LifelongMemory: Leveraging LLMs for Answering Queries in Long-form Egocentric Videos

Ying Wang, Yanlai Yang, Mengye Ren

New York University
{yw3076,yy2694,mengye}@nyu.edu
https://agenticlearning.ai/lifelong-memory

Abstract

In this paper we introduce LifelongMemory, a new framework for accessing long-form egocentric videographic memory through natural language question answering and retrieval. LifelongMemory generates concise video activity descriptions of the camera wearer and leverages the zero-shot capabilities of pretrained large language models to perform reasoning over long-form video context. Furthermore, LifelongMemory uses a confidence and explanation module to produce confident, high-quality, and interpretable answers. Our approach achieves state-of-the-art performance on the EgoSchema benchmark for question answering and is highly competitive on the natural language query (NLQ) challenge of Ego4D. Code is available at https://github.com/agentic-learning-ai-lab/lifelong-memory.

1 Introduction

Long-form egocentric video understanding has the potential to make a tremendous impact in real-life applications such as personalized AI assistants. Imagine awkward moments when you find yourself asking "where did I put my glasses" or "what is the person's name I just talked to." A personalized AI assistant with a video memory can help us search for answers to questions like these. It takes in a question in the form of a natural language, and outputs either an answer or a video playback of the exact moment when the event of interest took place.



Figure 1: LifelongMemory employs natural language descriptions to create an episodic memory. It uses an LLM to sift through past events, retrieving specific moments in response to queries.

However, despite the progress made on video and natural language understanding in deep learning, long-form egocentric video question answering remains challenging for two reasons. First, unlike short-form videos [17, 40, 14] that usually only contain one single scene and action, long-form egocentric videos can involve multiple scenes where the camera wearers perform numerous tasks and interact with different people and objects. The abundance of details and long-range temporal dependencies make successful information retrieval difficult. Previous methods develop better video features to capture low-level action and object information [11, 21, 22, 5, 31], yet fall short of long-form video understanding [27, 13]. Second, question answering may require sophisticated reasoning of events and oftentimes end-to-end models do not have enough supervision data to generalize and correctly understand different types of questions [30].

To address these two challenges simultaneously, we propose a unified framework, *LifelongMemory*, for long-form video question answering using large language models (LLMs). We compress long video inputs into concise text descriptions with our proposed *Caption Digest* component. The text format can be then augmented to the context of an LLM for answering the questions and locating the most relevant time window. The LLM is capable of general question answering, and unlike end-to-end models, it has zero-shot

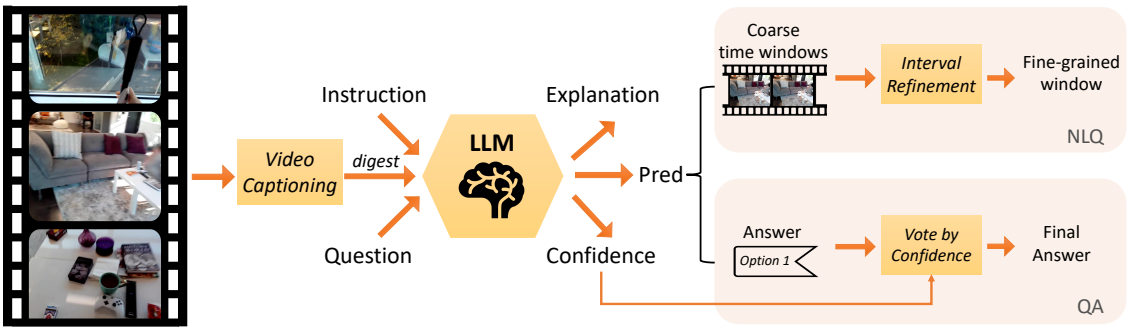


Figure 2: Our LifelongMemory Framework for Long-form Video Understanding. The video inputs are first converted into captions using a pretrained MLLM and then condensed via *Caption Digest*. Next, captions and queries are processed by an LLM to predict coarse temporal windows (for NLQ) or answers (for QA) with explanations and confidence levels. For the NLQ task, the predicted windows are further refined by a pre-trained NLQ model. For the video QA task, we ensemble the predictions of multiple runs and select the answers with the highest confidence.

generalization. Moreover, we also prompt the LLM to produce a confidence level with a textual explanation, both of which help refine the predictions and enhance the interpretability of the model outputs.

Our proposed framework achieves superior performance on two benchmarks for long-form egocentric video understanding, including multi-choice Video Question Answering (Video QA) and natural language query (NLQ). For zero-shot evaluation on the EgoSchema video QA benchmark [27], our method achieves the state of the art which doubles the accuracy of pretrained video QA models [5, 44, 12, 42] and significantly outperforms other LLM-based methods [48, 35]. In the Ego4D NLQ challenge, our method is able to increase the precision of pretrained NLQ models [30, 15] by providing coarse-grained candidate temporal windows in zero shot. In summary, our contributions are as follows:

- We propose a novel framework, *LifelongMemory*, that integrates pre-trained MLLMs to answer questions in long-form egocentric videos. It leverages the remarkable reasoning capabilities of LLMs to tackle the challenge of long-range temporal understanding.
- Our method significantly outperforms prior models and concurrent LLM-based solutions on EgoSchema, and remain highly competitive on Ego4D NLQ.
- Our framework enhances the interpretability and reliability of the results by providing a confidence level and textual explanation of its prediction, revealing the reasoning process of LLMs.

2 Related Work

Multimodal Large Language Models (MLLMs) have recently demonstrated their impressive capabilities in various downstream vision-language tasks [19, 18, 26]. In this paper, we discuss how to utilize frozen MLLMs for long-form video understanding by experimenting with two specific tasks—Video Question Answering (Video QA) and Natural Language Queries (NLQ)—both of which require comprehensive understanding and reasoning of texts and videos. In the following paragraphs, we survey prior works on MLLMs, Video QA, and NLQ.

Multimodal Large Language Models. Large language models (LLMs) [3, 51, 1, 33, 34, 2, 8, 7] have demonstrated their excellent ability to understand and reason with natural language inputs [4, 16, 52]. To extend this understanding and reasoning ability beyond text, many prior works have explored incorporating other modalities, especially visual perception, into LLMs. This leads to the rise of multimodal large language models (MLLMs)[36, 45, 46]. LLaVA [24, 23] connects the CLIP visual encoder [29] with the language decoder of an LLM and finetune them end-to-end on multimodal instruction-following data, achieving competitive performance in general-purpose visual and language understanding. LaViLa [55] adds visual

conditioning on the input to pre-trained LLMs, and finetunes them on Ego4D narrations [13] to create automatic video narrators. These MLLMs serve as critical components in a broad range of downstream applications, showcasing their strength in reasoning on multimodal data [43, 54, 10, 6, 9, 43]. Our work shares the successes in utilizing MLLMs for understanding and reasoning on text and image/video data. Although most current open-sourced MLLMs (such as LLaVA and LaViLa) only take one image or a very short video as inputs, our proposed framework can integrate those pretrained MLLMs and leverage them for the challenging task of long-form video understanding.

Video Question Answering with Multimodal LLMs. The success of LLMs in text QA [7, 33, 34, 1] leads to an increasing trend of applying MLLMs in video QA tasks [41, 42, 47, 37, 28, 35, 48, 53]. Due to the computational burden of large-scale pertaining, many prior works have explored leveraging pretrained (M)LLMs for zero-shot or few-shot QA. R2A [28] retrieves textual descriptions from an external text corpus based on the similarity of video frames and text features encoded by CLIP [29], then uses a pretrained LLM to generate answers given the question and the retrieved descriptions. However, retrieval from a pre-defined text corpus hinders scaling to unseen videos and results in vague or inaccurate video descriptions that can decrease the accuracy of the subsequent QA step. Instead of obtaining captions by retrieval, some prior works utilize a pretrained captioning model to generate high-quality descriptions of videos. VidIL [37] obtains frame-level captions from BLIP [20] and retrieves labels of objects, attributes, and events from pre-defined vocabularies using CLIP [29], then feed all these information into an LLM with a few labeled examples. Despite its good performance on short videos [39], this approach is not suitable for long videos in the wild because (i) pre-defined vocabularies limit the application of the approach and (ii) a large number of noisy and redundant low-level details can distract the LLM from the main task. Our proposed framework is more efficient and flexible: It is able to perform zero-shot video QA that only utilizes a concise list of distilled captions without requiring a fixed keyword vocabulary. Most relevant to our work, Socratic models [47] show qualitative examples of LLM's zero-shot performance on some toy examples, where the key moments of the input video are converted into a textual record by a captioning model. Concurrent work Vamos [35] and LLoVi [48] employ a similar approach of using a captioning model to bridge videos and LLMs for zero-shot video QA. Empirical evaluations show that our proposed framework significantly outperforms these two on EgoSchema [27], a zero-shot video QA benchmark designed for long-range temporal understanding.

Natural Language Queries in Egocentric Videos. The Natural Language Queries (NLQ) task involves localizing the temporal window corresponding to the answer to a question in a long video clip. This task is challenging for end-to-end supervised video localization models [49, 50] due to the sparsity of annotations and the length of videos in the dataset. Prior works have focused on constructing a hierarchical structure, augmenting the NLQ dataset and developing better video features through large-scale pretraining. ReLER [25] proposes a novel multi-scale cross-modal transformer architecture, a video frame-level contrastive loss, and two data augmentation strategies. InternVideo [5] improves the quality of video features by carefully pre-training and fine-tuning a VideoMAE-L Model [32], and ensemble the features and predictions. More recently, NaQ [30] introduces a data augmentation strategy to transform video narrations into training data for the NLQ task, alleviating the problem of sparse annotation. NaQ++ ReLER, obtained by training the ReLER model with NaQ data, was the previous state-of-the-art method for Ego4D NLQ. GroundNLQ [15] is the current state-of-the-art for this benchmark. It adopts a two-stage pre-training strategy to respectively train a video feature extractor and a grounding model on video narrations, and finally finetune the grounding model on annotated data. Our work is complementary to these prior works in that they can be used in the last stage of our proposed framework to produce more fine-grained predictions based on the predictions of the frozen LLM.

3 LifelongMemory

In this section, we introduce our proposed LifelongMemory framework. To tackle the challenge of long-form videos, we first transform egocentric videos into a comprehensive yet concise textual log and then further condense the information via *Caption Digest*. Then, we use an LLM to predict answers (for Video QA) or coarse temporal windows (for NLQ), along with confidence and explanation for interpretability. Finally, the predictions are further refined depending on the task. Figure 2 outlines the workflow and we describe different stages in detail below.

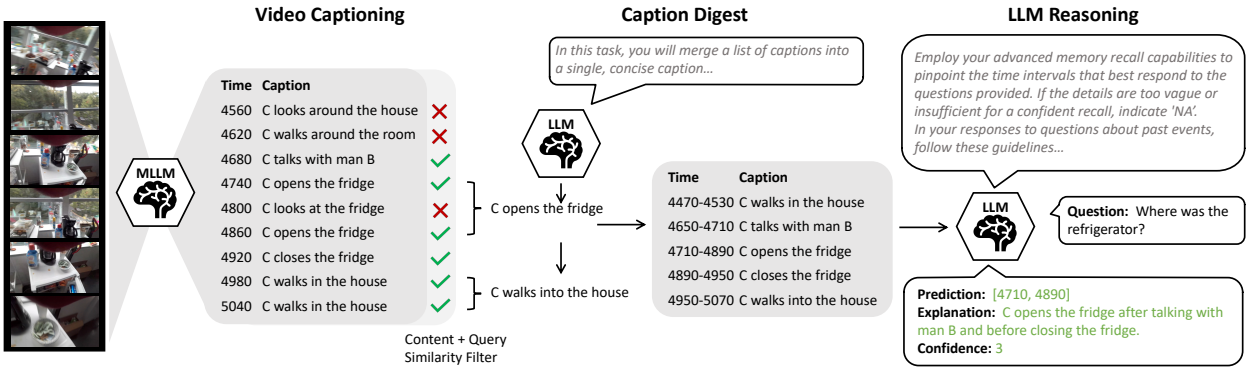


Figure 3: Example Caption Processing and LLM Reasoning for NLQ. 1) We use a multimodal LLM (MLLM) to produce captions from a list of short video clips. 2) Content and query similarity filters are then applied to remove redundant and irrelevant captions. Similar consecutive captions are merged by an LLM. 3) An LLM is instructed to take inputs from the list of condensed captions and retrieve the most relevant interval candidates. The same procedure is performed on the QA task.

3.1 Egocentric Video Captioning

We begin by summarizing the raw footage into a list of captions using pre-trained MLLMs (*e.g.* LaViLa [55]). We sample image frames or short video clips at a fixed interval, and produce a line of caption per clip. The text descriptions as a form of episodic memory enable the transformation of complex egocentric video footage into a coherent log of daily activities, capturing life's narrative in a more accessible and compressed format.

Caption Digest. Raw captions produced by MLLMs, however, can be rather verbose and repetitive, and consequently hinder the downstream reasoning process, especially for long-form videos. We propose to create a caption digest to condense the information. Moreover, we aim to increase the relevance of the captions in relation to the target queries. Figure 3 shows an example of the caption digest process. First, we remove uninformative captions (*e.g.* "looks around ..."). Second, we remove captions that are not relevant to the query by comparing the embedding similarity. Third, we gather adjacent captions that share a high similarity score and use an LLM to produce a single concise caption. The condensed list of captions then augments the context of the LLM for further reasoning and processing.

3.2 LLM Reasoning

With the list of condensed captions with their corresponding time interval from the previous stage, we leverage an LLM here for its impressive zero-shot context understanding and reasoning capability. We combine captions and queries into an instructive and contextualized prompt. A snippet of the instruction template is shown in Figure 3. The full prompt and a discussion of the prompt designs are in Section A.

We particularly instruct the LLM to aggregate information and imagine the visual scene underlying the given captions. The LLM is instructed to take into consideration the full context in the template and utilize different pieces of information to produce the most probable answer. For example, when asking "*Who did I interact with when I was shopping?*", the LLM is able to filter all captions and produce a list of intervals involving "*person x talking to C*" where C is the subject in the video and X refers to the other person. The LLM is also instructed to consider the loss of information when converting videos into concise captions. For example, one query asks "*What size of washer did I pick ?*" but there are no captions explicitly mentioning the washers. In this example, the LLM displays its capability to capture implicit information and infer based on context. The LLM answers "*choosing the time points where I picked items from the table or the floor, as these instances may provide more context about the objects and their locations.*" By grasping nuanced relationships and dependencies within the given context, LLM is able to filter out the most relevant information from the extensive video captions.

In addition to the predicted answers, we also ask the LLM to explain its predictions for more interpretability. Specifically, we ask the LLM to output a sentence of explanation to encourage introspective thinking and rate its confidence in the output out of three confidence levels. The verbalized confidence strategy [38] can help us control the precision of the output in later stages.

3.3 Vote by Confidence (Video QA)

To increase the reliability of LLM predictions for video QA, we ensemble the LLM's predictions using voting by confidence. We repeatly perform the LLM reasoning step where the LLM is prompted to generate predictions based on the same input in each run. From the pool of predictions, the answer with the highest confidence score is selected. In cases where multiple answers have the same highest confidence, a random selection is performed. By focusing on the most confident predictions, this ensemble step can further improve the accuracy and robustness of the results.

3.4 Fine-grained Interval Refinement (NLQ)

Since the time intervals are subsampled, to obtain a fine-grained interval prediction for the NLQ task, we revisit the video inputs and enhance our LLM interval predictions in the last stage. For this goal, we employ a pretrained NLQ model and feed in candidate intervals predicted by our previous stage. The intervals are padded with a small window of size α. Specifically, for each (s_i, e_i), the new start time is $s_i' = \max(s_i - \alpha, s)$ and the new end time is $e_i' = \min(e_i + \alpha, e)$ where s and e are the start and end time of the original clip. Then we extract video clips $[v_1, v_2 \dots v_n]$ according to the predicted intervals $[(s_1', e_1'), (s_2', e_2') \dots (s_n', e_n')]$.

When the prediction for a certain query contains multiple candidate intervals, we feed them along with the target query into a classifier that is trained on NLQ data to select the optimal candidate v^*. For queries without predictions (*i.e.* "NA"), we simply use the original full video. Localization within a coarse temporal window makes the NLQ task easier compared to doing it on the original full-length video.

4 Experiments

In this section, we evaluate our LifelongMemory framework in real-world egocentric video query tasks.

4.1 Experiment Setup

EgoSchema. The EgoSchema dataset [27] consists of over 5,000 question-answer pairs for 250 hours of Ego4D videos covering a wide range of human daily activities. For each question, the correct answer needs to be selected from 5 choices based on a three-minute egocentric video. The dataset is curated by human annotators to ensure all questions require long-term temporal understanding. We use the subset provided by EgoSchema, which contains 500 question-answer pairs, for ablation studies on prompt designs, captioning choices, and voting by confidence, then use the best setup for evaluation on the full benchmark.

Ego4D NLQ. The Ego4D dataset [13] is an egocentric video dataset including a wide variety of daily life activities recorded by individuals wearing cameras. The NLQ task, as one of the episodic memory tasks of Ego4D, requires localizing a temporal window of the video to answer a natural language query. The NLQ annotations are from 227 hours of videos, with a total of 19,200 queries spanning 13 query templates. The train/val/test split (60%, 20%, 20%) is composed of disjoint sets of video clips. The average video length is approximately 8.7 minutes, while the average duration of a response window is only 9.3 seconds, representing on average only 2% of the full video.

Evaluation Metrics. For the EgoSchema dataset, we use accuracy to evaluate our framework since it is a multi-choice QA task. For the NLQ dataset, we adopt different metrics for different stages as below. In the LLM Reasoning stage where we only have coarse-grained predictions, we evaluate on the validation set with metrics including **(i)** the ratio of predictions that overlap with the ground truth (denoted as *Overlap*), **(ii)** and the proportion of predictions where at least one candidate achieves an Intersection over Union (IoU)

greater than 0.3 with the ground truth (denoted as *IoU*@0.3*). During the refinement stage for NLQ, we obtain fine-grained predictions so we can evaluate the test dataset using the standard NLQ metrics – *R@1 IoU@0.3* and *R@1 IoU@0.5*, which is the recall of top one prediction having IoU with the ground truth larger than the threshold {0.3, 0.5}.

Caption Sources. We experiment with machine-generated captions and human-annotated captions and test the effect of text-conditioned captioning.

- **LLaVA**: LLaVa [23, 24] is a multimodal LLM pre-trained on a diverse set of 1.2M publicly available data, including various multimodal question-answering and reasoning tasks. To encourage LLaVA to generate captions that are relevant to the query while not introducing false positives, we follow the template proposed by LLaVA-1.5[23] and adopt the prompt *"If there are factual errors in the questions, provide a precise description of the image; if not, proceed answering the question. [queries]."*
- **LaViLa**: LaViLa [55] is a multimodal LLM pre-trained on the video-narration pairs from Ego4D and is thus capable of generating captions that mimic the ground-truth descriptions of the video. Each caption is generated using 4 frames uniformly taken from a two-second video clip.
- **Ego4D Narrations**: Ego4D [13] narrations include written sentence narrations in English from human annotators, describing a diverse set of activities in the dataset. The annotated narrations contain on average 13.2 sentences per minute of video, which is not as dense as the LLaVA and LaViLa captions that we sample every 2 seconds.

Caption Digest Details. The generated captions are distilled through filtering and merging in this step. We first remove ambiguous captions containing keywords that are associated with blurry and noisy frames. Then, we filter out irrelevant captions based on the similarity scores between the embedding of queries and captions encoded by LaViLa. Lastly, we identify groups of similar consecutive captions by calculating the similarity scores of the embeddings of neighboring captions and merge captions in the same group by querying GPT-3.5 with prompt *"In this task, you will merge a list of captions into a single, concise caption. Focus on clarity and brevity while ensuring no critical details are lost in the merging process."*

NLQ Refinement Details. For the refinement stage, we train a classifier on the NLQ train set to select the optimal candidate from multiple LLM predictions. To construct a video dataset similar to the real LLM predictions, we randomly shift and scale the ground-truth temporal windows. We then mark those intervals that have IoU with the ground truth larger than 0.5 as positives and randomly pick the same amount of negative samples from intervals with IoU less than 0.1. We utilize video features encoded by InternVideo [5] and EgoVLP [22] and adapt VSLNet [49], a span-based localization network, to this video classification task, where we replace the localization head with a classification head. After obtaining the optimal candidate temporal windows, we extend them by a window size of α to provide more context to the NLQ model and then feed them into the state-of-the-art NLQ model, NaQ++ReLER [30, 25] and GroundNLQ [15], which have been finetuned on the Ego4D NLQ dataset. This gives us fine-grained predicted temporal windows that can reflect the answers to the target queries.

4.2 Qualitative Results

We visualize *LifelongMemory* results in Figure 4. For the EgoSchema QA examples, we show that the captions capture objects and actions in the scene very well, and the LLM is then able to answer the question correctly, using its reasoning capability. For the NLQ examples, we show that many LLM predictions have high-quality overlaps with the ground-truth windows (without interval refinement). We note that many successful retrievals rely on high-quality captions and we expect there can be a large room for future improvement with a stronger captioning model. In both datasets, the LLM is able to explain its predictions with a confidence level, enhancing the interpretability of the results. We provide more qualitative examples in Section B.

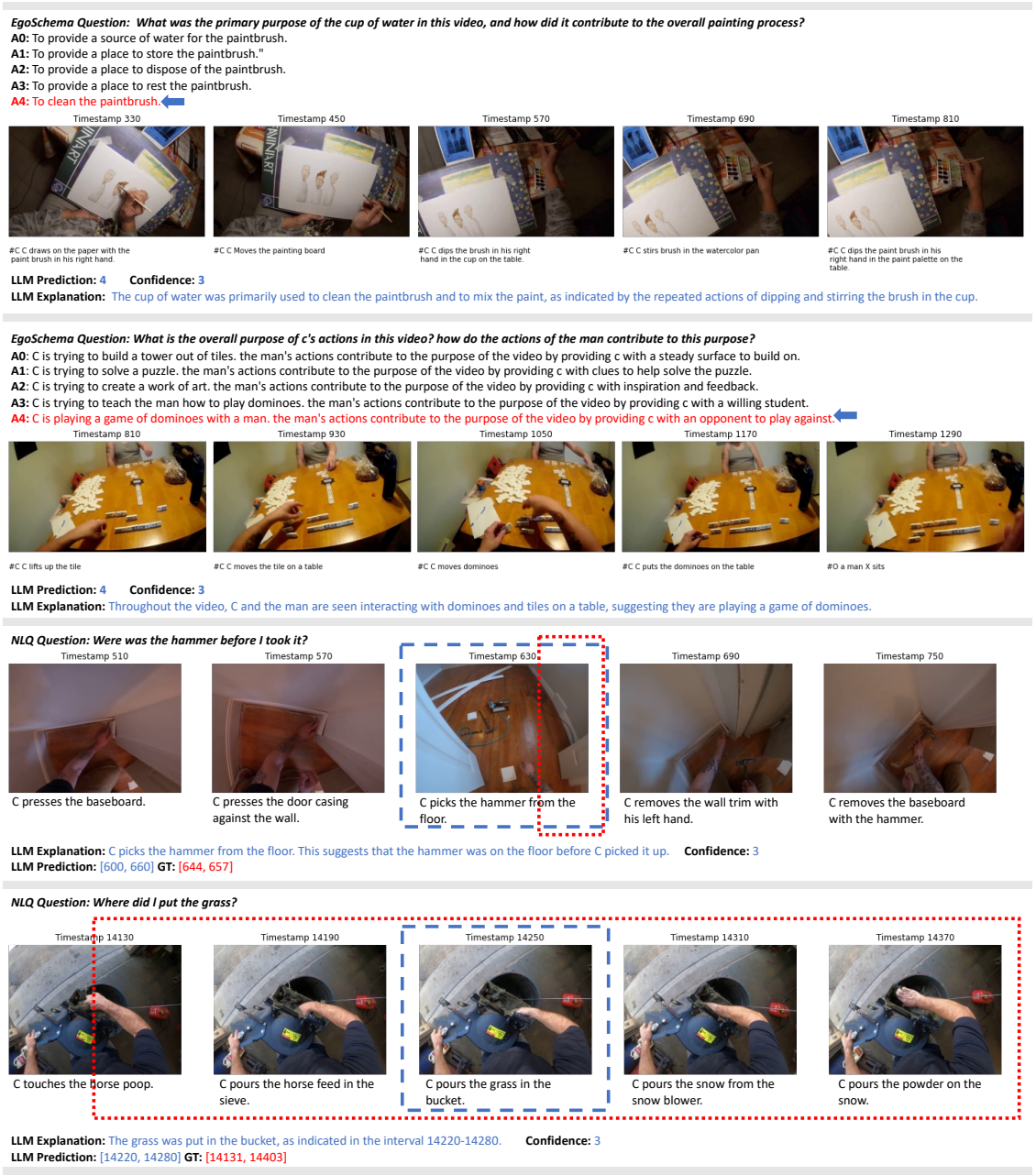


Figure 4: EgoSchema QA and Ego4D NLQ examples using LaViLa and GPT-4. The groud-truth answers are in red and the LLM predictions are in blue. The sampled frames are manually picked from the raw video input to show key events related to the query.

4.3 Quantitative Results

EgoSchema Benchmark Results. Our method achieves state-of-the-art performance on EgoSchema, as shown in Table 1. Due to the challenge of long-form videos, prior state-of-the-art video QA models [5, 42] struggle at this task with an accuracy not much better than random (20%). When compared with concurrent works – LLoVi [48] and Vamos [35] – that also leverage GPT-4, our approach outperforms them with a significant margin of over 10%. These empirical results confirm LifelongMemory is a simple yet effective framework that can reason and answer questions of very long egocentric videos.

Table 1: Zero-shot QA on EgoSchema with Different Models. * represents ensembled results using vote by confidence. Subset is 500 question-answer pairs provided by EgoSchema for validation.

Model	LLM	Input	Subset	Full
FrozenBiLM [42]	-	90 frames	-	26.9
InternVideo [5]	-	90 frames	-	32.1
Vamos [35]	GPT-4	mixed	51.2	48.3
LLoVi [48]	GPT-3.5	180 captions	57.6	50.3
LLoVi [48]	GPT-4	180 captions	58.3	-
Ours	Llama3-8B	90 captions	60.4	-
Ours	GPT-3.5	90 captions	64.0	-
Ours	Claude-3-Haiku	90 captions	64.8	55.2
Ours	GPT-4	90 captions	68.0	62.1
Ours*	GPT-4	90 captions	69.0	62.4
Ours	GPT-4o	90 captions	70.6	64.6
Ours*	GPT-4o	90 captions	**72.0**	**64.7**

Table 2: Ego4D NLQ benchmark results, using GPT-4. Our approach filters out noisy content in the video for the pretrained NLQ models, increasing the precision of the predictions of the NLQ models. Reported metrics all use predictions that rank the first.

Method	Set	Mean	IoU=0.3	IoU=0.5
NaQ++ [30]	val	20.20	25.00	15.40
Ours (LaViLa, NaQ++)	val	19.00	23.40	14.61
Ours (Ego4D, NaQ++)	val	**21.09**	**26.12**	**16.06**
NaQ++ [30]	test	17.67	21.70	13.64
Ours (LaViLa, NaQ++)	test	18.06	22.28	13.84
GroundNLQ [15]	test	20.08	23.43	16.71
Ours (LaViLa, GroundNLQ)	test	**20.27**	**23.68**	**16.86**

Ego4D NLQ Benchmark Results. We compare the performance of our method and two other competitive methods on the Ego4D NLQ benchmark[1] in Table 2. Our method with Ego4D ground-truth narrations achieves the best performance in the validation set and our method with GroundNLQ as the refinement model achieves the best performance in the test set. LifelongMemory is a flexible framework that can be plugged into any pretrained captioning model and video localization model, suggesting the potential of our method for future improvement using better pretrained MLLMs.

Captioning Model Choices. We compare the effect of different caption sources in Table 3. Although LLaVA generates longer captions conditioned on the queries, the performance of LaViLa is significantly better than LLaVA. This indicates the necessity of adopting an egocentric captioning model that focuses on the core activity of the individual. Despite the effectiveness of LaViLa in this task, we identify that LaViLa tends to generate false positive captions as it is finetuned on Ego4D data. We thus evaluate the ground-truth captions provided by the Ego4D Narrations data and observe that it achieves the best performance with significantly fewer captions. This confirms our assumption that an accurate well-crafted set of captions can effectively summarize the information of the camera wearer's activity in egocentric videos.

LLM Choices. We compare the effect of different LLMs for EgoSchema and NLQ respectively in Table 1 and Table 3. We observe that GPT-4 and GPT-4o significantly outperform GPT-3.5 and open-source models like Llama [33, 34, 2] and Vicuna [7] for both datasets. Note that the performance drop caused by weaker LLMs is much larger for the NLQ task because this task requires more precise instruction following capabilities: weaker models often misunderstand the prompt and output an answer instead of a list of

[1]Our results of GroundNLQ are slightly lower than their reported numbers. Since GroundNLQ has not released all checkpoints, we are unable to reproduce the results.

Table 3: NLQ performance using different caption and LLM components. The **bold** number denotes the highest and the underlined the second highest. † represents predictions with a confidence level of 3. *NA* represents the ratio of null predictions, and all other metrics do not include null predictions. *Count* represents the number of captions, and *Length* represents the average word count of captions.

Captions	*Count*	*Length*	LLM	*Overlap*	*Overlap* †	*IoU*@0.3*	*IoU*@0.3* †	*NA*
Ego4D	109	7.95	GPT-4	**51.73**	**53.98**	**15.99**	**27.38**	40.29
			GPT-3.5	31.27	33.15	0.91	4.34	94.13
LaViLa	186	6.40	GPT-4	36.61	38.35	9.74	19.22	47.04
			GPT-3.5	20.47	22.33	1.29	4.78	89.75
LLaVA	250	52.52	GPT-4	6.42	8.79	1.50	2.71	60.92

Table 4: Ego4D NLQ and EgoSchema QA performance using LaViLa + GPT-4, with different frame sampling intervals and digest strategy.

		Ego4D NLQ					EgoSchema	
Freq.	**Digest**	**# Captions**	***Overlap***	***Overlap*†**	***IoU*@0.3***	***IoU*@0.3*†**	**# Captions**	***Acc***
4s	Yes	70	33.55	36.99	6.46	14.64	39	26.4
2s	Yes	186	**36.61**	**38.35**	**9.74**	**19.22**	75	**68.0**
2s	No	250	23.71	23.89	4.92	11.28	90	**68.0**

temporal intervals, leading to a high NA ratio. As our framework is agnostic to LLMs—it's very easy to plug in a future version of LLMs to further boost the performance.

Caption Digest. We evaluate the effect of caption digest in Table 4. With *Caption Digest*, we discard ambiguous and irrelevant captions and use LLM to merge similar ones as described in Section 4.1. For NLQ, this technique significantly improves both metrics by around 10%, suggesting that a concise context leads to a much better retrieval performance. However, similar effects are not observed in EgoSchema as the original undigested context length is already relatively small (*i.e.* less than 100 captions). Since reduced context lengths save the computation costs, we adopt caption digest for both datasets.

Caption Sampling Interval. Given the same captioning models and preprocessing process, smaller caption intervals lead to higher performance as they provide richer contexts for the LLM. Since each Ego4D video contains a large amount of activities, coarse-grained captioning is very likely to miss key moments and results in a loss of information. Decreasing the sampling frequency of captioning leads to a large drop in the accuracy of predictions of both NLQ and EgoSchema, as shown in Table 4. It is worth noting that using very limited captions leads to a very low EgoSchema accuracy that is not much better than random guess (20%) due to the significant information loss.

Table 5: Effect of explanation and confidence levels.

	Ego4D NLQ		EgoSchema
Conf. Level	***Overlap***	***IoU*@0.3***	**Acc.**
≥ 1	36.46	9.63	68.0
≥ 2	36.49	17.52	69.7
≥ 3	**38.20**	**19.06**	**74.6**
Explanation	***Overlap***	***IoU*@0.3***	**Acc.**
No	32.73	8.65	64.2
Yes	**36.61**	**9.74**	**68.0**

Effect of Explanation. We also experiment with different prompts in Table 5. To encourage LLM reasoning step by step, we provide detailed instructions on how to retrieve the temporal windows and answer the queries while explicitly asking it to explain its prediction. The request for explanation encourages the LLM to reason step by step and improves the performance in both datasets. Moreover, providing textual explanations also increases the interpretability and reliability of the model outputs.

Effect of Confidence Levels. To encourage the LLM to make more reliable predictions, we also explicitly ask the LLM to predict a confidence level for each of its own outputs. We report the relationship between scores and their confidence levels in Table 5. The increase in confidence scores leads to an increase in accuracy in both datasets, suggesting the verbalized confidence scores are calibrated. For EgoSchema, we also use confidence level to vote during model ensembling, leading to a 0.1-0.3% increase in test accuracy as shown in Table 1.

4.4 Error Analysis

The majority of errors stem from the captioning step, where inevitable information loss occurs during the transformation from long video inputs into text, as shown in 5. For NLQ with insufficient information, we encourage the LLM to make null predictions and rely on the refinement stage to make the final prediction based on the full input video. On the contrary, we encourage the LLM to select the most plausible answer for EgoSchema when uncertain because we don't rely on a pretrained QA model in the refinement stage. Our prompts are included in Section A.

We also observe that sometimes the LLM proposes multiple temporal windows for NLQ that seem to be reasonable, but only one ground-truth answer is available, as shown in Section C. This suggests some NLQ queries are ambiguous and require more careful annotations.

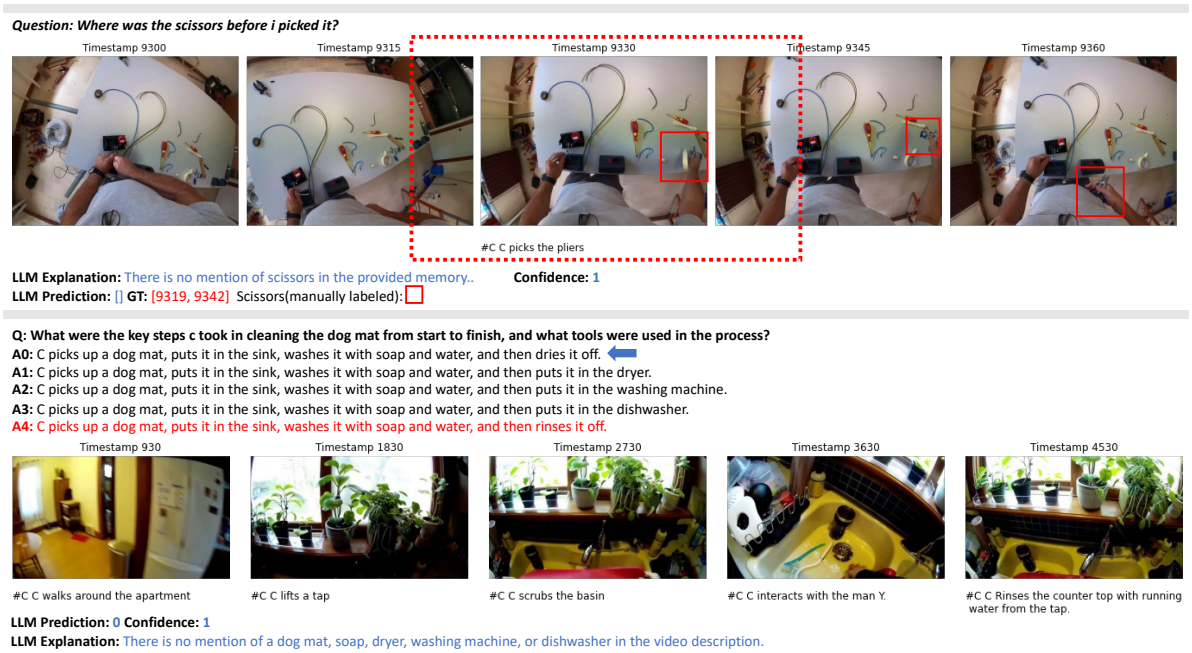


Figure 5: Error caused by insufficient captioning. The upper figure is an NLQ example and the lower figure is an EgoSchema example. LLM predictions are in blue boxes and the ground truth is in red.

5 Conclusion

In this paper, we propose LifelongMemory, a novel framework that leverages pre-trained MLLMs for answering natural language queries in long-form egocentric videos. To address the challenges of long-range temporal dynamics, we condense the input videos into a concise textual log and utilize an LLM to comprehend the context and answer the given queries. Our method achieves state-of-the-art performance on EgoSchema and remains highly competitive on Ego4D NLQ, with enhanced interpretability provided by verbalized confidence and explanation. LifelongMemory showcases the potential of leveraging LLMs in video understanding and opens up opportunities for personalized AIs that can answer daily queries for individuals requiring assistance.

Acknowledgment

We would like to thank the Microsoft Accelerating Foundation Models Research program for providing cloud compute credits for running some parts of our LLM experiments. This work was also supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

[1] Achiam, O. J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] AI@Meta. Llama 3 model card. *https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md*, 2024.

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[4] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[5] Chen, G., Xing, S., Chen, Z., Wang, Y., Li, K., Li, Y., Liu, Y., Wang, J., Zheng, Y.-D., Huang, B., et al. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *arXiv preprint arXiv:2211.09529*, 2022.

[6] Chen, L., Li, B., Shen, S., Yang, J., Li, C., Keutzer, K., Darrell, T., and Liu, Z. Large language models are visual reasoning coordinators. *arXiv preprint arXiv:2310.15166*, 2023.

[7] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *https://lmsys.org/blog/2023-03-30-vicuna/*, 2023.

[8] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[9] Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[10] Du, Y., Yang, M., Florence, P., Xia, F., Wahid, A., Ichter, B., Sermanet, P., Yu, T., Abbeel, P., Tenenbaum, J. B., et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023.

[11] Feichtenhofer, C., Fan, H., Malik, J., and He, K. Slowfast networks for video recognition. In *ICCV*, 2019.

[12] Fu, T.-J., Li, L., Gan, Z., Lin, K., Wang, W. Y., Wang, L., and Liu, Z. An empirical study of end-to-end video-language transformers with masked visual modeling. In *CVPR*, 2023.

[13] Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[14] Hendricks, L. A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., and Russell, B. Localizing moments in video with natural language. In *ICCV*, 2017.

[15] Hou, Z., Ji, L., Gao, D., Zhong, W., Yan, K., Li, C., Chan, W. K., Ngo, C.-W., Duan, N., and Shou, M. Z. Groundnlq @ ego4d natural language queries challenge 2023. *arXiv preprint arXiv:2306.15255*, 2023.

[16] Huang, J. and Chang, K. C.-C. Towards reasoning in large language models: A survey. In *Findings of ACL*, 2023.

[17] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. Kinetics400 dataset: The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[18] Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., and Shan, Y. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023.

[19] Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[20] Li, J., Li, D., Xiong, C., and Hoi, S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[21] Lin, K. Q., Wang, A. J., Soldan, M., Wray, M., Yan, R., Xu, E. Z., Gao, D., Tu, R., Zhao, W., Kong, W., et al. Egocentric video-language pretraining@ ego4d challenge 2022. *arXiv preprint arXiv:2207.01622*, 2022.

[22] Lin, K. Q., Wang, J., Soldan, M., Wray, M., Yan, R., Xu, Z., Gao, D., Tu, R.-C., Zhao, W., Kong, W., Cai, C., HongFa, W., Damen, D., Ghanem, B., Liu, W., and Shou, M. Z. Egocentric video-language pretraining. In *NeurIPS*, 2022.

[23] Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[24] Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023.

[25] Liu, N., Wang, X., Li, X., Yang, Y., and Zhuang, Y. Reler@ zju-alibaba submission to the ego4d natural language queries challenge 2022. *arXiv preprint arXiv:2207.00383*, 2022.

[26] Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., and Lin, D. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[27] Mangalam, K., Akshulakov, R., and Malik, J. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2312.17235*, 2023.

[28] Pan, J., Lin, Z., Ge, Y., Zhu, X., Zhang, R., Wang, Y., Qiao, Y., and Li, H. Retrieving-to-answer: Zero-shot video question answering with frozen large language models. *arXiv preprint arXiv:2306.11732*, 2023.

[29] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[30] Ramakrishnan, S. K., Al-Halah, Z., and Grauman, K. NaQ: Leveraging narrations as queries to supervise episodic memory. In *CVPR*, 2023.

[31] Shao, J., Wang, X., Quan, R., and Yang, Y. Action sensitivity learning for the ego4d episodic memory challenge 2023. *arXiv preprint arXiv:2306.09172*, 2023.

[32] Tong, Z., Song, Y., Wang, J., and Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

[33] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[34] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[35] Wang, S., Zhao, Q., Do, M. Q., Agarwal, N., Lee, K., and Sun, C. Vamos: Versatile action models for video understanding. *arXiv preprint arXiv:2311.13627*, 2023.

[36] Wang, Y., Chen, W., Han, X., Lin, X., Zhao, H., Liu, Y., Zhai, B., Yuan, J., You, Q., and Yang, H. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024.

[37] Wang, Z., Li, M., Xu, R., Zhou, L., Lei, J., Lin, X., Wang, S., Yang, Z., Zhu, C., Hoiem, D., et al. Language models with image descriptors are strong few-shot video-language learners. In *NeurIPS*, 2022.

[38] Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

[39] Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., and Zhuang, Y. Video question answering via gradually refined attention over appearance and motion. In *ACM MM*, 2017.

[40] Xu, J., Mei, T., Yao, T., and Rui, Y. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[41] Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Just ask: Learning to answer questions from millions of narrated videos. In *CVPR*, 2021.

[42] Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*, 2022.

[43] Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., and Wang, L. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

[44] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., and Huang, F. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[45] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2024.

[46] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.

[47] Zeng, A., Attarian, M., brian ichter, Choromanski, K. M., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M. S., Sindhwani, V., Lee, J., Vanhoucke, V., and Florence, P. Socratic models: Composing zero-shot multimodal reasoning with language. In *ICLR*, 2023.

[48] Zhang, C., Lu, T., Islam, M. M., Wang, Z., Yu, S., Bansal, M., and Bertasius, G. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023.

[49] Zhang, H., Sun, A., Jing, W., and Zhou, J. T. Span-based localizing network for natural language video localization. In *ACL*, 2020.

[50] Zhang, S., Peng, H., Fu, J., and Luo, J. Learning 2d temporal adjacent networks formoment localization with natural language. In *AAAI*, 2020.

[51] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[52] Zhang, Y., Mao, S., Ge, T., Wang, X., de Wynter, A., Xia, Y., Wu, W., Song, T., Lan, M., and Wei, F. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*, 2024.

[53] Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[54] Zhao, X., Li, M., Weber, C., Hafez, M. B., and Wermter, S. Chat with the environment: Interactive multimodal perception using large language models. In *IROS*, 2023.

[55] Zhao, Y., Misra, I., Krähenbühl, P., and Girdhar, R. Learning video representations from large language models. In *CVPR*, 2023.

A Prompting

We provide the complete prompt used for EgoSchema (Figure 6) and Ego4D NLQ (Figure 7). As the NLQ dataset contains multiple queries for one video clip, we avoid passing the same caption list multiple times by including all queries of the same clip in the prompt to reduce the cost of API calls. We provide an instructive prompt with detailed steps and ask the LLM to produce responses in the structured format to expedite post-processing. Note that we encourage the LLM to refuse to answer NLQ questions if the context is not informative so we can feed the full-length video into the refinement stage later. In contrast, we encourage the LLM to pick the most possible answer for EgoSchema because we must provide an answer to each question and there is no refinement stage for the QA task.

QA System Prompt

You are individual C, with others represented as O. In your responses to questions about past events, it is vital to provide not only the key You are presented with a textual description of a video clip. Your task is to answer a question related to this video, choosing the correct option out of five possible answers. It is crucial that you imagine the visual scene as vividly as possible to enhance the accuracy of your response. After selecting your answer, rate your confidence level in this choice on a scale from 1 to 3, where 1 indicates low confidence and 3 signifies high confidence. Please provide a concise one-sentence explanation for your chosen answer. If you are uncertain about the correct option, select the one that seems closest to being correct.

QA User Prompt

Memory:

timestamp	caption
0-60	#C C paints a picture
...	...

###

Question: What was the primary purpose of the cup of water in this video, and how did it contribute to the overall painting process?

Option 0: To provide a source of water for the paintbrush.
...

###

The dictionary with keys of prediction, explanation, confidence, where prediction is a number:

Figure 6: System prompt and user prompt for video QA (EgoSchema). The text in blue should be replaced by the captions and the corresponding question-answer pair.

NLQ System Prompt

You are individual C, with others represented as O. In your responses to questions about past events, it is vital to provide not only the key moment but also the relevant context. To enhance the clarity and reliability of your answers, please also indicate your confidence level in each response, with 1 being the lowest and 3 being the highest. Follow these guidelines:

1. Incorporate Context: Expand your answers to include not just the central event but also the context preceding and following it.
2. Unify Related Actions: When a question requires a sequence of actions, such as 'Where did I put the scarf after I closed the door?', merge all relevant events into a single interval that conveys the full story.
3. Opt for Broad Understanding: Favor comprehensive intervals that cover all relevant details over more precise but less informative ones. If the information is too vague, respond with 'NA' and include your confidence level to reflect the certainty of your response.
4. Assign a Confidence Level: After providing a time interval, add a confidence level to each response:
   - Level 1: The information is present, but the context is not clear or the captions are ambiguous.
   - Level 2: The information is fairly clear and context is somewhat discernible, but there is still some uncertainty.
   - Level 3: The information and context are clear and well-supported by the captions, ensuring high confidence.

NLQ User Prompt

As individual C, with others as O, employ your advanced memory recall capabilities to pinpoint the time intervals that best respond to the questions provided. Ensure each answer encompasses not just the event in question but also the relevant context before and after. If the details are too vague or insufficient for a confident recall, indicate 'NA'. Remember to merge related events into a continuous interval to reflect the full narrative.

###

Questions:

query_index	query
1	where was the jenga block?
...	...

###

Memory:

timestamp	caption
150-210	C converses with man X
...	...

###

Please provide a TSV with columns: query_index, predictions, explanation, confidence.

Figure 7: System prompt and user prompt for NLQ. The text in blue should be replaced by the queries and captions in the video clip.

B Additional Qualitative Results

We visualize the outputs of the LLM in Figure 8 and Figure 9. As shown by abundant qualitative examples, the LLM can produce high-quality answers in zero shot. It is worth noting that the machine-generated captions may contain objects that are not present in the video or miss critical information that can potentially answer the target query. Based on the imperfect captions, the LLM is still able to capture the key event and produce high-quality responses, suggesting a more powerful captioning model will further boost the performance.

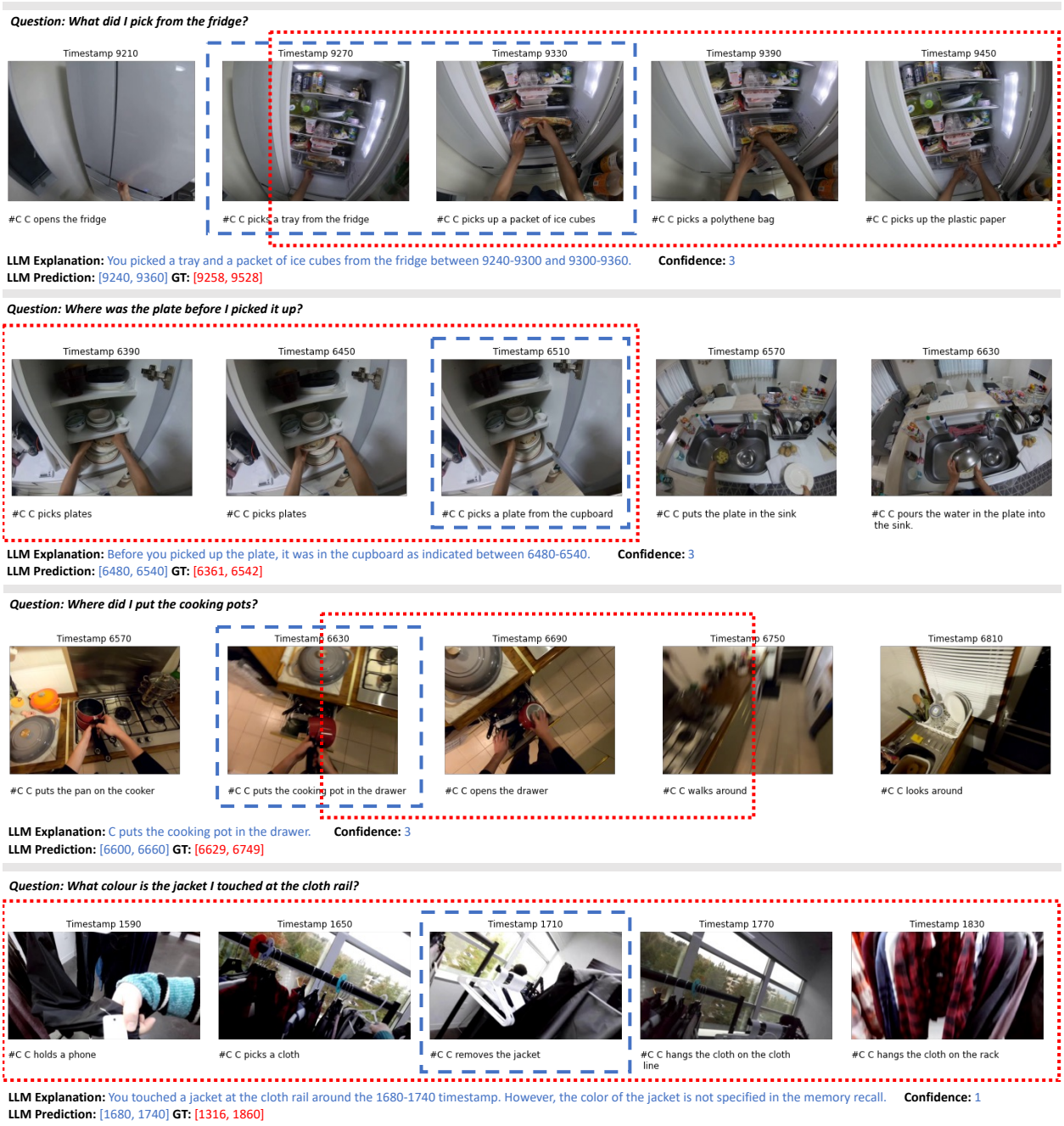


Figure 8: NLQ Examples. Each figure represents a two-second 30fps video clip (which is 60 frames). LLM interval predictions are denoted as blue boxes and the ground truth is in red. LLM engine here is GPT-4 and the captioning model is LaViLa. To illustrate the reasoning skills of LLMs, we show the raw LLM predictions without any refinement.

Q: Based on the repetitive actions throughout the video, what would you say is c's primary task or goal in the given environment?

A0: To move the chair.
A1: To clean the floor.
A2: To touch the shelf.
A3: To touch the cat.
A4: To move the guitar bag.



LLM Prediction: 1 **Confidence:** 3
LLM Explanation: C's primary task is to clean the floor as it is the most repeated action in the video log.

Q: Can you identify the primary activity that c is engaged in throughout the video, and how they incorporated breaks or pauses during this activity?

A0: C is working on a computer.
A1: Currently, c is spending time attentively watching an entertaining movie.
A2: Currently, individual c is enjoying and listening to their favorite music.
A3: C is playing a video game.
A4: Currently, c is peacefully sleeping and resting.



LLM Prediction: 3 **Confidence:** 3
LLM Explanation: Throughout the video, C is primarily engaged in playing a video game, with occasional breaks to look around and at one point to take a cup of water.

Q: In the context of the video, what is the primary purpose of the character's actions and how do they achieve this purpose?

A0: The character is making bread.
A1: In this scene, the main character is enthusiastically making a delicious pizza.
A2: The main character is joyfully making delicious cookies at home.
A3: The main character in the story is skillfully making a delicious cake.
A4: The character is making dough.



LLM Prediction: 4 **Confidence:** 3
LLM Explanation: The character is seen pouring water, using a mixer, and a weighing scale, which are actions typically associated with making dough.

Q: What is the primary purpose or task that c completes at the parking lot, and how does this compare to earlier actions in the video?

A0: You will need to carefully fold the receipt into a smaller size.
A1: Leisurely stroll to walk around the extensive parking lot area.
A2: In order to open the car boot effortlessly.
A3: To move the supermarket stroller.
A4: To put the shopping bags in the car boot.



LLM Prediction: 4 **Confidence:** 3
LLM Explanation: The primary task that C completes at the parking lot is putting the shopping bags in the car boot, as indicated by the captions at timestamps 2370, 2850, and 2970.

Figure 9: EgoSchema Examples. Each figure represents a two-second 30fps video clip (which is 60 frames). LLM predictions are denoted as blue boxes and the ground truth is in red. LLM engine here is GPT-4 and the captioning model is LaViLa.

C Ambiguity in NLQ Annotations

LLMs generate more than one interval when there are multiple temporal windows that can potentially answer the given query. We observe that some temporal windows proposed by the LLM seem reasonable despite only one ground-truth answer available in Ego4D NLQ annotations. These queries need to be filtered or modified to reduce ambiguity.

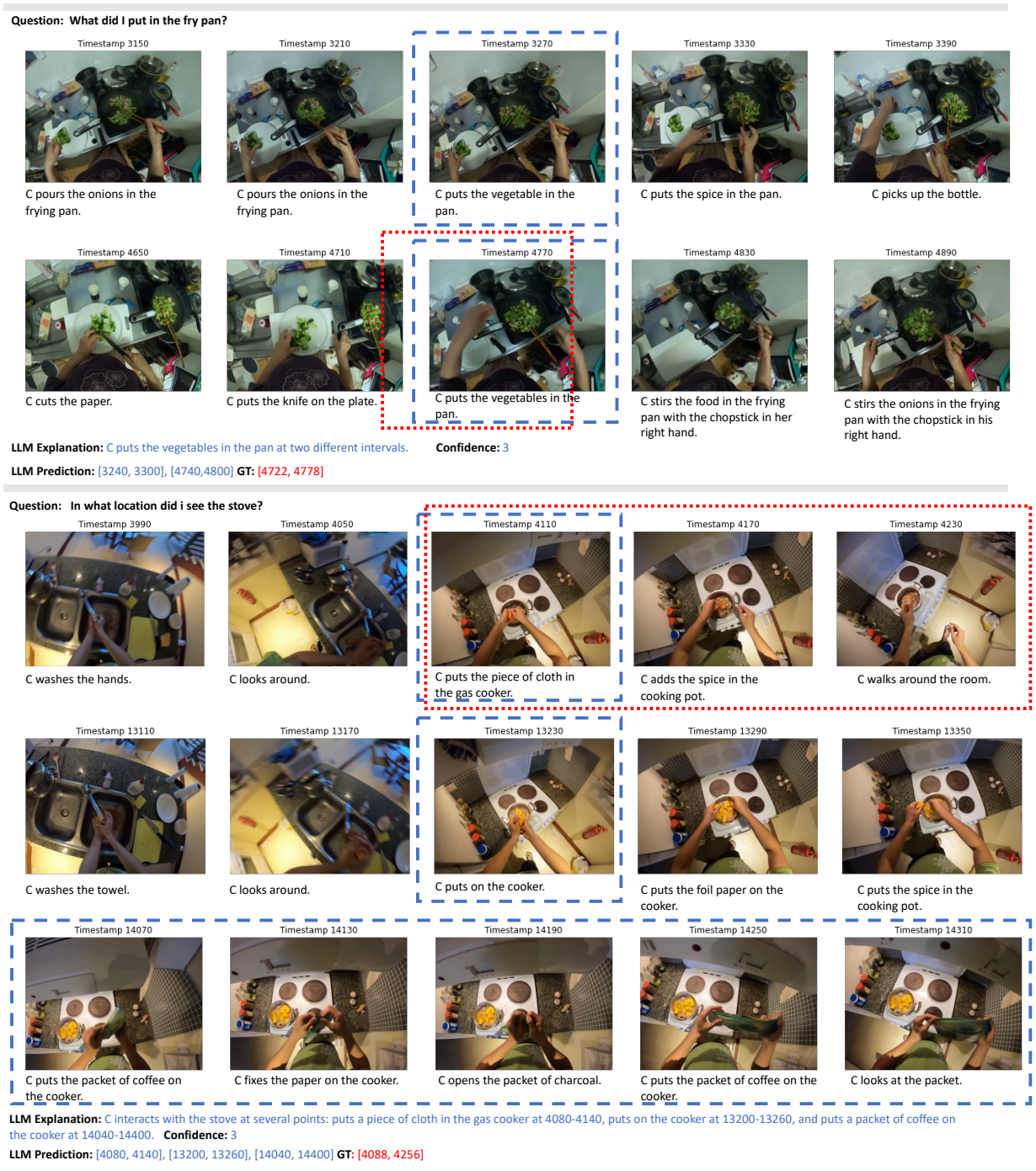


Figure 10: Examples of Ambiguous Queries. LLM predictions are in blue and the ground truth is in red.