RAVU: Retrieval Augmented Video Understanding with Compositional Reasoning over Graph

Sameer Malik, **Moyuru Yamada**, **Ayush Singh** and **Dishank Aggarwal**
Fujitsu Research of India Private Limited
{sameer.malik, yamada.moyuru, ayush.singh, dishank.aggarwal}@fujitsu.com

Abstract

Comprehending long videos remains a significant challenge for Large Multi-modal Models (LMMs). Current LMMs struggle to process even minutes to hours videos due to their lack of explicit memory and retrieval mechanisms. To address this limitation, we propose **RAVU** (**R**etrieval **A**ugmented **V**ideo **U**nderstanding), a novel framework for video understanding enhanced by retrieval with compositional reasoning over a spatio-temporal graph. We construct a graph representation of the video, capturing both spatial and temporal relationships between entities. This graph serves as a long-term memory, allowing us to track objects and their actions across time. To answer complex queries, we decompose the queries into a sequence of reasoning steps and execute these steps on the graph, retrieving relevant key information. Our approach enables more accurate understanding of long videos, particularly for queries that require multi-hop reasoning and tracking objects across frames. Our approach demonstrate superior performances with limited retrieved frames (5-10) compared with other SOTA methods and baselines on two major video QA datasets, NExT-QA and EgoSchema.

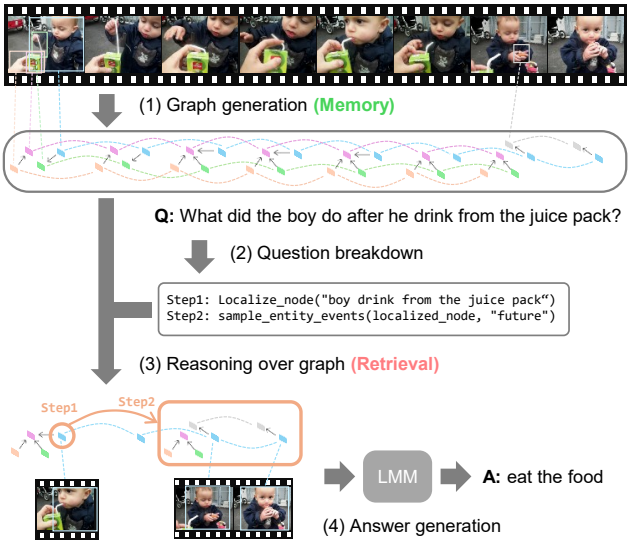

Figure 1: Overview of RAVU, memorizing a video as a spatio-temporal graph and retrieving key relevant parts by reasoning over the graph with a given query.

1 Introduction

Understanding videos inherently requires the ability to memorize multi-modal information and retrieve it according to a given task. Recent advancements in Large Multi-modal Models (LMMs) have shown promise in tackling this challenge [Song *et al.*, 2024a; He *et al.*, 2024; Wang *et al.*, 2024a]. However, comprehending long videos, particularly multi-hop reasoning tasks, remains a significant challenge, even for these powerful models.

This limitation primarily stems from the absence of explicit memory and retrieval mechanisms in current Transformer-based LMMs. The current LMMs represent each video frame as hundreds of tokens and thus have difficulty in processing hours of video content. Even at 64 tokens per frame an hour-long video could require over 200k tokens [Shen *et al.*, 2024]. Without retrieval mechanism, they need to take the entire video as input even for questions about a specific part of a video. While some studies have explored constructing a long-term memory from an input video [Song *et al.*, 2024a; He *et al.*, 2024], these approaches either sample key frames from the video or compress the video by grouping similar frames, regardless of the input query, potentially overlooking crucial details for the specific queries. Other methods [Song *et al.*, 2024b] have proposed constructing long-term memories based on query relevance but require recompressing the video for each query. Additionally, agentic approaches [Wang *et al.*, 2024a] have been explored, where relevant frames are iteratively retrieved until sufficient information is obtained to answer the query. These existing approaches highly rely on simple similarity between the query and individual frames, lacking the capability to track the identity of objects across consecutive frames. For instance, they may fail to correctly identify which man in the previous frame corresponds to the man performing a specific action in the current frame. Such temporal connections are essential for accurately understanding videos with complex queries, such as

"How does the girl in red react after being pulled backwards by the girl in blue?". These queries also necessitate multi-hop reasoning and may require more than simple frame-level relevance to identify important scenes in the long video.

The video is often represented as a graph in the fields of interaction detection [Yang *et al.*, 2023; Chen *et al.*, 2023], where nodes correspond to objects and edges represent the interaction between the objects. The graph then evolves through time. Some studies [Fei *et al.*, 2024a; Wang *et al.*, 2024b] have initiated explorations of using graph for video understanding. However their models are trained on a specific datasets and may struggle for generalization.

To address these limitations, this paper proposes **RAVU** (Retrieval Augmented Video Understanding), a novel framework for video understanding based on compositional reasoning over a spatio-temporal graph. We first construct a spatio-temporal graph from the video with the LMM. This graph is generated once per the video, independent of the queries, and serves as a memory. In this memory, the same entities (e.g., a man and a dog) are connected across frames, allowing us to track the actions of specific individuals over time. Unlike the conventional methods which simply retrieve the video frames based on the similarity between the query and frames, we first decompose the complex query into reasoning steps and then retrieve the necessary scenes for answering the query by performing each reasoning step on the graph sequentially. While we implement various reasoning steps to cover a wide range of queries, it is possible to employ neural networks for each step. Furthermore, our memory and retrieval framework can be readily applied to existing LMMs without fine-tuning or can be fine-tuned on open-source LMMs.

The main contributions of our paper can be summarized as follows:

- We propose a novel video understanding framework, which constructs a spatio-temporal graph as a memory from a video and retrieves key frames from the video by reasoning over the graph.
- We introduce a novel pipeline to generate the expressive spatio-temporal graphs from the video using a LMM.
- We also introduce key functions to perform multi-hop reasoning over the spatio-temporal graphs.
- Our comprehensive experiments and analysis demonstrate the effectiveness of our approach.

2 Related Work

The field of video understanding has seen significant advancements, particularly with the integration of MLLMs. This section reviews key contributions and methodologies that have shaped the landscape of video comprehension.

2.1 Large Multi-Modal Models

Recent advances in Large Multi-Modal Models (LMMs) have demonstrated remarkable competencies in various tasks such as captioning and visual question answering. The LMMs like GPT-4V [Achiam *et al.*, 2023], Gemini-1.5 [Team *et al.*, 2024], and LLaMA 3.2 [Dubey *et al.*, 2024] take a text prompt and a set of images as inputs and generate a rich text as an output which follows the input prompt as an instruction. Inspired by unprecedented capabilities of such LMMs, recent studies [Wang *et al.*, 2024a; Shen *et al.*, 2024; He *et al.*, 2024] have initiated explorations of extending LMMs for video understanding tasks. A primary challenge for the LMMs to understand the video contents is that it is impractical to process all the frames in the video since they typically convert a raw image into a sequence of tokens (visual tokens) using an image encoder [Dosovitskiy *et al.*, 2021] or vision-language models [Radford *et al.*, 2021; Li *et al.*, 2023]. Due to their limited context length, they mostly can handle only few minutes of videos. This paper proposes a novel framework to address this limitation.

2.2 Long-Term Memory for Video Understanding

To address the limitation, various methods have been proposed for compressing long videos. Some techniques merge similar frames to create a long-term memory [He *et al.*, 2024; Song *et al.*, 2024a], but they may miss crucial details and struggle with hallucinations. Later studies introduced query-aware memory [Song *et al.*, 2024b; Shen *et al.*, 2024], which adaptively merges frames based on their similarity to an input query. For example, MovieChat+ [Song *et al.*, 2024b] adjusts the compression ratio based on similarity, but this requires recompressing the video for each query, adding computational overhead. Other methods [Zhang *et al.*, 2024; Islam *et al.*, 2024] divide the video into segments, generate textual descriptions for each, and store these as long-term memory. However, these descriptions might not capture essential details. Unlike these methods, we use a spatio-temporal graph to represent the video as long-term memory. This graph is generated once per video and retrieves key frames regardless of video length, allowing efficient processing of long videos.

2.3 Graph as Structured Video Representation

Representing videos as graphs has proven effective in various tasks, particularly in interaction detection [Yang *et al.*, 2023; Chen *et al.*, 2023], where nodes represent objects and edges capture their interactions. This structured representation allows for capturing spatio-temporal relationships. Some studies [Fei *et al.*, 2024a; Wang *et al.*, 2024b] have explored using graphs for video understanding, but their models may struggle with long videos and generalization due to training on specific datasets. We propose a novel pipeline to generate expressive spatio-temporal graph from a video with a LMM. Instead of feeding the graph into LLM, we run pre-designed reasoning functions over the graph to retrieve the key frames.

2.4 Multi-Step Reasoning for Video QA

Some recent studies [Wang *et al.*, 2024a; Jeong *et al.*, 2025] also have explored approaches to retrieve key frames relevant to the input query from a long video instead of compressing the video. Agentic approaches like VideoAgent [Wang *et al.*, 2024a] iteratively retrieve relevant frames until sufficient information is gathered. However, these methods rely heavily on simple similarity between the query and individual frames, lacking the capability to track entities across frames. This limitation hinders their ability to handle complex queries

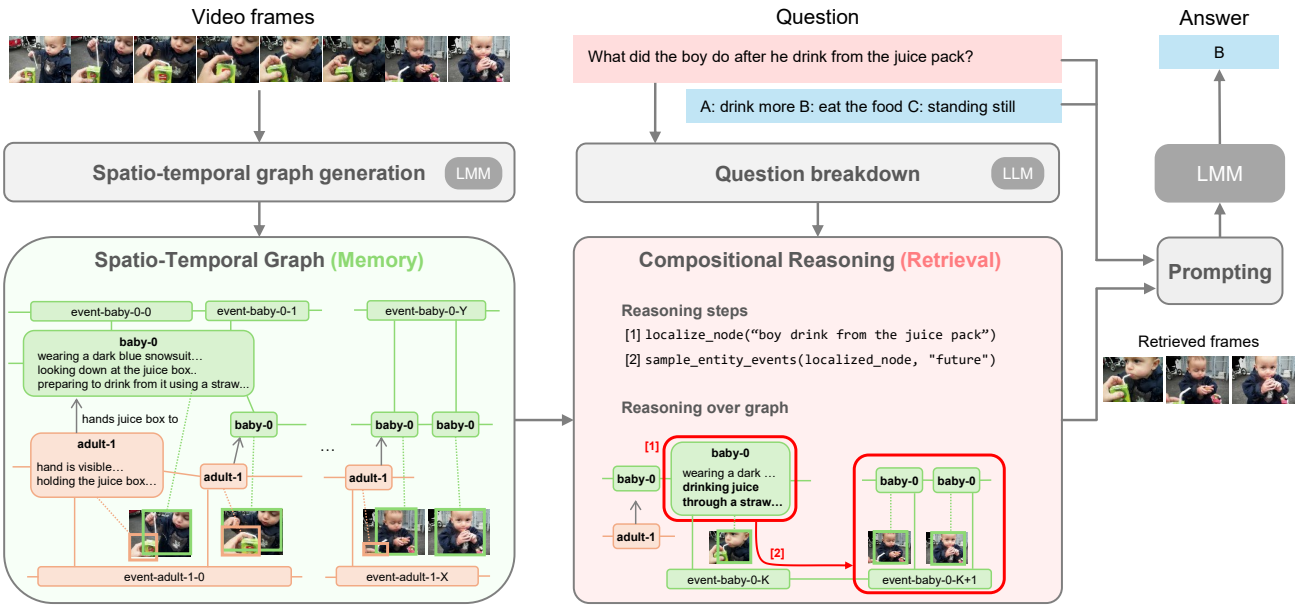

Figure 2: An entire pipeline of our RAVU consisting two core components: spatio-temporal graph generation (memory) and compositional reasoning (retrieval). These memory and retrieval mechanisms form the core of our proposed framework. Through the reasoning over the graph, the relevant frames can be identified and they are fed into the LMM to get a final answer for the given question.

that require multi-step reasoning to understand temporal relationships. To handle such multi-step reasoning, [Fei *et al.*, 2024a] has designed a specific reasoning steps. Unlike the existing works, we decompose the complex query into a sequence of reasoning steps and then execute each step with the reasoning function we designed for video understanding to retrieve the key frames.

3 Method

In this section, we introduce our novel video understanding framework called **RAVU** (**R**etrieval **A**ugmented **V**ideo **U**nderstanding) and discuss our approach to VideoQA through compositional reasoning over graphs in detail.

3.1 Overview

Figure 2 illustrates an overview of our proposed framework, RAVU which consists of two key components: (1) a spatio-temporal graph generation module that constructs a structured representation of the video content (memory), and (2) a compositional reasoning module that operates over these graphs to localize relevant segments in response to user queries (retrieval). These memory and retrieval mechanisms constitute the foundation of RAVU.

3.2 Spatio-Temporal Graph Generation

We represent a video as a spatio-temporal graph, where each frame is modeled as a sub-graph comprising entity nodes and their relationships as edges. The entity nodes contain their visual attributes and spatial location in the frame. Nodes corresponding to the same entity across consecutive sub-graphs are connected to track the entities through time and capture their temporal dynamics. While prior work has explored scene graph generation from images and videos [Zhu *et al.*, 2022; Yang *et al.*, 2023; Chen *et al.*, 2023], these methods often suffer from limited vocabulary and poor generalization due to the constraints imposed by the small size of the training data. To address these limitations, we employ an LMM (Large Multimodal Model) to generate frame-wise graphs and utilize object tracklets to establish temporal connections across frames. However, our preliminary experiment revealed two key challenges. First, directly prompting the LMM to identify entities, their attributes, and relationships within a frame often yields low-quality graphs, as the model tends to use the limited relation vocabulary. Second, establishing consistent correspondence between the entities across frames is not straightforward. This difficulty arises because visually similar entities might actually be distinct instances, leading the LMM to assign different IDs to the same entities across different frames.

We propose a multi-step approach to enhance the robustness of the spatio-temporal graph generation, as shown in Fig.3. Our approach first generates expressive descriptions of video frames and entities in each frame, then converts these descriptions into a spatio-temporal graph. Specifically, we begin by detecting entities in each frame and tracking them based on bounding box matching to assign consistent IDs across the frames. Next, we annotate the entities with distinct color-coded bounding boxes for every frame and prompt the LMM to generate the frame description, referring the entities by their IDs. We include a bounding-box color-to-ID map in the prompt, enabling the LMM to associate the bounding-box colors with the tracked identities. To generate rich and precise descriptions, we feed the LMM with a sequence of N frames, consisting of a target frame and its neighboring frames. Fi-

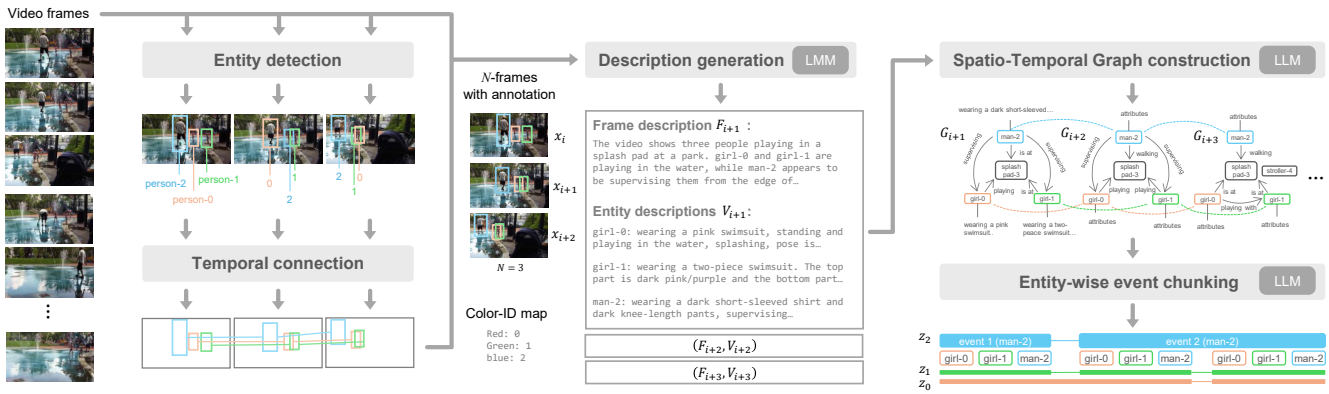

Figure 3: Spatio-temporal graph generation pipeline. Firstly, we detect entities in each frame and connect the detected entities across the frames, followed by LMM generates expressive descriptions for the frame and the entities with bounding box annotation. Then, we use LLM to construct the graph from them. The entity-wise events in the video are also generated to capture the segments of the video.

nally, we prompt a LLM (Large Language Model) to convert the consistent frame descriptions and entities into a graph for each frame to construct a spatio-temporal graph.

More formally, we process each frame x_i $(i = (0, ..., N-1))$ annotated with the entities using the LMM g_m with the instructions provided in the system prompt s_{fd}. This generates a tuple $(V_i, F_i) = g_m(s_{fd}, x_i)$, where, V_i is the set of entities $\{n_i^j\}$ $(j \in \{0, ..., M_i\})$ in the frame-i. Here, M_i denotes the cardinality of the set V_i. F_i is the frame description generated by referring to the entities by their IDs. Subsequently, for each frame F_i, LLM g_l converts the consistent frame descriptions and entities into a graph $G_i(V_i, E_i) = g_l(V_i, F_i, s_g)$. Here, E_i contains the edges representing the relationships between the entities from the frame description F_i. s_g represents the system prompt with instructions to construct the graph for each frame. Note that each node contains rich attributes and its location obtained in the first step. Additionally, the prompt in this step utilizes only the entities and the frame description in textual format, excluding the video frame itself.

To facilitate temporal reasoning, we further augment the graph by creating entity-wise events. This process involves chunking the spatio-temporal nodes for each entity into distinct events, thereby capturing significant behavioral and action changes. Specifically, for the entity-j, we generate the events z^j as $z^j = g_l(\{p_i^j\}_{i=0}^{N-1}, s_e)$, where the system prompt s_e contains instructions to segment the node into events, capturing the entity's actions and behaviors across the frames.

3.3 Compositional Reasoning over Graph

In this section, we present our methodology for localizing segments of a target video pertinent to answering a given question by reasoning over a spatio-temporal graph of the video. We employ dual-system models [Kahneman, 2011] in cognitive science to handle complex questions which require multi-step reasoning to identify the relevant segments. Our approach begins by breaking down the complex question into a sequence of reasoning steps. Then, we perform each reasoning step with a predefined function in sequence. Unlike existing works [Surís *et al.*, 2023; Ukai *et al.*, 2024], our reasoning functions operate on a spatio-temporal graph.

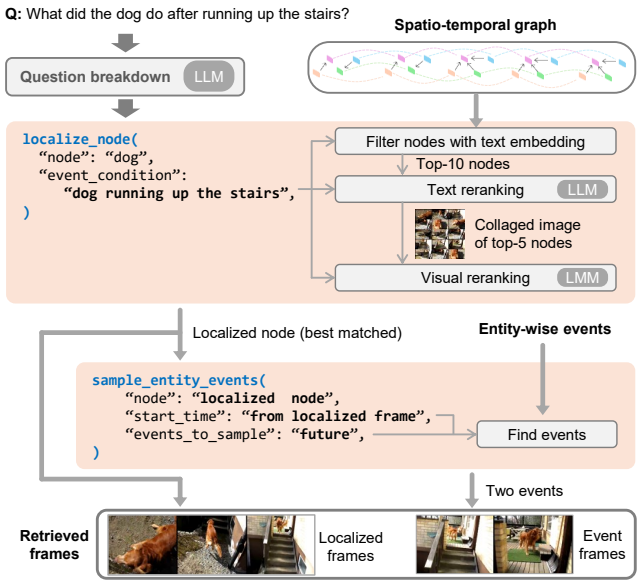

Figure 4: An example of our frame retrieval process. We execute reasoning steps sequentially and concatenate frames from each step.

We first break down the complex questions using a LLM. To facilitate this process, we manually create examples of question analysis and breakdown for various types of questions, including temporal, descriptive, and causal, then use these examples in-context when analyzing new questions. The primary functions within our set of predefined actions include *localize_node* and *sample_entity_events*. See our complete set of functions in Supplementary.

After breaking down the question into the sequence of reasoning steps, we perform them sequentially to retrieve the query-relevant key frames. We illustrate our retrieval process through an example in Figure 4. The question often re-

quires multi-step reasoning to identify the relevant frames. Specifically, if the question is "What did the dog do after running up the stairs?", the spatio-temporal entity node "dog" with the attribute "dog running up the stairs" must first be identified. The function *localize_node* identifies this spatio-temporal node, finding the best match of the phrase p_g=*"dog running up the stairs"* from the among the set of nodes $\{n_i^j\}$. We compute node embeddings by encoding concise textual description of the nodes with a text embedding model g_e. We first obtain the textual description p_i^j of the node n_i^j as,

$$p_i^j = g_l(n_i^j, \{e_i\}, s_d),$$

where, $\{e_i\}$ is the set of all edges in frame-i that include entity-j and s_d is the system prompt with instructions to compose a sentence that encapsulates the entity node's attributes and its relations to other entity nodes. We then compute embeddings of each node as $v_i^j = g_e(p_i^j)$. To localize the node, we first select the top-k spatio-temporal nodes whose embeddings exhibit the highest cosine similarity with the grounding phrase embeddings $v_g = g_e(p_g)$. Let $\eta = \{n_0, ..., n_{k-1}\}$ be the top-k filtered nodes and $P = \{p_0, ..., p_{k-1}\}$ be the corresponding node textual descriptions. Subsequently, we process the textual descriptions of the selected node to find the best matching node as $\hat{k} = g_l(P, p_g, s_r)$, where s_r is the system prompt with instructions to select the phrase that best matches the grounding phrase and $n_{\hat{k}}$ is the best matched node. Let $n_{\hat{k}}$ correspond to the node of entity-j in frame-i. Finally, the function *sample_entity_events* is invoked to retrieve frames from events of entity-j preceding the time index t_e.

4 Experimental Settings

In this section, we provide a comprehensive description of the datasets utilized for evaluating our method. Subsequently, we elaborate on the various baselines and the implementation details.

4.1 Evaluation Dataset

We benchmark our method and compare it with various baselines and SOTA methods on two popular datasets, Next-QA [Xiao *et al.*, 2021] and EgoSchema [Mangalam *et al.*, 2023].
NExT-QA: We follow [Wang *et al.*, 2024a] to focus on zero-shot evaluation on the validation set of the NExT-QA. It contains 570 videos and 5,000 multiple choice questions. NExT-QA provides 8 types of questions, including 2 types of causal questions, 3 types of temporal question, and 4 types of descriptive questions. Especially, this paper aims to address the temporal next/previous questions which are particularly difficult and require multi-hop reasoning to infer the past or future. These types of questions compose 29% of the dataset. The videos in this dataset average 45 and maximum 180 seconds in length.
EgoSchema: EgoSchema is a benchmark for zero shot comprehension of the long-form videos, containing 5,000 multiple-choice questions based on 5,000 egocentric videos. These videos capture a first-person perspective of individuals participating in various activities. Due to the focus on zero-shot evaluation, this dataset only comprises of the test set. Each video in this dataset is 3 minutes long. We follow [Wang *et al.*, 2024a] and evaluate on a subset of this dataset containing 500 questions corresponding to 500 videos having publicly accessible labels.

We employ accuracy as our evaluation metric since the datasets features multiple-choice questions.

4.2 Baselines and Other Methods

We compare RAVU with other state-of-the-art methods including both supervised and zero-shot methods, such as LongVU [Shen *et al.*, 2024], LLoVi [Zhang *et al.*, 2024], and VideoAgent [Wang *et al.*, 2024a]. However, it is not easy to compare our method with other existing zero-shot methods since they rely on different proprietary models (e.g., GPT-3.5 and GPT-4). Therefore, for a fair comparison, we conduct following exhaustive baseline experiments with a specific LMM as a fixed reasoning model across all the experiments.

- **BlindQA**: We just feed the multiple choice question (MCQ) and do not feed the video frames to the LMM.
- **All frames**: We feed all the frames (at 1fps) to the LMM to answer the question.
- **Image-based frame retrieval**: Here we employ a CLIP model to retrieve the top-5 frames most relevant to the query based on text-image similarity. These selected frames were passed to the LMM model for answering the question.
- **Text-based frame retrieval**: We use the frame descriptions as the retrieval key instead of the frame and compute the similarity to the query in the embedding space to find the top 5 relevant frames. These query relevant frames are then passed to the LMM for the question answering task.

4.3 Implementation Details

We use *gemini-1.5-flash-002* for detecting entities, generating the frame descriptions, constructing a graph, question breakdown, and inferring the final answer for the questions for all the experiments. This is a Gemini Flash model, and it is cheaper and faster than the Gemini Pro model, and thus can be used more often in the realistic scenarios. For entity tracking, we use SAM2 [Ravi *et al.*, 2024] as a tracker for EgoSchema and annotation in VidOR dataset [Shang *et al.*, 2019] for NExT-QA. We also use SAM2 for NExT-QA for comparison. For retrieval, we use a Sentence Transformer [Reimers and Gurevych, 2019] of *all-mpnet-base-v2* model. Video frames are uniformly sampled from the videos at 1 fps. For baseline experiments, we use *EVA02-CLIP-L-14* for text-image retrieval and Gemini *text-embedding-004* model for text retrieval. We set all the safety filters to the lowest option, while 33 videos were blocked.

Recent approaches [Zhang *et al.*, 2024; Wang *et al.*, 2024a] uses question options to retrieve the relevant frames or texts since these can serve as key words that directly correspond to the related images or the text, while this setting may not be a realistic scenario and thus we do not use them for retrieval.

Models	Acc_C	Acc_T	Acc_D	Acc
Human	87.61	88.56	90.4	88.38
Supervised				
HiTeA [Ye *et al.*, 2023]	62.4	58.3	75.6	63.1
VFC [Momeni *et al.*, 2023]	49.6	51.5	63.2	52.3
Vamos [Wang *et al.*, 2023a]	**77.2**	**75.3**	81.7	**77.3**
SeViLA [Yu *et al.*, 2023]	73.8	67.0	81.8	73.8
MotionEpic [Fei *et al.*, 2024b]	75.8	74.6	**83.3**	76.0
VLAP [Wang *et al.*, 2023b]	74.9	72.3	82.1	75.5
ViLA [Wang *et al.*, 2025]	75.3	71.8	82.1	75.6
Zero-shot				
AssistGPT [Gao *et al.*, 2023]	60.0	51.4	67.3	58.4
SeViLA [Yu *et al.*, 2023]	61.3	61.5	75.6	63.6
ViperGPT [Surís *et al.*, 2023]	-	-	-	60.0
LLoVi [Zhang *et al.*, 2024]	67.1	60.1	76.5	66.3
VideoAgent [Wang *et al.*, 2024a]	72.7	64.5	**81.1**	71.3
MotionEpic [Fei *et al.*, 2024b]	-	-	-	66.5
RAVU (non-blocked content)	**76.67**	**68.91**	76.11	**74.09**
RAVU (overall)	74.40	66.56	74.64	71.93

Table 1: Results on NExT-QA for Supervised and Zero-shot state-of-the-art methods. Acc_C, Acc_T, Acc_D and Acc represent accuracy on causal, temporal, descriptive subsets and overall accuracy, respectively. We bold the best results.

5 Results and Analysis

5.1 Comparison with State-of-the-arts

NExT-QA: In Table 1, we present the performance of our proposed RAVU model alongside other state-of-the-art (SOTA) supervised and zero-shot video understanding methods on the NeXT-QA dataset. It is important to note that our approach incorporates the proprietary Gemini model as a foundational component. Due to the non-configurable safety protocols embedded within Gemini, certain questions, frames, or videos identified as unsafe content are consequently blocked. Specifically, for the NeXT-QA dataset, our evaluation was conducted on 4,856 out of 5,000 questions, with the remaining questions being blocked by Gemini. To ensure a fair comparison with other SOTA methods, we report the accuracy of RAVU on both the non-blocked questions and on all questions, under the assumption that the blocked questions are incorrect. Notably, despite the limited number of frames (an average of 5 frames per video), RAVU demonstrates competitive performance relative to other methods.

EgoSchema: EgoSchema comprises global behavioral questions that necessitate reasoning over entire videos. For such questions, we employ a hierarchical retrieval approach. Our method utilizes a spatio-temporal knowledge graph, which includes entity node descriptions and event segmentations for each entity. Initially, we retrieve the most relevant entity node descriptions from each event based on the similarity between the embeddings of the question and the frame descriptions within that event. We then prompt the LMM with these retrieved descriptions to select the top 10 descriptions that best match the query. The frames corresponding to these top 10 descriptions are subsequently fed to the LMM along with the query to generate the answer. This event-based approach ensures diversity and relevance in the sampled frames.

Methods	Acc
Supervised	
LongViViT [Papalampidi *et al.*, 2024]	56.8
MC-ViT-L [Balažević *et al.*, 2024]	62.6
Zero-shot	
SeViLA [Yu *et al.*, 2023]	25.7
LLoVi [Zhang *et al.*, 2024]	52.2
VideoAgent [Wang *et al.*, 2024a]	60.2
RAVU (non-blocked content)	**67.41**
RAVU (overall)	66.60

Table 2: Results on EgoSchema 500 video subset as compared to state-of-the-art methods.We bold the best results.

In Table 2, we present the performance of RAVU and other state-of-the-art (SOTA) supervised and zero-shot video understanding methodologies on the EgoSchema dataset. For the EgoSchema dataset, six questions were blocked by the Gemini model, resulting in an evaluation on 494 questions for the non-blocked setting. We observe that RAVU demonstrates competitive performance with just 10 retrieved frames.

5.2 Comparison with baselines

NExT-QA: We present the performance of the baseline methods and our proposed approach in Table 3. For a fair evaluation, all methods in this table were assessed on 4,596 questions that were not blocked by the Gemini model for any of the methods. We note that the proposed RAVU approach demonstrates superior performance compared to other retrieval-based baselines, while utilizing a similar number of frames to answer the questions. Notably, we observe significant performance improvements in the temporal category of questions. This is anticipated, as temporal questions often involve queries about the state of entities following or preceding the event in question, which are challenging to address with similarity based retrieval methods. Additionally, we report the cost for each method in terms of the average number of tokens per question for the NExT-QA dataset. Our proposed approach incurs a higher cost than other methods due to the query breakdown process and use of LMM in the retrieval process. Query breakdown incurs more than half of the cost with 3465 tokens per question due to in-context query breakdown illustrations. However, this cost can be significantly reduced through finetuning the LMM for query breakdown which will remove the need of in-context examples. Further, the cost of reasoning in retrieval can also be reduced through system prompt compression techniques [Mu *et al.*, 2024].

EgoSchema: We present the performance for the EgoSchema dataset in Table 4. For the frame retrieval baseline, we employed the CLIP model to retrieve the top 10 frames for each video. To ensure a fair evaluation, all methods in Table 4 were assessed on 490 questions from an equal number of videos, which were not blocked by the Gemini model for any of the methods. We note that RAVU exhibits superior performance compared to the frame retrieval-based baseline. This underscores the efficacy of our approach in retrieving frames for global behavioral question types when compared to the similarity based retrieval approach.

Methods	Acc_C			Acc_T				Acc_D				Acc	Cost
	CW	CH	All_C	TP	TN	TC	All_T	DC	DL	DO	All_D		(10^3)
BlindQA (only MCQs)	35.3	38.1	36.0	13	17.3	21.6	18.9	1.1	8.1	19.0	10.6	26.6	0.2
All frames (1 fps)	80.1	81.0	80.3	69.6	71.6	79.6	75.1	64.9	87.7	89.1	82.5	79.0	11.8
Clip-based retrieval (k=5)	74.2	76.2	74.7	70.2	59.2	72.5	65.0	51.1	**90.9**	82.0	**77.7**	72.3	1.4
Text-based retrieval (k=5)	74.2	76.7	74.4	70.2	59.0	73.7	65.4	**52.0**	89.3	83.5	76.9	72.6	1.4
RAVU (ours)	**76.7**	**77.1**	**76.8**	**78.7**	**64.0**	**75.5**	**69.2**	42.8	89.8	**85.4**	76.5	**74.6**	5.9

Table 3: Zero-shot performance comparison with baselines on **NExT-QA** with Gemini 1.5 Flash as a reasoning LMM. Acc_C, Acc_T, and Acc_D are accuracy on causal, temporal, and descriptive subsets, respectively. CW/CH: causal-why/how, TP/TN/TC: temporal previous/next/current, DC/DL/DO: descriptive count/location/others.We bold the best results. **Cost** denotes the input token count per question.

Methods	Acc
All Frames (1 fps)	70.67
CLIP-based retrieval	63.88
RAVU (ours)	**67.76**

Table 4: Zero-shot performance comparison with baselines on EgoSchema 500 video subset. We bold the best results.

Methods	Acc_C	Acc_T
Proposed Reranking	**70.57**	**58.15**
Text Embedding	58.69	44.05
CLIP Embedding	60.87	44.47

Table 5: Frame localization accuracy with different ranking algorithms on a subset of NExT-QA on causal and temporal questions.

5.3 Localization Analysis

In this section, we evaluate the accuracy of our frame localization methodology within the $localize\_node$ function. Our approach, particularly for causal and temporal questions, involves initially localizing the entity and event referenced in the question. Subsequently, we sample from future, past, or neighboring events based on the requirements of the question. Therefore, assessing the localization performance is crucial, as subsequent processes depend on accurate localization. To this end, we manually annotated 381 questions from 49 randomly selected videos from the NExT-QA dataset. Specifically, for each question, we identified the frames containing the event mentioned in the question, excluding frames depicting events occurring before or after the specified event.

To evaluate our method on this data, we compared the frame indices predicted by the $localize\_node$ function to the ground truth frame indices. If the predicted frame indices fall within the ground truth, we consider the prediction correct, otherwise, it is deemed incorrect. We compare the localization performance of CLIP Embeddings, text embeddings of entity node descriptions and our proposed reranking approach for localization. We report the localization results in Table 5. We note that our proposed approach results in significant localization performance gains when compared to other approaches.

Methods	Acc@C	Acc@T	Acc@D	Acc@All
w/ GT tracklets	76.16	74.19	73.58	75.29
w/ SAM2 tracklets	77.51	70.86	73.58	74.58
w/ VidOR annotation	78.24	75.16	74.54	76.65

Table 6: Zero-shot performance ablations with different graphs on a subset of NExT-QA.

5.4 Impact of Generated Graphs

To evaluate our graph generation methodology, we measure QA accuracy on 424 questions from our subset of 49 videos our using three distinct graph variants: (1) a graph generated by our approach utilizing ground-truth tracklets from the VidOR dataset, (2) a graph generated by our approach using tracklets predicted by SAM2, and (3) a graph constructed from human-annotated scene graphs in the VidOR dataset. We report the results in Table 6. The highest accuracy is achieved using the spatio-temporal graphs from the VidOR dataset. This suggests that entity tracking performance may be a critical bottleneck. While our graph demonstrates greater expressiveness compared to the VidOR graph, which has a limited vocabulary, its performance is slightly inferior. This can be attributed to the presence of hallucinations in our graph, which is generated by the LLM, unlike the human-annotated graphs from VidOR.

6 Conclusion

We introduced RAVU, a novel retrieval-augmented video understanding framework that constructs spatio-temporal graphs for long-term memory and compositional reasoning. By leveraging these graphs for frame retrieval, RAVU excels in addressing complex temporal, causal, and global reasoning tasks. Evaluations on NExT-QA and EgoSchema show superior performance in answering multi-hop and object-tracking queries with minimal frame retrieval, highlighting its effectiveness in video understanding.

Ethical Statement

This study does not involve any ethical concerns. All datasets used in this work, including NExTQA and EgoSchema, are publicly available and and no sensitive or personal information is used.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Balažević *et al.*, 2024] Ivana Balažević, Yuge Shi, Pinelopi Papalampidi, Rahmadi Chaabouni, Skanda Koppula, and Olivier J. Hénaff. Memory consolidation enables long-context video understanding. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, 2024.

[Chen *et al.*, 2023] Siqi Chen, Jun Xiao, and Long Chen. Video scene graph generation from single-frame weak supervision. In *The Eleventh International Conference on Learning Representations*, 2023.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Dubey *et al.*, 2024] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Fei *et al.*, 2024a] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 13109–13125. PMLR, 21–27 Jul 2024.

[Fei *et al.*, 2024b] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning*, 2024.

[Gao *et al.*, 2023] Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn, 2023.

[He *et al.*, 2024] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[Islam *et al.*, 2024] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos, 2024.

[Jeong *et al.*, 2025] Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. Videorag: Retrieval-augmented generation over video corpus, 2025.

[Kahneman, 2011] Daniel Kahneman. *Thinking, Fast and Slow*. 2011.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023.

[Mangalam *et al.*, 2023] Karttikeya Mangalam, Raiymbek Akshkulakov, and Jitendra Malik. Egoschema: a diagnostic benchmark for very long-form video language understanding. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 2023.

[Momeni *et al.*, 2023] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models, 2023.

[Mu *et al.*, 2024] Jesse Mu, Xiang Li, and Noah Goodman. Learning to compress prompts with gist tokens. *Advances in Neural Information Processing Systems*, 36, 2024.

[Papalampidi *et al.*, 2024] Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joe Heyward, Viorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzdeh. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14386–14397, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[Ravi *et al.*, 2024] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP (1)*, pages 3980–3990. Association for Computational Linguistics, 2019.

[Shang *et al.*, 2019] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceed-*

ings of the 2019 on International Conference on Multimedia Retrieval, pages 279–287. ACM, 2019.

[Shen *et al.*, 2024] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv:2410.17434*, 2024.

[Song *et al.*, 2024a] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232, June 2024.

[Song *et al.*, 2024b] Enxin Song, Wenhao Chai, Tianbo Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *ArXiv*, abs/2404.17176, 2024.

[Surís *et al.*, 2023] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11888–11898, October 2023.

[Team *et al.*, 2024] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[Ukai *et al.*, 2024] Mahiro Ukai, Shuhei Kurita, Atsushi Hashimoto, Yoshitaka Ushiku, and Nakamasa Inoue. Adacoder: Adaptive prompt compression for programmatic visual question answering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 9234–9243, 2024.

[Wang *et al.*, 2023a] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding, 2023.

[Wang *et al.*, 2023b] Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Lou, Ming C. Lin, and Shan Yang. Vlap: Efficient video-language alignment via frame prompting and distilling for video question answering. *CoRR*, abs/2312.08367, 2023.

[Wang *et al.*, 2024a] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *European Conference on Computer Vision (ECCV)*, 2024.

[Wang *et al.*, 2024b] Yanan Wang, Shuichiro Haruta, Donghuo Zeng, Julio Vizcarra, and Mori Kurokawa. Multi-object event graph representation learning for video question answering. *ArXiv*, abs/2409.07747, 2024.

[Wang *et al.*, 2025] Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Lou, Ming C. Lin, and Shan Yang. Vila: Efficient video-language alignment for video question answering. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 186–204, 2025.

[Xiao *et al.*, 2021] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, June 2021.

[Yang *et al.*, 2023] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, and Ziwei Liu. Panoptic video scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18675–18685, June 2023.

[Ye *et al.*, 2023] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15405–15416, October 2023.

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

[Zhang *et al.*, 2024] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering, 2024.

[Zhu *et al.*, 2022] Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, and Mohammed Bennamoun. Scene graph generation: A comprehensive survey. *CoRR*, abs/2201.00443, 2022.

Functions	Arguments	Descriptions
$localize\_node$	query	retrieves the most relevant node and corresponding frame
$sample\_entity\_events$	node, sample_start_time, events_to_sample	sample frames from relevant entity events
$extract\_temporal\_part$	target_part	extracts relevant video segment (beginning, middle or end)
$count\_nodes$	node, event_condition	called for counting questions
$get\_global\_context$	-	samples frames uniformly
$analyze\_events$	query	LMM analyzes events for temporal reasoning
$identify\_node$	query	uses LMM to identify entity node based on given query

Table 7: A list of our reasoning functions.

A Reasoning Functions

Table 7 provides a list of reasoning functions we design to handle multi-hop reasoning over a spatio-temporal graph.