YTCommentQA: Video Question Answerability in Instructional Videos

Saelyne Yang[1*], Sunghyun Park[2], Yunseok Jang[3*], Moontae Lee[2,4]

[1]KAIST,
[2]LG AI Research,
[3]University of Michigan,
[4]University of Illinois Chicago
saelyne@kaist.ac.kr, {sunghyun.park, moontae.lee}@lgresearch.ai, yunseokj@umich.edu

Abstract

Instructional videos provide detailed how-to guides for various tasks, with viewers often posing questions regarding the content. Addressing these questions is vital for comprehending the content, yet receiving immediate answers is difficult. While numerous computational models have been developed for Video Question Answering (Video QA) tasks, they are primarily trained on questions generated based on video content, aiming to produce answers from within the content. However, in real-world situations, users may pose questions that go beyond the video's informational boundaries, highlighting the necessity to determine if a video can provide the answer. Discerning whether a question can be answered by video content is challenging due to the multi-modal nature of videos, where visual and verbal information are intertwined. To bridge this gap, we present the YTCommentQA dataset, which contains naturally-generated questions from YouTube, categorized by their answerability and required modality to answer – visual, script, or both. Experiments with answerability classification tasks demonstrate the complexity of YTCommentQA and emphasize the need to comprehend the combined role of visual and script information in video reasoning. The dataset is available at https://github.com/lgresearch/YTCommentQA.

1 Introduction

Instructional videos present a procedural guide on how to perform a task through visual demonstrations and verbal explanations (Miech et al. 2019). As people watch and follow the videos, they naturally have questions regarding video content, such as clarifying a certain part or adapting the instructional content to their own situations (Madden, Ruthven, and McMenemy 2013; Poché et al. 2017). Providing appropriate answers to these questions is crucial for people to comprehend and follow instructions effectively. However, addressing every question promptly can pose challenges for video creators.

At the same time, extensive research has been conducted in the domain of Video Question Answering (Video QA), which aims to provide answers to questions about video content (Colas et al. 2019; Zhao et al. 2021; Li et al. 2020; Yang et al. 2020a). However, the majority of models are trained on
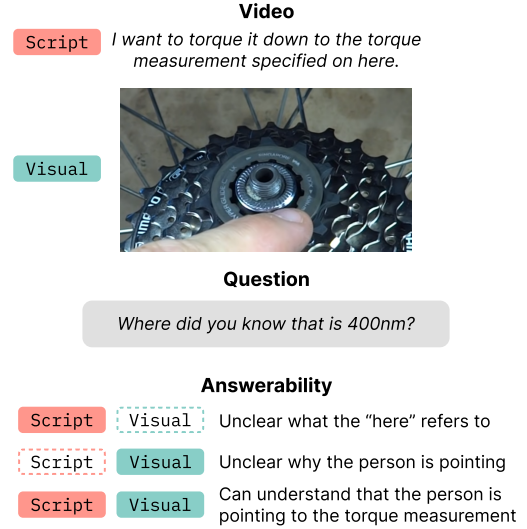
---

*Work done at LG AI Research

Figure 1: A question on video can be either (1) unanswerable by video, (2) answerable by visual, (3) answerable by script, or (4) answerable when both visual and script are present. The figure shows an example of (4), where the question is answerable with the understanding of both visual and script.

datasets labeled by crowd workers or domain experts who were instructed to create both questions and answers based on the video content (Colas et al. 2019; Zhao et al. 2021; Li et al. 2020), or those generated with automated methods (Yang et al. 2020a). Yet, questions from real-world users can include queries that go beyond the scope of what can be answered within the video while questions still remain relevant to the content. These queries may necessitate additional knowledge such as domain or creator-specific knowledge (Zhao et al. 2021). To avoid generating false information, it is essential to determine whether a given question can be answered within the video. However, to our knowledge, there has been limited prior research on video answerability.

The Video Answerability task poses significant challenges due to the multi-modal nature of videos, where information is conveyed through both visual and verbal channels. A simplistic approach to determine answerability is to ex-

Dataset	Domain	Question Collection	Question Answerability
TutorialVQA	Screencast Tutorial	Crowd workers	Answerable in video
PsTuts-VQA	Screencast Tutorial	Domain experts	Answerable with domain knowledge base
HowToVQA69M	General How-to	Automated method	Answerable in video
iVQA	General How-to	Crowd workers	Answerable in video
How2QA	General How-to	Crowd workers	Answerable in video
YTCommentQA(Ours)	General How-to	Real-world users	Answerable & Unanswerable (Answerable contains required modality)

Table 1: Comparison of existing instructional video QA datasets and YTCommentQA

tract frames and transcripts from videos and apply existing techniques for text answerability (Rajpurkar, Jia, and Liang 2018) and image answerability (Gurari et al. 2018). However, relying solely on these approaches is insufficient because in videos, visual and verbal information complement each other (Huang et al. 2018), and crucially, some questions can be answered only when both modalities are considered together (See Figure 1). Also, image answerability alone may not suffice since there are instances where a single frame does not provide an answer, but a series of frames can, such as repetition count (Jang et al. 2019).

This paper presents YTCommentQA, a comprehensive dataset containing naturally generated questions from real-world users in YouTube instructional videos. The dataset includes questions that naturally arise from real-world users in the comment sections of YouTube, demonstrating greater diversity and length than questions found in existing Video QA datasets. It also indicates whether the question is answerable within the video and, if so, provides the associated answer segments. The answerability information categorizes questions into four groups depending on the required modality to answer the questions: (1) Unanswerable, (2) Answerable with visuals, (3) Answerable with the script, and (4) Answerable only when both script and visuals are provided.

We design two tasks for the YTCommentQA dataset. The Segment Answerability Classification asks the model to predict the answerability of questions given the evidence segment, while the Video Answerability Classification task asks the model to predict answerability along with the required modality given the entire video. The results demonstrate the challenging nature of YTCommentQA in predicting the answerability of questions and its required modality. Our dataset highlights the need to understand the complementary role of visual and script information in video question answering.

2 Related Work

2.1 Video Question Answering

To enhance the comprehension of videos through question and answering, numerous computational approaches have been devised (Yang et al. 2020b; Ye et al. 2017; Yang et al. 2022) and a range of benchmark datasets have been proposed. As succinctly outlined by Zhong, Xiao, and Ji et al., Video QA datasets can be classified into three princi-

pal categories based on the information they engage: Plain, Multimodal, and Knowledge-based Video QA (Zhong et al. 2022): (1) Plain Video QA delves into the visual content of videos, encompassing understanding everyday activities (Yu et al. 2019; Castro et al. 2022). Given the intrinsic complexity arising from the spatial and temporal nature of videos, several datasets focus on the spatial, temporal, and causal reasoning within videos (Grunde-McLaughlin, Krishna, and Agrawala 2021; Jang et al. 2019; Xiao et al. 2021). (2) Multimodal Video QA extends its reach beyond visuals by incorporating supplementary information from videos, such as subtitles. This multifaceted approach empowers Video QA with a profound understanding encompassing both visual and verbal components (Castro et al. 2020; Lei et al. 2018; Tapaswi et al. 2015). (3) Lastly, knowledge-based Video QA draws upon external sources beyond the video itself. These sources might encompass, for instance, commonsense knowledge gleaned from a drama series (García et al. 2019) or software application-specific knowledge (Zhao et al. 2021). This Multimodal Video QA highlights the crucial role played by visual and verbal elements, and the knowledge-based Video QA highlights that there are questions that cannot be answered by the video that users are still interested in. We aim to foster this line of research by detecting which question needs additional knowledge in the first place by investigating the nuanced role played by each modality.

2.2 Instructional Video Question Answering

Focusing on instructional videos, TutorialVQA (Colas et al. 2019) and PsTuts-VQA (Zhao et al. 2021) have focused on screencast videos, aiming to cultivate a comprehensive understanding of software tutorial videos. Expanding the scope, HowToVQA69M (Yang et al. 2020a), iVQA (Yang et al. 2020a) and How2QA (Li et al. 2020) have harnessed questions extracted from the extensive HowTo100M dataset, an expansive repository of how-to videos spanning 12 distinct genres (Miech et al. 2019). However, their questions are artificially generated and might be far from what real-world users would have who follow the tutorial content. Some have gathered questions manually, presenting answer segments to crowdsourced workers and soliciting question generation (Colas et al. 2019; Li et al. 2020; Yang et al. 2020a). Others have relied on domain experts for crafting question-answer pairs (Zhao et al. 2021) or automated tech-
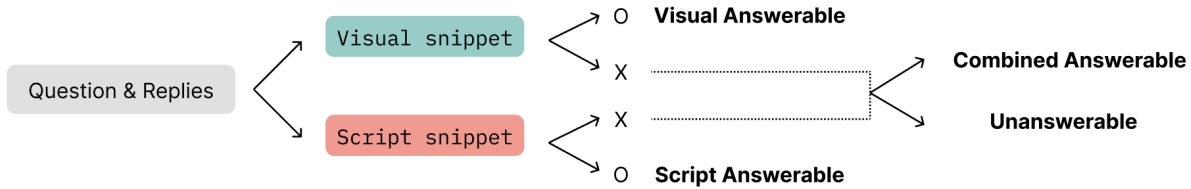
Figure 2: Annotation workflow for the video question answerability. Once an annotator identifies that the `timestamp` used in a reply to a given question suggests an answer in the video, they are provided with visual and script snippets centerd around the timestamp. For questions that could not be answered using visual or script snippets, the annotators are asked whether both were necessary to answer or if the question was unanswerable altogether.

niques (Yang et al. 2020a). Our dataset not only tackles the video question answerability problem but also is enriched by the incorporation of naturally generated questions, thus infusing a more human-like and authentic dimension to the landscape of Video QA.

2.3 Text and Image Question Answerability

Identifying whether a question can be answered or not from a given context is important for reliable and interpretable answer generation (Hu et al. 2018; Nishida et al. 2021). In the domain of machine reading comprehension with text-based context, numerous studies have contributed to understanding unanswerable questions. SQuAD 2.0 (Rajpurkar, Jia, and Liang 2018) introduces relevant unanswerable questions, challenging models to discern whether a question can be answered or not. Several other question-answering datasets have also incorporated unanswerable questions to improve reading comprehension (Choi et al. 2018; Reddy, Chen, and Manning 2018; Campos et al. 2016; Trischler et al. 2017). Additionally, data augmentation techniques, like generating relevant unanswerable questions (Zhu et al. 2019), have been introduced to enhance this line of research.

Question answerability has also been explored in the domain of visual question answering. A line of work augmented existing Visual Question Answering datasets with irrelevant questions to adequately handle unanswerable cases (Ray et al. 2016; Mahendru et al. 2017; Toor, Wechsler, and Nappi 2017). Additionally, VizWiz (Gurari et al. 2018) collected naturally occurring unanswerable questions from associated images taken by blind individuals. These images feature inadequate image quality or content, presenting real-world challenges in VQA. To predict answerability in VQA, several transformer-based models have been proposed (Le, Nguyen, and Nguyen 2021; Nguyen-Tran et al. 2022)

In light of these contributions to answerability in both text and image contexts, to our knowledge, there is no existing work on answerability in videos. To address this gap and further enhance question-answering systems, we propose a novel dataset for video question answerability that requires multi-modal reasoning skills. We believe that exploring answerability in videos can advance the understanding of question answerability in more diverse and complex scenarios.
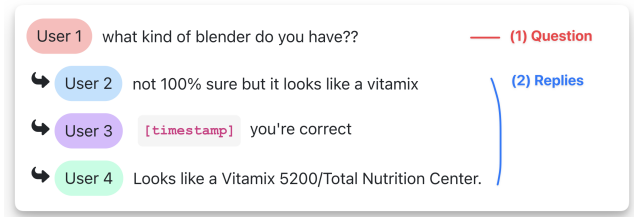

Figure 3: Example question and its replies that contain a timestamp.

3 YTCommentQA: Dataset Collection

We introduce YTCommentQA, a video question answerability dataset. It consists of 2,332 questions asked by real-world users on 2,004 YouTube videos and their associated answerability information. We describe our dataset collection procedure below.

3.1 Video Collection

We selected videos from the YT-Temporal-180M dataset (Zellers et al. 2021), which encompasses a wide range of domains and topics from public YouTube videos. We specifically focused on instructional videos present in the dataset, which overlaps with the HowTo100M (Miech et al. 2019) dataset. Then, we further refined our selection by filtering videos falling within the "Howto & Style" category on YouTube, resulting in 245,354 videos.[1] We obtained the transcripts of the videos and restored punctuations in the full transcript using an off-the-shelf BERT-based model (Nurmanbetov 2021), where completed sentences were segmented by full stops.

3.2 Question Collection

To gather a diverse range of questions relevant to the video content, we collected queries generated by actual users. To achieve this, we crawled the comment section on YouTube for each video utilizing the YouTube Data API (Google 2023). We employed a language detection library (Shuyo

---

[1]We filtered the videos according to the category because there were non-how-to videos in the initial selection, such as music videos.

Videos	2004
Duration (seconds)	524.32±262.8
Questions	2332
Questions per video	1.16±0.62
Question length (# words)	17.2±15.06
Evidence length (seconds)	27.32±17.1
Evidence length (# words)	67.34±37.13
(1) Unanswerable	326 (13.98%)
(2) Visual Answerable	1427 (61.19%)
(3) Script Answerable	1155 (49.53%)
(4) Combined Answerable	174 (7.46%)

Table 2: Data Statistics for YTCommentQA. Note that (2) and (3) have overlapping questions.

2014) to filter out non-English comments. After that, we utilized a question detection module (Khan 2021) to extract and retain only the comments that contained questions.

For the collection of real-world questions and answers that pertain to the video content, we devised the following process. First, we selected questions that had one or more replies and a timestamp (in the format of dd:dd) within the replies. The inclusion of timestamps in comments is a common practice where users refer to specific points in the video (Yarmand et al. 2019). We anticipated that timestamps used in replies would likely point to video relevant to the answers, streamlining the process for crowd workers to identify the answerability of questions. Next, we retained only those questions where the responses involved at least one user other than the original commenter. We excluded questions that had conversations involving more than five users to prevent excessively lengthy threads. This resulted in 3,260 questions from 2,912 videos, which were later narrowed down to 2,332 after another filtering in the Answerability Annotation phase (Section 3.3). In Appendix A, we provide a comparison between YTCommentQA questions and general questions that include questions that do not have replies with a timestamp.

3.3 Answerability Annotation

To annotate the answerability for each question, we developed an annotation system (Appendix I) and recruited annotators through Prolific (Prolific 2023). We designed a two-step process which we describe below.

First, since not all strings in timestamp format (dd:dd) necessarily means that the timestamp contains the answers, annotators were presented with the entire conversation (i.e., a question and its replies. See Figure 3 for an example) and asked if it *implied* that the answer to the question could be found around the timestamp. If the response was positive, we went to the second step, where we displayed a video snippet without audio source near the timestamp and queried whether the visual snippet provided the answer. Participants had two options: "yes" or "no". We then showed the corresponding transcript and inquired if the answer was available in the text. Participants had three response options: "yes",
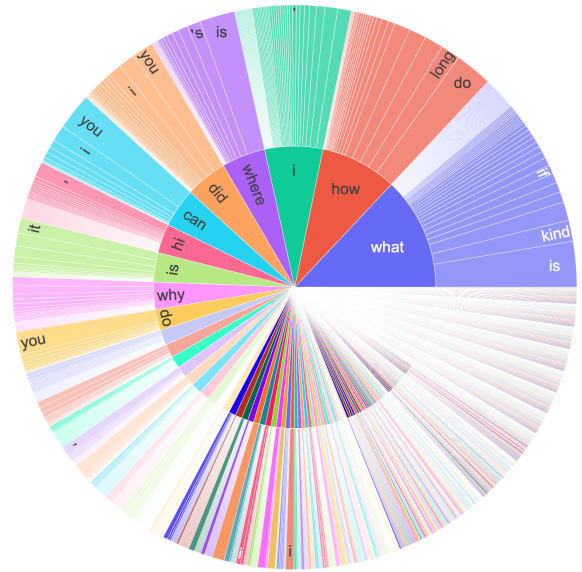

Figure 4: Distribution of the first two words for questions in YTCommentQA, which shows the diversity of the collected questions. The sequence of words begins from the center and extends outward. Words with small font sizes are omitted.

"no", or "requires both video and transcripts." Participants could extend the visual or transcript by up to three more sentences (corresponding video segments for visual snippets) if the snippet ended in the middle of the answer. If the response to the first stage was negative, we asked participants to indicate the alternate purpose of the timestamps being used. We did not include such questions in our dataset since the identified text as a timestamp did not relate to a potential answer segment.

In the entire process, to ensure quality control, we assigned two participants to each question, and in cases of disagreement, we sought one more participant's response until a single answer received at least two votes. Responses with less than 2/3 overlap with others were discarded. Participants were compensated with 0.17 GBP on average for annotating one question.

In summary, the annotated answerability information includes the following categories: (1) Unanswerable, (2) Answerable with visual, (3) Answerable with the script, (4) Answerable when both script and visual are provided. Note that (2) and (3) may co-exist. The complete annotation workflow is depicted in Figure 2. Ethical considerations are discussed in Appendix G.

4 YTCommentQA: Dataset Analysis

4.1 Analysis of Questions

As our questions naturally stem from real-world users rather than being crafted by hired workers, several distinctive traits emerge in our question set. First of all, our questions are roughly twice as long as those in existing Video QA datasets
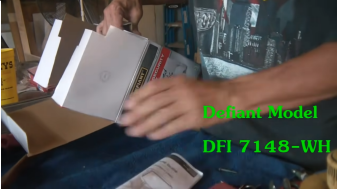
Answerability	Visual	Script	Question
Visual Answerable		you can see the damage on my bottle and the the top part comes off to the front rather than up.	i can't figure out how to take the top part off of the bottom in order to get the air vent back in. can you help me? thanks!!
		just get to some box, check it out, see what all there is.	no model number of the product only the name of manufacturer of the product your reviewing.
Script Answerable		right now I'm going to put it, already pre hit, at the air fryer for five minutes at 200 degrees Celsius.	what temperature did you cook it at?
		rice water can be used on all skin types, such as dry, oily or normal skin.	mam can we use dis for dry skin
Combined Answerable		There are a lot of different bits in this set, but a good place to start is with a ball shaped one.	Amazing! Can you tell me what bit did you use in this particular project?
		first of all, one cup of maida, half a cup of condensed milk, little bit more than one foot cup of butter.	can u pls give measurement of ingredients also

Table 3: Example questions of Visual answerable, Script answerable, and Combined answerable.

[2]. We believe this is partly because users often include context before or after their question to aid readers (i.e., content creators or other viewers) in better understanding the query.

Another notable feature is their diversity, illustrated in Figure 4 by the initial two words of each question. Notably, the most common starting word, "What", only consists of 13% of all questions, in contrast to other Video QA datasets where "What" is the clearly dominant initial word (Castro et al. 2022; Zhao et al. 2021; Li et al. 2020). Additionally, some questions start with casual words like "hi" or "great,"

reflecting users introducing their queries with greetings or expressions of appreciation in real-world situations. Some questions also include less polished or grammatically incorrect phrasing. We show that there is minimal bias in answerability to specific question types in Appendix F.

4.2 Analysis of Answerability Types

Table 2 presents the distribution of answerability types. Unanswerable questions make up 13.98% of the total, while the remaining 86.02% fall into the answerable category, encompassing different modalities required to answer: Visual, Script, and Combined (which requires both visual and script input). Below, we describe the characteristics of each an-

---

[2]Ours: 17.2 vs. How2QA: ~11, TutorialVQA: 9, HowToVQA69M: 8.7, iVQA: 7.6 words on average.

swerability type. Examples of questions can be found in Table 3.

(1) Unanswerable Approximately 13.98% of the questions are deemed unanswerable, both in terms of visual and script cues. Although timestamps are used in replies of the questions, they serve more to clarify the question rather than pinpoint the answer segment (e.g., "if you're talking about the tool at `timestamp`, it is ..."). Alternatively, replies contain relevant information related to the question, yet without providing a complete answer (e.g., "I talk about X at `timestamp` if that helps any.")

(2) Visual Answerable Approximately 61.19% of questions find answers through visual snippets from the video, without requiring verbal input. Examples include queries about visually demonstrable tasks (e.g., "how to take the top part off"). Another noteworthy instance occurs when visuals incorporate textual elements, such as names of depicted products. The practice of textual annotation is common in instructional videos, where it provides clarifications, corrections, or supplementary information that are not shown in the demonstration (Chi et al. 2013).

(3) Script Answerable Roughly 49.53% of questions find answers within the script. As the demonstrator visually guides through a process, they concurrently offer verbal explanations. Questions that seek specific details (e.g., "what temperature was used for cooking?") are answered as the demonstrator provides explanations on them in the video.

Note that (2) Visual Answerable and (3) Script Answerable share overlapping questions. This arises from instances where a question can be answered solely through visuals *and* solely through the script. Such cases include instances where the visual information closely mirrors the verbal content (or vice versa), such as when the creator verbally explains precisely what is occurring visually. It also includes cases where what the creator says are appear on visual through text. Approximately 32.16% of questions are answerable by both modalities.

(4) Combined Answerable Lastly, about 7.46% of questions necessitate both visual and script inputs to provide answers, as these modalities complement each other. This encompasses scenarios like resolving references, where the author uses verbal cues such as "here" or "this" alongside visual demonstrations showing what they are. It also includes instances where visuals solidify verbal explanations by offering concrete examples (e.g., "ball-shaped one"). Additionally, it includes cases where additional information is given in visuals through text annotations that complement what the author explains verbally.

5 Experiments

We benchmark the YTCommentQA across two tasks regarding question answerability: (1) Segment Answerability Classification and (2) Video Answerability Classification. Below we describe the task, experiment setting, and results.

Model	Token	F1 score
Finetuned LMs		
Llama-2 (7B)	4K	45.78
Llama-2 (13B)	4K	55.49
Zero-shot LMs		
GPT-3.5	16K	31.16
GPT-4	8K	33.02
Multimodal Model		
SeViLA	768	46.55

Table 4: The results of Segmentation Answerability Classification task on YTCommentQA. SeViLA selects 4 keyframes from 32 frames.

5.1 Segment Answerability Classification

Task Description Segment Answerability Classification asks the model to predict the answerability of questions given the evidence segment. Our evaluation methodology involves binary classification, where we categorize the answerability into 1) Unanswerable and 2) Answerable.

Experimental Setup We conducted two types of experiments: 1) employing language models by textualizing both visual and verbal content of video segments (Li et al. 2023) and 2) using multimodal models. For 1), we transcribed the visual content using an off-the-shelf image captioning model (Singh 2021) and used transcripts for the verbal content. Additionally, we observed that some questions that are Visual Answerable require the text in the visual content, such as text annotations detailing a product name (see Table 3 for examples). Thus, we extracted text from the answer segment using the Tesseract OCR (Tesseract OCR 2023) and included it in the visual descriptions. Any text displayed within the timeframe was extracted and combined, and the leading author refined the OCR results.

Evaluation Metric We used F1 score for the evaluation metric. Responses that do not follow the prompt (i.e., that are neither "Answerable" nor "Unanswerable") have all been marked as incorrect.

Baseline Models and Training Details We used both fine-tuning and zero-shot models for the language-based baseline models. We employed Llama2 (Touvron et al. 2023) for fine-tuning, which has demonstrated robust performance. We experimented with model sizes of 7B and 13B. As for the zero-shot models, we utilized two powerful LLMs: ChatGPT (OpenAI 2022) and GPT-4 (OpenAI 2023). For the multimodal baseline, we used SeViLA (Yu et al. 2023), which has demonstrated strong performance on video-language benchmarks.

We divided the collected data into training and evaluation sets in a 1:1 ratio to fine-tune the models. We addressed the class imbalance by augmenting unanswerable questions. This was achieved by integrating questions with descriptions of videos where the question had not been asked on. The prompts used for fine-tuning are outlined in Appendix B and training details in Appendix E.

Results The experimental results are presented in Table 4. The results demonstrate the challenge of determining answerability. In particular, GPT-4 exhibited a tendency to misclassify answerable cases as unanswerable, mainly when dealing with visual information. A significant portion (53%) of the false cases pertained to instances where GPT-4 mistakenly categorized Visual Answerable as Unanswerable. The result with the multimodal baseline closely resembles that of Llama-2 (7B), which highlights the challenging nature of the problem even with a multimodal model.

5.2 Video Answerability Classification

Task Description Video Answerability Classification task asks the model to predict the answerability of questions along with the required modality given the entire video. Our evaluation methodology involves multiple-choice question answering, where we categorize the answerability into five classes: 1) Unanswerable, 2) Visual Answerable, 3) Script Answerable, 4) Both Script and Visual Answerable, and 5) Combined Answerable.

Experimental Setup Similar to the Segment Answerability Classification task, we conducted experiments with both 1) language and 2) multimodal baselines. For 1), we first textualized both visual and verbal content of videos. However, as entire videos are often lengthy (8 min 44 sec on average), presenting the entire transcript, video caption and OCR text as input would exceed the model's capacity for training and inference. Therefore, we divided the video into segments corresponding to five transcript sentences. We then generated summaries for each segment using ChatGPT. For the verbal content, we provided the transcript, while for the visual content, we provided captions and OCR text. These verbal and visual summaries from each segment were then concatenated to create comprehensive visual and verbal descriptions for the entire video. More details about the prompts used for generating the summaries can be found in Appendix D.

Evaluation Metric We used accuracy for the evaluation metric. Responses that do not follow the prompt (i.e., that are not classified as one of the five classes) have all been marked as incorrect.

Baseline Models and Training Details We used the same baseline models as in the Segment Answerability Classification task: Llama2 (Touvron et al. 2023), ChatGPT (OpenAI 2022), GPT-4 (OpenAI 2023), and SeViLa (Yu et al. 2023). However, we extended its context window for Llama2 by using rotary positional embeddings (Su et al. 2022) and the FlashAttention algorithm (Dao et al. 2022) to incorporate the longer text inputs. For SeViLA, we processed 768 tokens from the transcript, truncating sequences that exceeded the limit.

We divided the collected data into training and evaluation sets in a 1:1 ratio for fine-tuning. We oversampled Unanswerable, Visual Answerable, and Combined Answerable classes by a factor of two to address the class imbalance in the training set. The prompts used for fine-tuning are outlined in Appendix C and training details in Appendix E.

Model	Token	Accuracy
Finetuned LMs		
Llama-2 (7B)	16K	36.24
Llama-2 (13B)	16K	37.70
Zero-shot LMs		
GPT-3.5	16K	18.42
GPT-4	8K	27.03
Multimodal Model		
SeViLA	768	35.27

Table 5: The results of Video Answerability Classification task on YTCommentQA. SeViLA selects 4 keyframes from 32 frames.

Results Table 5 shows the experimental results. The results highlight the difficulty in understanding extensive inputs encompassing video content and classifying the answerability with required modalities. In particular, GPT-4's false predictions predominantly leaned towards either Unanswerable or Script Answerable, indicating its heavy reliance on script information. Furthermore, a majority (85%) of Combined Answerable instances were inaccurately predicted, indicating the complexities associated with processing multimodal inputs. The result with the multimodal baseline falls slightly short of that achieved by Llama-2, highlighting the difficulty of the problem.

6 Conclusion

This paper presents the YTCommentQA dataset, which contains questions generated by real-world users and the answerability information containing the required modality to provide answers. Through experiments with two answerability classification tasks, we illustrate the inherent challenges in discerning the answerability of video questions and the required modality for the answer. We belive that our work will foster the development of Video QA systems capable of identifying answer sources, thereby enhancing answer reliability. Furthermore, we hope that our work will encourage research that necessitates a deeper understanding and reasoning of multimodality in videos, where the verbal and visual aspects complement each other.

While our study underscores the significance of video question answerability, it has a few limitations. Our annotations primarily focused on the segments centered around the provided timestamps, which could have led to an incomplete understanding of the entire video context. Future research can investigate scenarios where insights from different segments, external knowledge, or commonsense affect answerability. Moreover, our experimental approach involved utilizing summaries or partial chunks of the video due to input length constraints. This method might have overlooked specific video details. Subsequent efforts could explore incorporating complete and comprehensive video information to enhance long-form video question answerability.

References

Campos, D. F.; Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; Deng, L.; and Mitra, B. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *ArXiv*, abs/1611.09268.

Castro, S.; Azab, M.; Stroud, J.; Noujaim, C.; Wang, R.; Deng, J.; and Mihalcea, R. 2020. LifeQA: A Real-life Dataset for Video Question Answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4352–4358. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Castro, S.; Deng, N.; Huang, P.; Burzo, M. G.; and Mihalcea, R. 2022. In-the-Wild Video Question Answering. In *COLING*, 5613–5635. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Chi, P.-Y.; Liu, J.; Linder, J.; Dontcheva, M.; Li, W.; and Hartmann, B. 2013. DemoCut: Generating Concise Instructional Videos for Physical Demonstrations. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, 141–150. New York, NY, USA: Association for Computing Machinery. ISBN 9781450322683.

Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; tau Yih, W.; Choi, Y.; Liang, P.; and Zettlemoyer, L. 2018. QuAC: Question Answering in Context. In *Conference on Empirical Methods in Natural Language Processing*.

Colas, A.; Kim, S.; Dernoncourt, F.; Gupte, S.; Wang, D. Z.; and Kim, D. S. 2019. TutorialVQA: Question Answering Dataset for Tutorial Videos. In *International Conference on Language Resources and Evaluation*.

Dao, T.; Fu, D. Y.; Ermon, S.; Rudra, A.; and Ré, C. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. arXiv:2205.14135.

García, N.; Otani, M.; Chu, C.; and Nakashima, Y. 2019. KnowIT VQA: Answering Knowledge-Based Questions about Videos. In *AAAI Conference on Artificial Intelligence*.

Google. 2023. YouTube Data API. https://developers.google.com/youtube/v3. Accessed: 2023-08-12.

Grunde-McLaughlin, M.; Krishna, R.; and Agrawala, M. 2021. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. In *CVPR*.

Gurari, D.; Li, Q.; Stangl, A.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3608–3617.

Hu, M.; Wei, F.; Peng, Y.; Huang, Z.; Yang, N.; and Zhou, M. 2018. Read + Verify: Machine Reading Comprehension with Unanswerable Questions. In *AAAI Conference on Artificial Intelligence*.

Huang, D.-A.; Buch, S.; Dery, L. M.; Garg, A.; Fei-Fei, L.; and Niebles, J. C. 2018. Finding "It": Weakly-Supervised Reference-Aware Visual Grounding in Instructional Videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5948–5957.

Jang, Y.; Song, Y.; Kim, C. D.; Yu, Y.; Kim, Y.; and Kim, G. 2019. Video Question Answering with Spatio-Temporal Reasoning. *IJCV*.

Khan, S. 2021. question-vs-statement-classifier from huggingface. https://huggingface.co/shahrukhx01/question-vs-statement-classifier. Accessed: 2023-08-12.

Le, T. T.; Nguyen, H.-T.; and Nguyen, M. L. 2021. Vision And Text Transformer For Predicting Answerability On Visual Question Answering. *2021 IEEE International Conference on Image Processing (ICIP)*, 934–938.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. 2018. TVQA: Localized, Compositional Video Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1369–1379. Brussels, Belgium: Association for Computational Linguistics.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023. VideoChat: Chat-Centric Video Understanding. *arXiv preprint arXiv:2305.06355*.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020. Hero: Hierarchical Encoder for Video+Language Omni-representation Pre-training. In *Conference on Empirical Methods in Natural Language Processing*.

Madden, A.; Ruthven, I.; and McMenemy, D. 2013. A classification scheme for content analyses of YouTube video comments. *J. Documentation*, 69: 693–714.

Mahendru, A.; Prabhu, V.; Mohapatra, A.; Batra, D.; and Lee, S. 2017. The Promise of Premise: Harnessing Question Premises in Visual Question Answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 926–935. Copenhagen, Denmark: Association for Computational Linguistics.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Nguyen-Tran, D.-M.; Le, T.; Pho, K.; Nguyen, M. L.; and Nguyen, H.-T. 2022. RVT-Transformer: Residual Attention in Answerability Prediction on Visual Question Answering for Blind People. In *International Conference on Computational Collective Intelligence*.

Nishida, K.; Nishida, K.; Saito, I.; and Yoshida, S. 2021. Towards Interpretable and Reliable Reading Comprehension: A Pipeline Model with Unanswerability Prediction. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Nurmanbetov, D. 2021. Bert-restore-punctuation model from huggingface. https://huggingface.co/felflare/bert-restore-punctuation. Accessed: 2023-08-12.

OpenAI. 2022. ChatGPT. https://openai.com/chatgpt. Accessed: 2023-08-12.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Poché, E.; Jha, N.; Williams, G.; Staten, J.; Vesper, M.; and Mahmoud, A. 2017. Analyzing User Comments on YouTube Coding Tutorial Videos. In *2017 IEEE/ACM 25th International Conference on Program Comprehension (ICPC)*, 196–206.

Prolific. 2023. Prolific. https://www.prolific.co/. Accessed: 2023-08-12.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 784–789. Melbourne, Australia: Association for Computational Linguistics.

Ray, A.; Christie, G. A.; Bansal, M.; Batra, D.; and Parikh, D. 2016. Question Relevance in VQA: Identifying Non-Visual And False-Premise Questions. In *Conference on Empirical Methods in Natural Language Processing*.

Reddy, S.; Chen, D.; and Manning, C. D. 2018. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7: 249–266.

Shuyo, N. 2014. language-detection. https://github.com/shuyo/language-detection. Accessed: 2023-08-12.

Singh, A. 2021. vit-gpt2-image-captioning from hugging-face. https://huggingface.co/nlpconnect/vit-gpt2-image-captioning. Accessed: 2023-08-12.

Su, J.; Lu, Y.; Pan, S.; Murtadha, A.; Wen, B.; and Liu, Y. 2022. RoFormer: Enhanced Transformer with Rotary Position Embedding. arXiv:2104.09864.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2015. MovieQA: Understanding Stories in Movies through Question-Answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4631–4640.

Tesseract OCR 2023. 2023. Tesseract OCR. https://github.com/tesseract-ocr/tesseract. Accessed: 2023-08-12.

Toor, A. S.; Wechsler, H.; and Nappi, M. 2017. Question Part Relevance and Editing for Cooperative and Context-Aware VQA (C2VQA). In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, CBMI '17. New York, NY, USA: Association for Computing Machinery. ISBN 9781450353335.

Touvron, H.; Martin, L.; Stone, K. R.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D. M.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A. S.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I. M.; Korenev, A. V.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv*, abs/2307.09288.

Trischler, A.; Wang, T.; Yuan, X.; Harris, J.; Sordoni, A.; Bachman, P.; and Suleman, K. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 191–200. Vancouver, Canada: Association for Computational Linguistics.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 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)*, 9777–9786.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2020a. Just Ask: Learning to Answer Questions from Millions of Narrated Videos. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1666–1677.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022. Zero-Shot Video Question Answering via Frozen Bidirectional Language Models. In *Advances in Neural Information Processing Systems*.

Yang, Z.; Garcia, N.; Chu, C.; Otani, M.; Nakashima, Y.; and Takemura, H. 2020b. BERT Representations for Video Question Answering. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1545–1554.

Yarmand, M.; Yoon, D.; Dodson, S.; Roll, I.; and Fels, S. S. 2019. "Can You Believe [1:21]?!": Content and Time-Based Reference Patterns in Video Comments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–12. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359702.

Ye, Y.; Zhao, Z.; Li, Y.; Chen, L.; Xiao, J.; and Zhuang, Y. 2017. Video Question Answering via Attribute-Augmented Attention Network Learning. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yu, S.; Cho, J.; Yadav, P.; and Bansal, M. 2023. Self-Chained Image-Language Model for Video Localization and Question Answering. *arXiv:2305.06988*.

Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. In *AAAI*, 9127–9134.

Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J. S.; Cao, J.; Farhadi, A.; and Choi, Y. 2021. MERLOT: Multimodal Neural Script Knowledge Models. In *NeurIPS*.

Zhao, W.; Kim, S.; Xu, N.; and Jin, H. 2021. Video Question Answering on Screencast Tutorials. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20. ISBN 9780999241165.

Zhong, Y.; Ji, W.; Xiao, J.; Li, Y.; Deng, W.; and Chua, T.-S. 2022. Video Question Answering: Datasets, Algorithms and Challenges. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6439–6455. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zhu, H.; Dong, L.; Wei, F.; Wang, W.; Qin, B.; and Liu, T. 2019. Learning to Ask Unanswerable Questions for Machine Reading Comprehension. In *Annual Meeting of the Association for Computational Linguistics*.