UpstreamQA: A Modular Framework for Explicit Reasoning on Video Question Answering Tasks

Jason Nguyen
Lincoln North Star High School
jngyn1@icloud.com

Ameet Rao
The Charter School of Wilmington
ameet9034@gmail.com

Alexander Chang
Greenwich High School
28chang11@gmail.com

Ishaan Kumar
Santa Susana High School
kumarishaan2027@gmail.com

Erin Tan
UC Berkeley
tane@berkeley.edu

TL;DR

We introduce a UpstreamQA, a modular framework that inserts explicit upstream reasoning steps before downstream video question answering (VideoQA), and systematically study how it impacts VideoQA accuracy, interpretability, and flexibility.

Abstract

Video Question Answering (VideoQA) demands models that jointly reason over spatial, temporal, and linguistic cues. However, the task's inherent complexity often requires multi-step reasoning that current large multimodal models (LMMs) perform implicitly, leaving their internal decision process opaque. In contrast, large reasoning models (LRMs) explicitly generate intermediate logical steps that enhance interpretability and can improve multi-hop reasoning accuracy. Yet, these models are not designed for native video understanding, as they typically rely on static frame sampling. We propose UpstreamQA, a modular framework that disentangles and evaluates core video reasoning components through explicit upstream reasoning modules. Specifically, we employ multimodal LRMs to perform object identification and scene context generation before passing enriched reasoning traces to downstream LMMs for VideoQA. We evaluate UpstreamQA on the OpenEQA and NExTQA datasets using two LRMs (o4-mini, Gemini 2.5 Pro) and two LMMs (GPT-4o, Gemini 2.5 Flash). Our results demonstrate that introducing explicit reasoning can significantly boost performance and interpretability of downstream VideoQA, but can also lead to performance degradation when baseline performance is sufficiently high. Overall, UpstreamQA offers a principled framework for combining explicit reasoning and multimodal understanding, advancing both performance and diagnostic transparency in VideoQA in several scenarios.

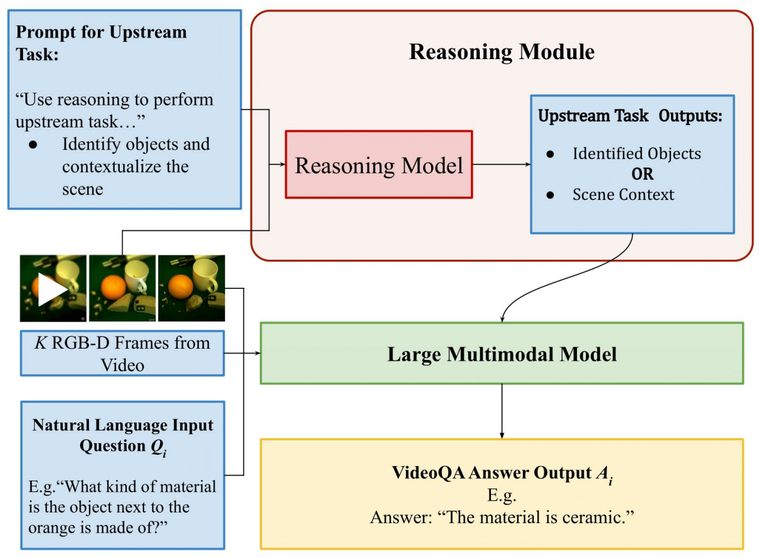
Motivation

- LMMs continue to face substantial limitations
- Traditionally, VideoQA has relied on end-to-end architectures, however their black-boxed nature hinders the transparency of their internal reasoning processes
- Large Reasoning Models (LRMs) can be leveraged in VideoQA to improve accuracy

Therefore, we introduce a framework for evaluating various upstream tasks processed by LRMs can influence the downstream VideoQA performance. Concretely, our contributions are as follows:

- We introduce UpstreamQA, a novel framework for evaluating explicit reasoning as upstream modules for VideoQA
- We perform experiments across two upstream tasks as well as two LRMs and two LMMs, reporting results of their effect on VideoQA performance
- We find that although explicit reasoning improves interpretability of logical decision making processes, performance differences are model- and dataset-dependent

Framework Overview



Overview of our UpstreamQA framework.

- First, the reasoning model is prompted to perform an upstream task based on the given RGB-D frames from the video.
- Next, the Large Multimodal Model (LMM) receives the output of the reasoning model, the same RGB-D frames, and a prompt asking it questions about the video.
- The LMM then outputs its answer and we evaluate the answer's accuracy

Methods

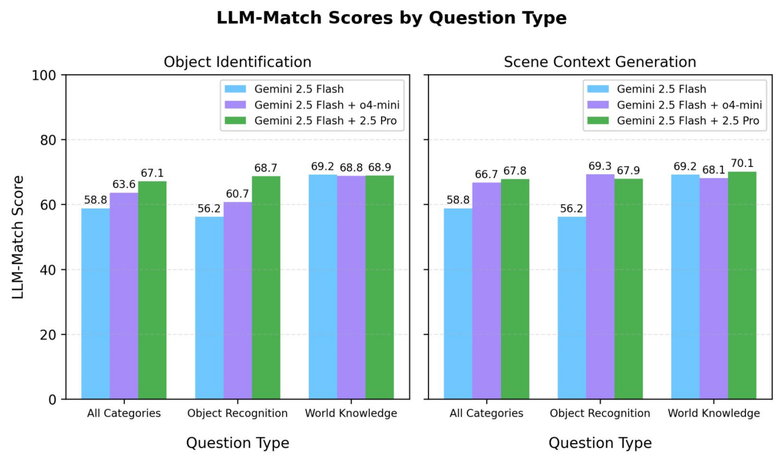
- Our method follows a two-stage pipeline.
  - First, we employ reasoning modules to perform distinct upstream video analysis tasks (object identification & scene context generation)
  - Second, the output is provided to an LMM, that performs the equivalent base VideoQA task but with additional upstream reasoning
- The object identification upstream task focuses on generating a structured inventory of the objects, their attributes (e.g., color, material, etc.), and spatial relationships with one another in a given video
- The scene context generation upstream task is aimed at recognizing the overall scene category (e.g., kitchen) and generating a comprehensive description of the environment (e.g., environmental details, ambiance, etc.)
- We test our framework on two datasets: OpenEQA and NExTQA
  - For OpenEQA, we utilize the same evaluation method and correctness metric (LLM-Match) introduced by OpenEQA. To evaluate outputs, an independent LLM (GPT-4) is used to score outputted answers compared to the ground truth
  - For NExTQA, we experiment on only the multiple-choice subset, and evaluate performance using accuracy (percentage of correct answers selected)

Experimentation Results

LMM	LRM	OpenEQA	NExTQA
GPT-4o	———	67.7	62.32%
Gemini 2.5 Flash	———	58.8	78.32%
Object Identification			
GPT-4o	o4-mini	55.7	**67.48%**
GPT-4o	Gemini 2.5 Pro	59.7	**67.08%**
Gemini 2.5 Flash	o4-mini	**63.6**	77.44%
Gemini 2.5 Flash	Gemini 2.5 Pro	**67.1**	78.00%
Scene Context			
GPT-4o	o4-mini	48.1	**67.68%**
GPT-4o	Gemini 2.5 Pro	47.8	**64.96%**
Gemini 2.5 Flash	o4-mini	**66.7**	77.20%
Gemini 2.5 Flash	Gemini 2.5 Pro	**67.8**	77.16%

Results on the OpenEQA and NExTQA datasets with distinct LMM and LRM pairs.

- Gemini 2.5 Flash with the addition of an upstream reasoning LRM generally outperforms its standalone counterpart on OpenEQA, whereas GPT-4o experiences diminished accuracy
- GPT-4o with the addition of an upstream reasoning LRM generally outperforms its standalone counterpart on OpenEQA, whereas Gemini 2.5 experiences no significant change in performance



LLM-Match Scores stratified by question category on OpenEQA using Gemini 2.5 Flash as the base model

Discussion

- The modularity of our framework allows for greater flexibility and interpretability
- The results of our experiments reveal the effect of our framework on VideoQA accuracy on certain tasks, while leading to performance degradation on others
- Our work presents promising preliminary findings for better understanding the role of explicit reasoning models in improving complex tasks like VideoQA
- Future expansions of this work may consider exploring:
  - usage of modular reasoning to encompass other core video reasoning components and their effect on VideoQA performance
  - why performance degradation occurs on certain models while significant performance improvements are observed on others