VidCtx: Context-aware Video Question Answering with Image Models

Andreas Goulas
CERTH-ITI &
Queen Mary University of London
Thessaloniki, Greece
agoulas@iti.gr

Vasileios Mezaris
CERTH-ITI
Thessaloniki, Greece
bmezaris@iti.gr

Ioannis Patras
Queen Mary University of London
London, UK
i.patras@qmul.ac.uk

***Abstract*—To address computational and memory limitations of Large Multimodal Models in the Video Question-Answering task, several recent methods extract textual representations per frame (e.g., by captioning) and feed them to a Large Language Model (LLM) that processes them to produce the final response. However, in this way, the LLM does not have access to visual information and often has to process repetitive textual descriptions of nearby frames. To address those shortcomings, in this paper, we introduce VidCtx, a novel training-free VideoQA framework which integrates both modalities, i.e. both visual information from input frames and textual descriptions of others frames that give the appropriate context. More specifically, in the proposed framework a pre-trained Large Multimodal Model (LMM) is prompted to extract at regular intervals, question-aware textual descriptions (captions) of video frames. Those will be used as context when the same LMM will be prompted to answer the question at hand given as input a) a certain frame, b) the question and c) the context/caption of an appropriate frame. To avoid redundant information, we chose as context the descriptions of distant frames. Finally, a simple yet effective max pooling mechanism is used to aggregate the frame-level decisions. This methodology enables the model to focus on the relevant segments of the video and scale to a high number of frames. Experiments show that VidCtx achieves competitive performance among approaches that rely on open models on three public Video QA benchmarks, NExT-QA, IntentQA and STAR. Our code is available at https://github.com/IDT-ITI/VidCtx.**

***Index Terms*—large language model, video question answering, large multimodal model, question-aware caption, context**

I. Introduction

Video-Language modeling is an inherently resource-intensive task due to the necessity of processing a large number of frames, in order to generate accurate representations of the contents of a video. Recently, a considerable number of Video Large Language Models have emerged [1] [2] [3] [4] [5] [6] [7], which combine a visual encoder with a Large Language Model (LLM). These models have achieved state-of-the-art performance on a wide variety of video understanding tasks. They are pre-trained on large video corpora, thus they are typically not considered zero-shot approaches. Furthermore, they are generally limited by the maximum number of input tokens, and therefore frames, that they can process in one pass.

This work was supported by the EU Horizon Europe programme under grant agreement 101070109 TransMIXR.

In order to tackle this problem, prior works have proposed sophisticated frame selection policies [8] [9] [10], with the goal of selecting the most salient frames of the input video. In a different direction, several works [11] [12] [13] have proposed extracting textual information in the form of captions from the input video and using an LLM to process the extracted captions. These approaches rely on the ability of LLMs to reason over long input prompts and produce accurate results. However, this has been proven to be a challenging task [14] [15]. Furthermore, by converting the input video to a textual representation the model no longer has access to the rich visual information contained in the video; this reduces the final performance of question answering.

In an orthogonal direction, a promising technique for improving the reasoning capabilities of Large Language Models is multi-step reasoning. Early studies in multimodal Chain-of-Thought (CoT), such as [16] [17] [18] [19], have shown that an intermediate rationale generation step can aid the multimodal understanding capabilities of language models. In the video domain, Video-of-Thought [20] achieved similar results with a multi-step reasoning framework for Video QA.

In this paper, we focus on the multiple-choice Video Question Answering task, which requires strong temporal reasoning and long-term understanding. To this end, motivated by the success of multi-step reasoning frameworks, we propose a novel video QA framework, VidCtx, which is based on the consensus of multiple Large Multimodal Model (LMM) decisions across different frames. The proposed VidCtx architecture is shown in Fig. 1, where we also illustrate the fundamental differences between our approach and the typical approaches of the literature. Prior works typically rely either on visual features aligned with a trainable layer ("Encoder with LLM") or on generated captions ("Caption-based VQA"), that they process with an LLM. In contrast, VidCtx utilizes both the extracted captions and the visual information for question answering. The temporal evolution of the video signal is handled by appending the description of distant frames to the LMM prompt; and, by combining the knowledge contained in distant parts of the video, the model can also give better answers to descriptive questions.

VidCtx is not tailored to a specific pre-trained LMM, and requires no additional training or fine-tuning. Furthermore, it

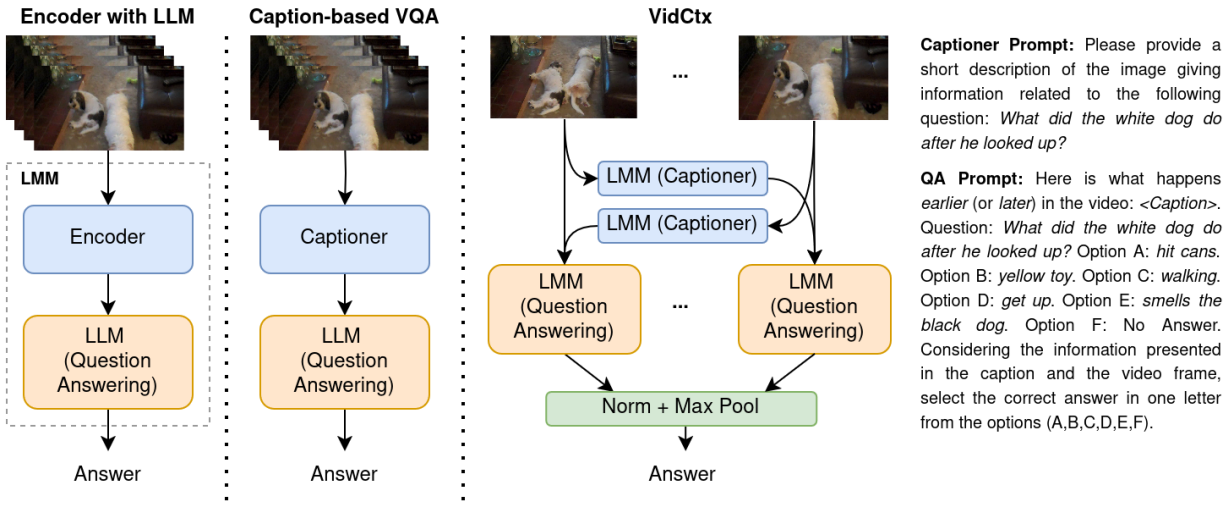


Fig. 1. **Comparison of video understanding frameworks.** Related works typically either project the input frames to the LLM input space or rely on extracted captions to answer questions about videos. We propose combining both modalities, processing the video frame-by-frame and inserting the relevant context (i.e., extracted question-aware captions) as part of the LLM prompt. In our architecture, we use the same LMM for both captioning and question answering.

can process any number of frames since it is not constrained by the maximum context length of the language model. Our experimental results on three publicly available Video Question Answering benchmarks show that our VidCtx framework achieves competitive performance among approaches that rely on comparable open models.

Our contributions can be summarized as follows:

1) We present a Video QA framework, VidCtx, that integrates both visual information and textual information in the form of captions to answer multiple choice questions.
2) We propose a frame-based architecture that can focus on the salient frames of the video and scale to a high number of frames.
3) We evaluate VidCtx on three public video QA benchmarks, NExT-QA, IntentQA and STAR, showing competitive performance among open models.

II. Related Works

A. Large Language Models for Video

The combination of a Large Language Model with a visual encoder or adapter module has resulted in a wide variety of models [1] [2] [3] [4] [5] [6] [7] that have achieved state-of-the-art performance on several video understanding tasks. These models primarily benefit from large-scale pre-training and instruction tuning, including training on video corpora, therefore their performance is strongly dependent upon the quantity and quality of the available training data.

To relax the requirement for large training corpora, another line of work proposes using pre-trained models as part of a video processing pipeline. For example, LLoVi [11] proposes a simple pipeline where captions are extracted from videos and an LLM is used to process the captions. Q-ViD [13] improves on this idea by extracting question-aware captions for further processing. LangRepo [12] introduces a textual repository that contains information and metadata about the contents of the video. In contrast to these works, we propose looking at both the visual information and extracted captions simultaneously, in order to capture a more complete representation of the input video.

Vamos [21] adopts a similar approach to ours, by concatenating the extracted video captions with visual features, which are aligned with a learnable projection layer. However, Vamos process the entire video in a single pass, relying on the long-context reasoning capabilities of LLMs, and, it contains trainable components. Remarkably, Vamos showed that the addition of visual features in combination with task-agnostic video captions had a minimal impact on its performance in the video QA task. In contrast, our method follows a training-free paradigm and we show that the two modalities are indeed complementary.

Finally, some works explore how to improve the frame selection mechanism, with the aim of selecting more relevant frames [8] [9] [10]. SeViLA [8] leverages a single Image-Language Model to perform both temporal localization and video question answering. Keyframe localization is achieved by asking the model whether a given frame contains the necessary information to answer the given query. Even though our proposed framework does not contain a dedicated keyframe selection component, inspired by the above work, we enable our model to ignore parts of the input video and focus only on the relevant segments.

B. Multi-step Reasoning in LLMs

Recent efforts in multi-step reasoning have shown promising results in advancing the reasoning capabilities of Large Language Models. Chain-of-Thought (CoT) reasoning refers to a group of techniques where LLMs are instructed to generate reasoning chains before answering a given query. Zero-shot CoT [22] instructs the language model to "think step-by-step",

thereby encouraging the model to utilize multi-step reasoning while generating the response. Few-shot CoT [23] provides reasoning chain demonstrations in-context, which can be either hand-crafted or automatically generated.

In multimodal understanding tasks, early studies of Chain-of-Thought reasoning [16] [17] [18] [19] have shown that the visual understanding capabilities of Large Multimodal Models can be improved by fine-tuning the LMMs on hand-crafted or automatically generated rationales and explanations. CCoT [24] proposes leveraging scene graph representations as an intermediate step, focusing on compositionality. In the video domain, Video-of-Thought [20] designs an elaborate multi-step reasoning framework that leverages spatio-temporal scene graphs for VQA. In contrast, we opt for a simpler intermediate step based on question-aware captions, which works well in practice to capture the temporal evolution of the video signal.

III. Method

A. Overall Architecture

The proposed VidCtx architecture is shown in Fig. 1. VidCtx utilizes both the extracted captions and the visual information contained in the video for question answering. Furthermore, VidCtx adopts a frame-based architecture that allows the model to focus on the relevant sections of the video and process a large number of frames. Specifically, we first perform frame sampling and we extract question-aware captions. We prompt the LMM with a standard VideoQA prompt and we insert the captions of distant frames in the prompt as additional context. This approach proves effective for modeling the temporal relations that are contained in the video and enables the model to combine information from distant parts of the video. Finally, the frame-level decisions are aggregated with a simple max pooling layer.

B. Question-aware Caption Extraction

The first step of our method is to extract question-aware captions from the input video. Given a long video V, we split V into N non-overlapping segments and we sample the central frame $V = \{v_i, i = 1, ..., N\}$ from each segment.

Following Q-ViD [13], we utilize a pre-trained LMM, denoted as ϕ, to extract question-aware captions $C = \{c_i, i = 1, ..., N\}$ from each frame. The captions are generated as follows:

$$c_i = \phi(v_i, I_{cap}) \tag{1}$$

where I_{cap} is the concatenation of the instruction prompt and the given question Q, as follows:

> "Please provide a short description of the image, giving information related to the following question: <Q>"

C. Context-aware Video Question Answering

The next step is to prompt the LMM with the contents of each video frame separately, thus generating N separate answers for the given question Q and video V pair. We instruct the LMM to output a special character if there is not enough information to generate an answer. This mechanism is crucial because it allows the LMM to ignore parts of the video that are not relevant to the given question.

The language model is able to reason about static events that are depicted in each separate frame. However, in order to be able to reason about long-term actions that span multiple frames, we insert the relevant context into its prompt. Specifically, we append the question-aware description of a distant frame to the model prompt as follows:

$$d_i = \phi(v_i, c_{r(i)}, I_{vqa}) \tag{2}$$

$$r(i) \equiv i + \frac{N}{2} \pmod N \tag{3}$$

where d_i is the i-th decision, $r(i)$ is the index of the distant frame and I_{vqa} is the concatenation of the instruction prompt and the given question Q. We choose to share information between frames that are exactly half the duration of the video apart, in order to ensure the greatest diversity between the contextual captions and the considered frame. We also insert the relevant temporal specifier ("earlier" or "later"), in order to inform the language model about the sequence of events in the video.

Following Q-ViD [13], we use the following instruction prompt I_{vqa} to generate the frame-level decisions:

> "Here is what happens earlier/later in the video: <Caption>
> Question: <Q> Option A: <A> Option B: <B> Option C: <C> Option D: <D> Option E: <E> 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)."

D. Video-level Decision

Finally, we aggregate the N separate frame-level decisions to generate a video-level decision. In order to speed up the answer generation process, we consider the probability of just the first token of each frame-level answer among the set of possible options, denoted as T. We opt for a simple max pooling mechanism to combine these probabilities across frames, which works well in practice and does not require any training:

$$y = \arg\max_{t \in T - \{F\}} \left[\max_{0 \le i < N} \frac{p(d_i = t)}{\sum_{k \in T} |p(d_i = k)|} \right] \tag{4}$$

where y is the final decision of the framework and $p(\cdot)$ is the log probability score of the given token. It is important that the per-frame scores are normalized across the target options before applying the max pooling operator. Empirically, we find that L1 normalization slightly outperforms the alternative softmax. For the purposes of generating the final decision, we ignore the special token *F*, which denotes that the LMM could not answer the question.

TABLE I

ZERO-SHOT RESULTS (TOP-1 ACCURACY %) ON NEXT-QA [25], INTENTQA [26] AND STAR [27]. WE DE-EMPHASIZE METHODS BASED ON PROPRIETARY MODELS WITH A SIGNIFICANTLY HIGHER NUMBER OF PARAMETERS, COMPARED TO OURS. AMONG COMPARABLE OPEN MODEL METHODS, WE **BOLD** THE BEST RESULTS AND UNDERLINE THE SECOND BEST. WE INDICATE WITH A STAR (*) METHODS THAT ARE (PRE-)TRAINED ON VIDEO DATASETS.

		NExT-QA					STAR				
Model	Params	Cau.	Tem.	Des.	All	IntentQA	Int.	Seq.	Pre.	Fea.	Avg
GPT-based Models											
LLoVi (GPT-3.5) [11]	N/A	67.1	60.1	76.5	66.3	-	-	-	-	-	-
LLoVi (GPT-4) [11]	N/A	73.7	70.2	81.9	73.8	67.1	-	-	-	-	-
VideoTree (GPT-4) [9]	N/A	75.2	67.0	81.3	73.5	66.9	-	-	-	-	-
VideoAgent (GPT-4) [10]	N/A	72.7	64.5	81.1	71.3	-	-	-	-	-	-
VFC [28] *	<1B	51.6	45.4	64.1	51.5	-	-	-	-	-	-
InternVideo [29] *	<1B	48.0	43.4	65.1	49.1	-	43.8	43.2	42.3	37.4	41.6
SeViLA [8] *	4B	61.5	61.3	75.6	63.6	60.9	48.3	45.0	44.4	40.8	44.6
LangRepo [12]	12B	64.4	51.4	69.1	60.9	59.1	-	-	-	-	-
Q-ViD [13]	12B	67.6	61.6	72.2	66.3	63.6	48.2	47.2	43.9	43.4	45.7
VideoChat2 [5] *	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 [28] is pre-trained on 0.5M video-caption pairs. SeViLA [8] contains a localizer component trained on 10K videos. InternVideo [29] and VideoChat2 [5] utilize large-scale pre-training on video datasets with over 10M video-caption pairs.

IV. EXPERIMENTS

A. Implementation Details

In our experiments, we use the 4-bit quantized version of LLaVa-1.6-Mistral-7B [30], which is an instruction-tuned LMM that can process a single image per forward pass. For NExT-QA [25] and IntentQA [26] we sample 64 frames per video. For STAR [27], we sample 32 frames per video, due to the smaller average video duration. We generate 200 tokens per caption. We ran all of our experiments on a single RTX4090 GPU. We report the zero-shot performance of our architecture, without performing any fine-tuning.

B. Datasets and Metrics

We conduct our experiments on the following publicly available video QA datasets:

1) NExT-QA [25] is a video QA dataset that contains 5,440 videos and 48K multiple-choice question-answer pairs. The average video length is 44 seconds. Following the relevant literature, we report the accuracy on the validation set that contains 570 videos and 5K questions. The questions are divided into Temporal, Causal and Descriptive questions.
2) IntentQA [26] is a video QA dataset, derived from NExT-QA, that focuses on intent reasoning, with an average video duration of 44 seconds. We perform our experiments on the test set that contains 2.1K question-answer pairs.
3) STAR [27] is a video QA dataset that contains 60K situated questions, divided into 4 categories (Interaction, Prediction, Sequence, Feasibility), with an average video duration of 12 seconds. Similarly to the relevant literature, we perform our experiments on the validation set, which contains 7K questions.

C. Results and Comparisons

Main Results. In Tab. I, we report the zero-shot top-1 accuracy of our architecture on NExT-QA, IntentQA and STAR. We compare VidCtx with the state-of-the-art approaches including both open models (VFC [28], InternVideo [29], SeViLA [8], LangRepo [12], Q-VID [13], VideoChat2 [5]) and GPT-based frameworks (LLoVi [11], VideoTree [9], VideoAgent [10]). We observe that our model achieves state-of-the-art performance among the open approaches on NExT-QA and IntentQA, outperforming Q-ViD by +4.4% on NExT-QA and by +3.5% on IntentQA. Our method even outperforms models that rely on closed proprietary LLMs, such as LLoVi with GPT-3.5 (+4.4% on NExT-QA). Furthermore, we observe that VidCtx outperforms the competition on all individual splits of NExT-QA. On the STAR dataset, our VidCtx method outperforms Q-ViD by +5.4% and is second only to VideoChat2. However, the latter on the one hand is pre-trained on large-scale video datasets, on the other hand performs poorly on NExT-QA (being outperformed not only by VidCtx but also by Q-ViD and SeViLA).

Computational Complexity. VidCtx has to process each frame of the input video twice: once for captioning and once for question answering. However, the second pass is significantly faster, since we only generate the first token of the answer (i.e., just one letter: A to F). The overall inference complexity of VidCtx is, therefore, linear with respect to the number of processed frames per video. In contrast, related works that concatenate the captions, such as LLoVi and Q-ViD, scale quadratically and have significantly higher GPU memory requirements, due to the $O(n^2)$ time and memory complexity of the attention mechanism used in LLMs.

D. Ablations

Choice of Context. In Tab. II, we compare different methods of inserting additional context in the form of text captions to our LMM prompts. We observe that distant captions (Equation (3)) provide the most helpful context, followed by "current captions", i.e. inserting the caption of the same frame that is provided to the LMM as input. In both cases, this indicates that the additional context can help the model make better

TABLE II
COMPARISON OF DIFFERENT APPROACHES OF INSERTING ADDITIONAL CONTEXT IN VIDCTX ON NEXT-QA [25].

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**

TABLE III
COMPARISON BETWEEN VIDCTX AND CAPTIONS-ONLY BASELINE ON NEXT-QA [25].

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

TABLE IV
EFFECT OF NUMBER OF FRAMES CONSIDERED BY VIDCTX ON NEXT-QA [25].

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

TABLE V
COMPARISON OF DIFFERENT AGGREGATION METHODS IN VIDCTX ON NEXT-QA [25].

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**

decisions. Furthermore, we observe that the question-aware captions outperform the static captions. This is consistent with the findings of Q-ViD [13].

Comparison with Captions-only Baseline. In Tab. III, we compare VidCtx with the baseline approach of concatenating the question-aware captions and utilizing a text-only LLM to process the captions; similarly to Q-ViD [13]. We run our experiments with the same set of captions and same model. We observe that our proposed framework outperforms the baseline approach by +3%.

Effect of Number of Frames. In Tab. IV, we observe a clear correlation between the number of frames and the performance of VidCtx. We note that our model can scale to an arbitrarily high number of frames, since each frame is processed separately, in contrast with most of the state-of-the-art approaches that are limited by the long-context capabilities of language models.

Choice of Aggregation Method. In Tab. V, we compare different methods of aggregating the frame-level decisions in order to generate a decision for the entire video. We observe that max pooling slightly outperforms mean pooling in our experiments. We find that normalizing the probability scores before applying the pooling operator provides a significant increase in accuracy. Finally, we find that L1 normalization is more suitable (by a small margin) to our task compared to the alternative of applying a softmax operator.

E. Qualitative Study

In Fig. 2 we present qualitative examples of VidCtx in action. We select one question from each question category of NExT-QA [25] and we present pairs of distant frames, along with their question-aware captions and normalized answer scores for each candidate answer. We observe that the utilization of captions of distant frames as additional context can help the model make better decisions. Even in cases where the context is misleading, such as the last example (i.e., the caption of Frame 25, which contradicts what can be seen in Frame 57), VidCtx can recover by generating higher scores for caption-frame pairs for which a more confident decision can be made (i.e., pairs presenting no such contradictions), thus leading to correct video-level answers.

V. CONCLUSION

We presented VidCtx, a training-free video question answering framework that integrates both visual features and text captions, in order to answer multiple-choice questions about videos. In VidCtx, the video is processed frame-by-frame by an LMM, and additional context is inserted in the form of captions. Our framework achieves competitive performance among open models on three public video QA datasets and even performs competitively to frameworks that rely on proprietary Large Language Models.

REFERENCES

[1] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv preprint arXiv:2306.05424*, 2023.

[2] Hang Zhang, Xin Li, and Lidong Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[3] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023.

[4] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li, "Video instruction tuning with synthetic data," *arXiv preprint arXiv:2410.02713*, 2024.

[5] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, et al., "Mvbench: A comprehensive multi-modal video understanding benchmark," in *Proc. IEEE/CVF CVPR*, 2024, pp. 22195–22206.

[6] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, et al., "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[7] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, et al., "Internvideo2: Scaling video foundation models for multimodal video understanding," *arXiv preprint arXiv:2403.15377*, 2024.

[8] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal, "Self-chained image-language model for video localization and question answering," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[9] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, et al., "Videotree: Adaptive tree-based video representation for llm reasoning on long videos," *arXiv preprint arXiv:2405.19209*, 2024.

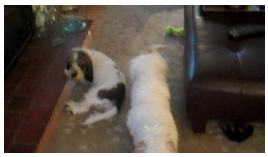
[10] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy, "Videoagent: Long-form video understanding with large language model as agent," in *Proc. ECCV*. Springer, 2025, pp. 58–76.

Question: What did the white dog do after he looked up? *(Category: Temporal)*

Frame 20 of 64

Caption: *The white dog (...) appears to have continued lying on the floor ...*

Frame 52 of 64

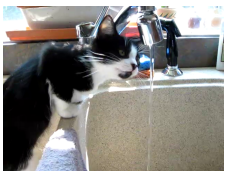


Caption: *The white dog (...) looks up and then stands up, possibly to investigate ...*

	No Context		VidCtx	
	F20	**F52**	**F20**	**F52**
A) hit cans	0.14	0.14	0.11	0.12
B) yelow toy	**0.23**	**0.22**	0.14	0.15
C) walking	0.18	0.19	0.11	0.13
D) get up	0.22	0.20	**0.53**	**0.32**
E) smells the black dog	0.18	0.20	0.03	0.16
F) No Answer	0.02	0.03	-0.06	-0.09

Question: Why does the cat suddenly move back at the start of the video? *(Category: Causal)*

Frame 7 of 64



Caption: *The cat in the image appears to be startled or surprised ...*

Frame 39 of 64

Caption: *The cat in the image is standing on the edge of a sink ...*

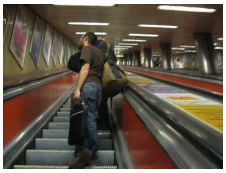
	No Context		VidCtx	
	F7	**F39**	**F7**	**F39**
A) waiting for owner	0.20	0.20	0.17	0.18
B) saw the person	0.20	0.19	0.17	0.17
C) attracted towards bubble	**0.30**	0.25	0.19	0.13
D) face got wet	0.24	**0.29**	**0.20**	**0.44**
E) clean itself	0.03	0.04	0.09	0.04
F) No Answer	0	0	0.15	-0.01

Question: How many bags is the man in black carrying? *(Category: Descriptive)*

Frame 25 of 64

Caption: *The man is carrying one bag.*

Frame 57 of 64



Caption: *The man is carrying two bags.*

	No Context		VidCtx	
	F25	**F57**	**F25**	**F57**
A) five	0.19	0.19	0.15	0.17
B) two	0.24	**0.26**	**0.45**	0.21
C) three	0.15	0.22	0.18	0.15
D) one	**0.34**	0.24	0.12	**0.39**
E) four	0.01	0.06	0.03	0.03
F) No Answer	0.01	0	-0.03	-0.03

Fig. 2. **Qualitative study of VidCtx on NExT-QA.** We select one question from each question category and we present the normalized answer scores for pairs of distant frames, along with their question aware captions.

[11] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, et al., "A simple llm framework for long-range video question-answering," *arXiv preprint arXiv:2312.17235*, 2023.

[12] Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo, "Language repository for long video understanding," *arXiv preprint arXiv:2403.14622*, 2024.

[13] David Mogrovejo and Thamar Solorio, "Question-instructed visual descriptions for zero-shot video answering," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 9329–9339.

[14] Mosh Levy, Alon Jacoby, and Yoav Goldberg, "Same task, more tokens: the impact of input length on the reasoning performance of large language models," *arXiv preprint arXiv:2402.14848*, 2024.

[15] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, et al., "Large language models can be easily distracted by irrelevant context," in *Int. Conf. on Machine Learning*. PMLR, 2023, pp. 31210–31227.

[16] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, et al., "Learn to explain: Multimodal reasoning via thought chains for science question answering," *Advances in Neural Information Processing Systems*, vol. 35, pp. 2507–2521, 2022.

[17] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola, "Multimodal chain-of-thought reasoning in language models," *arXiv preprint arXiv:2302.00923*, 2023.

[18] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang, "Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 5168–5191, 2023.

[19] Ning Pang, Wansen Wu, Yue Hu, Kai Xu, Quanjun Yin, and Long Qin, "Enhancing multimodal sentiment analysis via learning from large language model," in *Proc. IEEE ICME*, 2024, pp. 1–6.

[20] 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.

[21] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun, "Vamos: Versatile action models for video understanding," *arXiv preprint arXiv:2311.13627*, 2023.

[22] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22199–22213, 2022.

[23] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[24] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig, "Compositional chain-of-thought prompting for large multimodal models," in *Proc. IEEE/CVF CVPR*, 2024, pp. 14420–14431.

[25] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua, "Next-qa: Next phase of question-answering to explaining temporal actions," in *Proc. IEEE/CVF CVPR*, 2021, pp. 9777–9786.

[26] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan, "Intentqa: Context-aware video intent reasoning," in *Proc. IEEE/CVF ICCV*, 2023, pp. 11963–11974.

[27] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan, "Star: A benchmark for situated reasoning in real-world videos," *arXiv preprint arXiv:2405.09711*, 2024.

[28] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid, "Verbs in action: Improving verb understanding in video-language models," in *Proc. IEEE/CVF ICCV*, 2023, pp. 15579–15591.

[29] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, et al., "Internvideo: General video foundation models via generative and discriminative learning," *arXiv preprint arXiv:2212.03191*, 2022.

[30] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024, Accessed: Dec. 12, 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/.