A Simple LLM Framework for Long-Range Video Question-Answering

Anonymous ACL submission

Abstract

We present LLoVi, a simple yet effective **L**anguage-based **Lo**ng-range **Vi**deo question-answering (LVQA) framework. Our method decomposes short and long-range modeling aspects of LVQA into two stages. First, we use a short-term visual captioner to generate textual descriptions of short video clips (0.5-8s in length) densely sampled from a long input video. Afterward, an LLM aggregates the densely extracted short-term captions to answer a given question. Furthermore, we propose a novel multi-round summarization prompt that asks the LLM first to summarize the noisy short-term visual captions and then answer a given input question. To analyze what makes our simple framework so effective, we thoroughly evaluate various components of our framework. Our empirical analysis reveals that the choice of the visual captioner and LLM is critical for good LVQA performance. The proposed multi-round summarization prompt also leads to a significant LVQA performance boost. Our method achieves the best-reported results on the EgoSchema dataset, best known for very long-form video question-answering. LLoVi also outperforms the previous state-of-the-art by 4.1% and 3.1% on NExT-QA and IntentQA. Finally, we extend LLoVi to grounded VideoQA which requires both QA and temporal localization, and show that it outperforms all prior methods on NExT-GQA. We will release our code.

1 Introduction

Recent years have witnessed remarkable progress in short video understanding (5-15s in length) (Wang et al., 2022a; Ye et al., 2023; Fu et al., 2021; Yang et al., 2022a; Wang et al., 2023g). However, extending these models to long videos (e.g., several minutes or hours in length) is not trivial due to the need for sophisticated long-range temporal reasoning capabilities. Most existing long-range video models rely on costly and complex long-range temporal modeling schemes, which include memory queues (Wu et al., 2022; Chen et al., 2020; Lee et al., 2021, 2018), long-range feature banks (Wu et al., 2019; Cheng and Bertasius, 2022; Zhang et al., 2021), space-time graphs (Hussein et al., 2019b; Wang et al., 2021), state-space layers (Islam and Bertasius, 2022; Islam et al., 2023; Wang et al., 2023a) and other complex long-range modeling modules (Hussein et al., 2019a; Bertasius et al., 2021; Yang et al., 2023).

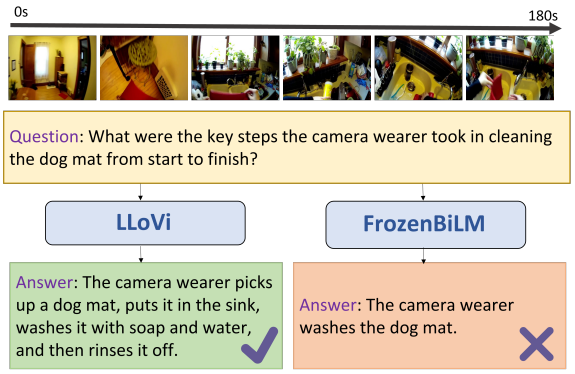


Figure 1: Comparison between LLoVi (ours) and the recent FrozenBiLM (Yang et al., 2022a) video QA method. Like most prior methods, FrozenBiLM is best suited for short-range video understanding. Thus, as illustrated in the figure, it fails to answer a question that requires reasoning about complex human activities in a long video. In comparison, our method effectively reasons over long temporal extents and produces a correct answer.

Recently, Large Language Models (LLMs) have shown impressive capability for long-range reasoning on a wide range of tasks such as document understanding (Sun et al., 2023; Wang et al., 2023e; Gur et al., 2023) and long-horizon planning (Liu et al., 2023a; Hao et al., 2023; Song et al., 2023a). Motivated by these results in the natural language and decision-making domain, we explore using LLMs for long-range video question answering

(LVQA). Specifically, we propose LLoVi, a simple yet effective language-based framework for long-range video understanding. Unlike prior long-range video models, our approach does not require specialized long-range video modules (e.g., memory queues, state-space layers, etc.) but instead uses a short-term visual captioner coupled with an LLM, thus exploiting the long-range temporal reasoning ability of LLMs. Our simple two-stage framework tackles the LVQA task by decomposing it into short and long-range modeling subproblems:

1. First, given a long video input, we segment it into multiple short clips and convert them into short textual descriptions using a pre-trained frame/clip-level visual captioner (e.g., BLIP2 (Li et al., 2023c), LaViLa (Zhao et al., 2023), LLaVa (Liu et al., 2023b)).
2. Afterwards, we concatenate the temporally ordered captions from Step 1 and feed them into an LLM (e.g., GPT-3.5, GPT-4, LLaMA) to perform long-range reasoning for LVQA.

To further enhance the effectiveness of our framework, we also introduce a novel multi-round summarization prompt that asks the LLM first to summarize the short-term visual captions and then answer a given question based on the LLM-generated video summary. Since the generated captions may be noisy or redundant, such a summarization scheme enables filtering out potentially distracting/irrelevant information and eliminating redundant sentences, which significantly improves the reasoning ability of the LLM for LVQA.

We also conduct an empirical study to investigate the factors behind our framework's success. Specifically, we study (i) the selection of a visual captioner, (ii) the choice of an LLM, (iii) the LLM prompt design, (iv) few-shot in-context learning, (v) optimal video processing configurations (i.e., clip length, sampling rate, etc.), and (vi) the generalization of our framework to other datasets and tasks. Our key empirical findings include:

- The multi-round summarization prompt leads to the most significant boost in performance (**+5.8%**) among the prompts we have tried (e.g., zero-shot CoT, Self-Consistency).
- GPT-4 as an LLM provides the best performance, while GPT-3.5 provides the best trade-off between the accuracy and the cost.
- LaViLa (Zhao et al., 2023) as a visual captioner produces best results (**51.8%**) followed by BLIP-2 (Li et al., 2023c) (**46.7%**) and EgoVLP (Qinghong Lin et al., 2022) (**46.6%**).
- Few-shot in-context learning leads to a large improvement on both the variant of our model with a standard prompt (**+4.7%**) and our best-performing variant with our proposed multi-round summarization prompt (**+4.1%**).
- Extracting visual captions from consecutive 1-second video clips of the long video input leads to the best results.
- LLoVi outperforms all prior approaches on EgoSchema, NeXT-QA, IntentQA and NeXT-GQA LVQA benchmarks.

Overall, our framework is simple, effective and training-free. Furthermore, it is agnostic to the exact choice of a visual captioner and an LLM, which allows it to benefit from future improvements in visual captioning and LLM model design. We hope that our work will encourage new ideas and a simpler model design in LVQA. We will release our code to enable the community to build on our work.

2 Related Work

Long-range Video Understanding. Modeling long-range videos (e.g., several minutes or longer) typically requires models with sophisticated temporal modeling capabilities, often leading to complex model design. LF-VILA (Sun et al., 2022) proposes a Temporal Window Attention (HTWA) mechanism to capture long-range dependency in long-form video. MeMViT (Wu et al., 2022) and MovieChat (Song et al., 2023b) adopt a memory-based design to store information from previously processed video segments. Several prior methods use space-time graphs (Hussein et al., 2019b; Wang et al., 2021) or relational space-time modules (Yang et al., 2023) to capture spatiotemporal dependencies in long videos. Lastly, the recently introduced S4ND (Nguyen et al., 2022), ViS4mer (Islam and Bertasius, 2022) and S5 (Wang et al., 2023a) use Structured State-Space Sequence (S4) (Gu et al., 2021) layers to capture long-range dependencies in the video. Unlike these prior approaches, we do not use any complex long-range temporal modeling modules but instead develop a simple and strong LLM-based framework for zero-shot LVQA.

LLMs for Video Understanding. The recent surge in large language models (LLMs) (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023; Raffel et al., 2020; Chung et al., 2022; Tay et al., 2022) has inspired many LLM-based applications in video understanding. Methods like Socratic

Models (Zeng et al., 2022) and VideoChat (Li et al., 2023e) integrate pretrained visual models with LLMs for extracting visual concepts and applying them to video tasks. Video ChatCaptioner (Chen et al., 2023) and ChatVideo (Wang et al., 2023b) leverage LLMs for video representation and dialog-based user interaction, respectively. VidIL (Wang et al., 2022b) employs LLMs for adapting image-level models to video tasks using few-shot learning. Beyond short-term video understanding, the works in (Lin et al., 2023a; Chung and Yu, 2023; Bhattacharya et al., 2023) explored LLMs for long-range video modeling. The work in (Lin et al., 2023a) uses GPT-4 for various long-range video modeling tasks but lacks quantitative evaluation. Meanwhile, (Chung and Yu, 2023) focuses on movie datasets, requiring limited visual analysis (Mangalam et al., 2023) and mostly relying on non-visual speech/subtitle inputs. In contrast to these prior methods, we focus on the LVQA task and provide an extensive empirical analysis of various design choices behind our LLM framework.

Video Question Answering. Unlike image question-answering, video question-answering (VidQA) presents unique challenges, requiring both spatial and temporal reasoning. Most existing VidQA methods, either using pretraining-finetuning paradigms (Cheng et al., 2023; Lei et al., 2021; Yu et al., 2023), zero-shot (Yang et al., 2022b; Surís et al., 2023; Lin et al., 2023b; Yu et al., 2023), or few-shot learning (Wang et al., 2022b), focus on short-term video analysis (5-30s). To overcome the limitations of short-term VidQA, new benchmarks have been proposed: ActivityNet-QA (Yu et al., 2019), TVQA (Lei et al., 2018), How2QA (Yang et al., 2021), MovieQA (Tapaswi et al., 2016), and DramaQA (Choi et al., 2021) ranging from 100s to several minutes in video duration. Despite longer video lengths, the analysis in (Mangalam et al., 2023; Yang et al., 2020; Jasani et al., 2019) found that many of these benchmarks can be solved by analyzing only short clips (i.e., not requiring long-range video modeling) or by using pure text-only methods that ignore visual content. To address these issues, the EgoSchema benchmark (Mangalam et al., 2023) was recently introduced, requiring at least 100 seconds of video analysis and not exhibiting language-based biases.

LLM Prompt Design. With the emergence of LLMs, there has been an increasing research emphasis on LLM prompt design. The recent works in (Wei et al., 2022; Zhou et al., 2023; Schick and Schütze, 2020; Chen et al., 2022; Yao et al., 2022) explored prompting strategy in few-shot learning settings. To eliminate the need for extensive human annotations, (Kojima et al., 2022; Wang et al., 2023c,f) proposed zero-shot prompting methods. Subsequent research (Zhou et al., 2022; Zhang et al., 2022; Pryzant et al., 2023) has concentrated on the automatic refinement of prompts. Instead, we propose a multi-round summarization LLM prompt for handling long, noisy, and redundant textual inputs describing video content for LVQA.

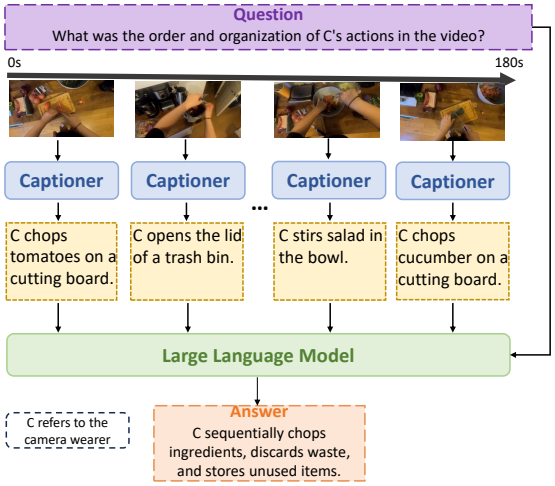


Figure 2: An illustration of LLoVi, our simple LLM framework for long-range video question-answering (LVQA). We use Large Language Models (LLMs) like GPT-3.5 and GPT-4 for their long-range modeling capabilities. Our method involves two stages: first, we use short-term visual captioners (e.g, LaViLa, BLIP2) to generate textual descriptions for brief video clips (0.5s-8s). Then, an LLM aggregates these dense, short-term captions for long-range reasoning required for LVQA. This simple approach yields impressive results, demonstrating LLMs' effectiveness in LVQA.

3 Method

Our method, named LLoVi, consists of two stages: 1) short-term video clip captioning and 2) long-range text-based video understanding using an LLM. Figure 2 presents a detailed illustration of our high-level approach. Below, we provide more details about each component of our framework.

3.1 Short-term Video Clip Captioning

Given a long untrimmed video input V, we first segment it into N_v non-overlapping short video clips $v = \{v_m\}_{m=1}^{N_v}$, where $v_m \in \mathbb{R}^{T_v \times H \times W \times 3}$ and T_v, H, W are the number of frames, height

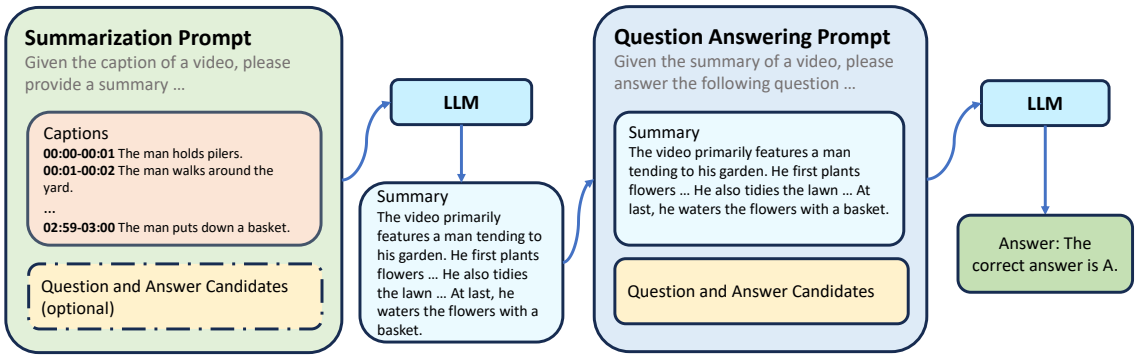


Figure 3: An illustration of our multi-round summarization prompt that first asks an LLM to summarize the noisy short-term visual captions (first round of prompting) and then answer a given question about the video based on the LLM-generated summary (second round of prompting). Our results indicate that such a multi-round prompting strategy significantly boosts LVQA performance compared to standard prompting techniques (**+5.8%**).

and width of a short video clip respectively. Afterward, we feed each video clip v_m into a pre-trained short-term visual captioner ϕ, which produces textual captions $c_m = \phi(v_m)$, where $c_m = (w_1, \ldots, w_{L_m})$ and w_i represents the i-th word in caption c_m of length L_m. Note that our model is not restricted to any specific visual captioning model. Our experimental section demonstrates that we can incorporate various video (LaViLa (Zhao et al., 2023), EgoVLP (Qinghong Lin et al., 2022), and image (BLIP-2 (Li et al., 2023d)) captioning models. Next, we describe how our extracted short-term captions are processed by an LLM.

3.2 Long-range Reasoning with an LLM

We want to leverage foundational LLMs for holistic long-range video understanding. Formally, given short-term visual captions $\{c_m\}_{m=1}^{N_v}$ for all N_v short video clips, we first concatenate the clip captions into the full video captions $C = [c_1, \ldots, c_{N_v}]$ in the same order as the captions appear in the original video. Afterward, the concatenated video captions C are fed into an LLM for long-range video reasoning. Specifically, given the concatenated video captions C, the question Q, and the answer candidates A, we prompt the LLM to select the correct answer using the following prompt template: *"Please provide a single-letter answer (A, B, C, D, E) to the following multiple-choice question* $\{Q\}$*. You are given language descriptions of a video. Here are the descriptions:* $\{C\}$*. Here are the choices* $\{A\}$*."*. The full prompt is included in the Supplementary Material.

Our experiments in Section 4.3 suggest that this simple approach works surprisingly well for LVQA. However, we also discovered that many modern LLMs (e.g., GPT-3.5, LLaMA) may struggle when provided with long (>1K words), noisy, and potentially redundant/irrelevant caption sequences. To address these issues, we investigate more specialized LLM prompts that ask an LLM first to summarize the noisy short-term visual captions (first round of prompting) and then answer a given question about the video (second round of prompting). Specifically, we formulate such a multi-round prompt as follows: given the video captions C, the question Q, and the answer candidates A, instead of directly feeding the $\{C, Q, A\}$ triplet into LLM for LVQA, we first ask the LLM to provide a summary of the captions in the first round, which we denote as S using the following prompt template: "*You are given language descriptions of a video: {C}. Please give me a {*N_w*} word summary.*" N_w denotes the desired number of words in the summary S. Afterward, during the second round of prompting, instead of using the captions C, we use the summary S as input for the LLM to select one of the answer candidates. Conceptually, such a prompting scheme is beneficial, as the LLM-generated summary S filters out irrelevant/noisy information from the initial set of captions C, making LLM inputs for the subsequent QA process more succinct and cleaner. A detailed illustration of our multi-round prompt is shown in Figure 3.

3.3 Implementation Details

For the experiments on EgoSchema, we use LaViLa (Zhao et al., 2023) as our captioner. We segment each video into multiple 1s clips with a stride of 1s, resulting in a list of consecutive clips that

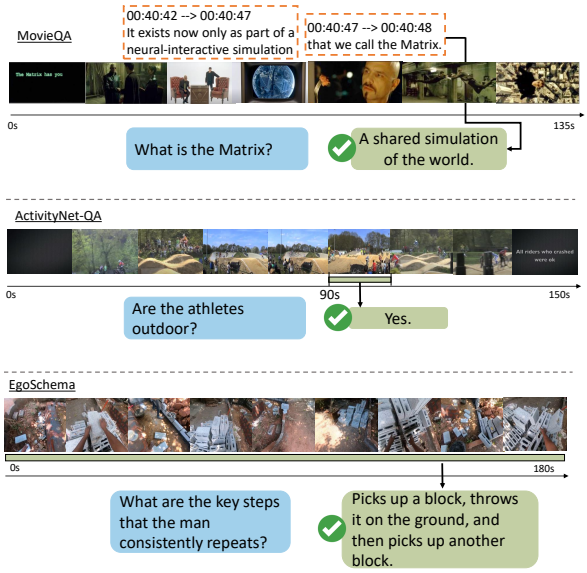


Figure 4: **An illustration of prior LVQA dataset limitations. Top**: An example from MovieQA (Tapaswi et al., 2016). The model can use the provided subtitle information to answer a question while ignoring visual cues in a video. **Middle**: An example from the ActivityNet-QA Dataset (Yu et al., 2019). Despite long video inputs, the model only needs to analyze a short 1s video clip to answer the question. **Bottom**: An example from the EgoSchema Dataset (Mangalam et al., 2023). The model must analyze visual cues from the video to answer a given question without relying on additional textual inputs (e.g., speech, subtitles).

cover the entire video. We use GPT-3.5 as the LLM on EgoSchema. For NeXT-QA, IntentQA, and NeXT-GQA, we use LLaVA-1.5 (Liu et al., 2023b) as the visual captioner and GPT-4 as the LLM. We downsample the videos to 0.5 FPS and prompt LLaVA to generate captions with roughly 30 words for each frame. More details are provided in the Supplementary Material.

4 Experiments

4.1 Datasets and Metrics

Unlike short-term video question-answering, long-range video question-answering (LVQA) lacks robust and universally agreed-upon benchmarks. As shown in Figure 4, many prior LVQA benchmarks either exhibit significant language biases, or do not require long-range video modeling capabilities. To address these limitations, recent work introduced **EgoSchema** (Mangalam et al., 2023), a new long-range video question-answering benchmark consisting of 5K multiple choice question-answer pairs spanning 250 hours of video and covering a wide range of human activities. By default, our experiments are conducted on the validation set of 500 questions (referred to as the EgoSchema Subset). The final comparison is done on the full test set of 5K EgoSchema questions. We use QA accuracy (i.e., the percentage of correctly answered questions) as our evaluation metric. Additionally, we also perform zero-shot LVQA experiments on three commonly-used LVQA benchmarks: **NExT-QA** (Xiao et al., 2021), **IntentQA** (Li et al., 2023a), and **NExT-GQA** (Xiao et al., 2023). Detailed dataset information and metrics can be found in the supplementary material.

Captioner	Caption Type	Ego4D Pre-training	Acc. (%)
VideoBLIP (Yu)	clip-level	✓	40.0
EgoVLP (Qinghong Lin et al., 2022)	clip-level	✓	46.6
BLIP-2 (Li et al., 2023d)	frame-level	✗	46.7
LaViLa (Zhao et al., 2023)	clip-level	✓	**51.8**
Oracle	clip-level	-	65.8

Table 1: **Accuracy of our framework with different visual captioners.** LaViLa visual captioner achieves the best results, outperforming other clip-level (e.g., EgoVLP, VideoBLIP) and image-level (e.g., BLIP-2) captioners. We also observe that the Oracle baseline using ground truth captions greatly outperforms all other variants, suggesting that our framework can benefit from the future development of visual captioners.

4.2 Empirical Study on EgoSchema

Before presenting our main results, we first study the effectiveness of different components within our LLoVi framework, including (i) the visual captioner, (ii) the LLM, (iii) the LLM prompt design, and (iv) few-shot in-context learning. The experiments are conducted on the EgoSchema Subset with 500 multi-choice questions. We discuss our empirical findings below. We also include additional experiments in the supplementary material.

4.2.1 Visual Captioning Model

In Table 1, we study the effectiveness of various clip-level video captioners, including LaViLa (Zhao et al., 2023), EgoVLP (Qinghong Lin et al., 2022), and VideoBLIP (Yu). In addition to video captioners, we also try the state-of-the-art image captioner, BLIP-2 (Li et al., 2023c). Lastly, to study the upper bound of our visual captioning results, we include the ground truth Oracle captioning baseline obtained from the Ego4D dataset. All baselines in Table 1 use similar experimental settings, including the same LLM model, i.e., GPT-3.5. The results are reported as LVQA accuracy on

LLM	Model Size	Acc. (%)
Llama2-7B (Touvron et al., 2023)	7B	34.0
Llama2-13B (Touvron et al., 2023)	13B	40.4
Llama2-70B (Touvron et al., 2023)	70B	50.6
GPT-3.5 (Brown et al., 2020)	175B	51.8
GPT-4 (OpenAI, 2023)	N/A	**58.3**

Table 2: **Accuracy of our framework with different LLMs.** GPT-4 achieves the best accuracy, suggesting that stronger LLMs perform better in LVQA. However, we use GPT-3.5 for most of our experiments due to the best accuracy and cost tradeoff.

Prompt Type	Standard	(C) → S	(C, Q) → S	(C, Q, A) → S
Acc. (%)	51.8	53.6	**57.6**	55.9

Table 3: **Different variants of our multi-round summarization prompt.** Our results indicate that the (C, Q) → S variant that takes concatenated captions C and a question Q for generating a summary S works the best, significantly outperforming (**+5.8%**) the standard prompt. This confirms our hypothesis that additional inputs in the form of a question Q enable the LLM to generate a summary S tailored to a given question Q.

the EgoSchema Subset.

The results in Table 1, suggest that LaViLa provides the best results, outperforming BLIP-2, EgoVLP, and VideoBLIP. We also observe that despite not being pre-trained on Ego4D (Grauman et al., 2022), BLIP-2 performs reasonably well (**46.7%**) and even outperforms other strong Ego4D-pretrained baselines, EgoVLP and VideoBLIP. Lastly, the Oracle baseline with ground truth captions outperforms LaViLa captions by a large margin (**14.0%**). This shows that our method can benefit from future improvements in captioning models.

4.2.2 Large Language Model

In Table 2, we analyze the performance of our framework using different LLMs while fixing the visual captioner to be LaViLa. Our results indicate that GPT-4 achieves the best performance (**58.3%**), followed by GPT-3.5 (**51.8%**). Thus, stronger LLMs (GPT-4) are better at long-range modeling, as indicated by a significant margin in LVQA accuracy between GPT-4 and all other LLMs (>**6.5%**). We also note that Llama2 performs reasonably well with its 70B variant (**50.6%**), but its performance drastically degrades with smaller capacity LLMs (i.e., Llama2-7B, Llama2-13B). Due to the tradeoff between accuracy and cost, we use GPT-3.5 for most of our experiments unless noted otherwise.

4.2.3 LLM Prompt Analysis

In this section, we (1) analyze several variants of our summarization-based prompt (described in Section 3), and (2) experiment with other commonly used prompt designs, including Zero-shot Chain-of-Thought (Zero-shot CoT) (Wei et al., 2022), Plan-and-Solve (Wang et al., 2023c), and Self-Consistency (Wang et al., 2023f). Below, we present a detailed analysis of these results.

Multi-round Summarization Prompt. Given a concatenated set of captions C, an input question Q, and a set of candidate answers A, we can use several input combinations to obtain the summary S. Thus, here, we investigate three distinct variants of obtaining summaries S:

- (C) → S: the LLM uses caption-only inputs C to obtain summaries S in the first round of prompting.
- (C, Q) → S: the LLM uses captions C and a question Q as inputs for generating summaries S. Having additional question inputs is beneficial as it allows the LLM to generate a summary S specifically tailored for answering an input question Q.
- (C, Q, A) → S: the LLM takes captions C, a question Q, and the answer candidates A as its inputs to produce summaries S. Having additional answer candidate inputs enables the LLM to generate a summary S most tailored to particular question-answer pairs.

In Table 3, we explore the effectiveness of these three prompt variants. Our results show that all three variants significantly outperform our standard LVQA prompt (described in Section 3). Specifically, we note that the variant (C) → S that uses caption-only inputs to obtain the summaries outperforms the standard baseline by **1.8%**. Furthermore, we observe that incorporating a given question as an input (i.e., the (C, Q) → S variant) leads to the best performance (**57.6%**) with a significant **5.8%** boost over the standard LVQA prompt baseline. This confirms our earlier intuition that having additional question Q inputs enables the LLM to generate a summary S specifically tailored for answering that question, thus leading to a big boost in LVQA performance. Lastly, we observe that adding answer candidates A as additional inputs (i.e., the (C, Q, A) → S variant) leads to a drop in performance (**-1.7%**) compared with the (C, Q) → S variant. This might be because the wrong answers in the candidate set A may mislead the LLM, leading to a suboptimal summary S.

We also investigate the optimal length of the

Number of words	50	100	300	500	700
Acc. (%)	55.6	57.4	55.8	**57.6**	55.0

Table 4: **Number of words in a generated summary.** We study the optimal number of words in an LLM-generated summary. These results suggest that the optimal LVQA performance is obtained when using 500-word summaries.

Prompting Technique	Acc. (%)
Zero-shot	
Standard	51.8
Zero-shot Chain-of-Thought (Wei et al., 2022)	53.2
Plan-and-Solve (Wang et al., 2023c)	54.2
Self-Consistency (Wang et al., 2023f)	55.4
Ours	**57.6**
Few-shot	
Standard	56.5
Ours	**61.7**

Table 5: **Comparison with commonly used prompting techniques.** The "Standard" means a standard LVQA prompt (see Section 3). We show that our framework benefits from more sophisticated prompting techniques. Our multi-round summarization prompt performs best in both zero-shot and few-shot learning settings.

generated summary S, and present these results in Table 4. Specifically, for these experiments, we ask the LLM to generate a summary S using a different number of words (as part of our prompt). We use the best performing (C, Q) $\rightarrow$ S variant for these experiments. Our results indicate that using a very small number of words (e.g., 50) leads to a drop in performance, indicating that compressing the caption information too much hurts the subsequent LVQA performance. Similarly, generating summaries that are quite long (e.g., 700 words) also leads to worse results, suggesting that the filtering of the potentially noisy/redundant information in the captions is important for good LVQA performance. The best performance is obtained using 500-word summaries.

Comparison with Commonly Used Prompts. Next, in Table 5, we compare our multi-round summarization prompt with other commonly used prompts such as Zero-shot Chain-of-Thought (Wei et al., 2022), Plan-and-Solve (Wang et al., 2023c), and Self-Consistency (Wang et al., 2023f). These results show that all of these prompts outperform the base variant of our model that uses a standard prompt. In particular, among these commonly used prompts, the self-consistency prompting technique achieves the best results (**55.4%**). Nevertheless, our multi-round summarization prompt performs best (**57.6%**).

Model	Acc. (%)
Zero-shot	
FrozenBiLM (Yang et al., 2022a)	26.9
mPLUG-Owl (Ye et al., 2023)	31.1
InternVideo (Wang et al., 2022a)	32.1
LongViViT (Papalampidi et al., 2023)	33.3
Vamos (Wang et al., 2023d)	48.3
LLoVi (Ours)	**50.3**
Few-shot	
LLoVi (Ours)	52.5

Table 6: **Results on the full set of EgoSchema.** The best-performing zero-shot variant of our LLoVi framework achieves **50.3%** accuracy, outperforming the previous best-performing InternVideo model by **18.2%**. For fair comparisons, we gray out our best few-shot variant.

4.2.4 Few-shot In-Context Learning

In-context learning with LLMs has shown strong few-shot performance in many NLP tasks (Brown et al., 2020; Wei et al., 2022). In Table 5, we evaluate the few-shot in-context learning capabilities of our LLoVi framework. Our results show that our LLoVi framework greatly benefits from few-shot in-context learning. Specifically, the few-shot in-context learning leads to a **4.7%** boost on the variant of our framework that uses a standard prompt and **4.1%** boost on our advanced framework using a multi-round summarization prompt. We used 6 few-shot examples as we found this configuration to produce the best performance.

4.3 Main Results on EgoSchema

In Table 6, we evaluate our best-performing LLoVi framework on the full EgoSchema test set containing 5K video samples. We compare our approach with prior state-of-the-art methods including InternVideo (Wang et al., 2022a), mPLUG-Owl (Ye et al., 2023), FrozenBiLM (Yang et al., 2022a), as well as the concurrent works of LongViViT (Papalampidi et al., 2023), and Vamos (Wang et al., 2023d). Based on these results, we observe that the best-performing zero-shot variant of our LLoVi framework achieves **50.3%** accuracy, outperforming the concurrent Vamos model (**+2.0%**). Additionally, we show that by using few-shot in-context learning, our best variant improves even further. These results validate our design choice of using the long-range modeling abilities of LLMs for LVQA. Furthermore, since our proposed LLoVi framework is agnostic to the visual caption-

Model	Cau. (%)	Tem. (%)	Des. (%)	All (%)
VFC (Momeni et al., 2023)	45.4	51.6	64.1	51.5
InternVideo (Wang et al., 2022a)	43.4	48.0	65.1	49.1
ViperGPT (Surís et al., 2023)	-	-	-	60.0
SeViLA (Yu et al., 2023)	61.3	**61.5**	**75.6**	63.6
LLoVi (ours)	**69.5**	61.0	**75.6**	**67.7**

Table 7: **Zero-shot results on NeXT-QA.** LLoVi achieves **67.7%** accuracy, outperforming previous best-performing model SeViLA by **4.1%**. Notably, LLoVi excels at causal reasoning outperforming SeViLA by **8.2%** in the causal question category.

ing model and an LLM it uses, we believe we could further improve these results by leveraging more powerful visual captioners and LLMs.

4.4 Results on Other Datasets

Next, we demonstrate that our simple framework generalizes well to other LVQA benchmarks.

NExT-QA. In Table 7, we evaluate LLoVi on the NExT-QA (Xiao et al., 2021) validation set in a zero-shot setting. We compare our approach with prior methods: VFC (Momeni et al., 2023), InternVideo (Wang et al., 2022a), ViperGPT (Surís et al., 2023), and SeViLA (Yu et al., 2023). We observe that LLoVi outperforms the previous best-performing method, SeViLA by **4.1%**. Notably, in the Causal category, LLoVi achieves **8.2%** improvement. We conjecture this improvement comes from the simple 2-stage design of our LLoVi framework: captioning followed by LLM reasoning. By captioning the video, we are able to directly leverage the reasoning ability of the powerful LLMs and thus achieve good causal reasoning performance.

IntentQA. In Table 8, we evaluate our method on the IntentQA (Li et al., 2023a) test set. In our comparisons, we include several supervised methods (HQGA (Xiao et al., 2022a), VGT (Xiao et al., 2022b), BlindGPT (Ouyang et al., 2022), CaVIR (Li et al., 2023b)) and the recent state-of-the-art zero-shot approach, SeViLA. From the results in Table 8, we observe that our method greatly outperforms all prior approaches, both in the fully supervised and zero-shot settings.

NExT-GQA. In Table 9, we extend our framework to the grounded LVQA task and evaluate it on the NExT-GQA (Xiao et al., 2023) test set. We compare LLoVi with the weakly-supervised methods: IGV (Li et al., 2022), Temp[CLIP](NG+) (Xiao et al., 2023), FrozenBiLM (NG+) (Xiao et al., 2023) and SeViLA (Yu et al., 2023). These baselines are first trained on NExT-GQA to maximize the QA accuracy, and then use ad-hoc methods (Xiao et al., 2023) to estimate a relevant video segment for question-answering. Although LLoVi is not trained on NExT-GQA, it still outperforms these weakly-supervised methods by a large margin according to all evaluation metrics. These results demonstrate that our framework can be used to temporally ground its predictions for more explainable long-range video understanding.

Model	Acc. (%)
Supervised	
HQGA (Xiao et al., 2022a)	47.7
VGT (Xiao et al., 2022b)	51.3
BlindGPT (Ouyang et al., 2022)	51.6
CaVIR (Li et al., 2023b)	57.6
Zero-shot	
SeViLA (Yu et al., 2023)	60.9
LLoVi (ours)	**64.0**

Table 8: **Results on IntentQA.** Our zero-shot framework outperforms previous supervised methods by a large margin (**6.4%**). LLoVi also outperforms the recent state-of-the-art zero-shot method, SeViLA, by **3.1%**.

Model	mIoP	IoP@0.5	mIoU	IoU@0.5	Acc@GQA
Weakly-Supervised					
IGV (Li et al., 2022)	21.4	18.9	14.0	9.6	10.2
Temp[CLIP](NG+)	25.7	25.5	12.1	8.9	16.0
FrozenBiLM (NG+)	24.2	23.7	9.6	6.1	17.5
SeViLA (Yu et al., 2023)	29.5	22.9	21.7	13.8	16.6
Zero-shot					
LLoVi (ours)	**37.3**	**36.9**	**20.0**	**15.3**	**24.3**

Table 9: **Grounded LVQA results on NExT-GQA.** We extend LLoVi to the grounded LVQA task and show that it outperforms prior weakly-supervised approaches on all evaluation metrics. For a fair comparison, we de-emphasize the models that were pretrained using video-language grounding annotations.

5 Conclusion

In this work, we present a simple, yet highly effective LLM-based framework for long-range video question-answering (LVQA). Our framework outperforms all prior models on the newly introduced EgoSchema benchmark. Furthermore, we demonstrate that our approach generalizes to other LVQA benchmarks such as NeXT-QA, IntentQA, and it can also be extended to grounded LVQA tasks. Lastly, we thoroughly evaluate various design choices of our approach and analyze the key factors behind the success of our method. We hope that our simple LVQA framework will help inspire new ideas and simplify model design in long-range video understanding.

Limitations

Our proposed framework used different short-term visual captioning models for egocentric and exocentric videos due to the domain difference. A unified captioner that works for all kinds of videos remains to be explored in the future. Additionally, our multi-round summarization prompt requires two rounds of prompting LLMs. Although it leads to a significant performance boost on LVQA, it also causes extra computational cost. Therefore, the trade-off between efficiency and high performance in our prompt design can be further improved.

References

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4.

Aanisha Bhattacharya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. 2023. A video is worth 4096 tokens: Verbalize story videos to understand them in zero shot. *arXiv preprint arXiv:2305.09758*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. 2023. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. 2020. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10337–10346.

Feng Cheng and Gedas Bertasius. 2022. Tallformer: Temporal action localization with a long-memory transformer. In *European Conference on Computer Vision*, pages 503–521. Springer.

Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. 2023. Vindlu: A recipe for effective video-and-language pretraining. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Min Su Lee, and Byoung-Tak Zhang. 2021. Dramaqa: Character-centered video story understanding with hierarchical QA. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 1166–1174. AAAI Press.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jiwan Chung and Youngjae Yu. 2023. Long story short: a summarize-then-search method for long video question answering. In *BMVC*.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2021. VIOLET: End-to-End Video-Language Transformers with Masked Visual-token Modeling. In *arXiv:2111.1268*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2023. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. 2019a. Timeception for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 254–263.

Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. 2019b. Videograph: Recognizing minutes-long human activities in videos. *arXiv preprint arXiv:1905.05143*.

Md Mohaiminul Islam and Gedas Bertasius. 2022. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer.

Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, and Gedas Bertasius. 2023. Efficient movie scene detection using state-space transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18749–18758.

Bhavan Jasani, Rohit Girdhar, and Deva Ramanan. 2019. Are we asking the right questions in movieqa? In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Sangho Lee, Jinyoung Sung, Youngjae Yu, and Gunhee Kim. 2018. A memory network approach for story-based temporal summarization of 360 videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1410–1419.

Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. 2021. Video prediction recalling long-term motion context via memory alignment learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3054–3063.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. In *EMNLP*.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023a. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11963–11974.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023b. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11963–11974.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023d. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023e. Videochat: Chat-centric video understanding.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2928–2937.

Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, Ce Liu, and Lijuan Wang. 2023a. Mm-vid: Advancing video understanding with gpt-4v(ision). *arXiv preprint arXiv:2310.19773*.

Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. 2023b. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023a. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*.

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. 2023. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15579–15591.

Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, Tri Dao, Stephen Baccus, and Christopher Ré. 2022. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Advances in neural information processing systems*, 35:2846–2861.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joe Heyward, Viorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzdeh. 2023. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. *arXiv preprint arXiv:2312.07395*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. 2022. Egocentric video-language pretraining. *arXiv e-prints*, pages arXiv–2206.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023a. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. 2023b. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*.

Simeng Sun, Yang Liu, Shuohang Wang, Chenguang Zhu, and Mohit Iyyer. 2023. Pearl: Prompting large language models to plan and execute actions over long documents. *arXiv preprint arXiv:2305.14564*.

Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. 2022. Long-form video-language pre-training with multimodal temporal contrastive learning. *Advances in neural information processing systems*, 35:38032–38045.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. 2022. Ul2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. 2023a. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397.

Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2023b. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023c. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. 2023d. Vamos: Versatile action models for video understanding. *arXiv preprint arXiv:2311.13627*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023e. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023f. Self-consistency improves chain of thought reasoning in language models. In *ICLR*.

Yang Wang, Gedas Bertasius, Tae-Hyun Oh, Abhinav Gupta, Minh Hoai, and Lorenzo Torresani. 2021. Supervoxel attention graphs for long-range video modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 155–166.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022a. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. 2022b. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35:8483–8497.

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2023g. Unified coarse-to-fine alignment for video-text retrieval. *arXiv preprint arXiv:2309.10091*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. 2019. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293.

Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 2022. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Junbin Xiao, Angela Yao, Yicong Li, and Tat Seng Chua. 2023. Can i trust your answer? visually grounded video question answering. *arXiv preprint arXiv:2309.01327*.

Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022a. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2804–2812.

Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. 2022b. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022a. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022b. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141.

Jianing Yang, Yuying Zhu, Yongxin Wang, Ruitao Yi, Amir Zadeh, and Louis-Philippe Morency. 2020. What gives the answer away? question answering bias analysis on video qa datasets. *arXiv preprint arXiv:2007.03626*.

Xitong Yang, Fu-Jen Chu, Matt Feiszli, Raghav Goyal, Lorenzo Torresani, and Du Tran. 2023. Relational space-time query in long-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6408.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Keunwoo Peter Yu. VideoBLIP.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-chained image-language model for video localization and question answering. *NeurIPS*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv*.

Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. 2021. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2023. Least-to-most prompting enables complex reasoning in large language models.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

Our appendix consists of Additional Datasets and Metrics (Section A), Additional Analysis (Section B), Additional Implementation Details (Section C) and Qualitative Analysis (Section D).

A Additional Datasets and Metrics

In this section, we provide detailed information about the datasets and the metrics we use.

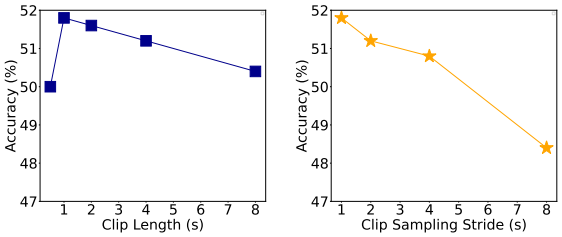
- **NExT-QA (Xiao et al., 2021)** contains 5,440 videos with an average duration of 44s and 48K multi-choice questions and 52K open-ended questions. There are 3 different question types: Temporal, Causal, and Descriptive. Following common practice, we perform zero-shot evaluation on the validation set, which contains 570 videos and 5K multiple-choice questions.
- **IntentQA (Li et al., 2023a)** contains 4,303 videos and 16K multiple-choice question-answer pairs focused on reasoning about people's intent in the video. We perform a zero-shot evaluation on the test set containing 2K questions.
- **NExT-GQA (Xiao et al., 2023)** is an extension of NExT-QA with 10.5K temporal grounding annotations associated with the original QA pairs. The dataset was introduced to study whether the existing LVQA models can temporally localize video segments needed to answer a given question. We evaluate all methods on the test split, which contains 990 videos with 5,553 questions, each accompanied by a temporal grounding label. The metrics we used include: 1) Intersection over Prediction (IoP) (Xiao et al., 2023), which measures whether the predicted temporal window lies inside the ground truth temporal segment, 2) temporal Intersection over Union (IoU), and 3) Acc@GQA, which depicts the percentage of accurately answered and grounded predictions. For IoP and IoU, we report the mean values and values with the overlap thresholds of 0.5.

B Additional Analysis

In this section, we provide additional analysis on the EgoSchema Subset using the standard prompt.

B.1 Video Sampling Configurations

In Figure 5, we investigate the sensitivity of LVQA performance on EgoSchem with different video sampling configurations. Specifically, in Subfigure 5a, we experiment with 4 different clip lengths: 0.5s, 1s, 4s, and 8s. For each clip length, we use the stride that would be sufficient to cover the entire long video input. For these experiments, we use a LaViLa visual captioner and a GPT-3.5 LLM. Our results indicate that LVQA performance is the best when the sampled clip length is 1s. We observe that using an even shorter video clip length (i.e., 0.5s) produces many repetitive/redundant captions, which leads to 2% drop in LVQA performance. Furthermore, we also note that increasing the clip length to longer durations (e.g., 2s-8s) makes the accuracy lower since the extracted captions start to lack detailed visual information needed to answer the question. In addition, in Subfigure 5b, we fix the clip length to 1s and experiment with 4 different stride values: 1s, 2s, 4s, and 8s. Note that the 1s clips sampled using a 1s stride will cover the entire video without overlap. Our results suggest a gradual decrease in LVQA accuracy when increasing the stride from 1s to 8s. This indicates that having gaps in long video coverage leads to suboptimal LVQA performance. Thus, for the rest experiments, we use 1s video clips sampled with a 1s stride.



(a) (b)

Figure 5: **The Analysis of Video Clip Sampling Strategy on EgoSchema**. These results show that sampling 1s video clips with a 1s stride leads to the best LVQA performance.

B.2 Accuracy on Different Question Types

To better understand the strengths and limitations of our LVQA framework, we manually categorize questions in the EgoSchema Subset into 5 categories: (1) Purpose/Goal Identification, (2) Tools and Materials Usage, (3) Key Action/Moment Detection, (4) Action Sequence Analysis, (5) Character Interaction (see Supplementary Materials for details). Note that some questions belong to more than one category. Based on this analysis, we observe that almost half of the questions relate to purpose/goal identification, which makes intuitive

Question Category	Category Percentage(%)	Acc.(%)
Purpose/Goal Identification	49.2	54.9
Tools and Materials Usage	21.8	50.5
Key Action/Moment Detection	21.6	43.5
Action Sequence Analysis	18.2	52.7
Character Interaction	9.4	63.8

Table 10: **Accuracy on different question categories of EgoSchema.** We manually categorize each question in the EgoSchema Subset into 5 categories. Note that each question may belong to one or more categories. Our system performs the best on questions that involve character interaction analysis or human purpose/goal identification. This is encouraging as both of these questions typically require a very long-form video analysis.

sense as inferring human goals/intent typically requires a very long video analysis. We also observe that a significant portion of the questions relate to tool usage, key action detection, and action sequence analysis. Lastly, the smallest fraction of the questions belong to character interaction analysis.

In Table 10, we break down our system's performance according to each of the above-discussed question categories. Our results indicate that our system performs the best in the Character Interaction category (**63.8%**). One possible explanation is that the LaViLa model, which we use as our visual captioner, is explicitly pretrained to differentiate the camera wearer from other people, making it well-suited for understanding various interactions between characters in the video. We also observe that our framework performs much worse in the Key Action/Moment Detection category (**43.5%**). We conjecture that this might be caused by the limitations in the visual captioning model, i.e., if the key action fails to appear in any of the visual captions, the question will be almost impossible to answer. Lastly, we note that our model performs quite well on the remaining categories (>**50%**). It is especially encouraging to see strong results (**54.9%**) in the Purpose/Goal Identification category since inferring human intentions/goals from the video inherently requires very long-form video analysis.

C Additional Implementation Details

C.1 Captioners

For most experiments on EgoSchema, we use LaViLa as the visual captioner. For other pre-trained visual captioners, we use off-the-shelf pre-trained models, e.g., BLIP2 (Li et al., 2023c), EgoVLIP (Qinghong Lin et al., 2022).

LaViLa is trained on the Ego4D dataset. The original LaViLa train set has 7743 videos with 3.9M video-text pairs and the validation set has 828 videos with 1.3M video-text pairs. The EgoSchema dataset is cropped from Ego4D. Since EgoSchema is designed for zero-shot evaluation and the original LaViLa train set includes EgoSchema videos, we retrain LaViLa on Ego4D videos that do not have any overlap with EgoSchema videos to avoid unfair comparison with other methods. After removing the EgoShema videos, the train set consists 6100 videos with 2.3M video-text pairs, and the validation set has 596 videos with 0.7M video-text pairs. We retrain LaViLa on this reduced train set to prevent data leakage. LaViLa training consists of two stages: 1) dual-encoder training and 2) narrator training. Below we provide more details.

Dual-encoder. We use TimeSformer (Bertasius et al., 2021) base model as the visual encoder and a 12-layer Transformer as the text encoder. The input to the visual encoder comprises 4 RGB frames of size 224×224. We randomly sample 4 frames from the input video clip and use RandomResizedCrop for data augmentation. The video-language model follows a dual-encoder architecture as CLIP (Radford et al., 2021) and is trained contrastively. Following LaViLa (Zhao et al., 2023), we use 1024 as batch size. We train at a 3×10^{-5} learning rate for 5 epochs on 32 NVIDIA RTX 3090 GPUs.

Narrator is a visually conditioned autoregressive Language Model. It consists of a visual encoder, a resampler module, and a text encoder. We use the visual encoder (TimeSformer (Bertasius et al., 2021) base model) from the pretrained dual-encoder (See the previous paragraph). The resampler module takes as input a variable number of video features from the visual encoder and produces a fixed number of visual tokens (i.e. 256). The text decoder is the pretrained GPT-2 (Radford et al., 2019) base model with a cross-attention layer inserted in each transformer block which attends to the visual tokens of the resampler module. We freeze the visual encoder and the text decoder, while only training the cross-attention layers of the decoder and the resampler module. Following the design in LaViLa (Zhao et al., 2023), we use a batch size of 256 and a learning rate of 3×10^{-5}. We use AdamW optimizer (Kingma and Ba, 2014) with $(\beta_1, \beta_2) = (0.9, 0.999)$ and weight decay 0.01. We train the model on 8 NVIDIA RTX 3090 GPUs for 5 epochs.

Narrating video clips. We use nucleus sampling (Holtzman et al., 2019) with $p = 0.95$ and return $K = 5$ candidate outputs. Then we take the narration with the largest confidence score as the final caption of the video clip.

For NExT-QA, IntentQA and NExT-GQA datasets, we use LLaVA1.5 as the visual captioner and GPT-4 as the LLM. Specifically, we use the `llava-1.5-7b-hf` variant with the prompt *"USER: <image>. Describe the image in 30 words. ASSISTANT: "*.

C.2 LLMs

For most experiments on EgoSchema we use GPT-3.5 as the LLM. Specifically, we use the `gpt-3.5-turbo-0613` variant which has 4K context. When the context length is not enough, we use the `gpt-3.5-turbo-16k` variant. We use 0 as temperature for all experiments.

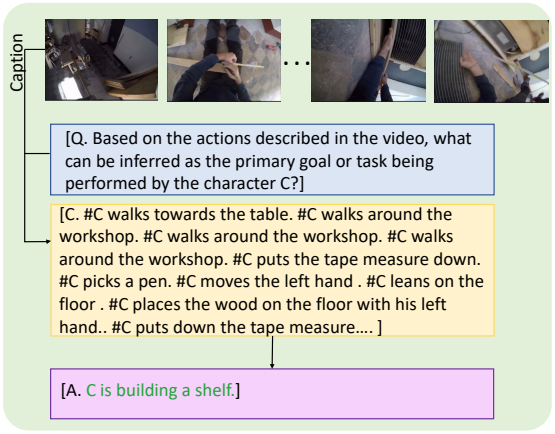
We use `Llama-2-7b-chat-hf`, `Llama-2-13b-chat-hf`, and `Llama-2-70b-chat-hf` variants as Llama2 models. For all Llama2 models, we use greedy sampling to generate the output.

For NExT-QA, IntentQA and NExT-GQA datasets, we use GPT-4 as the LLM with the variant `gpt-4-1106-preview`.

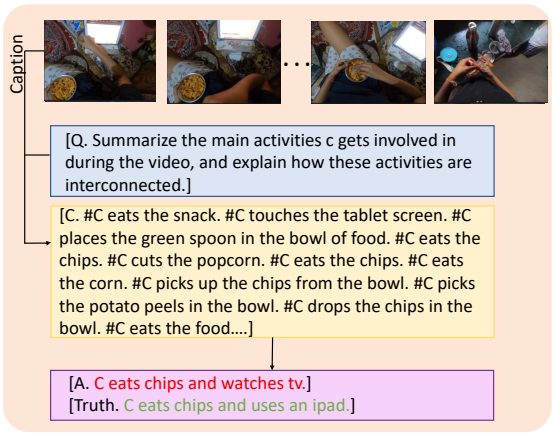
C.3 Prompting Techniques Implementation

Prompt Details. We provide detailed prompts for our standard prompt in Table 11, multi-round summarization-based prompt in Table 12, Zero-shot Chain of Thought in Table 13, and Plan-and-Solve prompting in Table 14. To implement Self-Consistency, we set the temperature of GPT3.5 to 0.7 and run Zero-shot Chain of Thought for 5 times following the design in Self-Consistency (Wang et al., 2023f). Each run of the model provides a result, and the final output is determined by a majority vote. The prompt for the grounded LVQA benchmark is shown in Table 15.

Output Processing. When answering multiple choice questions, GPT3.5 usually outputs complete sentences instead of a single-letter answer, i.e. A, B, C, D, or E. One way to obtain the single-character response is to perform post-processing on the output, which usually requires substantial engineering efforts. In our work, however, we observe that GPT3.5 is very sensitive to the starting sentences of the prompts. Therefore, we explicitly prompt it as in Table 11 to force GPT3.5 to generate a single character as response. In practice, we take out the first character of the output as the final answer.



(a) Success case



(b) Failure case

Figure 6: **Examples of our framework with a standard prompt on EgoSchema.** We show two examples, a successful one (a) and a failed one (b).

D Qualitative Analysis

D.1 Captioners

In Table 16 we compare different captions generated by BLIP2 and LaViLa on EgoSchema. LaViLa captions are generally more concise than BLIP2 captions, focusing more on the actions while BLIP2 focuses more on describing the objects. We also observe that LaViLa is better at differentiating the camera wearer and other person. As shown in the second image in Table 16, LaViLa tends to focus more on the action of the other person when the camera wearer and other person both appear in the video.

D.2 LLoVi with Standard Prompt

We show two examples of our method with standard prompt, including a successful one and a failed one in Figure 6. Our method performs long-range modeling from short-term video captions through

User
Please provide a single-letter answer (A, B, C, D, E) to the following multiple-choice question, and your answer must be one of the letters (A, B, C, D, or E). You must not provide any other response or explanation.
You are given some language descriptions of a first person view video. The video is 3 minute long. Each sentence describes a `clip_length` clip. Here are the descriptions: `Captions`
You are going to answer a multiple choice question based on the descriptions, and your answer should be a single letter chosen from the choices.
Here is the question: `Question`
Here are the choices. A: `Option-A`. B: `Option-B`. C: `Option-C`. D: `Option-D`. E: `Option-E`.

Assistant
`Answer`

Table 11: **LLoVi with Standard Prompt on EgoSchema.**

LLM to understand the video. In the success case demonstrated in Subfigure 6a, the captions describe the camera wearer's action in a short period of time, such as the interation with the tape measure and the wood. With the short-term captions, LLM understand the long video and answers the question correctly.

In the failure case shown in Subfigure 6b, although the video captioner identifies the object in the video correctly as a tablet, LLM understands the action of the camera wearer as watching TV rather than using an iPad. This might be caused by misguidance from the redundant captions that are not related to the question.

D.3 LLoVi with Multi-round Summarization-based Prompt

Figure 7 illustrates two EgoSchema questions that our framework with multi-round summarization-based prompt answers correctly. In Subfigure 7a, the question asks for the primary function of a tool that the video taker uses. However, shown in the first two images, the long video contains descriptions that are not related to the question, such as operating a machine and rolling a dough. As a result, the generated text captions would contain a large section that is not our direction of interest. By summarizing the captions with awareness to the question, LLM extracts key information and cleans redundant captions to provide clearer textual background for answering the question. The same pattern is observed in Subfigure 7b.

Figure 8 shows two questions that our method fails to answer. In the summarization stage, the LLM answers the question directly instead of using the question to guide the summarization. For example, in Subfigure 8a, all the frames show the camera wearer engaging in actions related to washing dishes, but LLM infers that the person is cleaning the kitchen in the summarization stage. This wrong inference further misdirects the following question answering stage, which leads to an incorrect answer. In Subfigure 8b, LLM concludes that the cup of water is used to dilute the paint because the camera wearer dips the brush into water before dipping it into the paint palette.

In Figure 9, we also show a question which the standard prompt fails to answer, but the multi-round summarization-based prompt answers correctly. In the video in the example question, we observe the camera wearer involving in activities related to laundry, such as picking up clothes from the laundry basket and throwing them into the washing machine. However, the short-term video captions shown in Subfigure 9a demonstrate the redundancy of actions. The repetitive actions complexes extracting and comprehending the information presented in the caption. For example, excessive captions on picking up clothes can make LLM think that the camera wearer is packing something. Our

User
You are given some language descriptions of a first person view video. Each video is 3 minute long. Each sentence describes a `clip_length` clip. Here are the descriptions: `Captions`
Please give me a `num_words` words summary. When doing summarization, remember that your summary will be used to answer this multiple choice question: `Question`.

Assistant
`Summary`

User
Please provide a single-letter answer (A, B, C, D, E) to the following multiple-choice question, and your answer must be one of the letters (A, B, C, D, or E). You must not provide any other response or explanation.
You are given some language descriptions of a first person view video. The video is 3 minute long. Here are the descriptions: `Summary`
You are going to answer a multiple choice question based on the descriptions, and your answer should be a single letter chosen from the choices.
Here is the question: `Question`
Here are the choices. A: `Option-A`. B: `Option-B`. C: `Option-C`. D: `Option-D`. E: `Option-E`.

Assistant
`Answer`

Table 12: **LLoVi with Multi-round Summarization-based Prompt on EgoSchema.** We show the variant (C, Q) → S, where we feed the question without potential choices to the summarization stage. **Top:** caption summarization prompt. **Bottom:** question answering prompt. In the first stage, GPT3.5 outputs a question-guided summary. In the second stage, GPT3.5 takes the summary without the original captions, then answer the multiple choice question.

multi-round summarization-based prompt mitigate this problem by first ask LLM to provide a summary of the captions. The summary shown in Subfigure 9b states clearly that the camera wearer is doing laundry. With the cleaner and more comprehensive summary, the LLM answer the question correctly.

D.4 Question Categories

We provide detailed descriptions of each question category in Table 17. Note that each question can be classified into multiple categories.

User
You are given some language descriptions of a first person view video. The video is 3 minute long. Each sentence describes a `clip_length` clip. Here are the descriptions: `Captions`
You are going to answer a multiple choice question based on the descriptions, and your answer should be a single letter chosen from the choices.
Here is the question: `Question`
Here are the choices. A: `Option-A`. B: `Option-B`. C: `Option-C`. D: `Option-D`. E: `Option-E`.
Before answering the question, let's think step by step.

Assistant
`Answer` and `Rationale`

User
Please provide a single-letter answer (A, B, C, D, E) to the multiple-choice question, and your answer must be one of the letters (A, B, C, D, or E). You must not provide any other response or explanation. Your response should only contain one letter.

Assistant
`Answer`

Table 13: **LLoVi with Zero-shot Chain of Thought Prompting on EgoSchema.**

User
You are given some language descriptions of a first person view video. The video is 3 minute long. Each sentence describes a `clip_length` clip. Here are the descriptions: `Captions`
You are going to answer a multiple choice question based on the descriptions, and your answer should be a single letter chosen from the choices.
Here is the question: `Question`
Here are the choices. A: `Option-A`. B: `Option-B`. C: `Option-C`. D: `Option-D`. E: `Option-E`.
To answer this question, let's first prepare relevant information and decompose it into 3 sub-questions. Then, let's answer the sub-questions one by one. Finally, let's answer the multiple choice question.

Assistant
`Sub-questions` and `Sub-answers`

User
Please provide a single-letter answer (A, B, C, D, E) to the multiple-choice question, and your answer must be one of the letters (A, B, C, D, or E). You must not provide any other response or explanation. Your response should only contain one letter.

Assistant
`Answer`

Table 14: **LLoVi with Plan-and-Solve Prompting on EgoSchema.**

User
I will provide video descriptions and one question about the video. The video is 1 FPS and the descriptions are the captions every 2 frames. Each caption starts with the frame number.To answer this question, what is the minimun frame interval to check? Follow this format: [frame_start_index, frame_end_index]. Do not provide any explanation.
Here are the descriptions: `Captions`
Here is the question: `Question`
Please follow the output format as follows: #Example1: [5, 19]. #Example2: [30, 60]. #Example3: [1, 10] and [50, 60]

Assistant
`Answer`

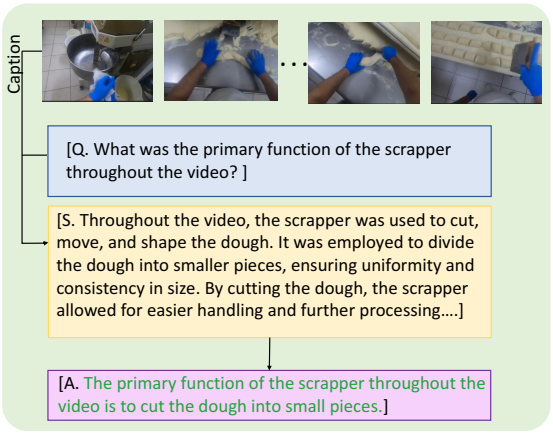
Table 15: **LLoVi Prompt on NExT-GQA.**

				
LaViLa	#C C drops the brick mould.	#O man X moves the cards.	#C C puts the cloth on the table.	#C C moves the dough in the tray.
BLIP2	A person is laying a brick in the dirt.	A child is playing a game of monopoly with a tray of paper plates.	A person is working on a tool.	Woman making dough in a kitchen.

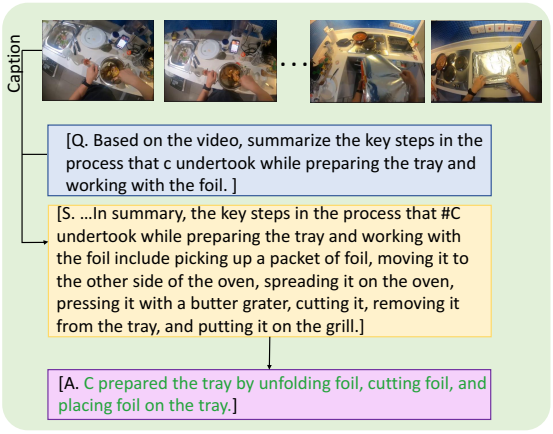
Table 16: **Comparison between different captioners. Top**: frames from EgoSchema videos. **Middle**: captions generated by LaViLa. **Bottom**: captions generated by BLIP2. LaViLa captions are more concise than BLIP2 captions. LaViLa is better at differentiating the camera wearer and other people.

Question Category	Description	Examples
Purpose/Goal Identification	primary goals, intentions, summary, or overarching themes of the video	1. Taking into account all the actions performed by c, what can you deduce about the primary objective and focus within the video content? 2. What is the overarching theme of the video, considering the activities performed by both characters?
Tools and Materials Usage	how the character engages with specific tools, materials, and equipment	1. What was the primary purpose of the cup of water in this video, and how did it contribute to the overall painting process? 2. Explain the significance of the peeler and the knife in the video and their respective roles in the preparation process.
Key Action / Moment Detection	identify crucial steps/actions, the influence/rationale of key action/moment/change on the whole task	1. Out of all the actions that took place, identify the most significant one related to food preparation and explain its importance in the context of the video. 2. Identify the critical steps taken by c to organize and prepare the engine oil for use on the lawn mower, and highlight the importance of these actions in the overall video narrative.
Action Sequence Analysis	compare and contrast different action sequences, relationship between different actions, how characters adjust their approach, efficacy and precision, expertise of the character	1. What is the primary sequence of actions performed by c throughout the video, and how do these actions relate to the overall task being performed? 2. Considering the sequence of events, what can be inferred about the importance of precision and accuracy in the character's actions, and how is this demonstrated within the video?
Character Interaction	how characters interact and collaborate, how their roles differ	1. What was the main purpose of the actions performed by both c and the man throughout the video, and how did their roles differ? 2. Describe the general activity in the room and how the different characters and their actions contribute to this environment.

Table 17: **Question categories of EgoSchema.** We manually categorize each question in the EgoSchema Subset into 5 categories. Note that each question may belong to one or more categories.

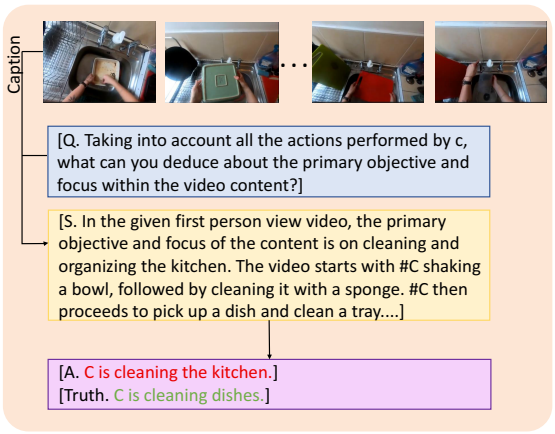


(a)

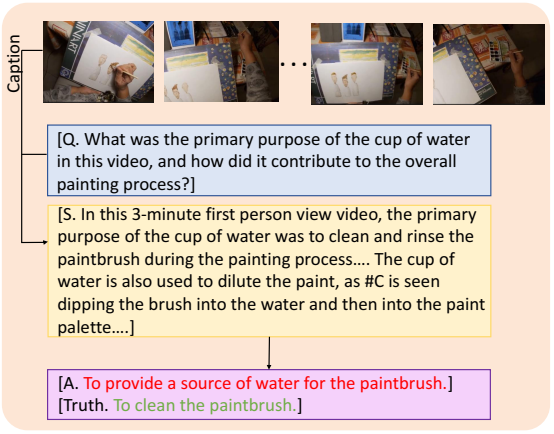


(b)

Figure 7: **Success cases of our multi-round summarization-based prompt.**

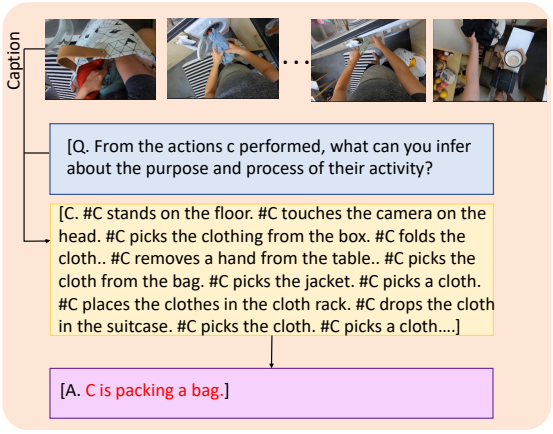


(a)

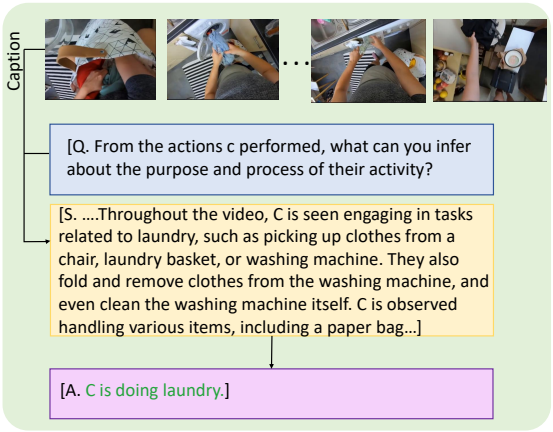


(b)

Figure 8: **Failure cases of our framework with multi-round summarization-based prompt.**



(a) Standard prompt (wrong answer).



(b) Multi-round summarization-based prompt (correct answer).

Figure 9: **Contrast between our standard prompt and our multi-round summarization-based prompt.** (a) demonstrates the process of answering the question with a standard prompt, and (b) shows answering the question with our multi-round summarization-based prompt.