Supplementary Material for: RAVU: Retrieval Augmented Video Understanding with Compositional Reasoning over Graph

A. RAVU Examples

In this section, we provide detailed visual examples of our RAVU framework applied to different question categories from NExT-QA dataset. These examples demonstrate how our method performs compositional reasoning over spatio-temporal graphs to retrieve relevant video segments and answer complex queries.

Our framework handles NExT-QA's five main categories of questions: (1) temporal-next questions that ask about future events, (2) temporal-previous questions that query past events, (3) temporal-current questions about simultaneous events, (4) causal-why questions that seek explanations for events, and (5) causal-how questions that ask about the manner of actions.

For each example, we show the complete pipeline including: the original question, the query breakdown into reasoning steps, the localized frames from our graph-based retrieval, the sampled frames from entity events, and the final answer generated by our system.

A.1. Temporal-Next Questions

Figure 1 demonstrates how RAVU handles temporal-next questions that require understanding of future events following a specific action in the video.

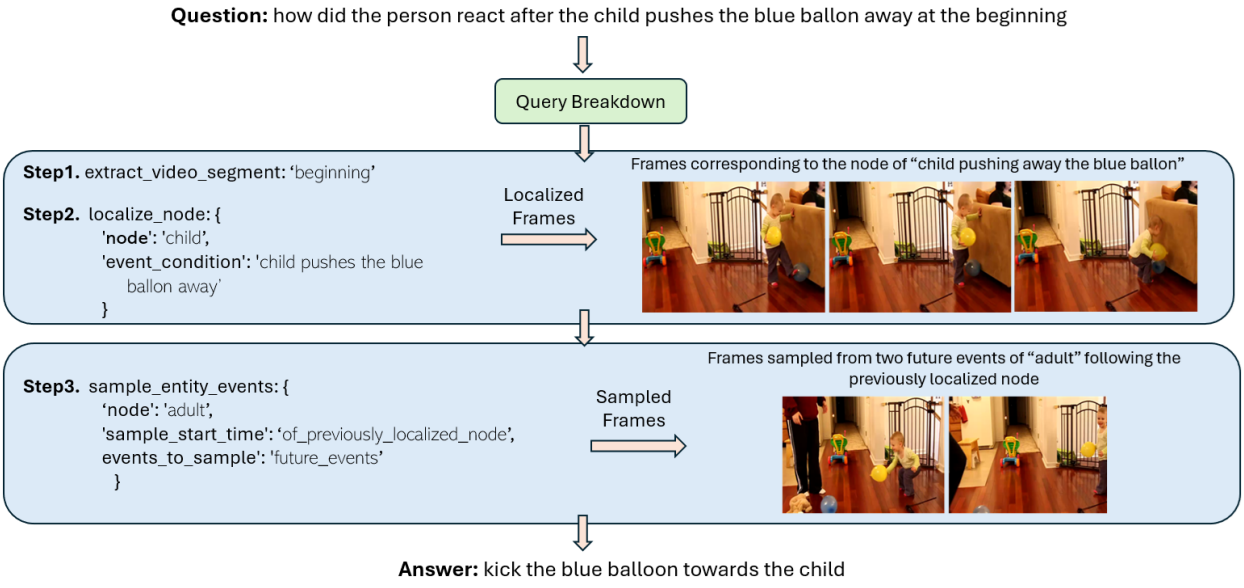

Figure 1. Example of temporal-next case: "How did the person react after the child pushes the blue balloon away at the beginning?" The system first extracts the video segment corresponding to "beginning", then localizes the node representing "child pushing away the blue balloon", and finally samples future events of "adult" following the previously localized node. The retrieved frames show the adult's reaction, leading to the answer "kick the blue balloon towards the child".

A.2. Temporal-Previous Questions

Figure 2 shows how our method processes temporal-previous questions that ask about events that occurred before a specific reference action.

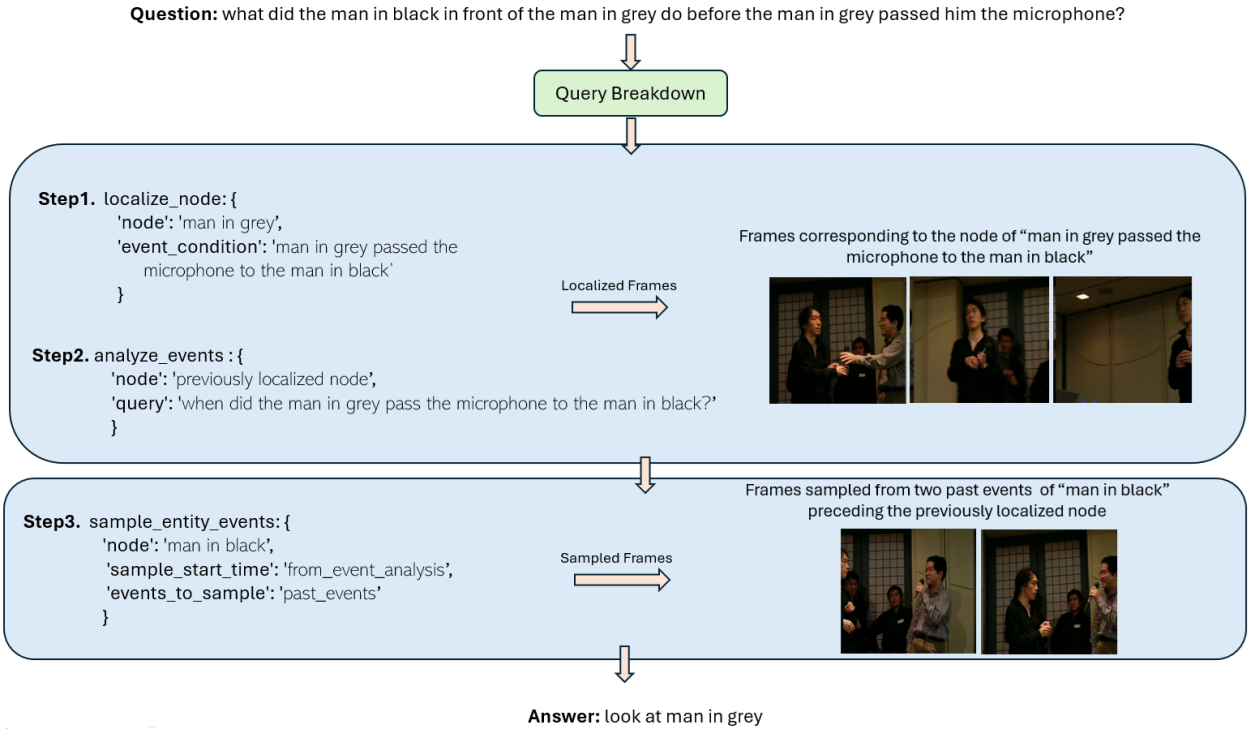

Figure 2. Example of temporal-previous case: "What did the man in black in front of the man in grey do before the man in grey passed him the microphone?" The system localizes the node of "man in grey" with the event condition "man in grey passed the microphone to the man in black", then analyzes events of the "man in black" preceding the previously localized node. The sampled frames from past events show the man looking at the other person, resulting in the answer "look at man in grey".

A.3. Temporal-Current Questions

Figure 3 illustrates how RAVU handles temporal-current questions that focus on simultaneous or concurrent actions.

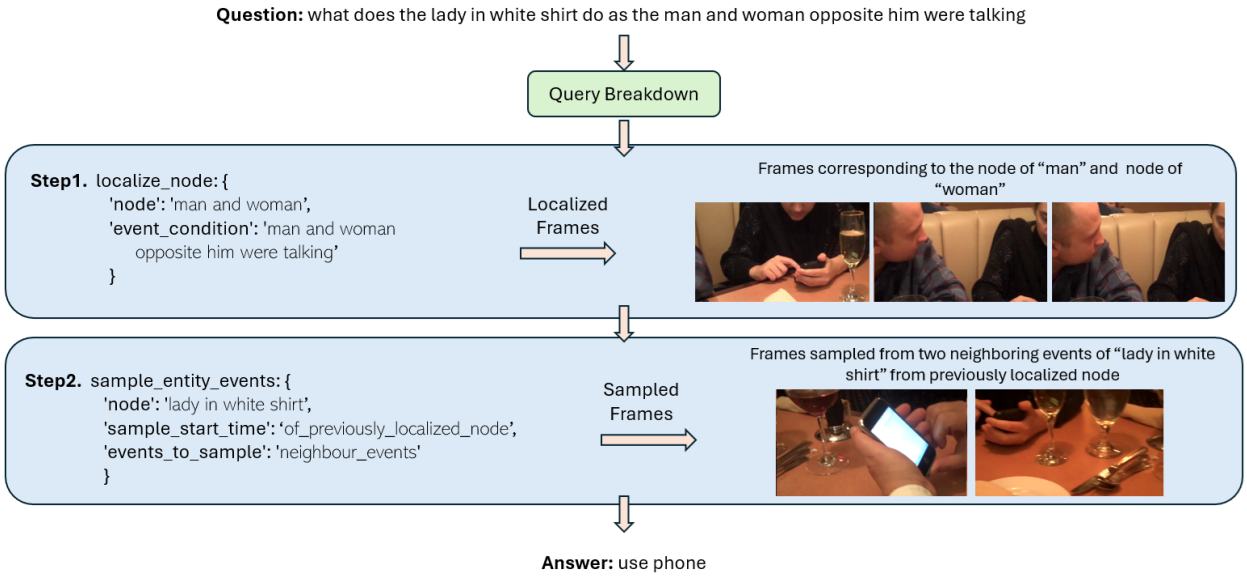

Figure 3. Example of temporal-current case: "What does the lady in white shirt do as the man and woman opposite him were talking?" The system localizes frames corresponding to the node of "man" and node of "woman" with the event condition "man and woman opposite him were talking", then samples neighboring events of "lady in white shirt" from the previously localized node. The retrieved frames show the lady using her phone, leading to the answer "use phone".

A.4. Causal-Why Questions

Figure 4 demonstrates how our framework addresses causal-why questions that seek explanations for why certain actions occur.

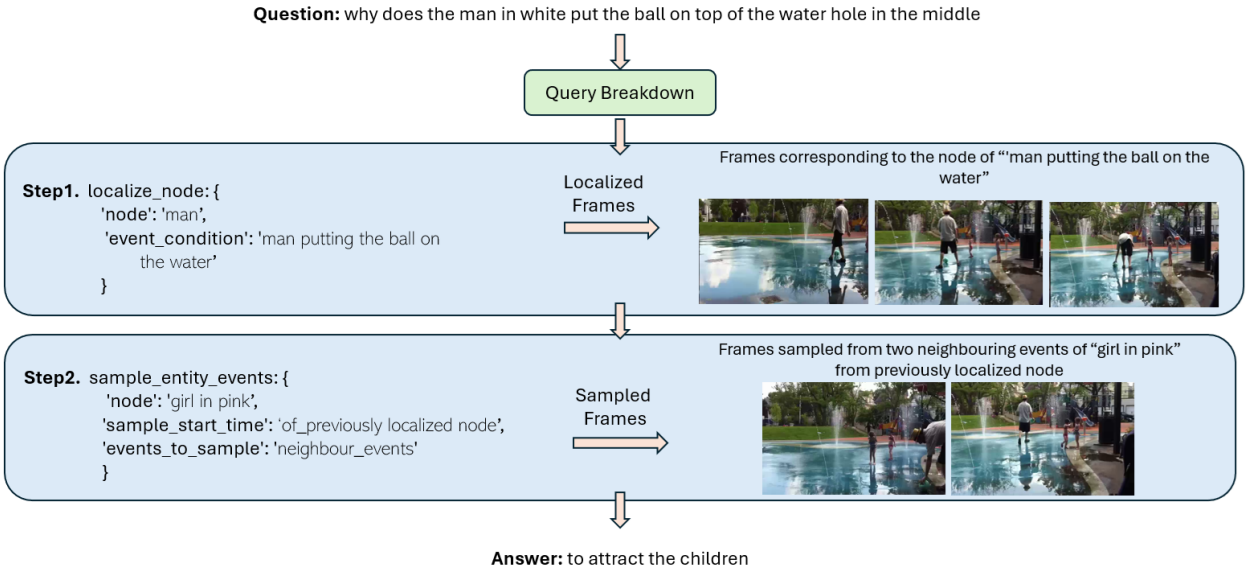

Figure 4. Example of causal-why case: "Why does the man in white put the ball on top of the water hole in the middle?" The system localizes the node of "man" with the event condition "man putting the ball on the water", then samples neighboring events of "girl in pink" from the previously localized node to understand the context. The retrieved frames show children around the water feature, leading to the answer "to attract the children".

A.5. Causal-How Questions

Figure 5 shows how RAVU processes causal-how questions that ask about the manner or method of performing actions.

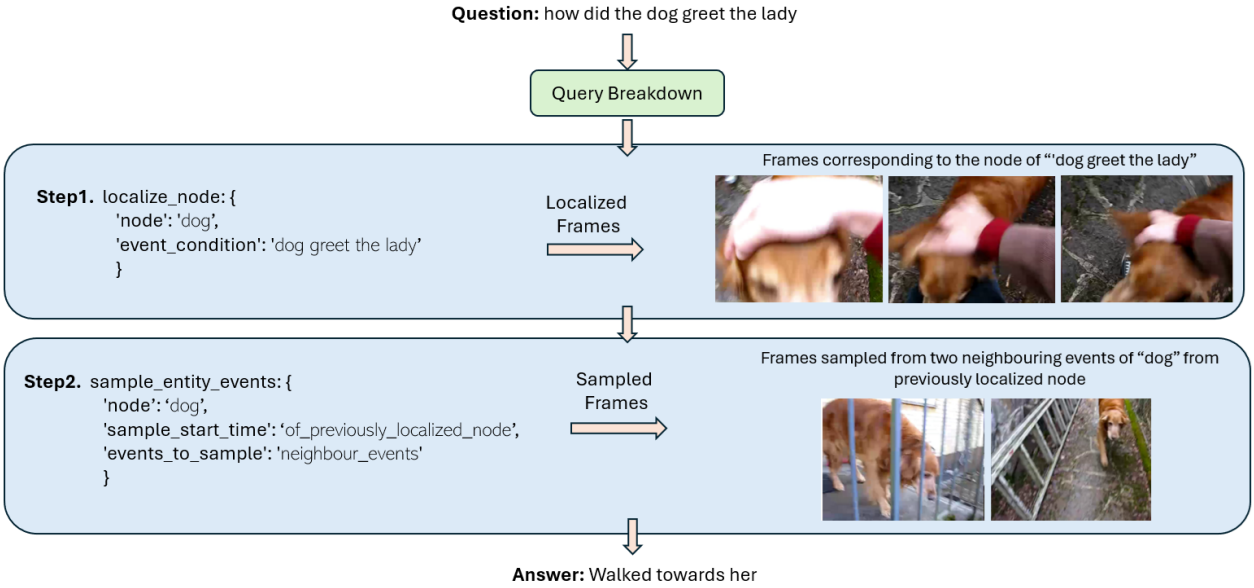

Figure 5. Example of causal-how case: "How did the dog greet the lady?" The system localizes the node of "dog" with the event condition "dog greet the lady", then samples neighboring events of "dog" from the previously localized node to capture the greeting behavior. The retrieved frames show the dog moving toward the lady, resulting in the answer "walked towards her".

The qualitative results shown in these examples validate our approach's effectiveness in handling diverse question types that require multi-hop reasoning and understanding of complex temporal and causal relationships in video content.

B. Reasoning Functions

Our RAVU framework employs a set of predefined reasoning functions that enable compositional reasoning over spatio-temporal graphs. These functions are the building blocks that allow our system to decompose complex video understanding queries into manageable reasoning steps. Table 1 provides a comprehensive overview of all reasoning functions used in our framework.

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_video_segment	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 1. A comprehensive list of reasoning functions used in our RAVU framework. These functions enable compositional reasoning over spatio-temporal graphs by providing modular components for entity localization, temporal reasoning, and context extraction. Each function can be combined in different sequences to handle various types of video understanding queries.

C. Discussion and Analysis

In this section, we provide additional insights and discussions about our RAVU framework, addressing key aspects of the method, design choices, and comparative analysis with existing approaches.

C.1. Design Choices and Implementation Details

Choice of Large Multimodal Model (LMM): We chose Gemini-1.5-Flash for our implementation due to several practical considerations. Compared to closed models like GPT-4o, Gemini-1.5-Flash offers significantly lower cost while maintaining superior video analysis capabilities compared to open-source alternatives like QWEN2-VL-7B. Its effectiveness in graph construction through consistent entity tracking via visual prompting, combined with API access enabling faster inference given our limited local compute resources, made it the optimal choice for our framework.

Dataset Selection: We selected NExT-QA and EgoSchema due to their strong alignment with our research objectives. NExT-QA focuses specifically on temporal and causal reasoning tasks, while EgoSchema features global behavioral questions—both requiring reasoning over multiple video frames. This contrasts with other popular datasets like MSRVTT-QA or ActivityNet-QA, which primarily contain descriptive questions often answerable using only a single frame. Additionally, NExT-QA provides dense tracklet annotations, allowing us to evaluate spatio-temporal graphs for question answering without the confounding factor of tracklet errors.

C.2. RAVU's STG vs. Traditional STGs.

While we do not focus on proposing entirely new STG topology, RAVU introduces a novel pipeline that leverages LMMs to generate expressive STGs with cross-frame identity resolution addressed through visual prompting. Compared to classical STGs, our graphs feature richer entity representations with detailed appearance and behavioral attributes, open-vocabulary relations, and are enhanced with entity-wise events to support temporal reasoning.

C.3. Sensitivity to External Object Tracker

Our analysis reveals that RAVU's performance is sensitive to tracking quality, particularly for temporal question types requiring temporal linking of entities. We observe accuracy drops ($74.19\% \rightarrow 70.86\%$) when using predicted tracklets compared to ground truth annotations, indicating that tracking errors negatively affect performance in cases requiring temporal consistency. However, question categories that do not rely heavily on temporal consistency show minimal impact from tracking errors.

C.4. Extension to Other Tasks

The STGs generated by our method can be applied to various video understanding tasks beyond question answering, including video captioning, summarization, and intent classification. The rich entity representations and temporal relationships captured in our graphs effectively model entity behaviors and interactions across videos, making them suitable for diverse applications requiring comprehensive video understanding.