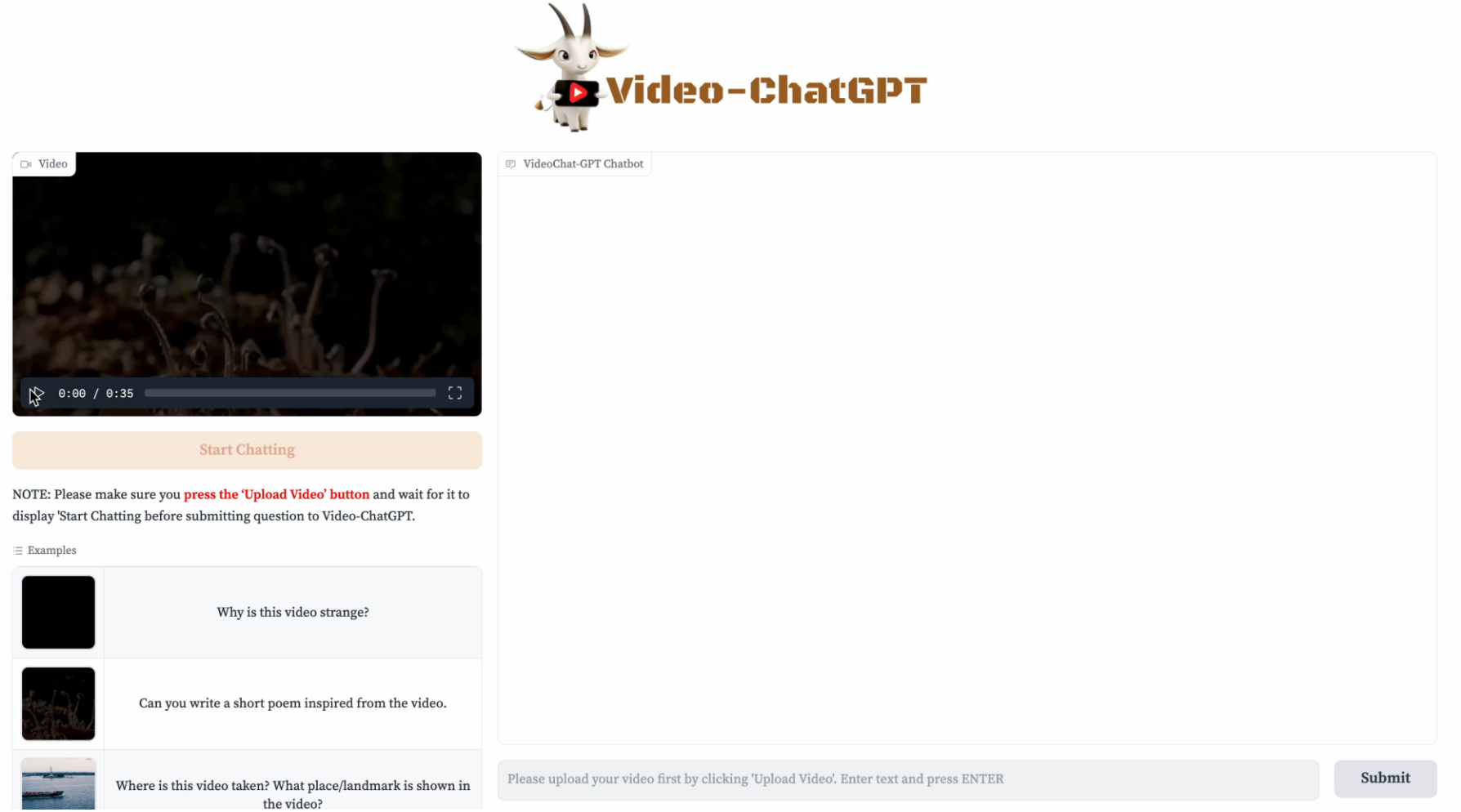




Video ChatGPT, PG-Video LLaVA

Video-ChatGPT: Qualitative Evaluation

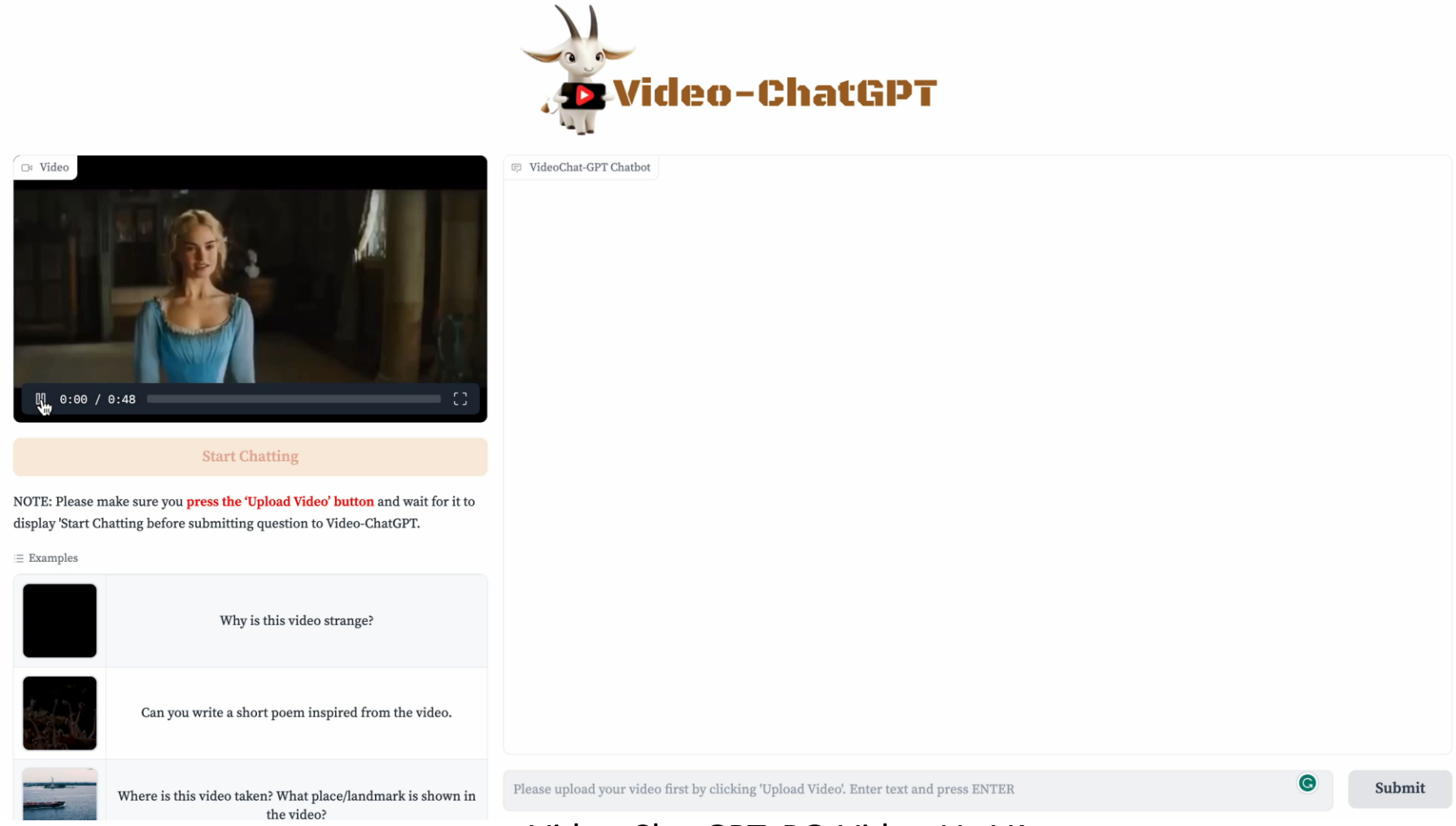
Creative and Generative Task:



Video ChatGPT, PG-Video LLaVA

Video-ChatGPT: Qualitative Evaluation

Creative and Generative Task:



Video ChatGPT, PG-Video LLaVA

Video-ChatGPT: Qualitative Evaluation

Spatial Understanding Task

Video-ChatGPT: Qualitative Evaluation

Spatial Understanding Task:



Video ChatGPT, PG-Video LLaVA

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, PG-Video LLaVA

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, PG-Video LLaVA

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, PG-Video LLaVA

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, growing 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, PG-Video LLaVA

Video-ChatGPT: Qualitative Evaluation

Temporal Understanding Task:



Video-ChatGPT: Qualitative Evaluation

Temporal Understanding Task:





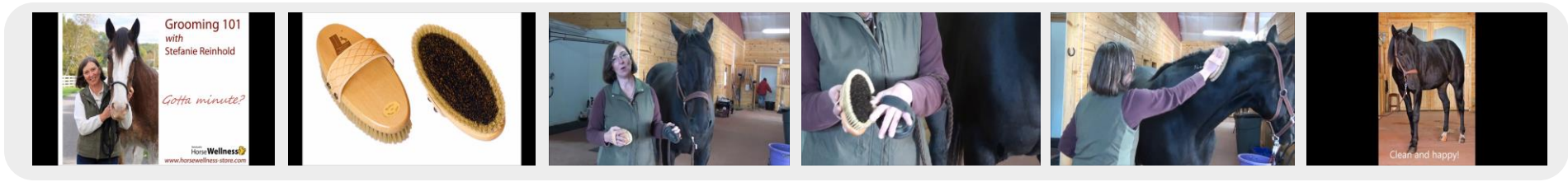
Step-by-step describe the cooking recipe shown in the video.

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, PG-Video LLaVA

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.



Video ChatGPT, PG-Video LLaVA

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.

arXiv:2311.13435v1 [cs.CV] 22 Nov 2023

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

Shehan Munasinghe[1*], Rusiru Thushara[1*], Muhammad Maaz[1], Hanoona Abdul Rasheed[1],
Salman Khan[1,2], Mubarak Shah[4], Fahad Khan[1,3]
[1]Mohamed bin Zayed University of AI, [2]Australian National University
[3]Linköping University, [4]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-ChatGPT, 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 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

*Equal Contribution

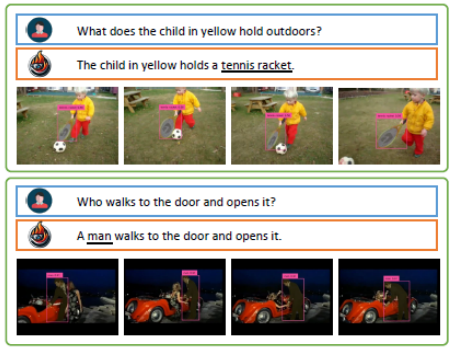


Figure 1. **Spatio-temporal video grounding** on example videos from Vid-STG [48] (above) and HC-STVG [34] (below) datasets. PG-Video-LLaVA can generate textual responses with referred objects grounded in the video content (*tennis racket* and *man* are localized in the top and bottom examples, respectively).

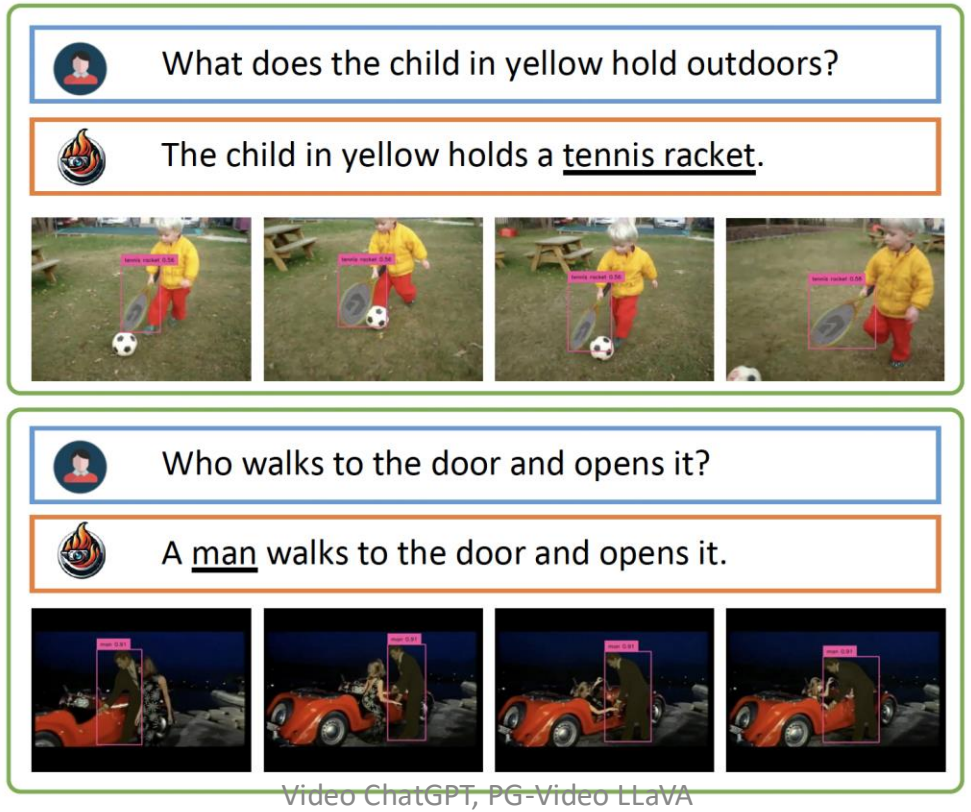
PG-Video-LLaVA: Pixel Grounding Large Video-Language Models

ArXiv Preprint, December 2023 (2 Citations)

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

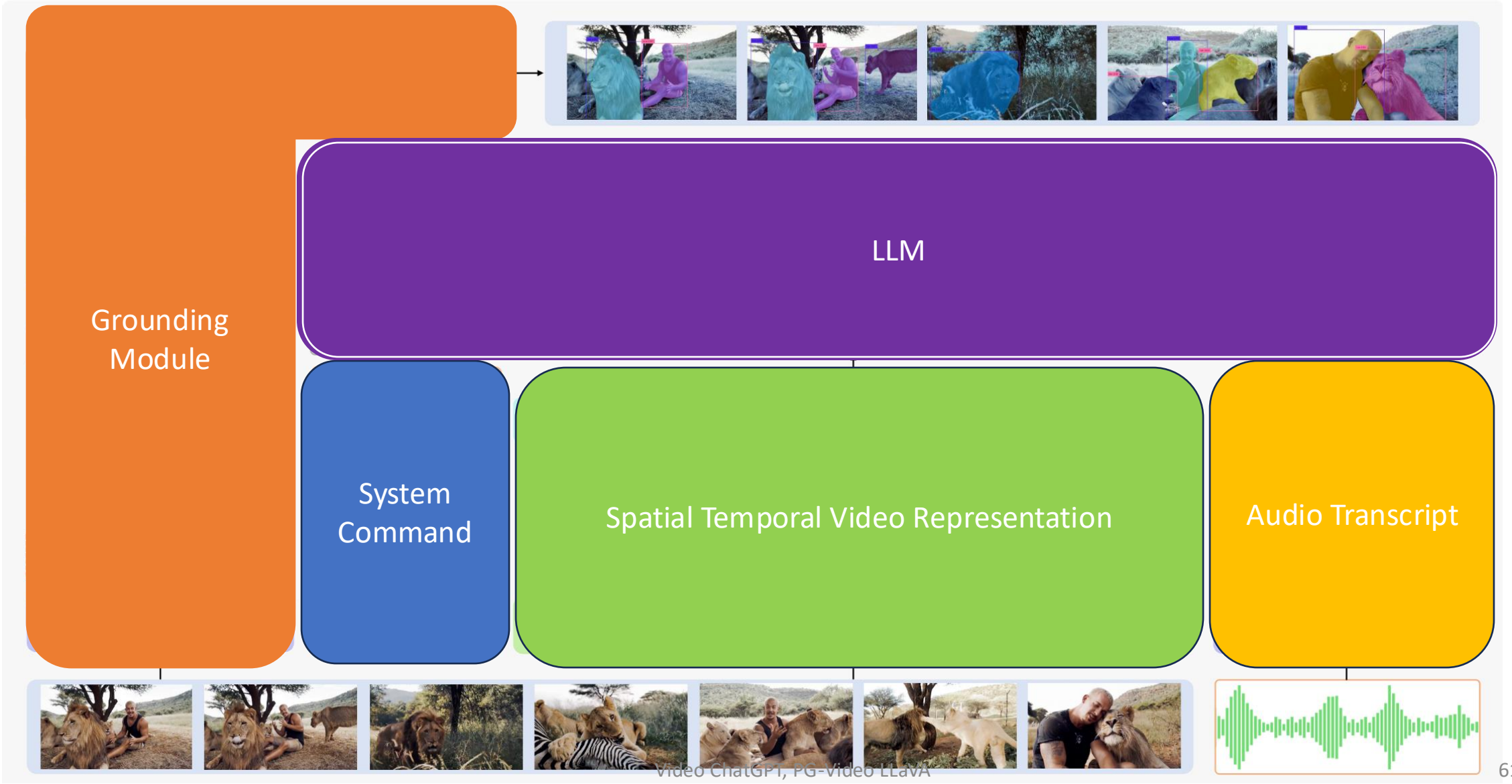
Presented by Group 3:

- Tyler VanderMate
- Ashton Frias
- Nicholas Gray
- Wen-Kai Chen
- Abhinav Kotta



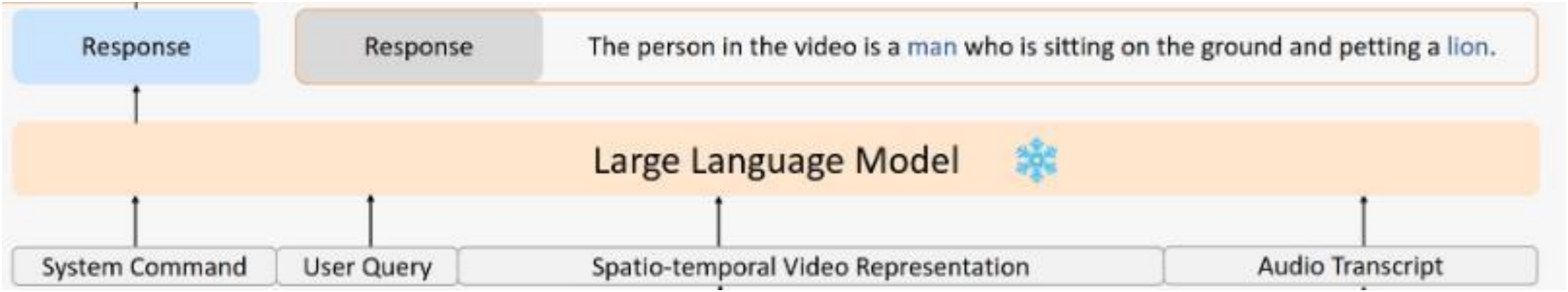
PG-Video-LLaVA Architecture

PG-Video-LLaVA

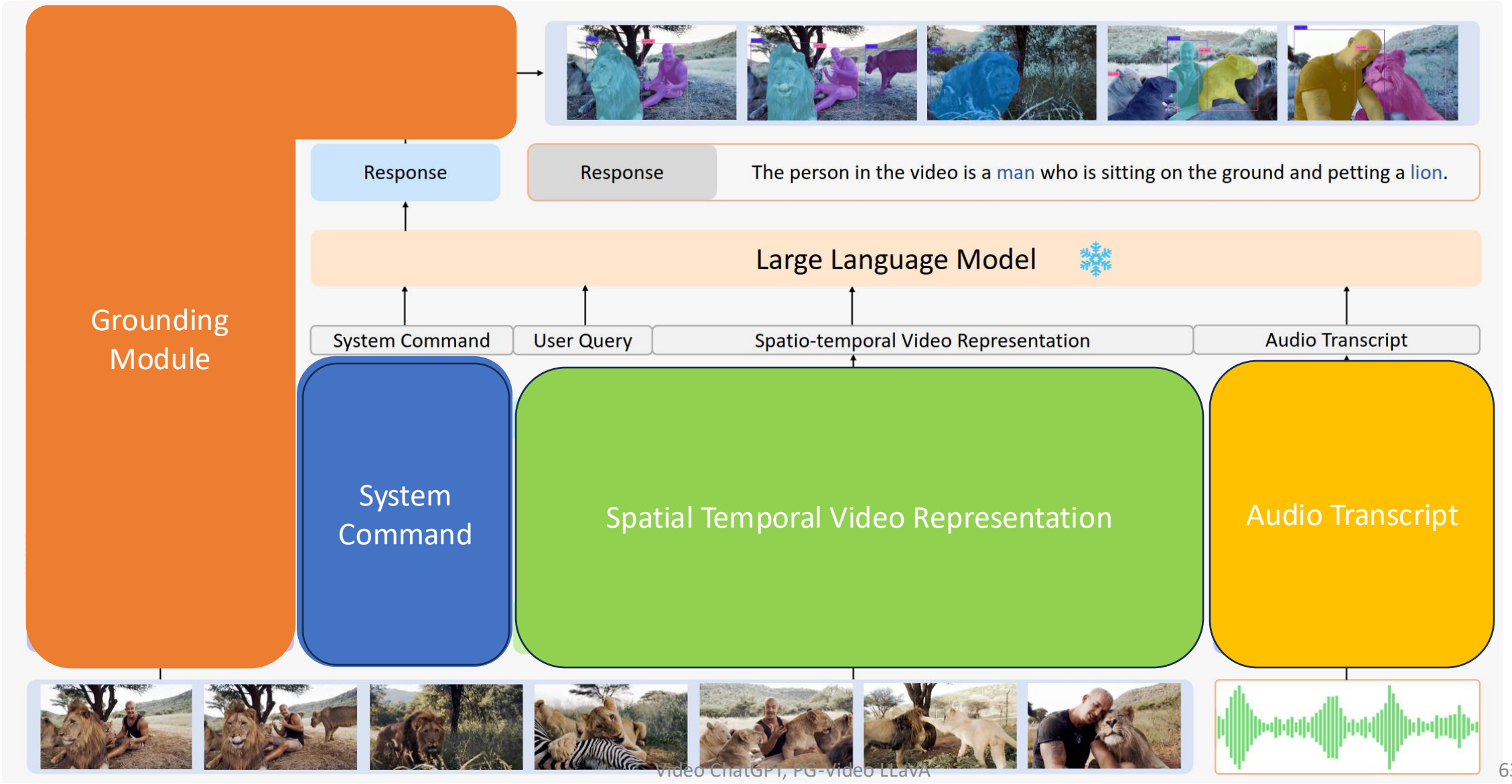


Frozen LLM (Vicuna-13b-v1.5)

- 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

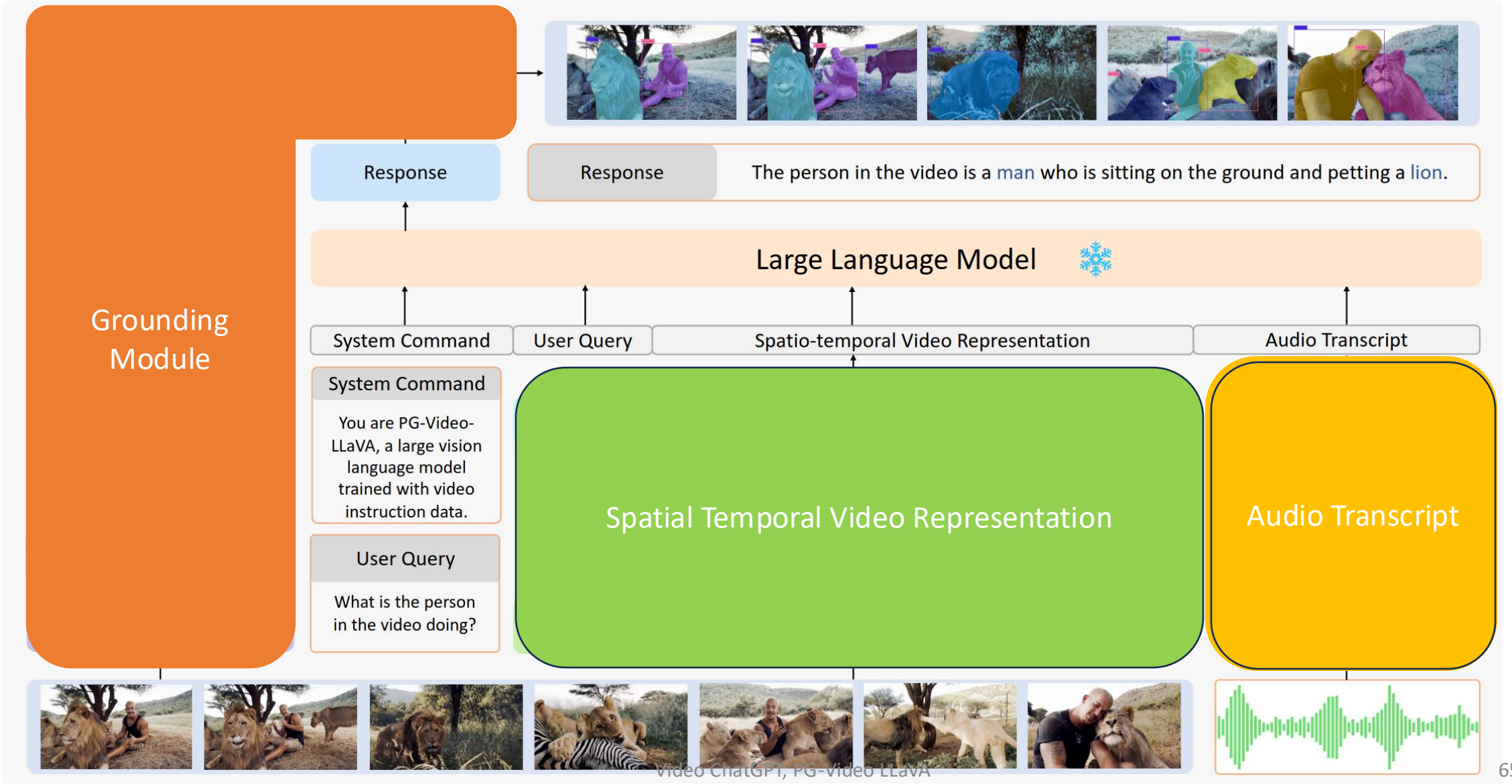


System Commands & User Query

- "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."

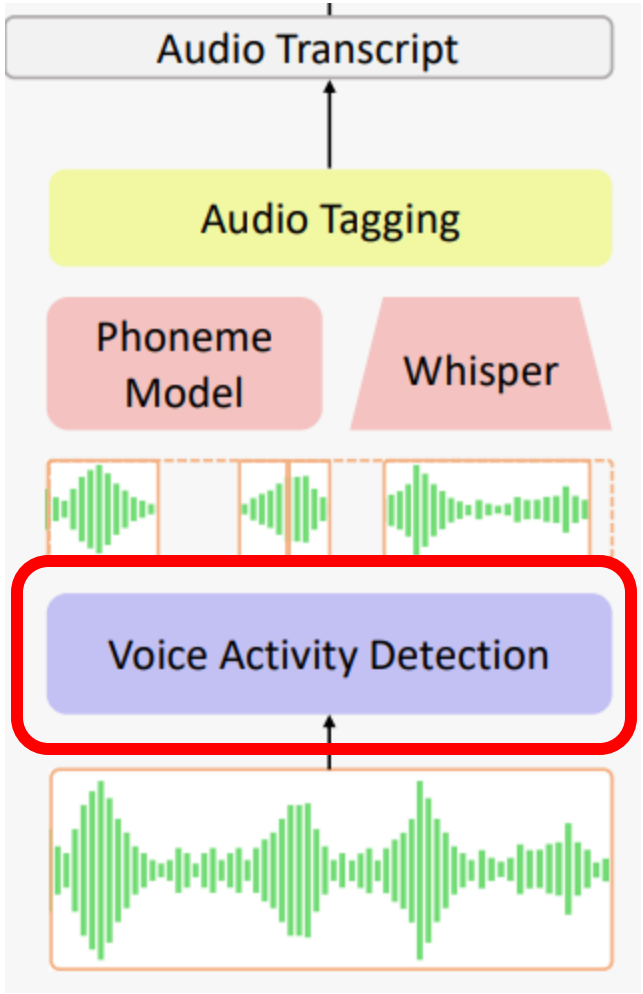
- User Query
  - What is the person doing in the video?

PG-Video-LLaVA



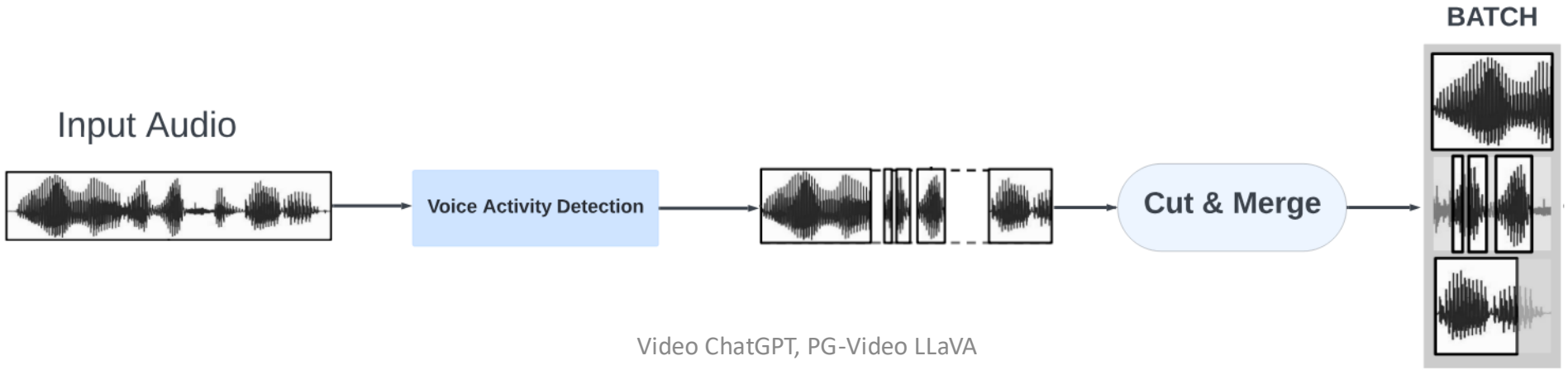
Audio Modality Integration

1. Voice Activity Detection



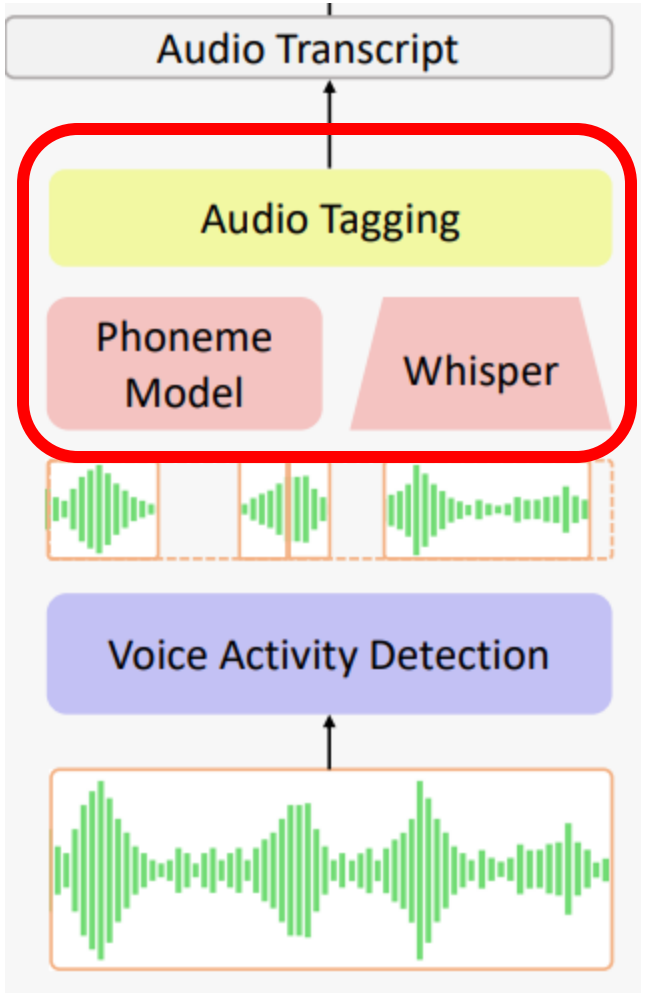
Voice Activity Detection & Cut/Merging

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



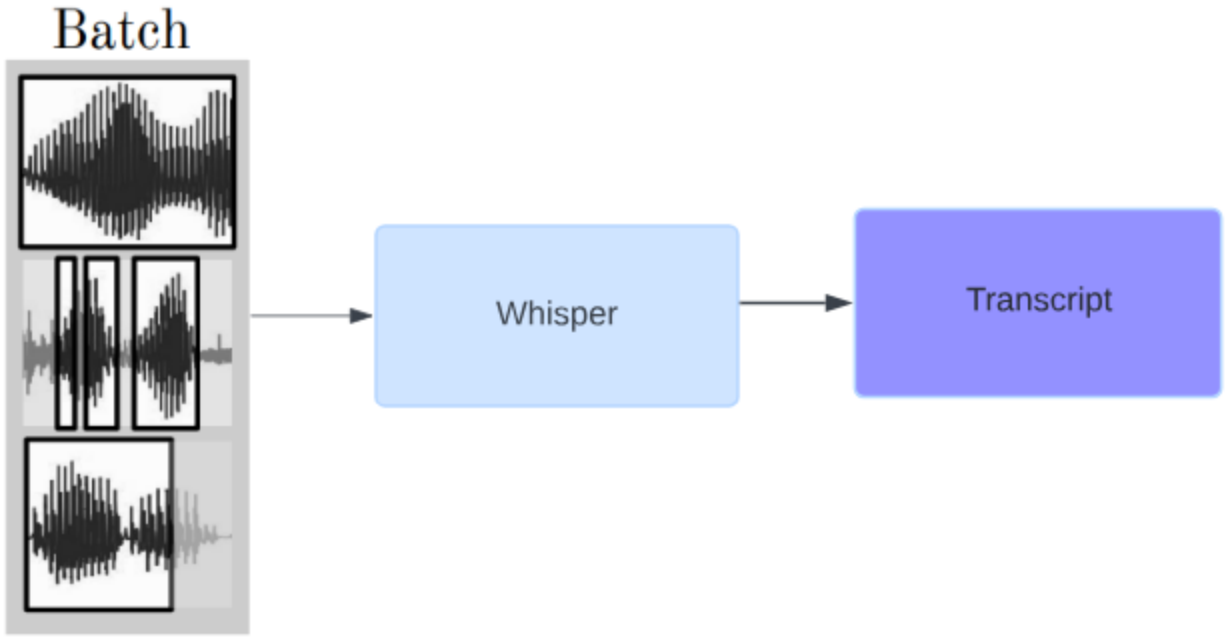
Audio Modality Integration

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



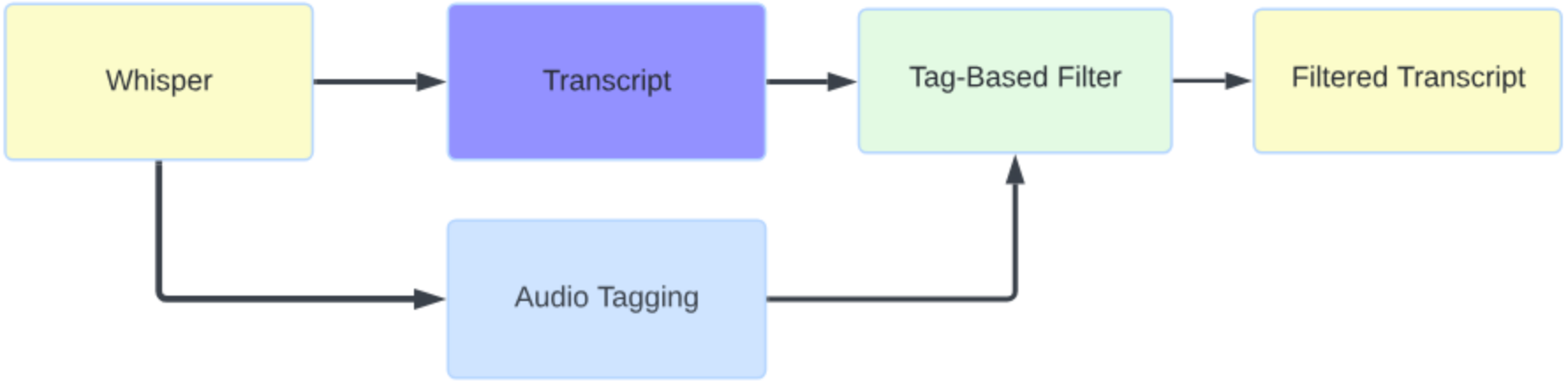
Whisper

- Whisper Model
  - Speech to transcript generation



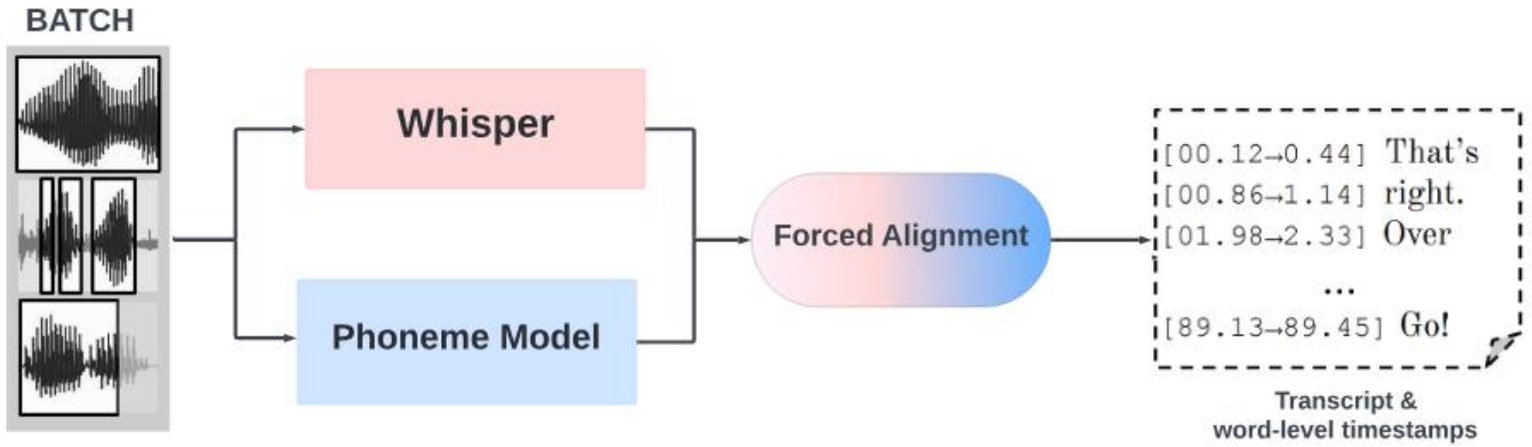
Audio Tagging (Whisper-AT)

- Produce audio-tagging output from the sliced original audio
- Consider the top 3 audio classes predicted
  - If “speech” is not among the top 3 predictions, the segment is ignored.
  - If P[music] > P[speech] and P[music] − P[speech] > threshold, the segment is



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 here 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.

Hallucination

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

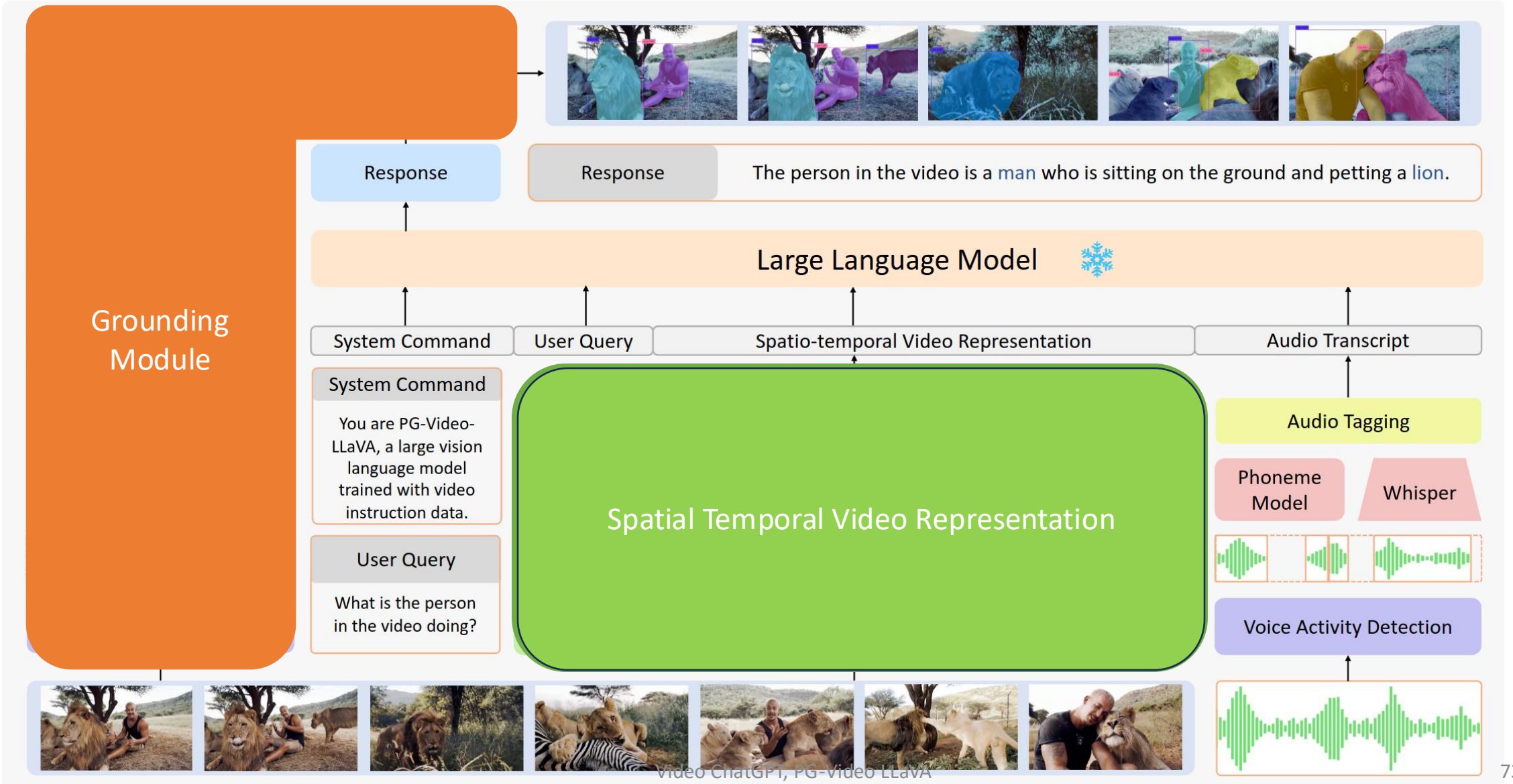
Music

Transcript Obtained from Whisper

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

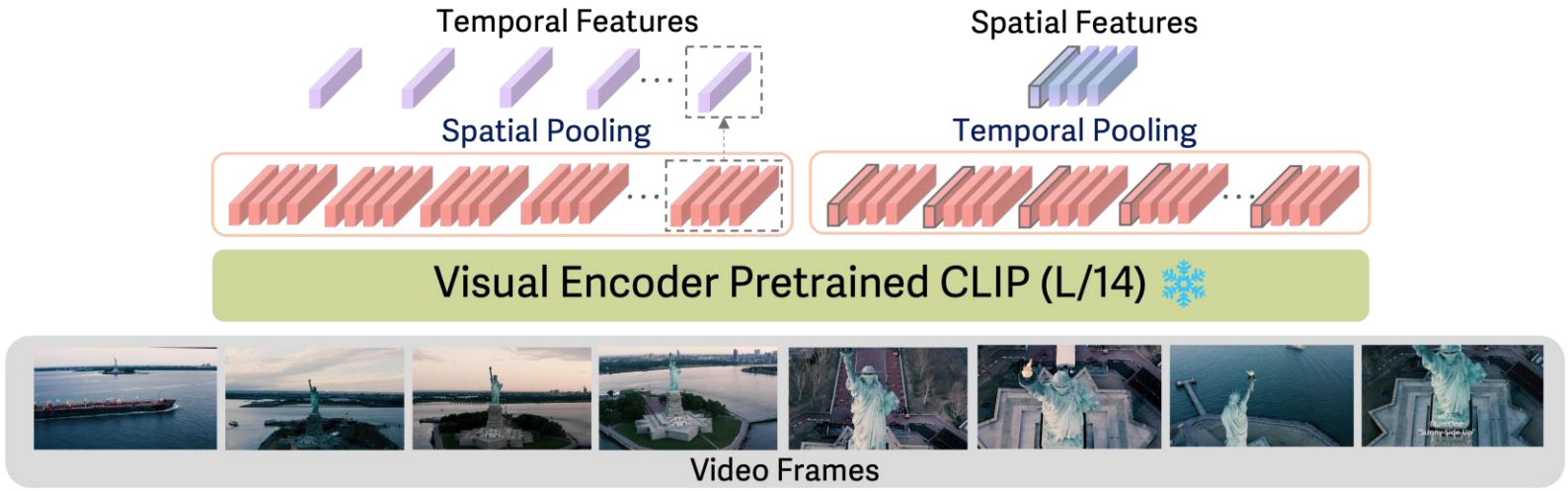
Garbage

PG-Video-LLaVA



Video-ChatGPT: Architecture

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



Video ChatGPT, PG-Video LLaVA

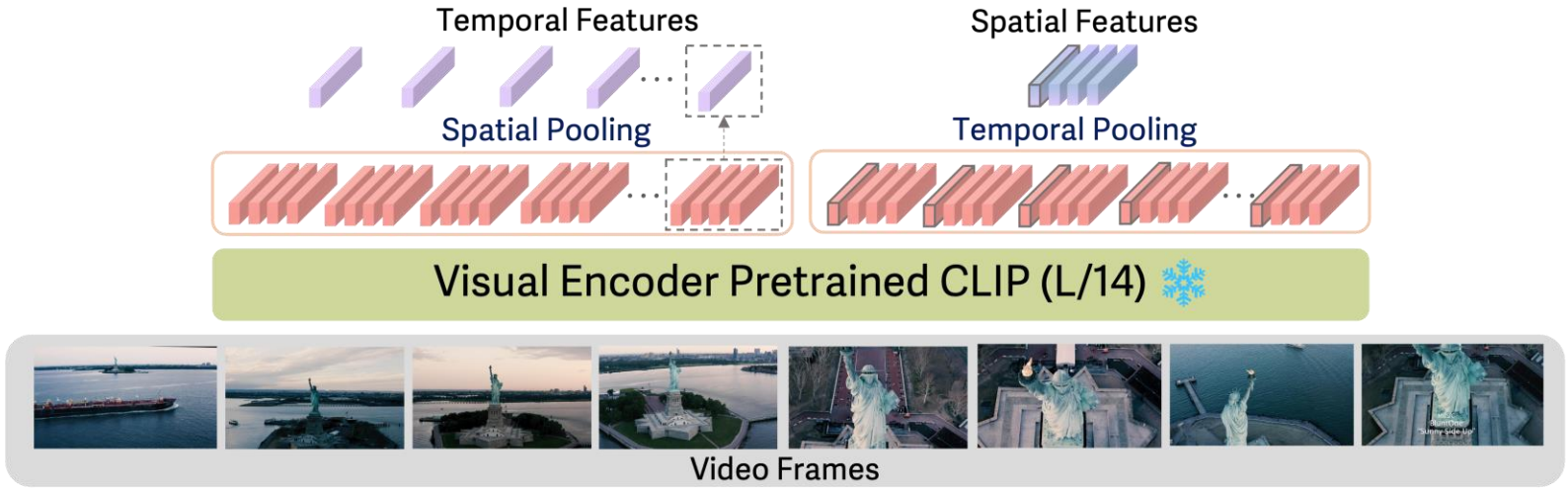
Video-ChatGPT: Architecture

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

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

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

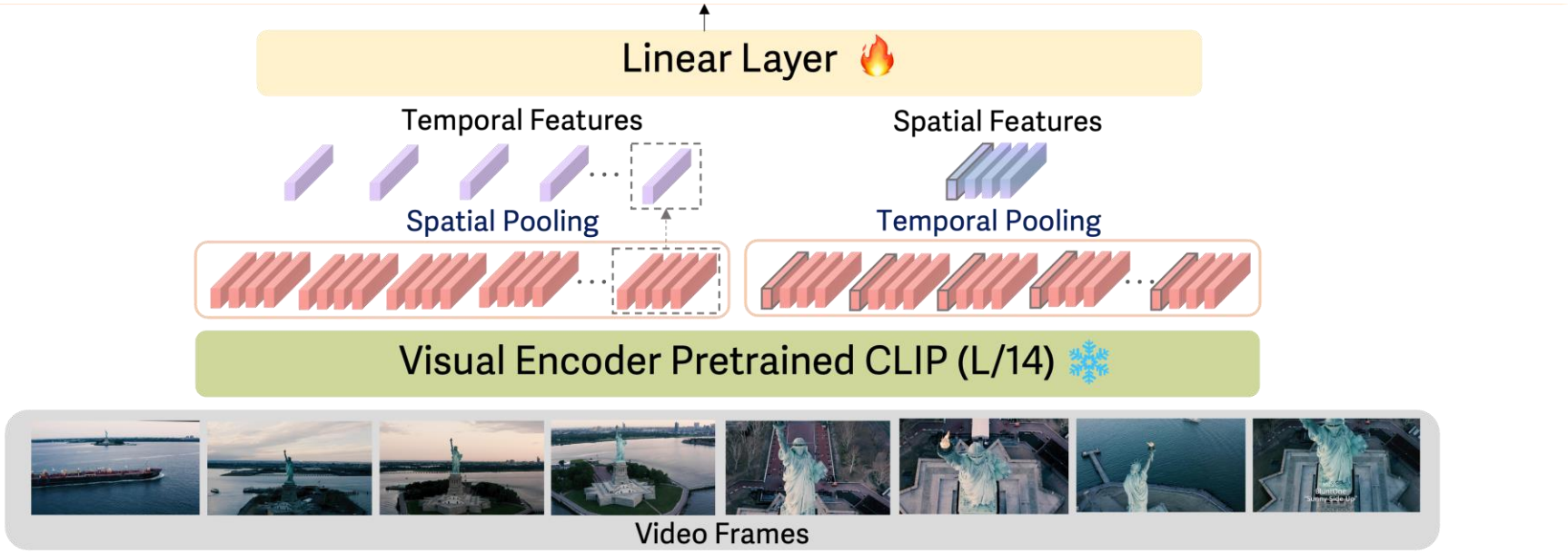
$v_i = [t_i \quad z_i] \in \mathbb{R}^{(T+N) \times D}$



Video ChatGPT, PG-Video LLaVA

Video-ChatGPT: Architecture

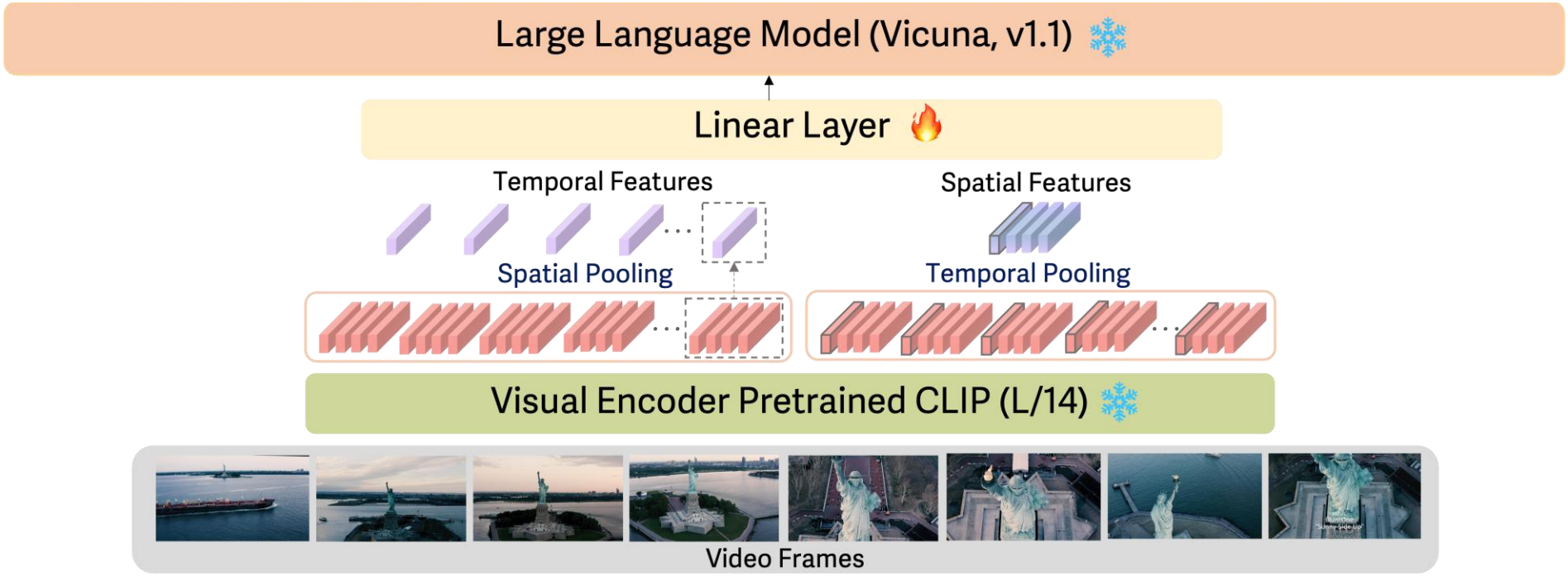
- Linear layer projects spatiotemporal video features to LLM input space



Video ChatGPT, PG-Video LLaVA

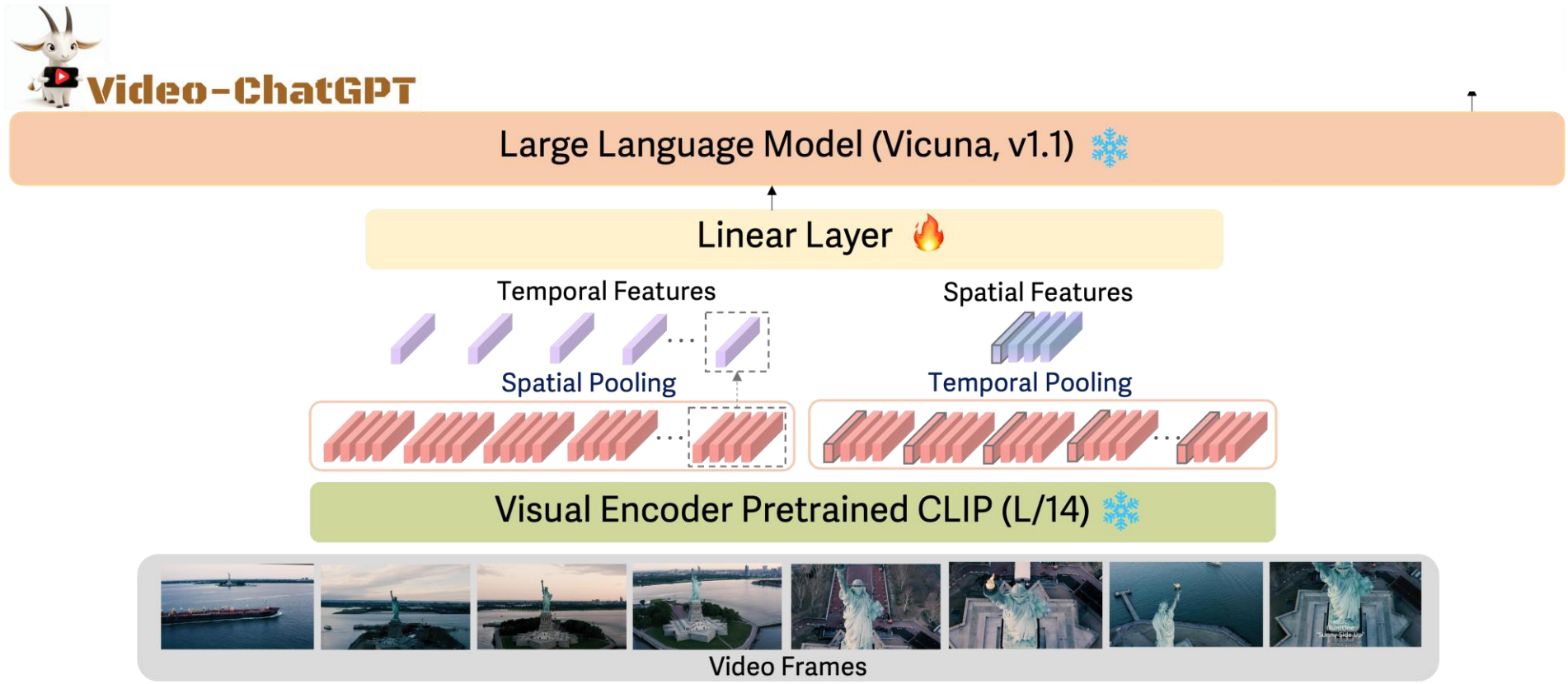
Video-ChatGPT: Architecture

- Linear layer projects spatiotemporal video features to LLM input space



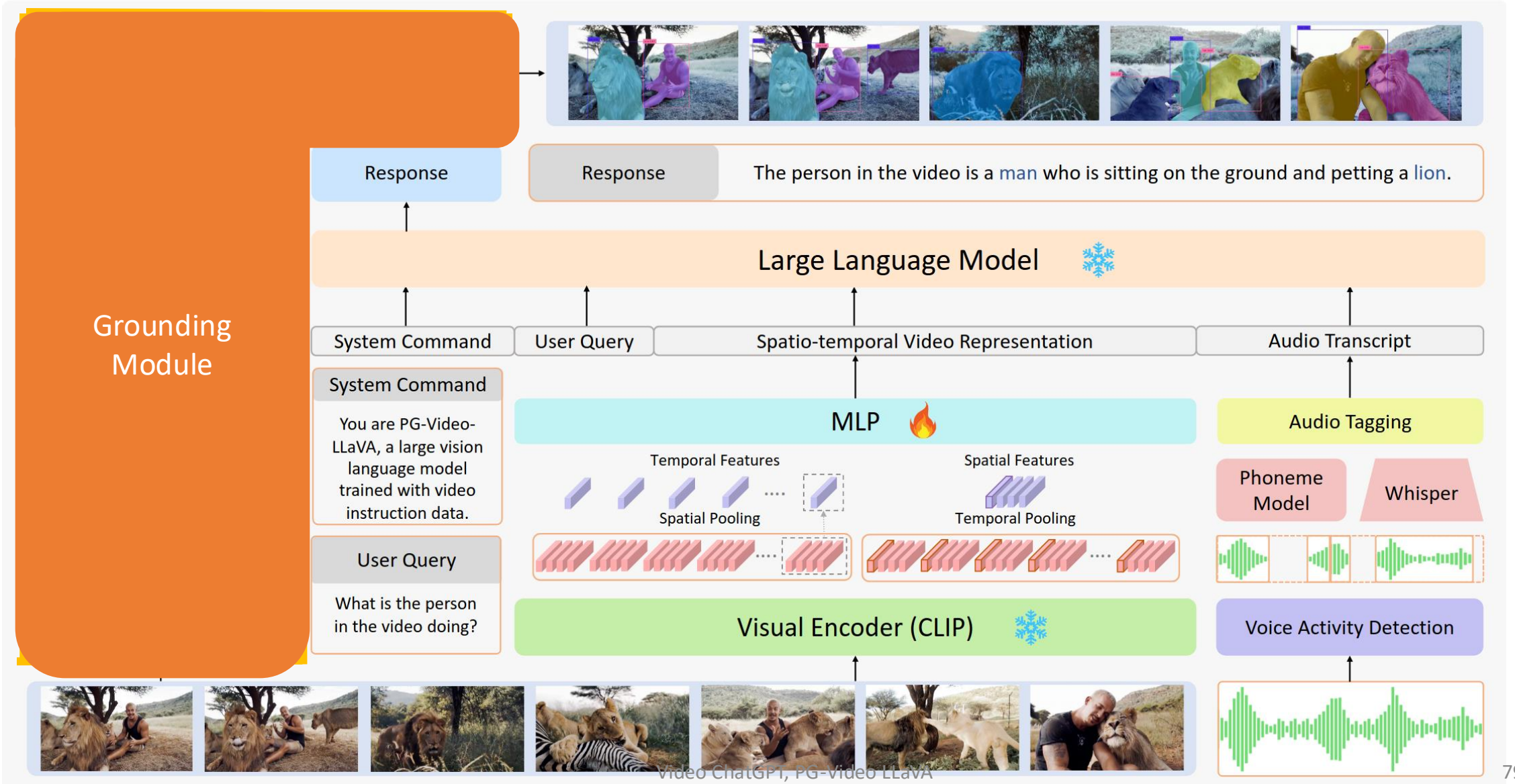
Video ChatGPT, PG-Video LLaVA

Video-ChatGPT: Architecture

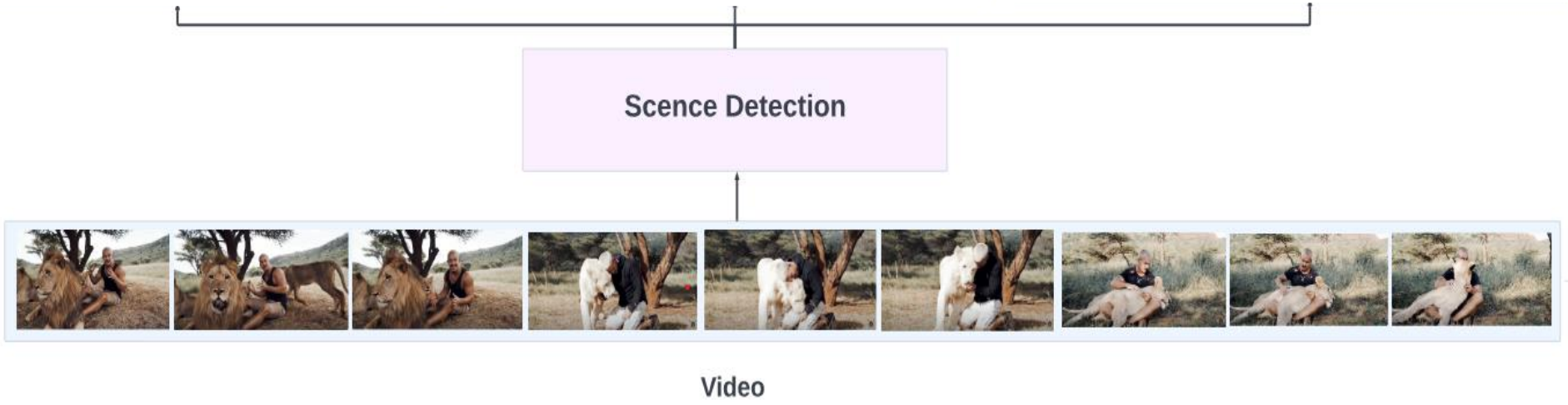


Video ChatGPT, PG-Video LLaVA

PG-Video-LLaVA



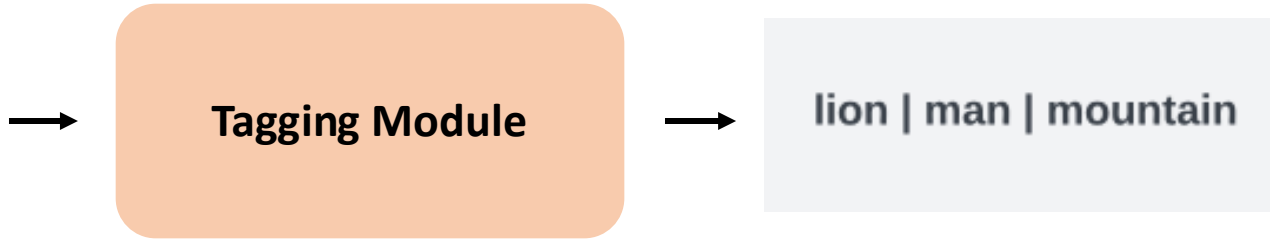
Scene Detection (PySceneDetect)

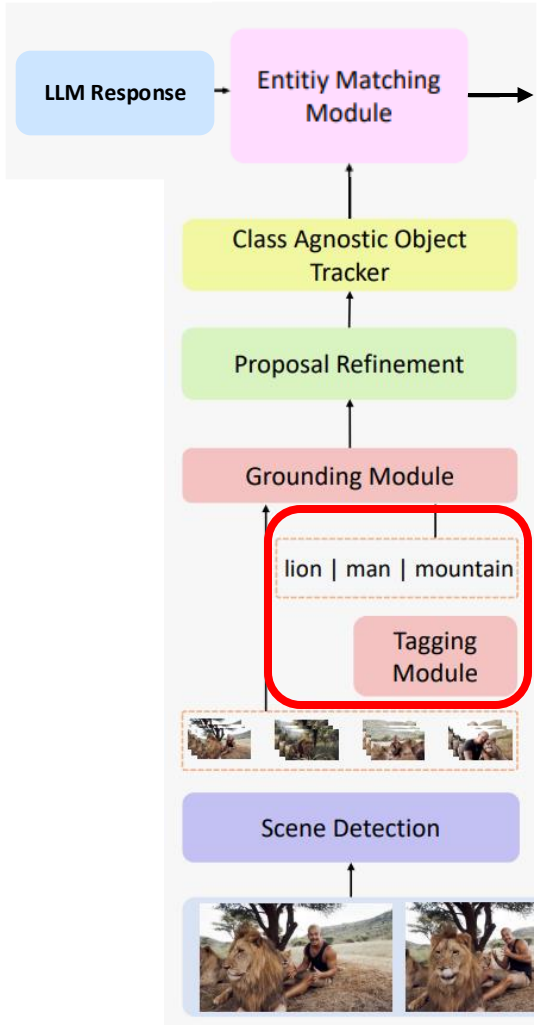




Tagging Module (RAM SWIN Large Variant)

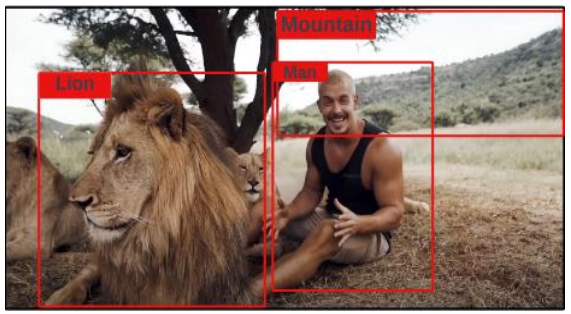






Grounding Module (GroundingDINO-T)

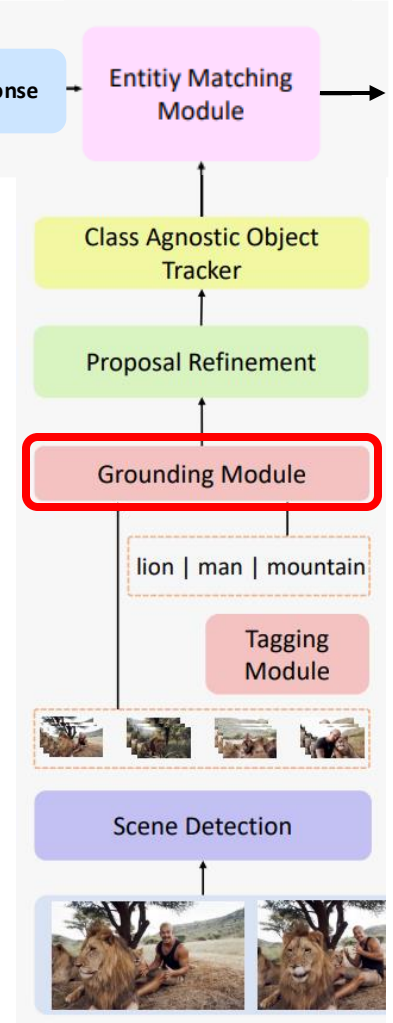




Grounding Module

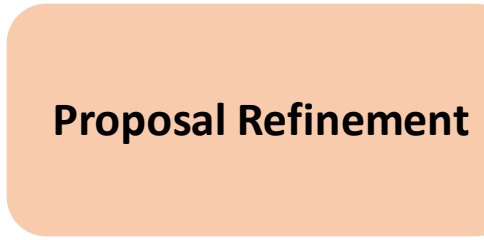
"lion, man, mountain"

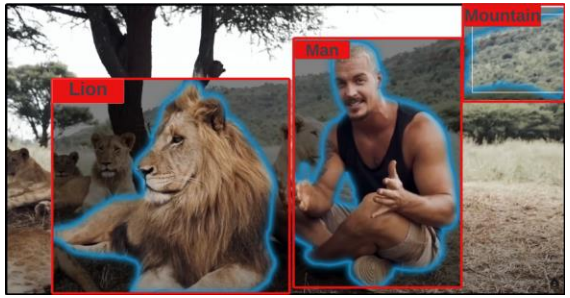
lion | man | mountain

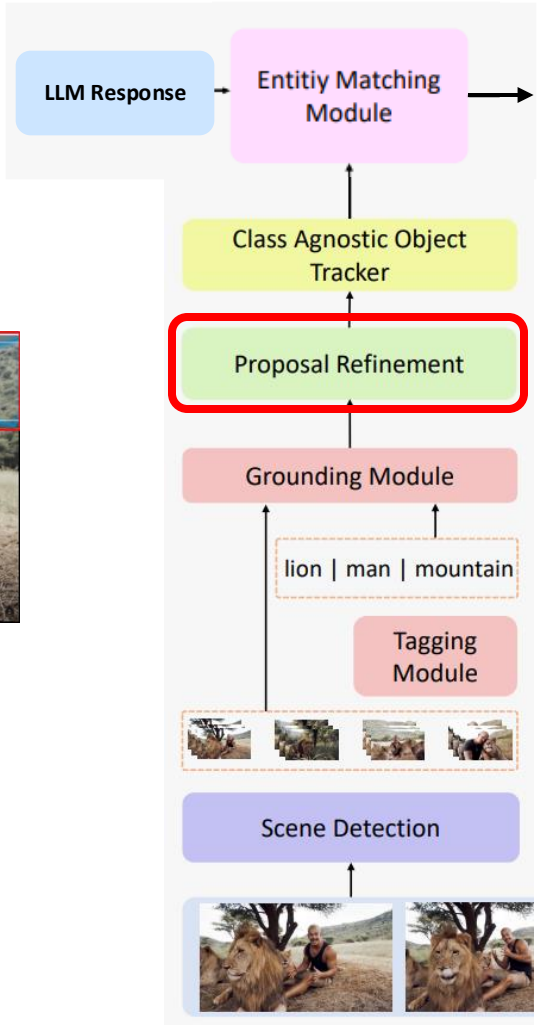


Proposal Refinement (SAM)



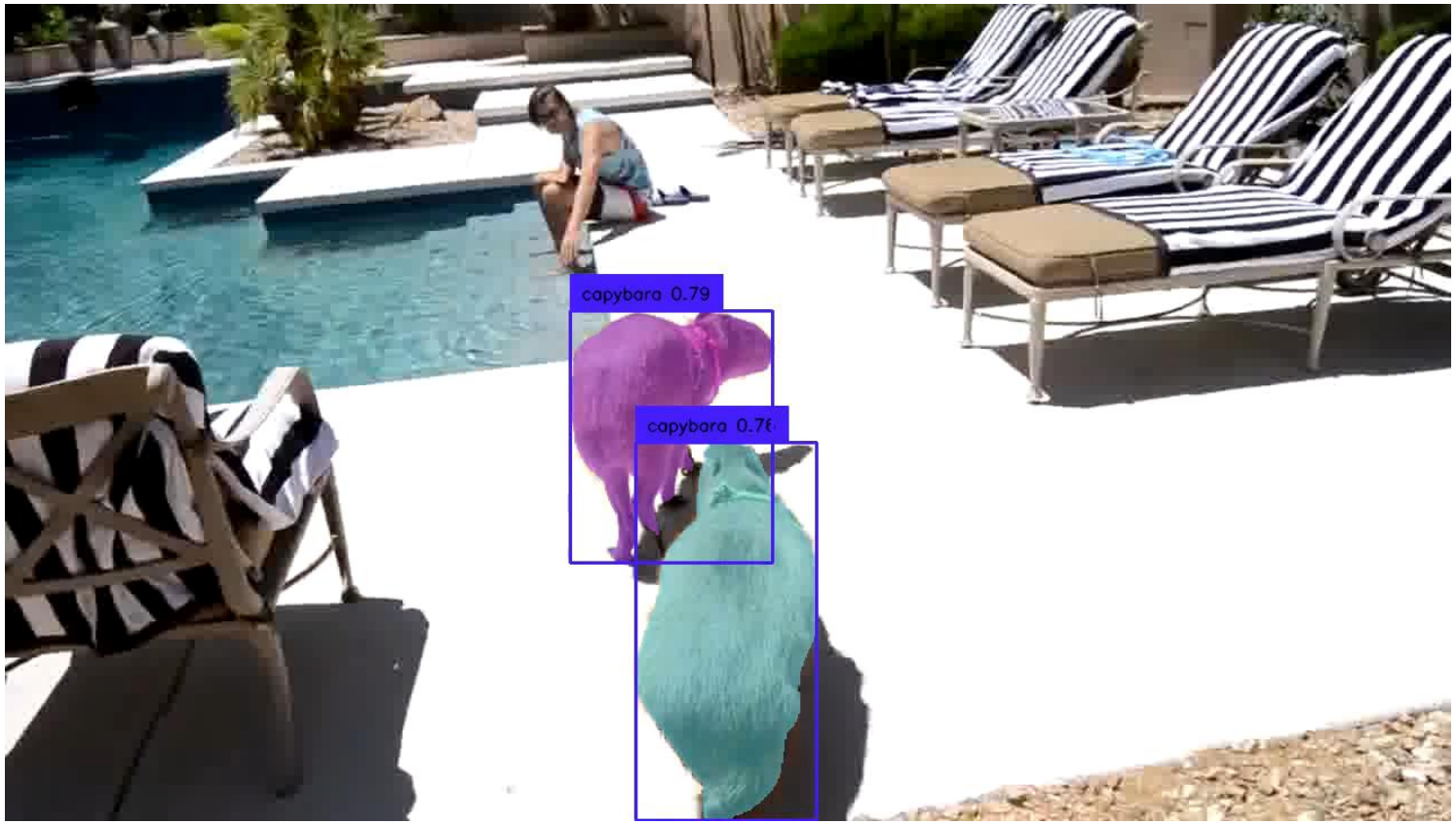


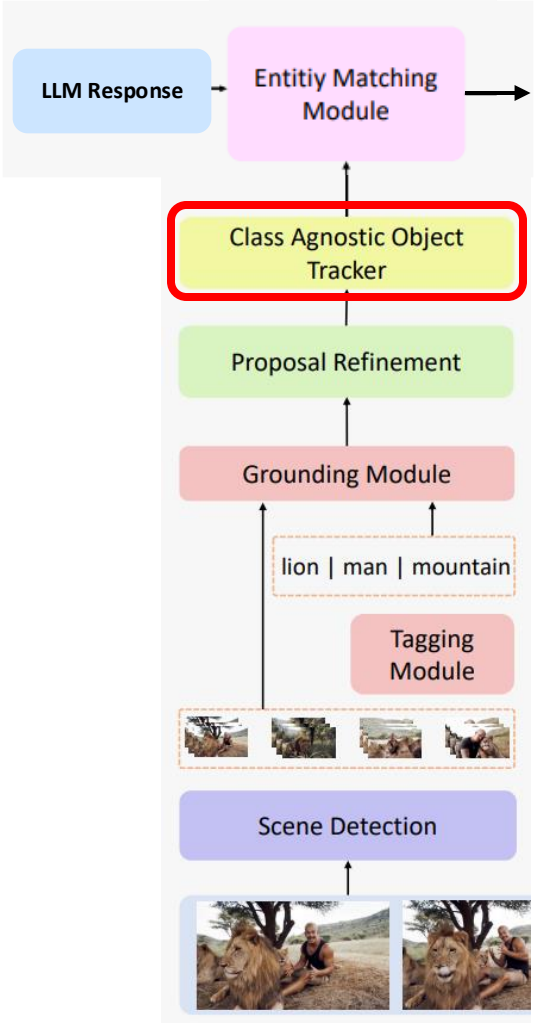




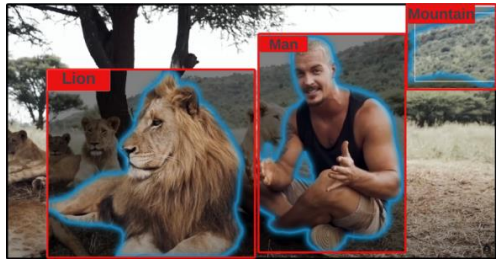
Class Agnostic Object Tracker (DEVA Tracker)

- Decouples segmentation and tracking tasks to be class agnostic
- Assigns agnostic IDs to individual objects through the video





Entity Matching Module (Vicuna 1.5)

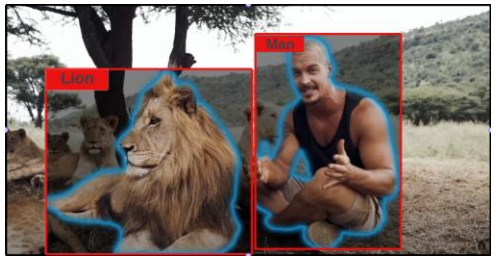


"lion, man, mountain"

"<List> Lion, Man, Mountain </List>
<Text> The person in the video is a man who is sitting on the ground and petting a lion. </Text>

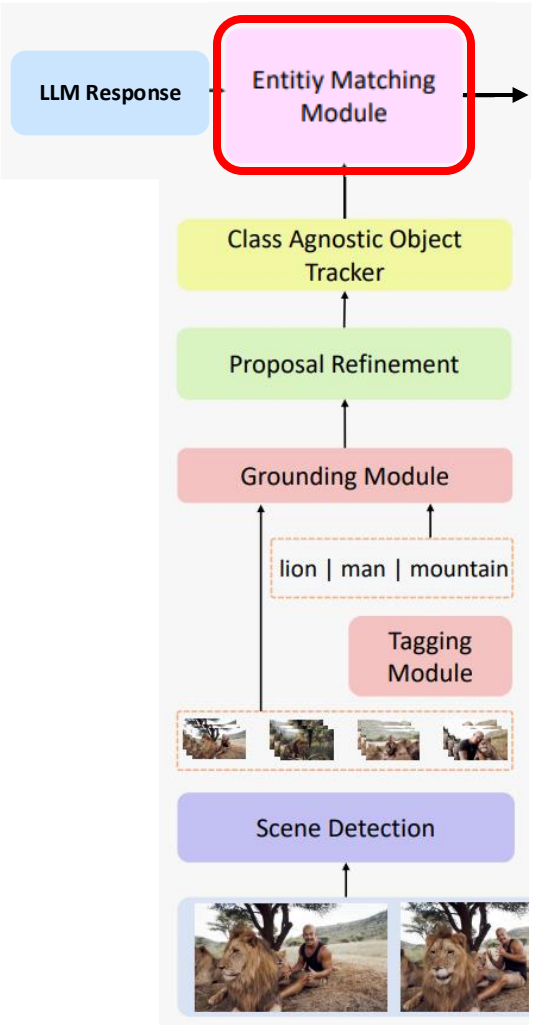
Vicuna 1.5
(Entity Matching)

"lion, man"

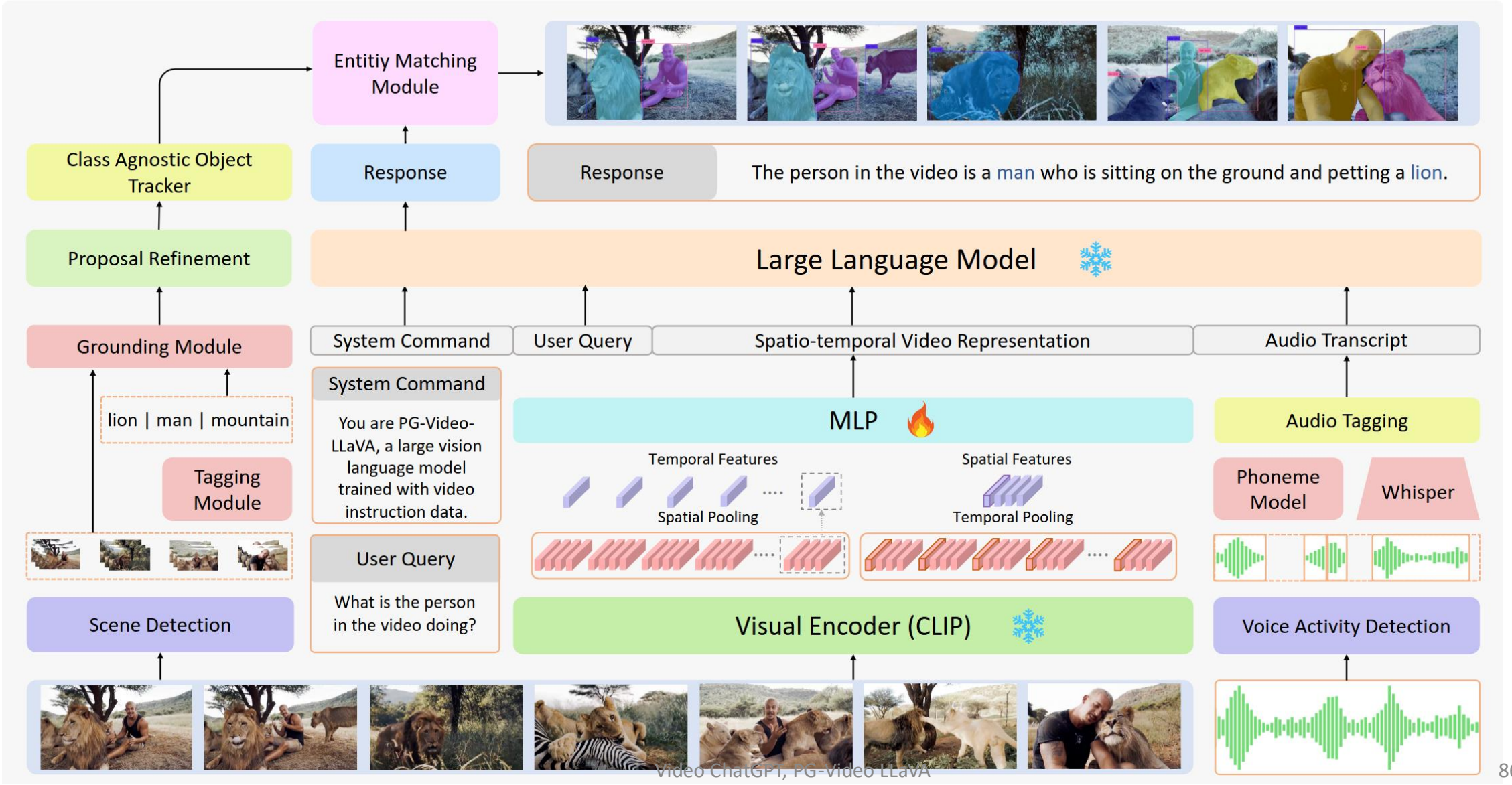


"The person in the video is a man who is sitting on the ground and petting a lion."

LLM Response



PG-Video-LLaVA



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
  - 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:
  - Correctness of Information
  - Detail Orientation
  - Contextual Understanding
  - Temporal Understanding
  - Consistency

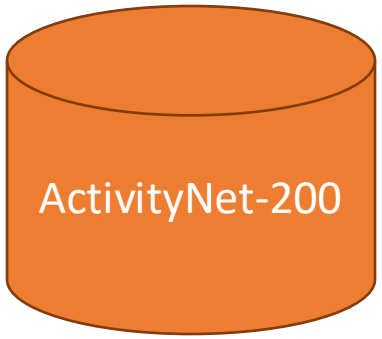
Conversational Benchmark Implementation

- 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.

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.\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}."
        }
    ]
)
```

Conversational Quantitative Benchmarks

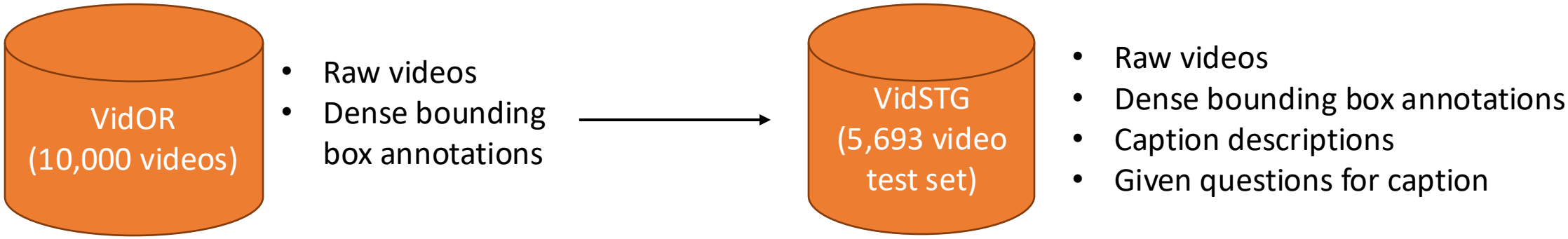


- Distinct video segments
- Human enriched ground truth annotations

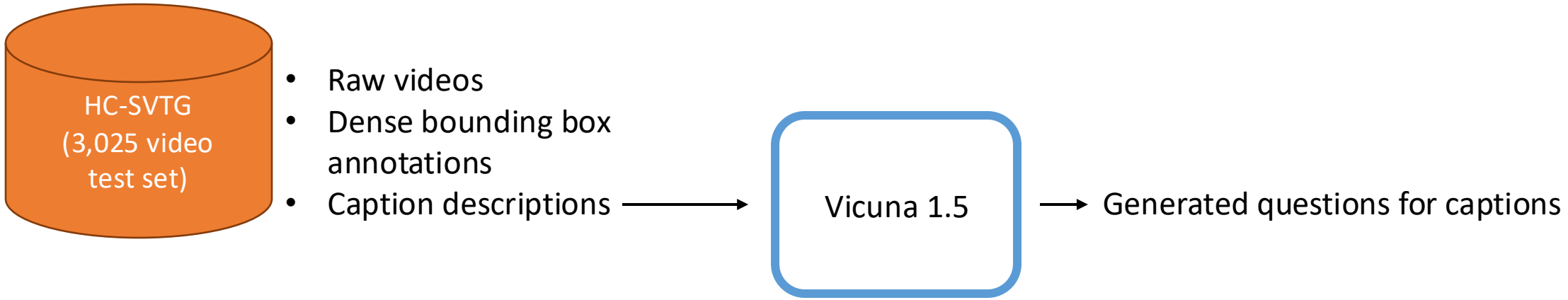
Model	Evaluation Metrics				
	Correctness	Detail Orientation	Contextual Understanding	Temporal Understanding	Consistency
LLaMA Adapter [9]	2.34	2.43	2.65	2.20	3.02
Video Chat [15]	2.49	2.82	2.92	2.27	3.11
Video-LLaMA [45]	2.29	2.61	2.68	2.17	2.87
Video-ChatGPT [22]	2.51	2.53	2.85	2.32	3.10
PG-Video-LLaVA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

Spatial Grounding Evaluation Datasets

VidSTG



HC-SVTG



Spatial Grounding Evaluation



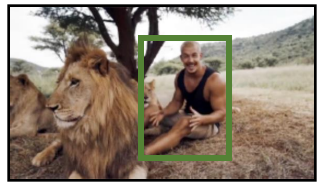


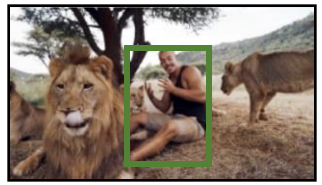
Grounding Module

Ground Truth









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



Q: Who is sitting next to the lions?

A: The man is sitting next to the lions.

Vicuna 1.5 (Referring Expression Extraction)

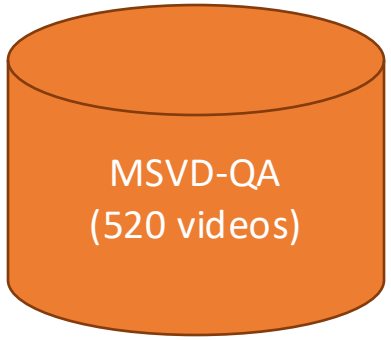
"man"

$$mean\ IoU = \frac{1}{|v|}\left(\sum_{\substack{v \in object\ is \\ present\ in\ GT}} IoU(v)\right)$$

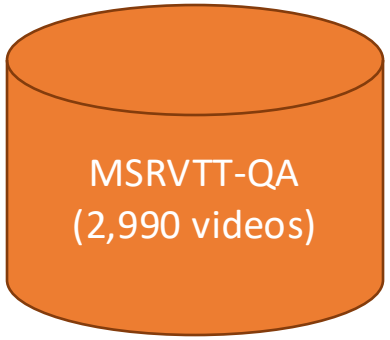
Spatial Grounding Quantitative Performance

Model	VidSTG [48]	HC-STVG [34]
Grounding DINO [20]	25.3	19.5
Video-LLaMA [45]	28.6	26.1
Video-ChatGPT [22]	32.8	20.8
PG-Video-LLaVA (7B)	34.2	**28.3**
PG-Video-LLaVA (13B)	**35.1**	27.3

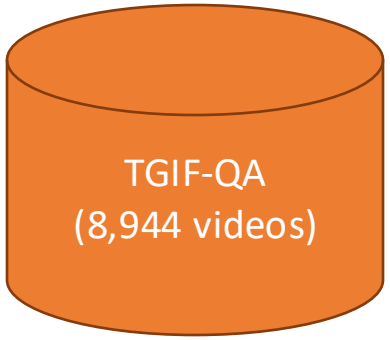
Zero Shot Video VQA Datasets



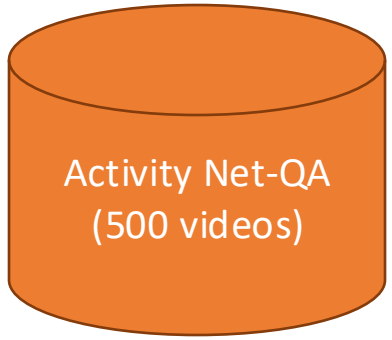
- Based on MSVD dataset
- QA pairs generated from video descriptions



- Based on MSR-VTT dataset
- QA pairs generated from video descriptions

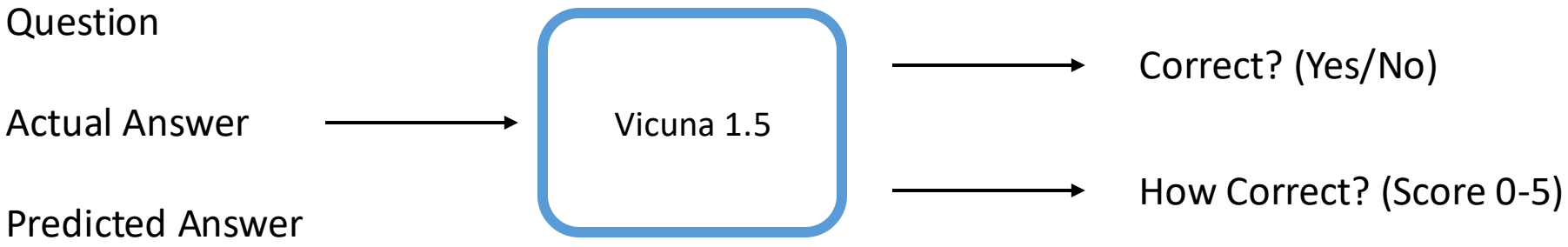


- Based on TGIF dataset
- GIFs, no audio present
- Crowdsourced QA pairs



- Based on ActivityNet-200
- Human annotated QA pairs

Zero Shot Video VQA Benchmarks



Model	MSVD-QA [39]		MSRVTT-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	**40.8**	**3.3**
PG-Video-LLaVA	**64.1**	**3.7**	**51.6**	**3.3**	**66.8**	**3.8**	39.9	**3.3**

Conversational Qualitative Performance



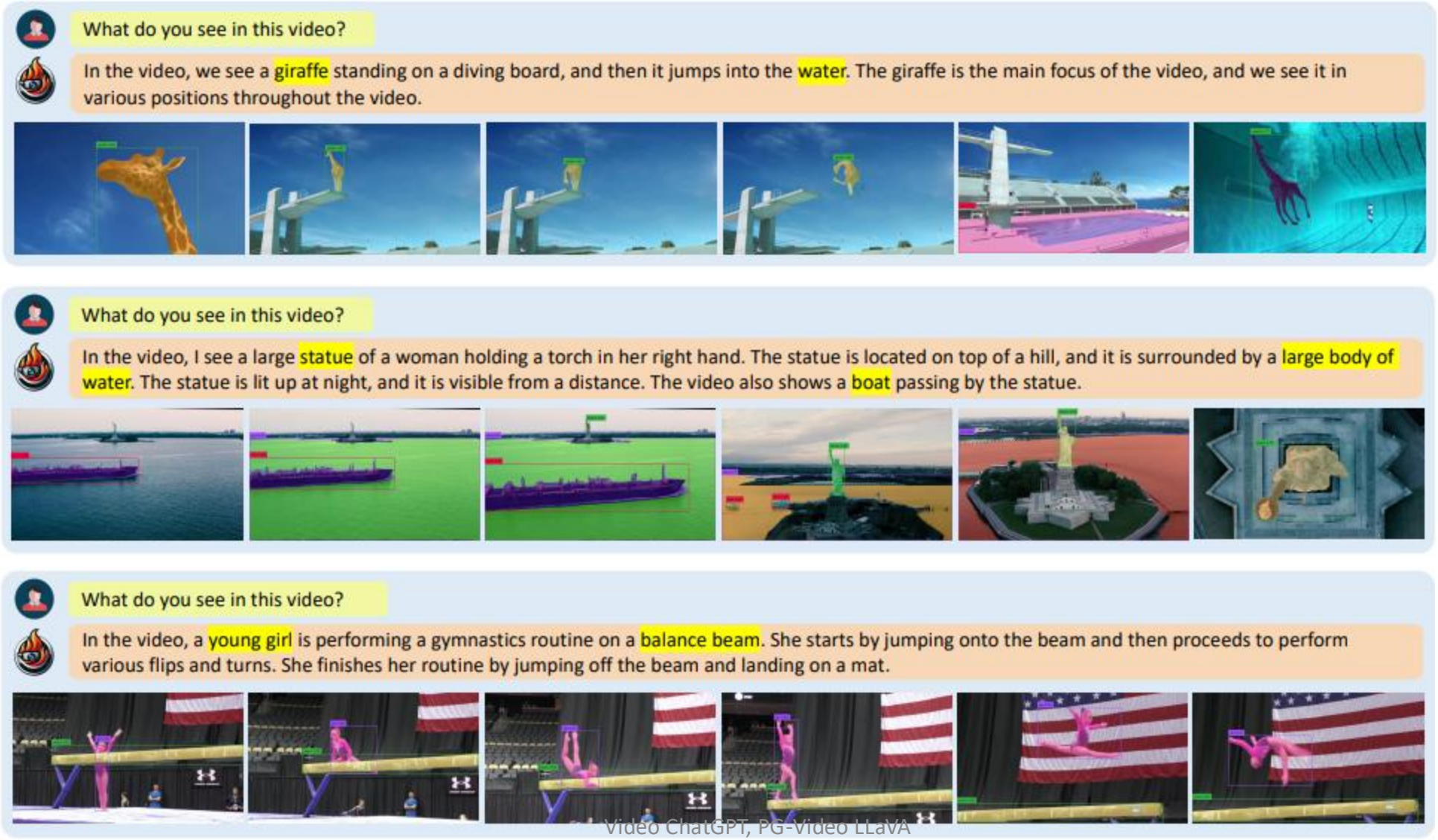
Describe this Video.

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.

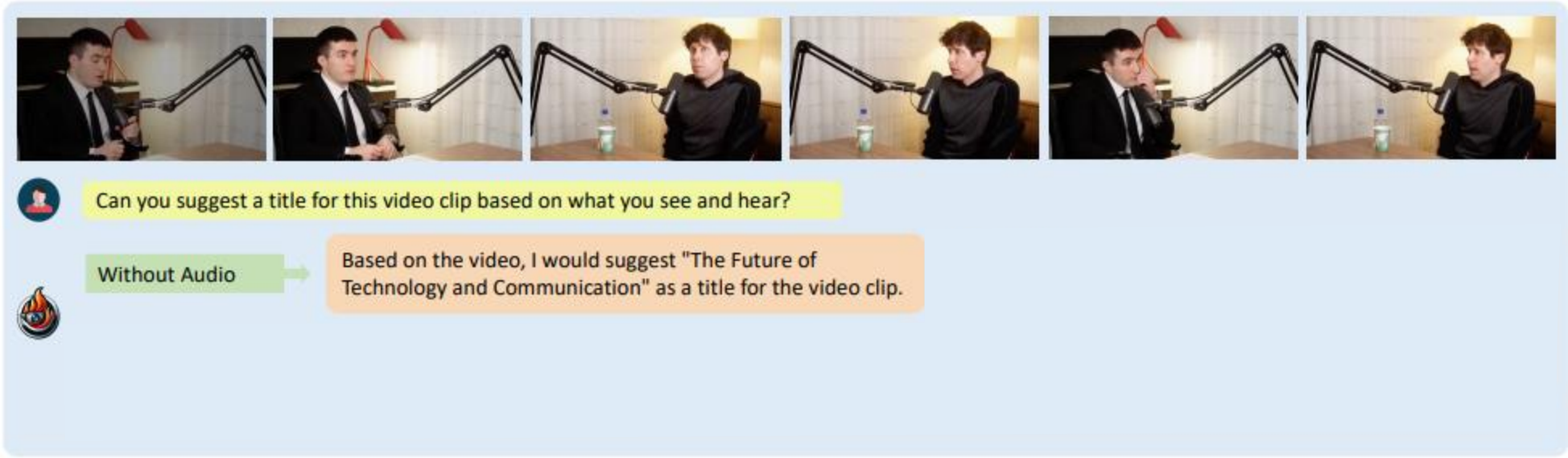
PG-Video-LLaVA (13B) → The video features a giraffe standing on a diving board and jumping into a pool. The giraffe is seen swimming in the pool and then jumping out of it. The video is set in a pool area with a blue sky in the background. The giraffe's movements are graceful and fluid, and it appears 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.

Spatial Grounding Qualitative Performance



Video ChatGPT, PG-Video LLaVA

Conversational + Audio Transcript Performance





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

Thank You!