Do Video Language Models really understand the video contexts?

Jeongwan Shin[*]
Kyungpook National University
jwshin@knu.ac.kr

Jinhyeong Lim[*]
AI Center,
HD Korea Shipbuilding
Offshore Engineering
jinhyeong.lim@hd.com

Hyeyoung Park[†]
Kyungpook National University
hypark@knu.ac.kr

Abstract

This paper examines how well visual language models (VLMs) understand video question answering (VideoQA) tasks and generate responses accordingly. Recently, VLMs based on Large Language Models (LLMs) have shown remarkable performance, but the processes of understanding and reasoning in VLMs remain under-explored. To tackle this challenge, we propose **V**ideo **U**nderstanding and **R**esponse **C**onsistency **A**ssessment, **VURCA**, a framework that incorporates a fine-grained question generation and answering process to measure how well the responses generated by VLMs align with what the model understands. In addition, we introduce an extended benchmark dataset, FgNExT-QA, which builds upon NExT-QA by incorporating more fine-grained VideoQA tasks. FgNExT-QA is designed to evaluate fine-grained understanding in video question answering. Through experiments, we found that despite the strong overall QA performance of VLMs, their understanding of both the video content and the question remains limited. In particular, they exhibit poor video comprehension in fine-grained VideoQA tasks.

1 Introduction

Video Question Answering (VideoQA) (Fei et al., 2024; Min et al., 2024) serves as a critical benchmark for evaluating the capabilities of foundational Visual Language Models (VLMs) (Zhang et al., 2023; Liu et al., 2024), particularly those trained on large-scale multi-modal datasets (Ye et al., 2023). Despite recent advancements in VideoQA performance, several fundamental concerns remain under-explored. A key question is whether these models accurately comprehend video and question to enable robust multi-modal reasoning, or if they merely mimic learned patterns from the training dataset (Xiao et al., 2024). Responses based on

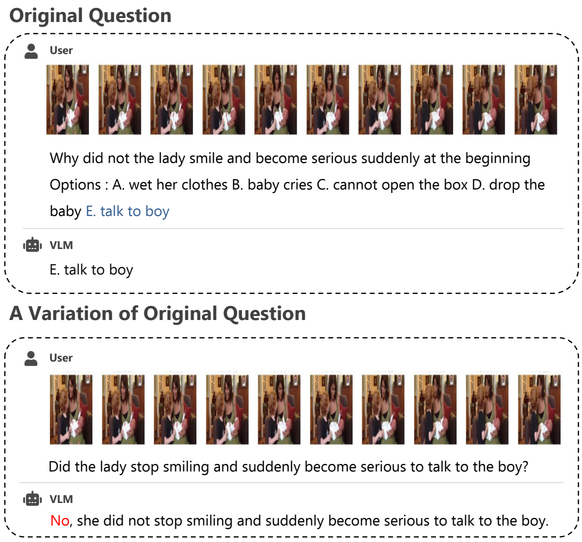


Figure 1: Example responses generated by VLMs(LLaVA-OneVision) on the NExT-QA dataset.

incomplete understanding can lead to significant issues in real-world applications, emphasizing the need for efforts to evaluate and address these limitations.

Figure 1 illustrates that VLMs often struggle to answer a variation of the original question, which are derived from the original question and its corresponding ground truth answer, even though VLMs generate correct answer. This observation demonstrates that VLMs can choose correct answer even without a precise understanding. If the answer is chosen based on accurate understanding, it should generate a consistent response to the variation. From the observation, VLMs for VideoQA still fall short in accurately understanding video contents and remain under-resourced in terms of the evaluation metrics and datasets required to assess trained models effectively. Existing research has primarily explored the estimation of consistency between generated textual outputs and image inputs in VLMs (Khan and Fu, 2024; Geng et al., 2024). However, we aim to evaluate the under-

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 408–417

standing of video content by VLMs. This marks a novel attempt to measure the consistency between responses and understanding in the domain of VideoQA.

As a novel approach, we propose the **V**ideo **U**nderstanding and **R**esponse **C**onsistency **A**ssessment, **VURCA**, a framework designed to investigate the understanding of VLMs through the process of generating fine-grained verification questions, integrating answer of the VideoQA, evaluating the consistency between fine-grained answers and initial response. First, VLMs generate an initial response by taking a video and original question as input. Based on the initial response and the original question, fine-grained verification questions are generated using an LLM. If VLM's answer is generated under through understanding on Video context, it should consistently generate responses to the variation of original questions that are semantically equivalent to the initial response. To investigate this, we input the fine-grained verification questions along with the video into the VLMs again to derive verification responses. Then, the verification responses are aggregated to quantitatively evaluate the VLM's understanding.

Moreover, our approach also enables the automatic expansion of VideoQA datasets, which are costly and time-intensive to construct. By extending the NExT-QA dataset, we construct FgNExT-QA, a fine-grained question-answering dataset with binary gold answer labels. FgNExT-QA allows us to verify that VLMs specifically understand the questions and can determine the correct answers. It can also be used as an independent benchmark for VideoQA performance evaluation.

In the experiments, we conduct a comprehensive analysis of how well state-of-the-art VLMs understand and response correct answers in VideoQA. Despite achieving high accuracy on VideoQA, VLMs exhibit inconsistencies when responding to semantically identical but rephrased questions. This observation highlights the challenges VLMs still face in aligning visual evidence with linguistic semantics, revealing areas that require further improvement. To encapsulate our contributions: 1) Introducing VURCA framework: We propose a novel framework for evaluating the alignment between video understanding and responses generated by VLMs; 2) Fine-Grained VideoQA dataset generated automatically: We present a fine-grained VideoQA dataset generated automatically, containing binary gold answer labels to systematically assess VLM understanding and response consistency; 3) Comprehensive Analysis of VLM Performance: Through experiments on various VLMs, we analyze their current challenges and interpret these issues in terms of understanding and response alignment.

2 Related Work

In videoQA tasks, a primary objective is to ensure that the model accurately comprehends video data and generates appropriate responses. Previous research has focused on building models for video action and dynamics recognition (Lei et al., 2018; Bertasius et al., 2021). However, most of these efforts fall under the category of simple perceptual-level understanding, such as handling straightforward video (Zolfaghari et al., 2018; Lin et al., 2019). Recent advancements in Transformer-based language models (Vaswani et al., 2017; Brown et al., 2020) have been accompanied by substantial progress in visual-language models (VLMs), leading to significant improvements in video question answering performance. Ko et al. (2023) integrated visual encoders and LLaMA-Adapter (Zhang et al., 2024) into LLMs to enable video understanding, training the model to process both textual and visual inputs effectively. Min et al. (2024) and Wang et al. (2024) demonstrated remarkable performance improvements by first generating image captions using a VLM, selecting frames directly relevant to the question from the video, and then integrating these captions with the reasoning process of an LLM, such as ChatGPT (OpenAI, 2024). (Fei et al., 2024) extended this approach by applying Chain-of-Thought (CoT) (Wei et al., 2024) reasoning capabilities from LLMs to VLMs. Recently, Xiao et al. (2024) critically questioned the degree to which the answers generated by such techniques are truly grounded in the relevant visual content. However, research verifying the alignment between understanding and response in VLMs has yet to be extensively explored.

3 Fine-Grained NExT-QA benchmark

Data Source. We introduce an extended benchmark dataset, FgNExT-QA, which builds upon NExT-QA (Xiao et al., 2021) to better align with fine-grained VideoQA tasks. Most existing VideoQA datasets (Yu et al., 2019; Mangalam et al., 2023) either consist of trimmed, short videos or lack closed-ended answers, making them unsuit-

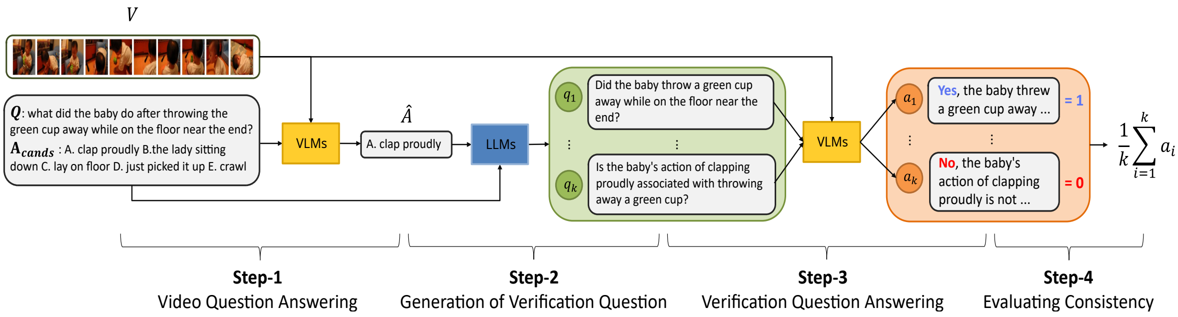


Figure 2: An overview of the proposed framework.

able for our purpose. In contrast, we chose the closed-ended VideoQA subset of NExT-QA, which provides five possible answer choices for each question, with one correct answer. NExT-QA comprises 1,000 videos and includes a total of 8.56k existing question-video pairs. For each question-option pair, we generated five variations. As a result, we created 42.82k newly generated binary-answerable questions. Furthermore, we include a comprehensive discussion of the core limitations inherited from NExT-QA in Appendix. D

Fine-grained Question Generation. To facilitate fine-grained question generation, we adopt open-source LLMs. First, a question and a option are input into the LLMs, along with few-shot examples, to generate fine-grained questions. Detailed prompts and examples are provided in Appendix A. For each question-option pair, up to five questions are generated. Empirically, we observed that generating more than five questions often results in duplicate questions. The generated questions are closed-ended questions (Xiao et al., 2021; Mangalam et al., 2023) that can be answered with "Yes" or "No," which facilitates verifying consistency with the original answer. However, due to the sampling characteristics of LLMs, unintended types of questions are occasionally generated, and such questions are excluded from the results. Detailed statistics on the generated questions are provided in Appendix B.

4 Video Understanding and Response Consistency Assessment

To investigate the understanding of VLMs, we present VURCA, a framework designed to quantify the consistency between video understanding and responses by integrating VideoQA with fine-granined questions generation. As illustrated in Figure 2, the process of the proposed framework consists of four main steps: video question answering, generation of verification question, verification question answering and evaluating consistency

Step-1: Video Question Answering

In the first step of our framework, we instruct the VLMs to respond to the closed-set VideoQA task. Specifically, a video V, the original question Q, and a set of candidate options $\mathbf{A}_{\text{cands}} = \{A_1, A_2, \dots, A_5\}$, are used as inputs for VLMs to generate an initial response $\hat{A}$, which is represented as:

$$\text{VLM}(V, Q, \mathbf{A}_{\text{cands}}) \mapsto \hat{A}.$$

The VLMs utilize their multimodal abilities to derive $\hat{A}$ through an integration of visual and textual reasoning. However, the processes underlying visual and textual reasoning remain a black box and cannot be directly observed.

Step-2:Generation of Verification Question

In this step, we generate fine-grained verification questions to investigate the understanding demonstrated by VLMs in their responses. Q and $\hat{A}$ are input into the LLM, generating a set of fine-grained questions $\mathbf{q}_{\text{fg}}$:

$$\text{LLM}(\mathcal{E}_{\text{few-shot}}, Q, \hat{A}) \mapsto \mathbf{q}_{\text{fg}}.$$

Using a few-shot example set $\mathcal{E}_{\text{few-shot}} = \{(Q^1, \hat{A}^1, \mathbf{q}^1_{\text{fg}}), (Q^2, \hat{A}^2, \mathbf{q}^2_{\text{fg}}), \dots, (Q^n, \hat{A}^n, \mathbf{q}^n_{\text{fg}})\}$, where n represents the number of examples provided to LLMs, we generated fine-grained questions $\mathbf{q}_{\text{fg}} = \{q_i\}_{i=1}^k$, where k is the number of questions. Each q_i is generated as a closed-ended question form with a "Yes" or "No" response.

Step-3:Verification Question Answering

Each fine-grained question q_i in $\mathbf{q}_{\text{fg}}$ is individually input into the VLMs along with V to generate a binary verification response a_i. This process can be expressed as follows:

$$\text{VLM}(V, q_i) \mapsto a_i \in \{1, 0\}.$$

Note that we encode the verification responses of VLMs as binary numbers: 1 for "Yes" and 0 for "No". These binary responses ensure a simple, objective, and accurate evaluation, minimizing ambiguity and streamlining the verification process.

Step-4:Evaluating Consistency

Finally, $\{a_i\}_{i=1}^{k}$ are aggregated to compute a consistency score for VLM's understanding for given Q. The consistency score is calculated as the ratio of the number of "Yes" responses to the number of the fine-grained questions. "Yes" responses indicate that the model demonstrates the same understanding for rephrased questions based on $\hat{A}$. Formally, the consistency score S_{cons} is defined as:

$$S_{\text{cons}} = \frac{1}{k}\sum_{i=1}^{k} a_i.$$

By evaluating the consistency for all the questions in VideoQA datasets, proposed framework provides an objective score to reflect its interpretative reliability.

5 Experiments

5.1 Overview

Our study aims to address three key research questions to evaluate and comprehensively analyze video comprehension and response consistency in VLMs. **Q1:** To what extent do VLMs exhibit consistent comprehension with the initial responses? Specifically, how does the comprehension manifest in cases where the response is correct versus when the response is incorrect? **Q2:** What is VLMs' level of understanding of other options not selected in the initial response? **Q3:** Do VLMs perform well even on fine-grained questions? To investigate these questions, we conduct the proposed framework to obtain S_{cons} and then perform additional comparative analyses to answer the key questions.

5.2 Experimental Settings

Our experiments, based on the close-ended videoQA tasks of the NExT-QA benchmark, were conducted using the proposed framework with state-of-the-art VLMs, including Llava-OneVision 0.5b, Llava-OneVision 7b, and Llava-Video 7b. For all VLMs, we uniformly sample 32 frames from the videos and input them, along with the corresponding questions and options, into the models. The generation of fine-grained questions, which is a part of the proposed framework, is carried out using the microsoft/Phi-3.5-mini-instruct (Abdin et al., 2024) LLM model. We implemented greedy search decoding by selecting the highest-probability token at each step in a fully deterministic manner. To this end, we set the temperature to 0 and disabled sampling-based parameters such as top-k and top-p.

Model	NExT-QA **Acc**	Consistency Score $S_{\text{cons}}^{\text{Total}}$	$S_{\text{cons}}^{\hat{A}=A^*}$	$S_{\text{cons}}^{\hat{A}\neq A^*}$
Llava-ov 0.5b	0.572	0.903	0.918	0.884
Llava-ov 7b	0.794	0.924	0.935	0.881
Llava-video 7b	0.832	0.924	0.936	0.878

Table 1: Evaluation of VLMs understanding of the initial responses.

5.3 Result and Analysis

5.3.1 Q1: Understanding of the Initial Responses

In this experiment, we investigate understanding exhibited by VLMs in the initial responses. To conduct this, we calculate $S_{\text{cons}}^{\text{Total}}$ which is the average S_{cons} over 8,564 question-video pairs in the test data of the NExT-QA benchmark using state-of-the-art VLMs. Additionally, we analyze the differences in S_{cons} between cases where the initial response $\hat{A}$ was correct ($\hat{A} = A^*$) and those where $\hat{A}$ was incorrect ($\hat{A} \neq A^*$), where A^* denotes the gold answer in $\mathbf{A}_{\text{cands}}$. The results are summarized in Table 1.

All VLMs show scores above 0.9 for $S_{\text{cons}}^{\text{Total}}$, indicating that the models provided a high consistent responses. Furthermore, each model showed a higher $S_{\text{cons}}^{\hat{A}=A^*}$ score when generating correct answers, while exhibiting a lower $S_{\text{cons}}^{\hat{A}\neq A^*}$ when VLMs fail to generate correct answers. These results suggest that when VLMs generate initial responses based on uncertain understanding of the video content, VLMs generate inconsistent response to fine-grained verification questions. This behavior becomes more pronounced as model size increases. $S_{\text{cons}}^{\hat{A}\neq A^*}$ score shows the largest gap of 0.057 in the 7B model, indicating that as the size and performance of VLMs increase, the consistency between the fine-grained verification answer and the initial response decreases when generating incorrect answers.

Model	$S_{\text{cons}}^{\text{Total}}$	S_{cons}^{-}
Llava-ov 0.5b	0.903	0.241
Llava-ov 7b	0.924	0.425
Llava-video 7b	0.924	0.437

Table 2: Comparison of VLMs overall consistency score $S_{\text{cons}}^{\text{Total}}$ and negative consistency score S_{cons}^{-}.

5.3.2 Q2: Evaluation of understanding to unselected options

In this experiment, we generate additional fine-grained questions based on randomly selected options different from $\hat{A}$ to investigate whether the VLM can generate negative responses for the options excluding $\hat{A}$. Specifically, the VLMs understanding about the original question is considered higher when it generates more "No" responses for the fine-grained questions that conflict with its initial response. To quantify this, the negative consistency score S_{cons}^{-} is defined as:

$$S_{\text{cons}}^{-} = \frac{1}{k}\sum_{i=1}^{k}(1 - a_i).$$

As shown in Table 2, S_{cons}^{-} is significantly lower than the $S_{\text{cons}}^{\text{Total}}$ across all VLMs. A low S_{cons}^{-} indicates that VLMs fail to demonstrate a clear understanding of why unchosen options were excluded. In other words, the results suggest that VLMs do not accurately understand the video content well enough to make a clear and justified choice among the options. In particular, for the 0.5b model, S_{cons}^{-} was 0.662 lower than the corresponding S_{cons}. For the 7b models, the differences were 0.499 and 0.487. These results indicate that scaling the model size leads to increased consistency score in its responses, reflecting enhanced certainty in its comprehension and decision-making processes.

5.3.3 Q3: Evaluation of Fine-Grained Question Responses

For the final experiment, we generated fine-grained questions for all options, covering both the gold answer and the other options and evaluated the accuracy $\textbf{Acc}^{\text{Total}}$. We also measured separately the accuracy on questions for the gold answers ($\textbf{Acc}^{+}$) and the accuracy on questions for other options ($\textbf{Acc}^{-}$). The results, compared to those of the original questions in NExT-QA, are summarized in Table 3.

Model	NExT-QA	Fine-grained QA		
	Acc	$\textbf{Acc}^{+}$	$\textbf{Acc}^{-}$	$\textbf{Acc}^{\text{Total}}$
Llava-ov 0.5b	0.572	0.895	0.242	0.373
Llava-ov 7b	0.794	0.916	0.435	0.529
Llava-video 7b	0.832	0.921	0.444	0.537

Table 3: Evaluation of VLMs performance on fine-grained questions for all options.

The 0.5b and 7b models showed a 0.253 difference in $\textbf{Acc}^{+}$, but a significantly larger performance gap of 0.201 in $\textbf{Acc}^{-}$, demonstrating superior performance by the larger model. Despite this improvement, the 7B model still exhibits insufficient performance. These results highlight that video comprehension is not only about accurately identifying the correct answer but also about understanding objects or actions that are irrelevant or unsuitable for the VideoQA task. Furthermore, even though Llava models with 7B parameters achieve around 80% performance on the NExT-QA dataset, they exhibit low performance in $\textbf{Acc}^{-}$. The results also suggest that relying solely on accuracy in multiple-choice VideoQA is not sufficient to evaluate the understanding of VLMs, emphasizing the need for further advancements to address the current limitations of VLMs.

6 Conclusion

This paper explores how visual language models understand VideoQA tasks and generate appropriate responses. However, evaluating whether these models truly understand both video and language inputs remains a challenging task. To address the challenge of evaluating VLMs comprehension, we propose VURCA, a framework to assess the alignment between the initial responses and VLMs understanding. VURCA achieves this by generating verification questions and comparing the subsequent responses with its initial answers. Additionally, we introduce FgNeXT-QA, a benchmark dataset designed for fine-grained VideoQA tasks, which offers more fine-grained assessment scenarios. Our experimental results indicate that despite their impressive performance in QA tasks, VLMs often fail to adequately understand video content and the corresponding questions. These results provide valuable insights for the development of advanced evaluation frameworks, the design of more robust model architectures, and the refinement of training methodologies. Future research should aim to enhance the reasoning capabilities of VLMs through improved pre-training strategies that integrate a more comprehensive understanding of video content and question semantics.

7 Limitation

While our proposed framework and dataset extension provide valuable insights into fine-grained VideoQA evaluation, several limitations remain. The generation of verification questions in FgNExT-QA relies on large language models (LLMs), which may introduce noise or bias in the reformulated binary questions. This could potentially affect the reliability and objectivity of the evaluation process. Therefore, we guided the model through few-shot examples to generate questions that can be answered with a simple "yes" or "no" which significantly reduced errors. However, since the model also tended to generate questions starting with The Five Ws (what, where, who, when, why, how), we excluded those from our final set. A detailed discussion of these inherited issues is provided in Appendix B.

Moreover, our research specifically focuses on the VideoQA task, which is important but may limit its applicability to broader multimodal or general video understanding research. Therefore, as a next step, we plan to expand our work to tackle more challenging benchmarks such as VideoMME and explore tasks like description generation and open-ended question answering.

8 Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2021-II212068, Artificial Intelligence Innovation Hub). We would like to thank YoungOk Kim, Chunhwan Jung, and Yubin Ock for their support to this project. This research was supported by the AI Center at HD Korea Shipbuilding & Offshore Engineering.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, and Ahmed Awadallah et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, and Dhariwal et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Zaid Khan and Yun Fu. 2024. Consistency and uncertainty: Identifying unreliable responses from blackbox vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10854–10863.

Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. 2023. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4316, Singapore. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7082–7092.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. 2024. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2024. Videotree: Adaptive tree-based video representation for LLM reasoning on long videos.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13204–13214.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. 2023. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15405–15416.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. 33:9127–9134.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. 2024. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*.

Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. 2018. Eco: Efficient convolutional network for online video understanding. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, page 713–730, Berlin, Heidelberg. Springer-Verlag.

A Prompt Example

This section introduces the prompts used for VLMs and LLMs. Table 4 shows the prompts used in VideoQA. By providing a simple task description along with the video, video question, and answer options, the VLMs generate the system output, which is the final answer. In this paper, NExT-QA data was used, where there are a total of five options, A to E. The final answer is to select one of them. Table 5 presents the prompt for fine-grained question generation. It begins with a simple task instruction, followed by a few-shot example to produce the desired context as output. The few-shot example consists of a question, an answer, and five atomic questions. After the few-shot example, the target question and answer are provided to the LLMs, which then generate the corresponding atomic questions. Table 6 shows the prompt for generating verification responses using atomic questions as input. Similar to Table 4, this prompt excludes the answer options and instead focuses solely on inputting the atomic question to guide the output generation.

B FgNExT-QA statistics

We generated five atomic questions for each of the 8.56k question-video pairs in NExT-QA, with 5 answer options per pair, resulting in 21.41k atomic questions. Due to the characteristics of the LLM, we excluded potential questions that could be generated starting with The Five Ws (what, where, who, when, why, how). These excluded questions accounted for approximately 0.78% of the total. After this filtering process, 21.24k questions were retained for the experiments.

C Qualitative Example from FgNExT-QA

In this section, we perform a qualitative analysis based on actual output examples. Figure 3 illustrates a case where the VLM correctly identified the answer. For the fine-grained questions, the VLM responded with "Yes" to all questions generated for the correct option, while it generated responses including "No" for fine-grained questions generated for other options. In contrast, Figure 4 shows a case where the VLM generated a response different from the target. In this case, the VLM demonstrated a slightly higher proportion of "Yes" responses for the answer it generated. This suggests that the model tends to provide answers consistent with its earlier response, even in fine-grained questions. The similar distribution of responses across diverse questions indicates a lack of understanding and confidence in its answers.

D Inheriting Limitations from NExT-QA

Although FgNExT-QA reformulates the original NExT-QA into binary QA format by decomposing multiple-choice questions into individual question-option pairs, it inherits limitations from NExT-QA, as it is built upon the same question-answer pairs and video contexts. For example, NExT-QA has been shown to contain biases in its answer distribution, and some questions may rely more on textual commonsense knowledge rather than visual evidence from the video. These aspects can limit the effectiveness of evaluating true video-text understanding. While converting to a binary QA format allows for more granular evaluation of model understanding for each candidate option, the quality of distractor options in NExT-QA still affects the difficulty level and diagnostic power of the dataset.

As a result, although FgNExT-QA enhances the evaluation granularity by shifting to binary QA, its reliance on NExT-QA's original structure and content means that certain dataset-level limitations—such as superficial distractors or a lack of visually grounded reasoning—may still affect the robustness and generalizability of model performance.

System You are a helpful assistant.
User *<video>* {input question} A. {option} B. {option} C. {option} D. {option} E. {option}
System assistant:

Table 4: Examples of instructions used in the video question answering.

User After reviewing the questions and answers, generate atomic questions to clarify and verify your responses, ensuring that the questions can be answered with a simple ’yes’ or ’no’. Question:how does the girl show her excitment over the toy? Answer:By turning it around repeatedly Atomic Questions: -Can it be said that the girl is excited about the toy, as shown by her turning it around repeatedly? -Is the girl turning the toy around repeatedly to show her excitement? -Can the girl’s excitement be inferred from her action of turning the toy around repeatedly? -Is the toy the reason for the girl’s excitement, as evidenced by her turning it around repeatedly? -Does the girl’s action of turning the toy around repeatedly suggest she is excited? *<Few shot examples>* Question:{input question} Answer:{input answer}
System Atomic Questions:

Table 5: Examples of instructions and few-shot examples used in the fine-grained question generation.

System You are a helpful assistant.
User *<video>* {input atomic question}
System assistant:

Table 6: Examples of instructions used in the fine-grained question answering.

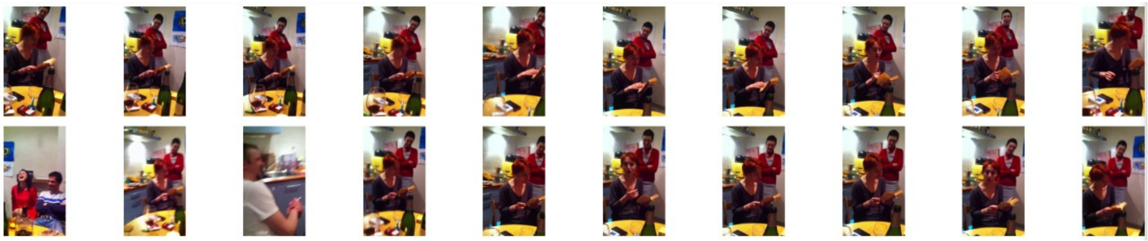
- **Video** :

- **Question** : where is this place?
- **Options** : A. restaurant B. theatre C. farm D. kitchen E. living room
- **Target** : E. living room
- **VLM Answer** : E. living room

- **Fine-grained questions andanswers**:
  - C. farm
    - Is this place located on a farm? No.
    - Can it be inferred that the location of this place is a farm? No, it cannot be inferred that the location is a farm.
    - Is the farm the setting for this place? No, the setting is a toy train set.
    - Does this place belong to or take place on a farm? No, the place is a cozy living room.
    - Is 'farm' the correct term to describe where this place is situated? No, 'farm' is not the correct term; it should be 'home'.
  - E. living room
    - Is the place being referred to as a living room? Yes
    - Can it be said that the location in question is a living room? Yes
    - Is the place described as a living room? Yes
    - Does this place have the designation of a living room? Yes
    - Is 'living room' an accurate description for this location? Yes

Figure 3: Example of a correct prediction by the VLM.

- **Video** :



- **Question** : why did the woman in red sitting down lean forward in the middle?
- **Options** : A. stabilize her B. service staff C. for support to writeD. observe the childrenE. laughing
- **Target** : E. laughing
- **VLM Answer** : A. stabilize her

- **Fine-grained questions andanswers**:
  - A. stabilize her
    - Is the woman in red leaning forward to stabilize herself while sitting down in the middle? No, the woman in red is standing with her arms crossed.
    - Does the woman in red sit down and then lean forward for stabilization purposes? No, the woman in red stands with her arms crossed and leans forward for stabilization purposes.
    - Is the action of leaning forward by the woman in red observed for stabilization reasons? Yes, the action of leaning forward by the woman in red is observed for stabilization reasons.
    - Does the woman in red use her body position (leaning forward) to achieve stability while sitting down in the middle? Yes, the woman in red uses her body position (leaning forward) to achieve stability while sitting down in the middle.
    - Can it be inferred that the woman's leaning forward is a method of stabilizing herself while seated at this location? Yes, it can be inferred that the woman's leaning forward is a method of stabilizing herself while seated at this location.
  - E. laughing
    - Is the woman in red leaning forward because she is laughing? No, the woman in red is leaning forward with her arms crossed.
    - Did the woman in red sit down and then lean forward due to laughter? No, the woman in red leaned forward due to laughter before sitting down.
    - Can it be inferred that the woman's laughter caused her to lean forward while sitting down in the middle ? Yes, it can be inferred that the woman's laughter caused her to lean forward while sitting down in the middle.
    - Is laughing a reason for the woman in red to lean forward while seated in the middle? Yes, laughing is a reason for the woman in red to lean forward while seated in the middle.
    - Does the act of laughing explain why the woman in red leans forward while sitting down in the middle? No, the act of laughing does not explain why the woman in red leans forward while sitting down in the middle.

Figure 4: Example of an incorrect prediction by the VLM.