

LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models *Supplementary Material*

Yanwei Li^{1*} Chengyao Wang^{1*} Jiaya Jia^{1,2}

CUHK¹ SmartMore²

A Experimental Details

In this section, we delve into the experimental details of the proposed LLaMA-VID framework. Generally, we adopt a similar training strategy with that in [7], except that we freeze the proposed text decoder during pretraining and subsequently optimize it during the finetuning phase. As outlined in Table 10, we employ customized training settings for the distinct stages illustrated in Figure 3 of the main paper. Specifically, during the modality alignment phase (stage 1), we keep the text decoder fixed and limit the maximum token length to 2K. Moving on to instruction tuning (stage 2), we unfreeze and optimize the text decoder, accommodating a larger maximum token length of 2K to facilitate video tuning. If stage 3 is adopted for long video tuning, we once again freeze the text decoder to conserve memory resources while increasing the maximum token length to a substantial 64K, catering to the demands of hour-long movie content. Detailed training settings are further explicated in Table 10.

For results on Something-Something-v2 (SSv2) dataset, we evaluate the temporal understanding ability of the model in the main paper. Because the SSv2 is an action recognition dataset with over 170 categories, we conduct the zero-shot evaluation and match the open-ended description of LLaMA-VID with the annotated caption using GPT3.5, same to that in MSRVTT-QA benchmark. We do not find such an evaluation with open-ended results in previous work, so we set up a baseline with Video-ChatGPT using the same manner.

B Instruction Tuning Details

Context Extension. To support hour-long video understanding, we conduct context extension of the language model in LLaMA-VID to accommodate inputs of up to 64K tokens. In line with prior research on long context LLMs [2, 3], we employ position interpolation techniques [2] to scale the rotary position encoding [11] from 4K to 64K, thereby enabling the processing of extended sequences. Subsequent to this adaptation, we perform supervised fine-tuning using Long-VideoQA, a dataset specifically designed for long-duration video instruction-following tasks.

* equal contribution

Table 10: Training settings of LLaMA-VID.

Settings	Stage 1	Stage 2	Stage 3
Batch size	256	128	8
Learning rate	1e-3	2e-5	2e-5
Learning schedule	Cosine decay		
Warmup ratio	0.03		
Weight decay	0		
Epoch	1		
Optimizer	AdamW		
DeepSpeed stage	2		
Vision encoder	Freeze		
Text decoder	Freeze	Open	Freeze
Max token	2048	2048	65536

Long-VideoQA Dataset. To improve the ability for understanding hour-long videos, we develop a specialized dataset for supervised fine-tuning purposes, named Long-VideoQA. The Long-VideoQA dataset comprises 15K question-and-answer (QA) pairs, with 9K of these pairs derived from movie scenes and the remaining 6K pairs sourced from LongAlpaca [3]. This dataset is designed to improve the performance of models on long-duration videos by providing them with relevant and diverse instruction-following scenarios. Inspired by previous works on image and short video instruction-following data collection [6, 8, 9], we leverage advanced language models like GPT-4 [10] and Claude-2 [1] as the strong teacher, to create instruction-following data that incorporates video content. To effectively prompt these language models, we utilized two types of symbolic representations to encode the video content:

- *Synopses.* It offers comprehensive narratives detailing the plot of a movie. These synopses give a broad overview of the storyline, encompassing major events and character developments.
- *Scripts.* It provides a more granular representation, including the storyline, character dialogues, and specific actions. Scripts convey the content in a format that closely follows the actual sequence of scenes and dialogues.

By utilizing movies from MovieNet [5] along with their corresponding synopses and scripts, we were able to create a rich dataset. It serves as a foundation for finetuning models to understand and respond to instructions in the context of long-form video content. As illustrated in Figure 7, we prompt GPT-4/Claude2 to generate three types of instruction-following data:

- *Video summary.* Utilizing synopses as prompts, we engage GPT-4 to generate two types of summaries for each movie: one brief summary and one detailed summary. This approach is designed to equip the model with a comprehensive understanding of the video content, providing both a concise encapsulation and an in-depth exposition of the narrative. The brief

summary offers a quick snapshot of the movie’s overarching plot, while the detailed summary delves into the nuances, including character arcs, thematic elements, and key plot points.

- *Plot-level understanding.* Leveraging the synopses as input prompts, we employ GPT-4 to generate both plot-related and character-related QA pairs. This process is tailored to yield a deeper insight into the movie’s narrative structure and the intricacies of character development. The plot-related QA pairs focus on the events, settings, and conflicts within the movie, enabling the model to grasp the story progression and the causal relationships between different events. Simultaneously, the character-related QA pairs delve into the personalities, motivations, and relationships of the characters, offering a nuanced perspective on how each character influences and is influenced by the unfolding story. In this way, we ensure that the model acquires a well-rounded comprehension of both the plot dynamics and the character profiles, which is crucial for accurately understanding and responding to complex instructions related to hour-long video content.
- *Detail description and reasoning.* By inputting the extensive movie scripts into Claude-2, we generate two distinct sets of data for each film: 5 plot-related reasoning pairs and 5 detail-related descriptions. This methodology serves to enhance the model’s capacity for learning and understanding fine-grained details. The plot-related reasoning pairs are crafted to challenge the model ability to make inferences and understand the logical flow of the narrative. These pairs require the model to connect different plot points, reason about cause and effect, and predict outcomes based on earlier events in the story. On the other hand, the detail-related descriptions focus on capturing the subtleties within the movie, such as specific descriptions of scenes, nuanced character expressions, or particular dialogue exchanges. These descriptions are crucial for the model to learn about intricacies that are often pivotal for a comprehensive understanding of video content.

C Additional Analyses

We provide more examples for long videos and images, as well as the visualization for response in context attention.

Examples of Long Videos. In Figure 8, we further demonstrate the effectiveness of LLaMA-VID by interacting with a broader range of hour-long movies, including genres like Romance, Adventure, and Sci-Fi. This diverse selection of films allows us to evaluate the performance across different narrative structures and thematic elements. It demonstrates a proficient ability to summarize storylines and engage in plot-related reasoning by synthesizing information from both video frames and subtitles. In particular, we compare with LLaMA 2 [12] and LongLoRA [3] by feeding the movie name and all subtitles, respectively. As depicted in Figure 9, the result reveals that LLaMA-VID outperforms its counterparts in tasks involving character understanding and plot-related reasoning.

The nuanced understanding of complex narratives and the integration of multi-modal data underscore its potential for advanced applications in video analysis and interaction.

Examples of Images. In Figure 10, we present additional interactions with LLaMA-VID, showcasing the model’s adeptness in knowledge-based perception and reasoning. The examples illustrate the model’s capability to discern and interpret object details within a given environment and to utilize environmental cues to effectively respond to user inquiries. For instance, when provided with a description or an image of the surrounding environment, LLaMA-VID is able to identify specific objects and their attributes. It can then leverage this information to answer questions posed by the user that may relate to object functions, spatial relationships, or contextual relevance. This performance is indicative of the sophisticated understanding. It is not limited to mere recognition but extends to a deeper cognitive level where it can process and integrate environmental information to engage in informed and accurate dialogue. Such demonstrations reinforce the potential of LLaMA-VID as a tool for complex interaction involving visual data and natural language processing.

Response in Context Attention. In Figure 11, we extend the visualization of the attention mechanism by presenting additional results of the response in context attention. This visualization is in alignment with that in Figure 6 of the main paper, where we highlight the areas with the highest attention scores. Specifically, we focus on the top 20 scores while retaining the first five queries in \mathbf{Q}_t . These visualizations serve to underline the effectiveness of the text-guided query \mathbf{Q}_t on areas that are most pertinent to input questions. It is dynamically adjusted based on the content of input questions, demonstrating its ability to discern and prioritize different regions depending on the context. For example, in the last two rows of Figure 11, where different questions are posed regarding the same image, LLaMA-VID adjusts its focus accordingly, directing its attention to specific regions that are relevant to each question. This capability signifies a sophisticated level of contextual understanding and adaptability, showcasing the potential of LLaMA-VID in processing complex visual and textual inputs.

 **Video Frame:** Only for illustration here, not used to produce instruction data.




 **Synopsis for the whole movie**
Synopsis: Young Rose, angry and distraught that her mother has apparently arranged the marriage, considers committing suicide by jumping from the stern. Jack manages to pull her back over the rail after she loses her footing.....

 **Script for the whole movie**
Script: Rose runs along the B deck promenade. She is dishevelled, her hair flying. She is crying, her cheeks streaked with tears. But also angry, furious! She is shaking with emotions she doesn't understand... hatred, self-hatred.....


 **Instruction pairs for movie summary**
 **User:** You are an artificial intelligence visual assistant that can understand and analyze long videos. You will receive the synopsis of a movie, which provide a detailed description of the plot of the movie.
The task is to use the provided movie plot synopsis, create the plot summary of the given movie.
Create 2 plot summary of the movie.
The first one should be a brief summary written in one paragraph.
The second one should be a detail summary written in multiple paragraphs.
Please write the summary strictly according to the provided synopsis. Do not add additional information. Do not modify the storyline of the movie.
Here is an example for plot summary. Write summary in this format:
Brief Summary: Dorothy Gale, a young girl living on a Kansas farm, dreams of escape before a tornado whisks her away to the magical land of Oz
Detail Summary: Dorothy Gale is an orphaned teenager living with her Aunt Em and Uncle Henry on a Kansas farm. Her life takes a turn when
 **GPT-4:** Brief Summary:....., Detail Summary:.....

 **Instruction pairs for plot level understanding**
 **User:** You are an artificial intelligence visual assistant that can understand and analyze long videos. You will receive the synopsis of a movie, which provide a detailed description of the plot of the movie.
The task is to use the provided movie plot synopsis, create some plausible questions about the movie, and provide answer for each questions.
Create 5 questions about the movie plot, including plot understanding, plot description, plot analysis, etc.
Create 5 questions about characters, including character relationship, character personality, character behavior, etc.
Only include questions that have definite answers. Do not ask any question that cannot be answered confidently with the information from the provided synopsis.
Understand and memorize the plot of the movie from plot synopsis first, then answer the question.
When using the information from the synopsis, directly explain the movie content, and do not mention that the information source is the synopsis. Always ask and answer the question as if you are directly watching the movie.
 **GPT-4:** Question:..... Answer:.....

 **Instruction pairs for movie reasoning**
 **User:** You are an artificial intelligence visual assistant that can understand and analyze long videos. You will receive the movie script, which includes outlines of the story, setting, dialogue, and actions of characters in a movie.
The task is to use the provided movie script, create 5 plausible question about the movie, and provide answer for these questions in detail.
Create 5 complex questions about plot reasoning rather than simply describe the plot.
To answer such a question, you first need to fully understand the movie, and then explain why a certain plot or scene in the movie happened the way it did based on the plot. To make the question challenging, the answer needs to be from some plot of the movie and different from the episode mentioned in the question. The user needs to fully understand the movie to answer the question.
Only include questions that have definite answers. Do not ask any question that cannot be answered confidently with the information from the provided script.
When using the information from the script, directly explain the movie content, and do not mention that the information source is the script. Always ask and answer the question as if you are directly watching the movie.
Here is an example for question and answer. Generate question and answer pairs in this format:
Question: Why doesn't Rose get in the lifeboat with her mother when she has the chance?
Answer: Even though Rose has the opportunity to safely get on Lifeboat with her mother Ruth and other women, she decides not to
 **Claude-2:** Question:..... Answer:

 **Instruction pairs for movie details**
 **User:** You are an artificial intelligence visual assistant that can understand and analyze long videos. You will receive the movie script, which includes outlines of the story, setting, dialogue, and actions of characters in a movie.
The task is to use the provided movie script, create some plausible questions about the movie, and provide answer for each questions in detail.
Create 5 complex questions about detail scene and activity description.
To answer such questions, you first need to fully understand the movie, and then describe the scene and person activity for a specific scene in detail. To make the question challenging, make sure these questions can only be answered correctly using the visual information in the movie, rather than reading lines from the script only. The user needs to fully understand the movie to answer the question.
Only include questions that have definite answers. Do not ask any question that cannot be answered confidently with the information from the provided script.
When using the information from the script, directly explain the movie content, and do not mention that the information source is the script. Always ask and answer the question as if you are directly watching the movie.
Here is an example for question and answer. Generate question and answer pairs in this format:
Question: Describe how Rose returns to the cabin to save Jack after she found out Jack was innocent in detail.
Answer: After realizing that Cal and Lovejoy framed Jack by putting the diamond necklace in his pocket, Rose runs back into the sinking Titanic
 **Claude-2:** Question:..... Answer:

Fig. 7: Details in constructing instruction pairs for the movie Titanic. It is detailed version of that in Figure 4 of the main paper.



3-hour Video
Titanic

User: Describe the storyline of this movie.

Assistant: In 1912, a young artist named Jack Dawson wins a ticket to the first class of the RMS Titanic on a game show. He is accompanied by his best friend, Fabrizio De Rossi, and they both board the ship with their mother, Ruth. Jack is immediately attracted to Rose DeWitt Bukater, the daughter of a wealthy steel magnate. Rose is engaged to Cal Hockley, but she is not happy with her life and is having an affair with Jack. Jack and Rose spend the first night on the ship together, and Jack saves Rose from falling off the ship. They become close friends, and Jack is invited to dinner with Rose and her family. Rose's mother and Cal are not happy about Jack and Rose's relationship, and they try to keep them apart. Jack and Rose continue to see each other, and Jack even asks Rose to marry him. Rose is attracted to Jack, but she is engaged, and she is not ready to give up her life of luxury. Jack and Rose spend the rest of the voyage together, and they share a romantic moment on the ship's bow. Jack and Rose are the last two passengers to board the ship when it departs from Southampton. They are the last two passengers to leave the ship, and they share a passionate kiss before the ship strikes the iceberg and sinks.

User: Who is the hero and heroine of this movie?

Assistant: The hero and heroine of this movie are Jack Dawson and Rose DeWitt Bukater.

User: How does Jack get the ticket to the Titanic?

Assistant: Jack won the ticket to the Titanic in a poker game.

User: Why did Rose board the Titanic?

Assistant: Rose boarded the Titanic to travel to America to marry her wealthy fiancé, Cal Hockley.

User: How was Jack and Rose meet?


Assistant: Jack and Rose meet each other when Jack saves Rose from falling off the ship after she slips on the deck.

User: Why did Rose's mother hope Rose to marry Cal Hockley?

Assistant: Rose's mother hoped Rose would marry Cal Hockley because he was a wealthy steel tycoon who could provide for Rose and help her family, who were financially ruined after losing their fortune.

User: What is Cal's engagement gift to Rose?

Assistant: Cal's engagement gift to Rose is a rare diamond known as the Heart of the Ocean.



3-hour Video
Interstellar

User: Write a review of this movie.

Assistant: Interstellar is a thought-provoking and visually stunning film that explores the mysteries of space and time. The story follows Cooper, a former astronaut, as he searches for a new home for humanity after Earth becomes uninhabitable due to blight. His journey leads him through a wormhole and into a black hole, where he encounters a mysterious force that communicates through gravity. The film effectively uses the concepts of time dilation and the effects of gravity on time to drive the plot forward. The visual effects and the performances of the cast, particularly Anne Hathaway and Jessica Chastain, are outstanding. The film concludes with a sense of hope and optimism, as Cooper finds a new home for humanity, leaving behind a legacy for his daughter. Overall, Interstellar is an engaging and thought-provoking film that explores complex scientific concepts in a compelling way, while also telling a poignant and emotional story.

User: What is the function of the wormhole? And where is it located?

Assistant: The wormhole is a shortcut through space-time that leads to a distant galaxy, and it is located near Saturn.

User: Who might have placed the wormhole near Saturn, and what was their purpose?

Assistant: It is possible that an advanced civilization from the future placed the wormhole near Saturn as a way to guide humanity to new planets for survival.

User: How does Cooper transmit data from the black hole to Murphy?


Assistant: Cooper transmits data from the black hole to Murphy by using a timer watch with the data encoded into the second hand and as a Morse code.

User: Why does time run faster on Miller planet than on Earth? And how fast it is.

Assistant: Time runs faster on Miller planet due to the proximity of the black hole Gargantua, which has a gravitational pull that slows time relative to Earth. Every hour spent on Miller planet is equivalent to 7 years on Earth. This is because the closer one is to a black hole, the slower time runs due to the extreme gravitational forces pulling time into the future....

User: Write a brief summary of the plot of this movie.

Assistant: In a post-apocalyptic future, Cooper, a former astronaut and farmer, is tasked with piloting a mission to find a new home for humanity after Earth becomes uninhabitable. He and his daughter Murphy uncover anomalies that lead them to a secret facility housing an advanced space station....



2-hour Video
Forrest Gump

User: Briefly describe storyline of this movie.

Assistant: "Forrest Gump" follows the life of a man with a low IQ, played by Tom Hanks, who embarks on a journey of self-discovery and love. The film begins with Forrest's childhood in Alabama, where he forms a close friendship with Jenny, and his life is shaped by his mother's wisdom and the lessons he learns from his experiences. Forrest's life takes unexpected turns, including his time in the Vietnam War, where he becomes a hero and is awarded the Medal of Honor. After returning home, he finds success in the shrimping business and becomes a ping-pong champion, eventually meeting President Nixon. His life takes another turn when he embarks on a cross-country running spree, which inspires the nation and leads him to reconnect with Jenny, who has a son named Forrest Jr. The film concludes with Forrest reflecting on his life and the lessons he has learned, as he continues to live a simple but fulfilling life.

User: Who is Jenny, and what is her significance in Forrest's life?

Assistant: Jenny is Forrest's childhood friend who becomes his girlfriend and later his wife. She is significant in Forrest's life as she is the person he loves and cares for the most, and their relationship is a source of happiness and comfort.

User: Why Forrest Gump Was Awarded the Medal of Honor After the Vietnam War?

Assistant: Forrest Gump was awarded the Medal of Honor for his actions during the Vietnam War, specifically for his heroism in rescuing his platoon mates from an ambush.

User: Why did Lieutenant Dan Taylor blame Forrest for rescuing him?

Assistant: Lieutenant Dan Taylor blamed Forrest for rescuing him because he felt that Forrest had cheated him out of his destiny to die in battle.

User: Why did Forrest Gump get into shrimping after he retired?

Assistant: Forrest Gump got into shrimping after he retired because he had promised his friend Bubba that he would become a shrimping boat captain, and he wanted to honor Bubba's memory by keeping his promise.

User: How many American presidents did Forrest Gump meet in the movie? What is the reason for each meeting?

Assistant: Forrest Gump met three American presidents: President Kennedy, President Johnson, and President Nixon. He met President Kennedy at the White House after being named an All-American football player, President Johnson at the White House after being awarded the Medal of Honor, and President Nixon at the White House after being named to the All-American Ping-Pong team.

Fig. 8: More examples of LLaMA-VID with hour-long movies using Vicuna-7B based models.



LLaMA-VID

User: How did Jack Dawson and Rose DeWitt Bukater meet each other in this movie?

Assistant: Jack Dawson and Rose DeWitt Bukater met each other when Rose was about 10 years old. They were both on the Titanic, and Jack saved her from drowning in the ship's swimming pool.

User: Does Jack Dawson and Rose DeWitt Bukater both survive the sinking of the Titanic?

Assistant: No, only Rose survives the sinking. Jack is not shown to survive the sinking, but he is later seen in a dream that Rose has.



LLaMA-VID

User: How did Cooper transmit data from the black hole to Murphy?

Assistant: Cooper transmitted data from the black hole to Murphy by using a watch that had been given to Murphy by Cooper.

User: Why does time pass slower on Miller's planet than on Earth? How slowly does it go?

Assistant: Time passes slower on Miller's planet because it is in orbit around a black hole, which has a strong gravitational pull. The stronger the gravitational pull, the slower time moves. On Miller's planet, time moves at about 7 years per hour, compared to Earth's 1 hour per hour. This is because the black hole's gravity is so strong that it distorts space-time, causing time to run more slowly.

LLaMA 2

User: How did Jack Dawson and Rose DeWitt Bukater meet each other in this movie?

Assistant: In the movie Titanic, Jack Dawson and Rose DeWitt Bukater met each other on the ship during its ill-fated maiden voyage. Rose, a high-society passenger, was being pursued by her fiancé, Cal Hockley, but she felt suffocated by their relationship and longed for something more authentic. One night, Jack, a penniless artist, won passage on the ship in a poker game and was able to enter the first-class section, where he met Rose on the deck. They struck up a conversation and quickly formed a connection, despite their different social backgrounds.

User: Does Jack Dawson and Rose DeWitt Bukater both survive the sinking of the Titanic?

Assistant: No, in the movie Titanic, Jack Dawson (played by Leonardo DiCaprio) does not survive the sinking of the Titanic. He dies in the icy waters of the North Atlantic after helping Rose DeWitt Bukater (played by Kate Winslet) escape on a life raft. Rose, on the other hand, does survive and is rescued by the crew of the RMS Carpathia.

LLaMA 2

User: How did Cooper transmit data from the black hole to Murphy?

Assistant: There is no character named Murphy in the movie Interstellar, and there is no scene where a character named Cooper transmits data from a black hole. The movie Interstellar is a work of science fiction and does not depict any real events or technology. Therefore, it is not possible for Cooper to transmit data from a black hole to Murphy, as there is no Murphy character in the movie and no such technology exists.

User: Why does time pass slower on Miller's planet than on Earth? How slowly does it go?

Assistant: In the movie Interstellar, time runs slower on Miller planet than on Earth due to the planet's strong gravitational pull. The stronger the gravitational force, the slower time passes, according to Einstein's theory of general relativity. On Miller planet, the gravitational force is much stronger than on Earth, causing time to slow down significantly. The exact slowing down of time on Miller planet is not explicitly stated in the movie, but it is implied to be around 1.5 to 2 times slower than on Earth. This means that for every 1 hour that passes on Earth, only around 45 minutes would pass on Miller planet. The more pronounced the gravitational force as you approach the planet's surface, making time appear to slow down even more.

LongLoRA

User: How did Jack Dawson and Rose DeWitt Bukater meet each other in this movie?

Assistant: Jack Dawson and Rose DeWitt Bukater met each other on the RMS Titanic during its maiden voyage in 1912. Rose was traveling first class and Jack was traveling third class. They were drawn together after Jack saved Rose from drowning in the ship's swimming pool. Their chance meeting and subsequent romance was cut short when the Titanic sank, killing Jack and separating Rose from her fiancé, Cal Hockley. However, Rose never forgot Jack and the time they spent together, carrying a flame of love for him for the rest of her life.

User: Does Jack Dawson and Rose DeWitt Bukater both survive the sinking of the Titanic?

Assistant: According to the story, Jack Dawson and Rose DeWitt Bukater both survive the sinking of the Titanic. After the ship hits the iceberg and begins to sink, Jack saves Rose from the freezing water by pulling her onto the ship's rail. They are then able to escape the sinking ship together, with Jack helping Rose onto a life boat. Though the ship is destroyed, Jack and Rose both live on through their love.

LongLoRA


User: How did Cooper transmit data from the black hole to Murphy?

Assistant: Cooper realizes that TARS, the robot, is the only one standing between him and being able to transmit data from the black hole to Murphy. TARS has the ability to transmit data through radio waves, so Cooper comes up with a plan to have TARS transmit the quantum data he has collected while inside the black hole. Cooper programs TARS to transmit the data in Morse code, which TARS is able to understand and decode. Through this method, Cooper is able to successfully transmit the data from the black hole to Murphy, who is able to receive and understand it.

User: Why does time pass slower on Miller's planet than on Earth? How slowly does it go?

Assistant: Miller planet orbits a black hole, which has a huge gravitational pull. This causes time to run slower on Miller planet compared to Earth. Specifically, time runs 7 years slower on Miller planet for every 1 year spent on Earth. The effect is more pronounced the closer one gets to the black hole. So time runs significantly slower on Miller planet, especially in the vicinity of the black hole.

Fig. 9: Comparisons with LLaMA 2 [12] and LongLoRA [3] given the same questions.




Single Image

User: What is unusual about this image?

Assistant: The unusual aspect of this image is that a man is standing on the back of a yellow taxi cab while holding a clothes iron. This is not a typical sight, as people usually do not use clothes irons while traveling in a vehicle. The man's actions suggest that he might be trying to iron clothes on the go, which is an unconventional and potentially dangerous way to handle such a task.

User: Can you explain this meme in detail?

Assistant: The meme features a man standing on the back of a yellow taxi cab while holding a clothes iron. This unusual scene is meant to be humorous, as it is not a typical sight to see someone using a clothes iron while traveling in a vehicle. The meme likely aims to entertain viewers by presenting an unexpected and absurd situation.



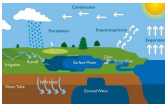
Single Image

User: Describe both the image and logo in details.

Assistant: The image features two cans of Coca-Cola placed on a surface, which appears to be a frozen surface, possibly ice or a cold surface. The cans are sitting next to each other, with one can on the left and the other on the right. The Coca-Cola logo is prominently displayed on both cans, indicating that they are part of the same brand. The logo is a well-known and recognizable symbol for the Coca-Cola brand, featuring a red and white color scheme with a white letter "C" and a red letter "O" on the front.

User: What's the difference between two bottles?


Assistant: The difference between the two bottles is that one is a can and the other is a bottle. Both bottles are of the same brand, Coca-Cola, and are placed next to each other on a frozen surface. However, the can is a more common and convenient packaging option for the brand, while the bottle is a traditional and iconic packaging option that has been used for many years. The can is generally lighter and more portable than the bottle, making it a popular choice for on-the-go consumption.



Single Image

User: Suppose you are a teacher, please use the figure to explain the distinction between evaporation and evapotranspiration.

Assistant: In the image, there is a diagram illustrating the water cycle, which includes precipitation, surface water, ground water, and evapotranspiration. The diagram shows the process of evapotranspiration, which is the combination of evaporation and transpiration. Evaporation is the process by which water vapor is released from the surface of water, such as a pond or a lake. This occurs when the water surface is exposed to the atmosphere, and the temperature difference between the water and the air is sufficient to create a vapor pressure difference. The water vapor then turns into a gas and rises into the air. Transpiration, on the other hand, is the process by which water vapor is released from the leaves of plants. This occurs when the plant's stomata open, allowing water vapor to escape into the atmosphere. The water vapor then cools and condenses, forming droplets of water on the surface of the leaf or the surrounding area. In the image, the diagram shows the process of evapotranspiration, which is the combination of these two processes. It is the process by which water is released from the surface of the water and the leaves of plants, contributing to the overall water cycle.



2-hour Video
Wonka

User: In the movie, what are the characteristics of Wonka's chocolate?

Assistant: Wonka's chocolate is unique and magical, with the ability to make its consumers fly and experience a variety of emotions.

User: In the movie, why does Wonka need to work at the laundromat?


Assistant: Wonka needs to work at the laundromat to earn enough money to pay for his room at the Scrubbitts' boarding house.

User: Why does Wonka love chocolate?

Assistant: Wonka loves chocolate because it represents his mother's love for him.

User: What rivals does Wonka face in the movie? And what is the name of the group formed by these people?


Assistant: Wonka faces three rivals: Slugworth, Ficklegrubber, and Prodnose. They form a group called The Chocolate Cartel.



Single Image

User: After seeing the below image, how people might emotionally feel and react?

Assistant: In the image, a man and a woman are embracing each other in front of a stage, with a crowd of people in the background. The couple appears to be enjoying a special moment together, possibly celebrating a significant event or sharing a heartfelt embrace. The presence of the stage suggests that there might be a performance or a celebration taking place. People in the crowd might feel a mix of emotions, such as happiness, excitement, or nostalgia, as they witness the couple's affectionate gesture. Some may also feel a sense of connection or camaraderie with the couple, as they share a similar experience in the same environment. Overall, the scene evokes a sense of warmth, love, and togetherness among the people present.



Single Image

User: After seeing the below image, how people might emotionally feel and react?

Assistant: In the image, a group of people is gathered around a man who is being held by police officers. The man appears to be resisting the officers, and there is a sense of tension and conflict in the scene. People might feel a mix of emotions, such as concern, fear, or sympathy for the man being held by the police. Some individuals might also feel a sense of injustice or outrage, depending on their personal beliefs and experiences. The situation could lead to various reactions, including protests, calls for justice, or discussions about the use of force by law enforcement. The emotional response to this image would depend on the viewer's perspective and their understanding of the context and circumstances surrounding the event.

Fig. 10: More examples of LLaMA-VID using Vicuna-7B based models. Images are sampled from [13].

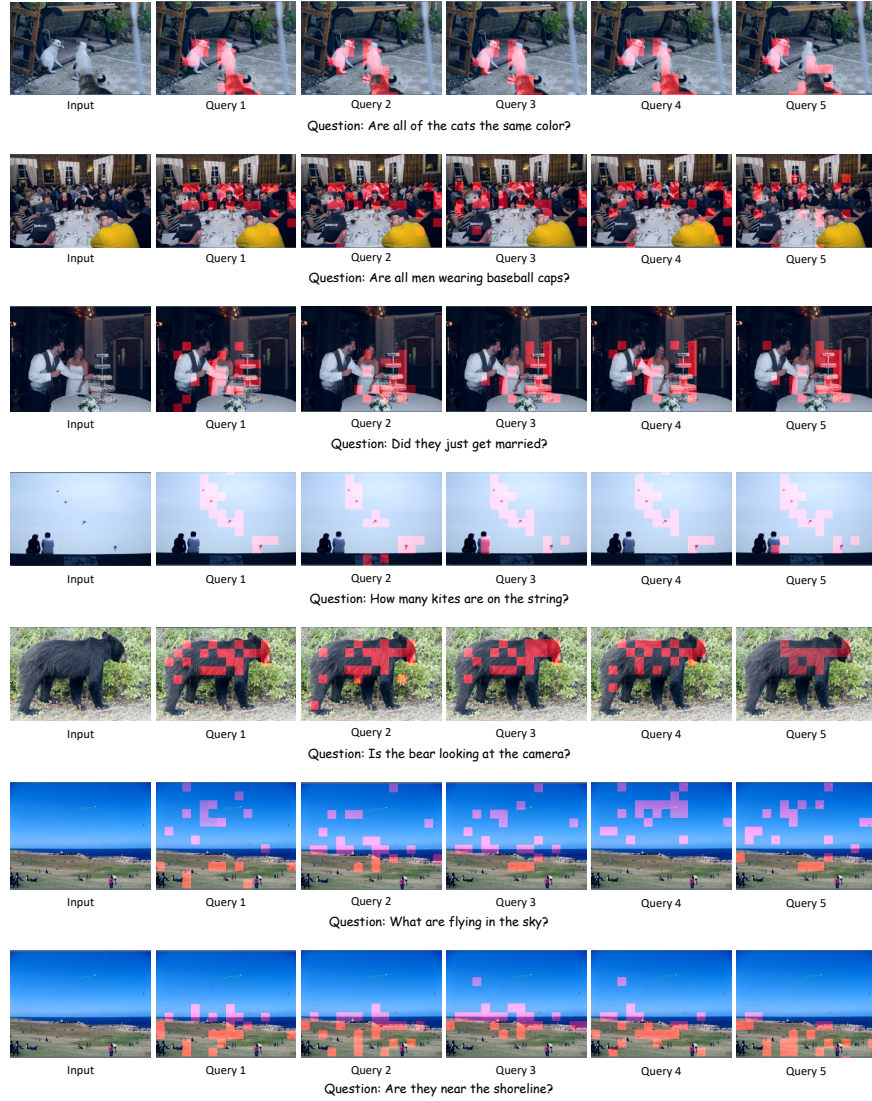


Fig. 11: High response areas with top scores to input question in Equation 1 of the main paper. We present the response of first five queries in Q_t . Images are randomly sampled from VQA V2 [4] *test-dev* set.

References

1. Anthropic: Claude 2. <https://www.anthropic.com/index/claude-2> (2023) **2**
2. Chen, S., Wong, S., Chen, L., Tian, Y.: Extending context window of large language models via positional interpolation. arXiv:2306.15595 (2023) **1**
3. Chen, Y., Qian, S., Tang, H., Lai, X., Liu, Z., Han, S., Jia, J.: Longlora: Efficient fine-tuning of long-context large language models. arXiv:2309.12307 (2023) **1, 2, 3, 7**
4. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017) **9**
5. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: ECCV (2020) **2**
6. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv:2305.06355 (2023) **2**
7. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv:2310.03744 (2023) **1**
8. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeruIPS (2023) **2**
9. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv:2306.05424 (2023) **2**
10. OpenAI: Gpt-4 technical report. arXiv:2303.08774 (2023) **2**
11. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. arXiv:2104.09864 (2021) **1**
12. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288 (2023) **3, 7**
13. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv:2309.17421 (2023) **8**