Language Instructed Temporal-Localization Assistant (Supplemenatry Materials)

1 Task Prompts

We discuss the training tasks for LITA in Section 3.4. We use task prompts to convert dense video captioning and event localization to natural language question answering. For both of the tasks, there are two components of the prompts: (1) task description and (2) timestamp instruction. The task description instructs the model about the task, and the timestamp instruction shows how to format timestamps in the response. For dense video captioning, we use the following task descriptions:

- "Provide a detailed description of the given video."
- "Describe the provided video in detail."
- "Summarize the visual content of the video."
- "Write a informative summary of the video."

For timestamp instructions we use:

- "Each sentence should begin with the start and end timestamps."
- "At the beginning of each sentence, include the start and end timestamps."
- "Prepend each sentence with its start and end timestamps."

We then randomly sample one task description and one timestamp instruction as the full task prompt. For event localization, we use the following task descriptions:

- "When does "`SENTENCE`" happen in the video?"
- "At what point in the video does "`SENTENCE`" happen?"
- "When is "`SENTENCE`" depicted in the video?"
- "At what time in the video does "`SENTENCE`" take place?"

Here, `SENTENCE` refers to sentence describing the event of interest. Similarly for timestamp instruction, we use:

- "Answer the question only using start and end timestamps."
- "Provide a response using only start and end timestamps."
- "Convey your answer using start and end timestamps exclusively."

2 Dataset Generation Prompt

We discuss our dataset curation in Section 4. We build our ActivityNet-RTL dataset from the ActivityNet-Captions dataset. We use annotated dense video captioning as context and ask GPT-4 to generate reasoning temporal localization questions. For our training split, the GPT-4 outputs are directly used. For our evaluation split, the GPT-4 outputs are manually processed. The prompt is shown in Table 1.

3 Explanation Evaluation Prompt

As discussed in Section 4.3, we evaluate the explanations using GPT-4's assistance. The prompt for evaluation is shown in Table 2. Most of the prompt follows LLaVA [24]. The main differences are changing "images" to "videos" and instructing GPT-4 to ignore timestamps in explanations.

4 Limitations

One limitation of LITA is its spatio-temporal resolution. For temporal resolution, we only uniformly sample T frames from a video, which limits our temporal localization resolution to $\frac{1}{T}$ of the video length. The fixed number of frames also limit our adaptability to videos of various length. For spatial resolution, our slow tokens perform spatial downsampling of visual tokens, which reduces the spatial information and can potentially hurt spatial reasoning. This is a trade-off for better temporal understanding. Nevertheless, we believe LITA provide a good starting point to advance temporal understanding, especially temporal localization, for Video LLMs.

5 Standard Video Tasks

We further conduct experiments on standard video tasks. This includes video question answering (Table 3) and event temporal localization (Table 4) on ActivityNet, and Temporal Grounding on Charades-STA (Table 5).

6 Applicability of GPT4V to RTL

. We apply GPT4V to samples in our ActivityNet-RTL dataset and observe that it does not give the desired answers for temporal localization. This is with reasonable prompts that explain the context for video temporal localization. We observe two shortcomings: First, GPT4V analyzes each frame individually when asked with a RTL question. The answer thus misses the overall video context. For example, GPT4V fails to capture the most relevant action "falling off the beam" in Figure 4 (right), but instead focus on how gymnastics is related to resilience for each frame. Second, GPT4V does not directly answer temporal localization questions. Even with prompts that specifies the length of videos, GPT4V still does not give the desired answer (*i.e.* timestamps of the event of interest).

Table 1: Prompt used to generate our ActivityNet-RTL dataset. The prompt starts with system message instructions, followed by few-shot examples. We use three examples in our data generation.

You are an AI visual assistant, and you are seeing a single video. What you see are provided with sentences describing the same video you are looking at. Each sentence begins with the start and end timestamps of the event described by the sentnece: ⟨start timestamp⟩ ⟨end timestamp⟩ SENTENCE. Answer all questions as you are seeing the video.

Design a conversation between you and a person asking about this video. The answers should be in a tone that a visual AI assistant is seeing the video and answering the question. Ask 2 to 5 diverse questions and give corresponding answers. Be assertive in the questions. Avoid words like: possibly, likely.

Only include questions asking about the temporal extent of events. Do not ask about events that are already described by the provided sentences. Do not ask binary questions. Do not ask for further description. Ask complex questions beyond the provided sentences.

To answer such complex questions, one should first understand the visual content of the video, then derive answers based on background knowledge or reasoning. The answer should again begin with start and end timestamps of the event, along with detailed explanations about the answer: ⟨start timestamp⟩ ⟨end timestamp⟩ EXPLANATION. The EXPLANATION should include the reasoning process. Use ⟨t⟩ when mentioning a timestamp t even in the explanation. Generate the answers and explanations as if you are seeing the video instead of basing it on the provided sentences. Do not refer to the provided description in the explanation. Only include questions that have definite answers.

Example 1

Provided Sentences
⟨0⟩ ⟨7.49⟩ We see a hallway with a wooden floor.
⟨7.49⟩ ⟨18.09⟩ A dog in socks walks slowly out onto the floor as a lady films him.
⟨19.37⟩ ⟨36.55⟩ The dog turns around and goes back to the other room.

Conversation
Question: When is the hallway empty?
Answer: ⟨0⟩ ⟨7.49⟩ The hallway is empty between ⟨0⟩ and ⟨7.49⟩. During this time, we only see a hallway with a wooden floor.
Question: When is the dog the sole focus of the video?
Answer: ⟨7.49⟩ ⟨18.09⟩ The dog, in socks, is the sole focus of the video between ⟨7.49⟩ and ⟨18.09⟩. During this time, the dog is featured walking slowly out onto the floor while a lady films him.

Example 2
...

Table 2: Prompt used to evaluate explanations in reasoning temporal localization. We instruct GPT-4 to ignore the accuracy of timestamps in the explanation since it is evaluated separately. Only the system message is included.

> We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above. The user asks the question on observing a video. For your reference, the visual content in the video is represented with a few sentences describing the same video. Each sentence begins with the start and end timestamps of the event described by the sentence: ⟨start timestamp⟩ ⟨end timestamp⟩ SENTENCE. Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.
>
> **IMPORTANT GUIDELINE** Timestamps in the responses are represented as ⟨t⟩ for timestamp t. These timestamps should not be considered in the ratings, as they are evaluated separately. Please provide ratings based solely on the content of the responses, excluding the timestamps. Correct identification of time periods from the context is irrelevant to the rating and should not be addressed. Adherence to this system guideline is crucial.
>
> Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Table 3: ActivityNet question answering results.

	VideoChat	Video-LLaMA-v2	Video-ChatGPT	LITA
Acc.	26.5	31.2	35.2	**43.0**
Score	2.2	2.5	2.7	**2.9**

Table 4: ActivityNet event localization. * indicates CLIP encoder without audio and HowTo100M training for a fair comparison.

	LITA	MFT [37]	Vid2Seq*	PDVC*
F1 Score	<u>47.1</u>	33.0	46.5	**53.9**

Table 5: Charades-STA temporal grounding results.

	VideoChat	Video-LLaMA-v2	Video-ChatGPT	LITA
R@0.5	3.3	3.8	7.7	**34.8**
R@0.7	1.3	0.9	1.7	**18.4**