

A. More Examples

We showcase additional examples of video dialogues across various tasks, encompassing a creative task (Figure 1), a fine-grained understanding task (Figure 2), and a video reasoning task (Figure 3). In the creative task (Figure 1), our VTimeLLM demonstrates a remarkable capacity to comprehend visual information and subsequently craft a poem inspired by it. This achievement is attributed to we freeze the LLM at all three stages of training, thereby preserving its ability for engaging in creative dialogue. In the fine-grained understanding task (Figure 2), our VTimeLLM comprehends multiple events within the video, as well as the specific visual content within individual events. This demonstration underscores its proficiency in grasping temporal and spatial details, a capability attributed to our three-stage training strategy. In the video reasoning task (Figure 3), our VTimeLLM responds to several questions requiring inference, showing its capacity to engage in reasoning based on a comprehensive understanding of visual content.

B. Templates and Prompts

In Stage 2, we need to transform events $\{s_i, e_i, T_i\}$ into template-based QA, where s_i and e_i represent the start and end timestamps of a segment, ranging from 00 to 99. T_i corresponds to its textual description. For a given sequence of events, there is a 20% probability of transformation into single-turn QA, completing a dense caption task where all events are described within a single answer. Conversely, there is an 80% probability of transformation into Multi-turn QA. In this scenario, each event is individually queried and answered within a dialogue, in the form of two tasks, event captioning or temporal grounding. We provide 10 templates for each task, as shown in Box 3.

In Stage 3, we need to transform events into high-quality dialogue. This is accomplished by providing a prompt to a text-based LLM (Vicuna-7B v1.5). The prompt can be found in Box 4. In the prompt, specific timestamps are not provided because their inclusion does not enhance the LLM’s comprehension of temporal relationships. On the contrary, they may introduce errors into the dialogue. Consequently, events are presented in a sequential order, accompanied by specific symbols e.g., $\langle s1 \rangle$, $\langle e1 \rangle$ in the box, denoting the timestamps. The generated dialogue is expected to integrate temporal perception and reasoning.

C. Evaluation Process

In this section, we provide a detailed process on the evaluation of temporal grounding and dense captioning tasks for VTimeLLM and other Video LLMs.

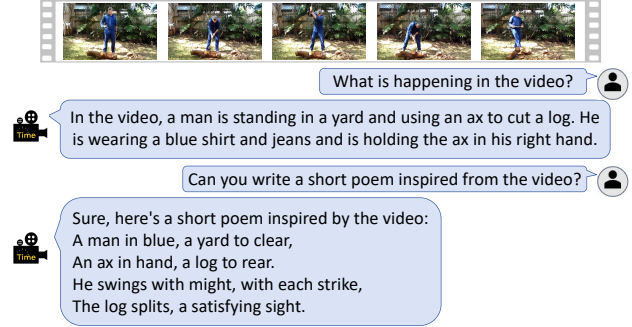


Figure 1. Creative Task. Our VTimeLLM comprehends the visual information and crafts a poem inspired by it.

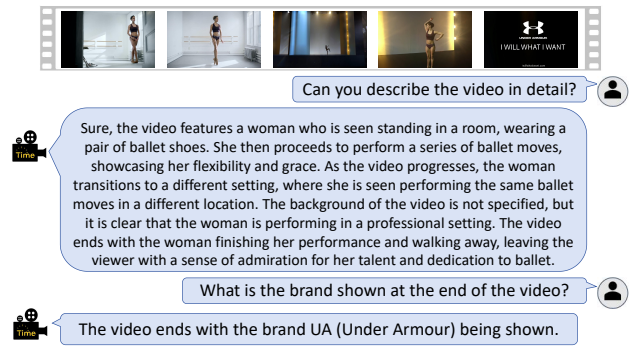


Figure 2. Fine-grained Understanding Task. Our VTimeLLM comprehends multiple events within the video, as well as the specific visual content within individual events.

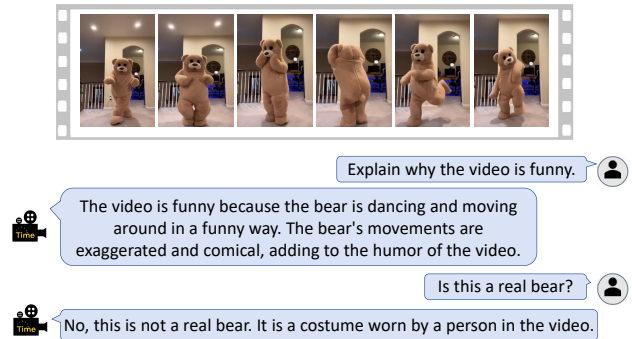


Figure 3. Video Reasoning Task. our VTimeLLM responds to several questions requiring reasoning based on a comprehensive understanding of visual content.

C.1. Evaluation of VTimeLLM

For VTimeLLM that has undergone only stages 1, 2 training without stage 3, the input and output formats remain entirely consistent with the template. Consequently, we can directly employ the templates in Box 3 as queries. Specifically, for the dense captioning task, we employ Q_{D1} , i.e.,

“Could you please detail the events that took place during different time segments in the video?” as the query. For the temporal grounding task, we employ queries Q_{T1} , Q_{T2} , and Q_{T3} to compute IoU for their respective outputs, and we report the average metrics. The performance obtained from different queries is similar.

VTimeLLM that has undergone stage 3 training demonstrate commendable instruction-following ability, and the performance may vary with different queries. For example, the inclusion of the phrase “in detail” in the query leads to a more detailed description of the video. For the dense captioning task, we utilize the following query: “Could you please describe the events in the video in detail? Be specific about the activities of individuals, their surroundings, and interactions with others. The output should be in JSON format, structured as follows: {‘event’: ‘xx’, ‘timestamps’: ‘from xx to xx’}.” We find that this query outperforms Q_{D1} across various metrics by approximately 10%. For the temporal grounding task, we continue to report the average results of queries Q_{T1} , Q_{T2} , and Q_{T3} , with metrics for each query remaining consistently close. Notably, even with the adoption of a simpler query such as “When does T_i happen?”, we achieve comparable results, underscoring the stability of outputs in this task.

C.2. Evaluation of other Video LLMs

For other Video LLMs (VideoLLaMA, VideoChat, and VideoChatGPT) that we test in our study, we try our best to assess their optimal performance as they were not trained on these tasks. Our testing methodology follows several principles: First, we include video duration D in the query. Second, as these models often fail to adhere to our prompt for outputting in JSON format, we apply multiple regular expressions to format the output. This successfully handles over 70% of the outputs. For these outputs cannot be processed, we exclude the corresponding data from metric calculations. Third, we design multiple queries and select the one yielding the best performance as the final result. For example, in our experiment, we find that the best query for VideoChatGPT in the dense captioning task is: “This video has a duration of D seconds. From which second to which second in the video, what event happens? Be specific about the activities of individuals, their surroundings, and interactions with others. List the events in the format: 1. From $x1$ second to $y1$ second: event1. \n 2. From $x2$ second to $y2$ second: event2.\n ...”

Box 3: Templates to transform events $\{s_i, e_i, T_i\}$ into QA dialogues, 10 templates for each task.

Dense Captioning (task of Single-turn QA):

- Q_{D1} : Could you please detail the events that took place during different time segments in the video?
 Q_{D2} : I'm curious about what happened at different points in the video. Could you please describe the events?
 Q_{D3} : Could you provide a summary of the incidents that occurred at various timestamps in the video?
 Q_{D4} : I'd like to know what events transpired during specific time intervals in the video. Could you please elaborate?
 Q_{D5} : Can you give me a breakdown of the occurrences at different time stamps in the video?
 Q_{D6} : I'm interested in understanding the events that unfolded at different points in the video. Could you please specify?
 Q_{D7} : Could you outline the incidents that happened during different time periods in the video?
 Q_{D8} : I'm trying to grasp the sequence of events in the video. Could you please outline what happened at different times?
 Q_{D9} : Can you go through the video and describe what took place at different time intervals?
 Q_{D10} : I'd appreciate it if you could provide a detailed account of the events that occurred at different timestamps in the video.
 A_D : T_1 , from s_1 to e_1 . T_2 , from s_2 to e_2 . T_3 , from s_3 to e_3

Event Captioning (One task in Multi-turn QA):

- Q_{E1} : Can you describe what occurred from s_i to e_i in the video?
 Q_{E2} : Could you tell me what happened from s_i to e_i in the video?
 Q_{E3} : What transpired from s_i to e_i in the video?
 Q_{E4} : Describe what took place from s_i to e_i in the video.
 Q_{E5} : Tell me about the events from s_i to e_i in the video.
 Q_{E6} : What was going on from s_i to e_i in the video?
 Q_{E7} : Please recount what occurred from s_i to e_i in the video.
 Q_{E8} : Explain what happened from s_i to e_i in the video.
 Q_{E9} : Provide details about the events from s_i to e_i in the video.
 Q_{E10} : Share what transpired from s_i to e_i in the video.
 A_E : T_i .

Temporal Grounding (One task in Multi-turn QA):

- Q_{T1} : During which frames can we see T_i happening in the video?
 Q_{T2} : Between which frames is T_i visible in the video?
 Q_{T3} : At what point in the video can we observe T_i taking place?
 Q_{T4} : Between which two frames can we witness T_i occurring in the video?
 Q_{T5} : During which frames in the video can we observe T_i happening?
 Q_{T6} : At which time interval in the video can we see T_i occurring?
 Q_{T7} : Between which frames can we find T_i taking place in the video?
 Q_{T8} : At what point in the video can we witness T_i happening?
 Q_{T9} : Between which two frames in the video can we observe T_i taking place?
 Q_{T10} : During which frames does T_i occur in the video?
 A_T : From s_i to e_i .

Box 4: Prompt for LLM in Stage 3, designed to transform a series of events into high-quality dialogue.

You are an AI visual assistant with the task of analyzing a single video. Craft a conversation between yourself and a user discussing the video's content. Develop responses that embody the persona of an active visual AI assistant, capable of observing the video and providing insightful answers. Include inquiries about temporal perception and reasoning, like events preceding or succeeding specific occurrences, or requesting timestamps for particular actions or events. Ensure that the questions can be definitively answered based on the observable video content or confidently ascertainable absence from the video. Utilize the timestamps $\langle s? \rangle$ and $\langle t? \rangle$ to create contextual questions considering the temporal relationships between events. The conversations should be concise.

Here's an illustrative example:

=== example start ===

Events:

from $\langle s1 \rangle$ to $\langle e1 \rangle$: A man and woman play rock paper scissors, the woman wins and smiles.

from $\langle s2 \rangle$ to $\langle e2 \rangle$: The woman puts a blindfold on.

from $\langle s3 \rangle$ to $\langle e3 \rangle$: The woman continues playing rock-paper-scissors with the man and wins again.

from $\langle s4 \rangle$ to $\langle e4 \rangle$: The woman gives the man a hug.

Dialogue:

User: Could you provide a brief overview of the video's content?

Assistant: Certainly! In the video, a man and a woman engage in a game of rock-paper-scissors. The woman emerges victorious and shares a smile. Subsequently, she places a blindfold on. She then proceeds to win another round of rock-paper-scissors against the man. The video concludes with the woman embracing the man warmly.

User: Can you pinpoint when the woman achieved victory in the game twice?

Assistant: Certainly. The first victory occurs from $\langle s1 \rangle$ to $\langle e1 \rangle$, while the second triumph takes place from $\langle s3 \rangle$ to $\langle e3 \rangle$.

User: I'm curious about the interaction between $\langle s4 \rangle$ and $\langle t4 \rangle$. Could you elaborate?

Assistant: Absolutely. During the interval from $\langle s4 \rangle$ to $\langle t4 \rangle$, the woman conveys her emotions through a heartfelt embrace, demonstrating her genuine affection for the man.

User: What might be the underlying reason for the woman's affectionate hug?

Assistant: The woman's affectionate hug likely stems from her desire to uplift the man's spirits after his loss in the rock-paper-scissors game.

=== example end ===

Events:

from $\langle s1 \rangle$ to $\langle e1 \rangle$: T_1 .

from $\langle s2 \rangle$ to $\langle e2 \rangle$: T_2 .

from $\langle s3 \rangle$ to $\langle e3 \rangle$: T_3 .

.....

Dialogue: