Retrieval Augmented Generation for Natural Language Query in Egocentric Video

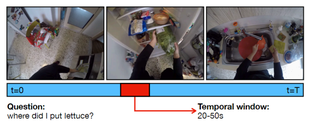
Jiahao Nick Li, Li Gu, Omid Reza Heidari

Université Concordia University

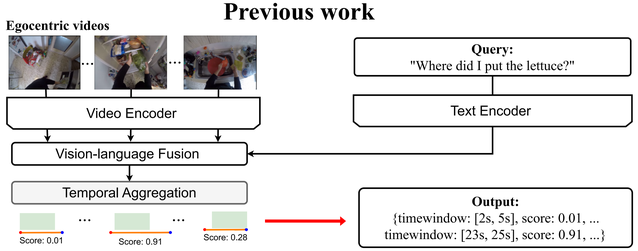
Concept

Natural language query (NLQ)

- Predict the temporal window in an egocentric video that answers the question

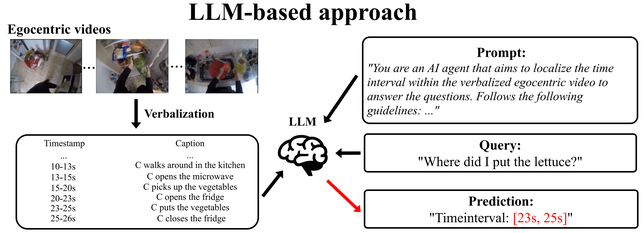


Previous work



***Issue**: Fail to handle long-form video sequence temporal reasoning due to lack of NLQ annotation*

LLM-based approach



***Motivation**: Leverage LLM to perform long-form temporal reasoning*

Contributions

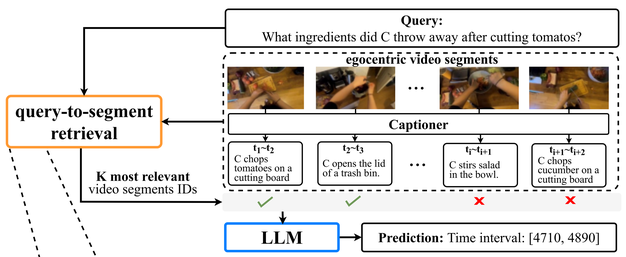
- Propose a retrieval augmented generation (RAG) framework that leverages LLM's temporal reasoning for NLQ
- Demonstrate that the effective query-to-segment retrieval can improve the LLM's prediction
- Establish a connection between RAG and Video understanding

Method

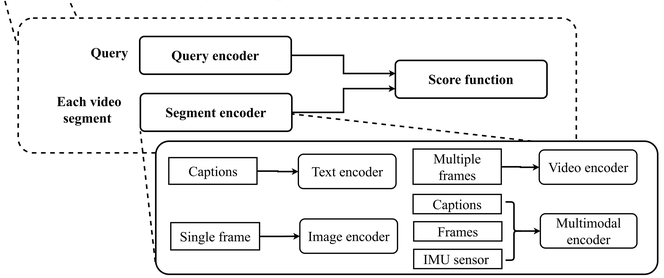
Our retrieval-augmented framework

Decompose the challenging video grounding task into two:

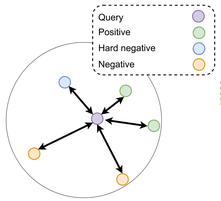
query-to-segment retrieval and temporal reasoning on selected segment captions

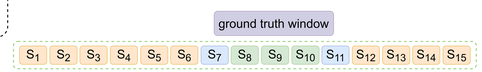


Query-to-segment Retrieval



Query-to-segment Constrastive Loss





Positive: Overlapping with Ground truth window > threshold_A
Negative: Overlapping with Ground truth window < threshold_B
Hard Negative: threshold_B < Overlapping < threshold_A

EgoNCE Loss: $\mathcal{L}_{\text{q2s}} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\sum_{j\in\mathcal{P}_i}\exp(\mathbf{q}_i^T\mathbf{s}_j/\tau)}{\sum_{j\in\mathcal{P}_i}\exp(\mathbf{q}_i^T\mathbf{s}_j/\tau)+\sum_{j\in\mathcal{N}_i}\exp(\mathbf{q}_i^T\mathbf{s}_j/\tau)}$

Qualitative Results

Prediction **w/ and w/o retrieval:**

***Query:** What colour is the jacket I examined at the clothes rack?*

Input video segments

Ground Truth

Prediction w/ retrieval (success)

Token cost per query: ~1K

Retrieved segments

Prediction w/o retrieval (failure)

Token cost per query: ~10K (10x)

Contains redundant information (noise) for LLM to reason

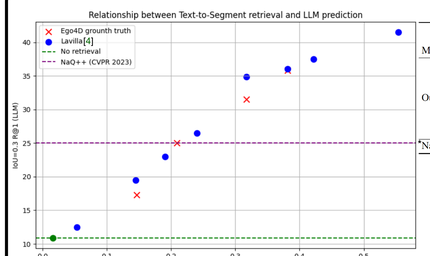
Failure case of query-to-segment retrieval:

***Query:** Who did I talk to when I was watching the television?*

- Low data annotation quality
- Need to see multiple frames to know whether the user talks with other people

✕ Even impossible for human to tell the difference between positives and negatives.

Quantitative Results



Method	Query-to-segment	Video Segment	LLM	Retrieval Recall	Retrieval Precision	NLQ IoU=0.3 R@1	NLQ IoU=0.5 R@1
Ours	No retrieval	-	GPT-4	1.0	0.008	10.88	6.84
	Zero-shot CLIP [1]	Single frame		0.335	0.051	11.38	6.36
	Zero-shot CLIP (text)	Caption		0.353	0.045	12.29	6.88
	Zero-shot CLIP (mixed)	Caption + Single frame		0.372	0.057	13.17	7.21
	Finetuned CLIP	Single frame		0.451	0.061	17.73	9.67
	Finetuned BLIP-2 [2]	Single frame		0.512	0.102	19.95	11.01
*NaQ++[3]	-	-	-	-	-	25.00	15.40

$$Recall = \frac{\sum T_{intersection}}{T_{GT}} \qquad Precision = \frac{\sum T_{intersection}}{\sum T_{segment}}$$

*: The state-of-the-art method involves training the model on both NaQ and NLQ train-set, where NaQ contains 80x more examples than the NLQ train-set.

[1]Alec Radford *et al.*, Learning Transferable Visual Models From Natural Language Supervision, in *ICML 2021*. [2] Junnan Li *et al.*, BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, in *ICML 2023*. [3] Santhosh Ramakrishnan *et al.*, NaQ: Leveraging Narrations as Queries to Supervise Episodic Memory, in *CVPR 2023*. [4] Yue Zhao *et al.*, Learning Video Representations from Large Language Models, in *CVPR 2023*.