VidCtx: Context-aware Video Question Answering with Image Models

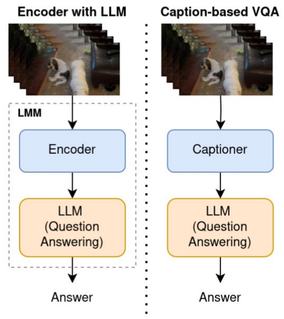
Andreas Goulas[1,2], Vasileios Mezaris[1], Ioannis Patras[2]

[1] Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece
[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK
{agoulas, bmezaris}@iti.gr, i.patras@qmul.ac.uk

Zero-shot Video Question Answering

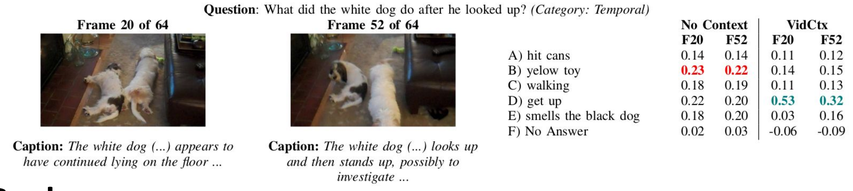
Motivation

- **Video LLMs** achieve state-of-the-art Video QA performance
- High demand for **memory** and **compute**
- **O(N^2)** space and time complexity
- Typically require **large-scale** video pre-training



Existing Video QA approaches

- **Encoder plus LLM:** reason over visual features
- **Caption-based:** reason over textual representations (e.g. captions)
- Usually rely on a single representation and ignore the other modality



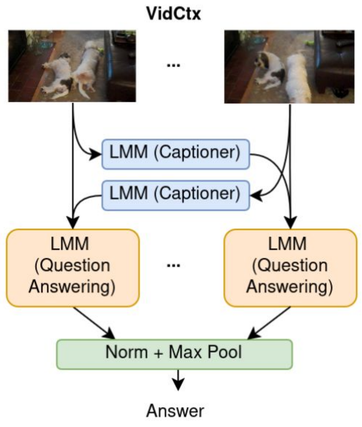
Goals

- Address **scaling** and **memory limitations** of Video LLMs
- Adopt a training-free paradigm by re-using pre-trained image LLMs and combine visual-text representations

Proposed method: VidCtx

Approach

- Integrate **both text and visual modalities**
- Process video **frame-by-frame** with image models
- Concatenate **helpful context** with prompts and visual features
- **Aggregate** decisions across frames with a pooling layer



Captioner Prompt: Please provide a short description of the image giving information related to the following question: *What did the white dog do after he looked up?*

QA Prompt: Here is what happens *earlier* (or *later*) in the video: *<Caption>*. Question: *What did the white dog do after he looked up?* Option A: *hit cans.* Option B: *yellow toy.* Option C: *walking.* Option D: *get up.* Option E: *smells the black dog.* Option F: No Answer. Considering the information presented in the caption and the video frame, select the correct answer in one letter from the options (A,B,C,D,E,F).

Architecture

- **Question-aware Captions:** pre-trained image LLM generates captions related to the given question
- **Context-aware QA**: image LLM answers question based on a) the current frame and b) the caption of a distant frame (N/2 frames apart)
- **Aggregation:** frame-level decision scores are aggregated with max pooling and L1 normalization

$$y = \underset{t \in T-\{F\}}{\arg\max} \left[\max_{0 \le i < N} \frac{p(d_i = t)}{\sum_{k \in T} |p(d_i = k)|} \right]$$

Frame-level scores

- Extract **scores** from **logits** of **first token**
- Ignore **special** No Answer token

Experimental results

Experimental setup

- **Datasets:** NExT-QA (5K), IntentQA (2K), STAR (7K)
- **Base Model:** LLaVa-1.6-Mistral-7B

Comparison of VidCtx with zero-shot open-model approaches

- **Competitive performance** on all datasets
- **Scales linearly** w.r.t. the number of frames
- **7B model even outperforms** some models that rely on proprietary LLMs (e.g. LLoVi with GPT-3.5)

Model	Params	NExT-QA Cau.	NExT-QA Tem.	NExT-QA Des.	NExT-QA All	IntentQA	STAR Int.	STAR Seq.	STAR Pre.	STAR Fea.	STAR Avg
GPT-based Models											
LLoVi (GPT-3.5) [9]	N/A	67.1	60.1	76.5	66.3	-	-	-	-	-	-
LLoVi (GPT-4) [9]	N/A	73.7	70.2	81.9	73.8	67.1	-	-	-	-	-
VideoTree (GPT-4) [7]	N/A	75.2	67.0	81.3	73.5	66.9	-	-	-	-	-
VideoAgent (GPT-4) [8]	N/A	72.7	64.5	81.1	71.3	-	-	-	-	-	-
VFC [27] *	<1B	51.6	45.4	64.1	51.5	-	-	-	-	-	-
InternVideo [28] *	<1B	48.0	43.4	65.1	49.1	-	43.8	43.2	42.3	37.4	41.6
SeViLA [6] *	4B	61.5	61.3	75.6	63.6	60.9	48.3	45.0	44.4	40.8	44.6
LangRepo [10]	12B	64.4	51.4	69.1	60.9	59.1	-	-	-	-	-
Q-ViD [11]	12B	67.6	61.6	72.2	66.3	63.6	48.2	47.2	43.9	43.4	45.7
VideoChat2 [1] *	7B	61.9	57.4	69.9	61.7	-	**62.4**	**67.2**	**57.5**	**53.9**	**63.8**
VidCtx (Ours)	7B	**71.7**	**65.1**	**79.2**	**70.7**	**67.1**	53.9	54.3	51.4	44.7	51.1

* VFC [27] pre-train their model on 0.5M video-caption pairs. SeViLA [6] trains a localizer component on 10K videos. InternVideo [28] and VideoChat2 [1] utilize large-scale pre-training on video datasets with over 10M video-caption pairs.

Effect of number of frames (NExT-QA)

# Frames	1	2	4	8	16	32	64
Top-1 (%)	63.5	66.8	68.8	69.2	70.1	70.3	**70.7**

Comparison with captions-only baseline (NExT-QA)

Method	Top-1 (%)
Captions Only (32 frames)	67.3
VidCtx (32 frames)	**70.3**

Ablation study

- **Distant captions** provide the most helpful context
- +3% improvement over **captions-only baseline**
- **Normalization** prior to pooling is important
- Performance **improves** with higher **number of frames**

Choice of context (NExT-QA)

Method (use of context)	Top-1 (%)
No Context	67.9
Concat 16 Captions (Q-Aware)	68.3
Current Caption (Q-Aware)	69.9
Distant Caption (Static)	69.5
Distant Caption (Q-Aware)	**70.7**

Choice of aggregation method (NExT-QA)

Method (aggregation)	Top-1 (%)
Voting	69.7
Mean Pooling	69.3
Max Pooling	69.2
Softmax + Mean Pooling	70.2
Softmax + Max Pooling	70.6
L1 Norm + Max Pooling	**70.7**

Software available at: https://github.com/IDT-ITI/VidCtx

Supported by: TRANSMIXR