Towards Multimodal Question Answering in Educational Domain

Anonymous ACL submission

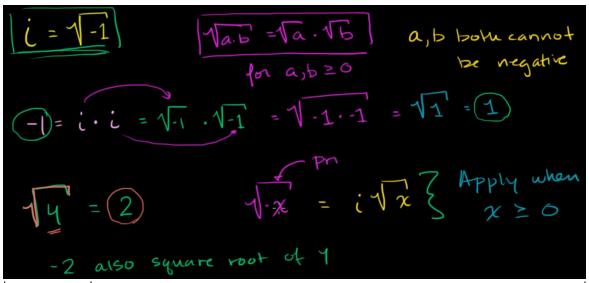
Abstract

The proliferation of educational videos on the Internet has changed the educational landscape by enabling students to learn complex concepts at their own pace. Our work outlines the vision of an automated tutor – a multimodal question answering (QA) system to answer questions from students watching a video. This can make doubt resolution faster and further improve learning experience. In this work, we take first steps towards building such a QA system. We curate and release a dataset named EDUVIDQA, with 3,158 videos and 18,474 QA-pairs. However, building and evaluating an educational QA system is challenging because (1) existing evaluation metrics do not correlate with human judgments, and (2) a student question could be answered in many different ways; training on a single gold answer could confuse the model and make it worse. We conclude with important research questions to develop this research area further.

1 Introduction

Academic videos enable better understanding, retention and overall learning for students by presenting the concepts via text, audio and visual modalities. Students can learn using recorded educational lectures at their own pace and revisit content as needed. Multiple online platforms host academic videos but lack effective doubt-solving support. Typically, learners must wait for an instructor or use discussion forums to resolve doubts, which delays doubt resolution, leading to stalled learning, reduced engagement and potentially confidence issues. This underscores the need for timely, accurate responses to student inquiries.

We believe an *automated tutor*, one that quickly responds to learner doubts, as they watch videos, has an enormous potential to improve their experience, by immediate doubt resolution, increasing concept retention, and aiding overall learning.



Question	I think Sal make a mistake on (vid @ 5:11) when he write the greater than sign! it should be Less than
Ground truth Answer	No, Sal is correct. If he had: i sqrt(x) where X<0, then X is negative. Backup thru Sal steps. If X is negative * (-1) = +X. And he would have started with sqrt(x), not sqrt(-x). He is also trying to highlight that if you had something like: sqrt(12), you would not make this into i sqrt(-12). The imaginary number is not needed if the radical contains a positive number to start with.

Figure 1: An example of a QA pair from our EDUVIDQA dataset for the video https://www.youtube.com/watch?v=rYG1D5lUE4I

Such a tutor will necessitate developing a system for multimodal question answering (QA) in education, where content from both videos and text (via speech transcripts) will be combined to answer questions. As first steps towards building such a tutor – we curate a multimodal QA dataset, identify good evaluation metrics, and train initial models based on existing large language models (LLMs) and vision LMs (VLMs).

We first curate a novel dataset (EDUVIDQA) by extensively filtering LearningQ (Chen et al., 2018). It has 7,297 videos and 201,398 QA pairs, however, many questions do not require visual information for answering. So, we create a more challenging subset of 3,158 videos and 18,474 QA pairs, where each question has an associated video timestamp. Fig. 1 shows an example.

Secondly, we explore the standard natural language generation (NLG) metrics like BLEURT and BERTScore for our task. We find that their scores (for systems with differing performance) are quite similar and correlate rather weakly with human judgments. The only exception is GPT-4 based evaluation, but that is expensive and closed source.

Thirdly, we benchmark existing LLMs (Vicuna (Chiang et al., 2023), LLaMA-3 (AI@Meta, 2024), GPT-4 (OpenAI et al., 2023)) and VLMs (LLaVA (Liu et al., 2023a), GPT-4V) on EDUVIDQA. Not surprisingly, we find that GPT-4V outperforms smaller models.

Overall, we make the following contributions in this paper: (1) We curate a novel dataset, EDUVIDQA with 7,297 videos and 201,398 QA pairs in education domain. (2) We propose GPT-4 based evaluation metrics using prompts specifically designed for the educational domain. (3) We benchmark 3 existing LLMs and 2 VLMs with and without transcript for this task. We make the code[1] and data[2] publicly available.

2 Related Work

Multimodal Question Answering. Visual Question Answering (VQA) aims at answering a text question in the context of an image (Antol et al., 2015). Several VQA datasets have been proposed (Antol et al., 2015; Geman et al., 2015; Singh et al., 2021). Multiple extensions have also been proposed like Text VQA (Singh et al., 2019), Visual Dialog (Das et al., 2017), Video QA (Zeng et al., 2017) and knowledge-based VQA for videos (Garcia et al., 2020). Most VQA methods use multimodal fusion of language and image embeddings (Kembhavi et al., 2017), attention-based multimodal fusion (Yang et al., 2016) or neural module networks (Hu et al., 2017). Recently, VLMs like LLaVA and GPT-4V have been shown to provide state-of-the-art results across several VQA datasets using in-context learning. In this paper, we aim to benchmark such VLMs for video QA for education domain.

Multimodal Educational Datasets. Recently, some datasets have been proposed for VQA and visual question generation (VQG) in education domain. These include TutorialVQA (Colas et al., 2020), SlideVQA (Tanaka et al., 2023), ScienceQA (Lu et al., 2022) for VQA and LearningQ (Chen et al., 2018) for VQG. However, TutorialVQA just identifies a span of a video segment as an answer while SlideVQA has only factual questions which can be simply answered by reading data from a slide. ScienceQA (Lu et al., 2022) contains image-based MCQs but do not have any videos or temporal understanding and hence fail to represent real-world variety of questions. In real world, learners ask diverse questions, from simple inferential queries to complex, open-ended ones.

3 EDUVIDQA Dataset

Dataset Curation. The proposed task requires a dataset where a video is associated with a QA pair. While many platforms offer lecture videos, they typically separate discussion forums from the lecture videos, making it difficult to map questions to videos. Fortunately, the following websites host video-specific questions: Khan Academy[3], TED-Ed[4], YouTube[5] and MathTV.[6] Khan Academy specifically provides a doubt-clarification platform with concise human-written answers, and so best aligns with our task.

Fortunately, the LearningQ dataset (Chen et al., 2018) is an open dataset with data crawled from Khan Academy and TED-Ed. Although the dataset was originally released for Question Generation, we repurpose it for our multimodal QA task. We extract the questions from discussion files, named with a YouTube video-id making it possible for us to map videos and questions. The dataset includes over 338,000 domain-wise question-answer pairs along with unique video-ids and subtitles. The videos have an average duration of 7 minutes, with most questions focused on remembering, understanding and analysis.

We filter this dataset by removing irrelevant or incorrectly answered QA pairs in two phases. In the first phase, we provide the question, answer and domain information to obtain a relevance score using GPT3.5 on a scale of 1-10. We then retain QA pairs with GPT3.5-score ≥ 6. But manual inspection revealed that this subset still had $\sim$15% irrelevant samples. Hence, in the second phase, we leverage GPT-4 and retain QA pairs with score ≥ 9 resulting in $\sim$95% relevant QA pairs. Appendix A.1 lists the scoring prompt. This leads to our EDUVIDQA dataset with 7,297 videos and 201,398 QA pairs.

Further, we observe that several QA pairs could be answered without the video in context (Appendix B). Hence, we retain questions only from Math and Science domains since questions from other domains are less likely to require visual content. To make the task truly multimodal, we sample

[1] https://anonymous.4open.science/r/EduVidQA-D86B/
[2] https://huggingface.co/datasets/vqamaster/EduVidQA
[3] https://www.khanacademy.org/
[4] https://ed.ted.com/
[5] https://www.youtube.com/education
[6] https://www.mathtv.com/

only those QA pairs where the question has a timestamp referring to a time point in the video. This ensures that most of the samples will require the model to apply multimodal reasoning skills over the input video and the question to generate an accurate answer. This results into a more challenging subset of our EDUVIDQA dataset, with 3,158 videos and 18,474 QA pairs. On average, the questions and answers contain 37 and 49 words respectively. We share our dataset publicly.[2] We split this dataset into 13,329 for train, 4,645 for validation and 500 for test.

Dataset Analysis. We perform an analysis to assess the cognitive complexity of questions using BloomBERT (Lau, 2023) on EDUVIDQA. We find that while most questions are understanding (65.3%) and remembering (16.3%) based, many questions also need creativity (6.5%), application (5.5%), analysis (4.6%) & evaluation (1.6%) skills.

4 Evaluation Metrics

Since multimodal QA is an NLG application, we leverage the most widely used NLG metrics. These include unsupervised metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang* et al., 2020), BARTScore (Yuan et al., 2021), and a supervised metric, BLEURT (Sellam et al., 2020). As recommended, we use the CNNDM version of BARTScore[7] and BLEURT-20 model.[8]

To check if these metrics are suitable for the proposed task, we manually annotate 70 samples and compute correlation with human judgments. Manual annotations covered 4 aspects of quality: (i) correctness of answer, (ii) coherence (ease of following an answer), (iii) visual grounding and (iv) relevance to the question on a scale of 0 to 5. Inspired by recent advancements in NLG evaluation using GPT as evaluator (Zheng et al., 2023; Liu et al., 2023b), we also obtain scores using GPT-4. Detailed definitions of these aspects are provided as prompts in Appendix A.3.

We found these correlations for various evaluation metrics with human: BLEU (0.130), ROUGE-1 (0.216), ROUGE-L (0.245), BERTScore (0.318), BARTScore (0.299), BLEURT (0.488), our GPT-4 eval (0.782), Inter-human (0.909). N-gram based metrics like ROUGE and BLEU perform word-level syntactic matches, which makes them unfit for capturing overall semantics. Neural methods (BERTScore, BARTScore and BLEURT) align better since they compare semantics of the entire text. Despite the advantages of automated metrics in terms of definitive scores and speed of computation, they demonstrate only moderate correlation with human judgments. Given that ground truth answers in the Khan Academy dataset are created by multiple humans, they often reflect a diversity of knowledge and creativity. Further, there can be multiple possible answers/explanations to the questions and since these metrics compute similarity of generated text with reference answers, they show poor correlations. Therefore, it is not the best strategy to directly compare the predicted answers with the single ground truth using standard NLG metrics. Overall, we find that our proposed GPT-4 eval metric provides the highest correlation score of 0.782 with human judgments.

5 Benchmarking Results on EDUVIDQA

5.1 Experiments

We experiment with 3 LLMs (Vicuna, LLaMA-3 and GPT-4) and 2 VLMs (LLaVA-v1.5 and GPT-4V). Note that Vicuna, LLaMA-3 and LLaVA-v1.5 have 7B, 7B and 8B parameters respectively. GPT-4 has 1.76T parameters; number of parameters for GPT-4V are unknown. LLaVA is based on aligning CLIP embeddings of the image with text by using a multimodal projection layer which brings the image and text embeddings into the same space.

As input to these models, we either just provide the question or provide a combination of text and video transcript. Since our dataset contains a timestamp mentioned in every question, we provide transcript of the video corresponding to $\pm$ 1.5 minutes around the timestamp as additional context from the video. This partial transcript is chosen to avoid large input lengths. For VLMs, we also pass the video frame corresponding to the timestamp in the question. Rather than passing just one frame as input, we also experimented with multiple frames (equidistant sampled from video or sampled close to the timestamp) but did not see any improvements.

Note that our dataset differs from text-only QA data since each sample also has an accompanying video. Our dataset also differs from standard visual QA since we use a transcript (which spans temporally as long as the video). Hence, EDUVIDQA is a dataset for QA on educational videos.

[7] https://github.com/neulab/BARTScore

[8] https://github.com/google-research/bleurt

				Std. NLG Metrics			GPT-4 evaluations				
Modality	Model	Model Size	Transcript provided?	BERT-Score	BART-Score	BLEURT	Correctness (2)	Coherence (1)	Visual Grounding (1)	Relevance (1)	Total (5)
Text	Vicuna	7B	No	0.821	-4.817	0.387	0.388	0.290	0.268	0.262	1.208
	Vicuna	7B	Yes	0.837	-4.137	0.439	1.118	0.542	0.576	0.612	2.848
	LLaMA-3	8B	No	0.832	-4.512	0.445	1.312	0.746	0.464	0.702	3.224
	LLaMA-3	8B	Yes	0.843	-4.216	0.445	1.568	0.916	0.774	0.852	4.110
	GPT-4	1.76T	No	0.843	-4.283	0.421	1.354	0.940	0.394	0.704	3.392
	GPT-4	1.76T	Yes	0.851	-4.080	0.439	1.832	0.986	0.772	0.930	4.520
Image	LLaVA-v1.5	7B	No	0.833	-4.416	0.423	0.796	0.332	0.394	0.530	2.052
	LLaVA-v1.5	7B	Yes	0.837	-4.193	0.435	1.030	0.458	0.528	0.650	2.666
	GPT-4V	-	No	0.842	-4.412	0.442	1.644	0.920	0.646	0.840	4.050
	GPT-4V	-	Yes	0.847	-4.289	0.449	1.884	0.974	0.734	0.962	4.554
Human (without video)*		-	-	0.862	-4.188	0.436	1.250	0.970	0.540	0.740	3.500
Human (with video)*		-	-	0.870	-3.934	0.469	1.900	0.990	0.940	0.920	4.750

Table 1: Main results on EDUVIDQA test set (500 questions). *Human perf. is on 100 random samples from test.

All GPT-4 evaluations were done using Azure AI GPT-4 deployments with temperature = 1e-9 and seed = 42. Human annotations in this study were done by two of the authors; conflicts were resolved by discussions and led to improvements in guidelines and prompts over iterations.

5.2 Results

Main Results: Table 1 shows zero shot inference results for various models on EDUVIDQA test set. We observe that providing transcript in input always improves performance across all the metrics and models. When no transcript is provided LLaVA (which uses Vicuna as the LM) and GPT-4V perform better than Vicuna and GPT-4 respectively. However, the performance improvement is marginal when both the image and transcript are provided as context. As expected the humongous GPT-4 and GPT-4V models perform the best. Of course, human baseline results are better especially when the video is provided along with the question. We also finetuned Vicuna, LLaMA-3 and LLaVA using the EDUVIDQA train data, but finetuning did not help improve the results. Perhaps, this arises because when model tries to learn the reference answer on train data, it performs worse on test data since multiple possible answer styles can be correct, making the model's learning ineffective. This is also consistent with findings in Ahn et al. (2024).

GPT-4 vs GPT-4V: We also perform a manual comparison between the outputs from GPT-4 and GPT-4V for 100 samples. For both models, we passed transcript in the input. In 71/100 cases, both models were found to be equally good. In 26 cases, GPT-4V was preferred compared to GPT-4 while GPT-4 was preferred over GPT-4V in only 3 cases.

Qualitative Analysis: Appendix C (Tables 2, 3, 4 and 5) shows a few case studies of predictions from various models where the models perform accurately as well as examples where models generate erroneous predictions. Without the visual information, even LLMs and VLMs like GPT-4 and GPT-4V express their inability to answer by generating responses like "... The transcript does not provide the specific details regarding what was written at timestamp 5:11 in the video ..." or "... I would need to actually watch the video, which I cannot do ..." (Table 2). Further, Table 3 shows how GPT-4 attempts (but fails) to generate a reasonable answer because it cannot align transcript words with what is mentioned in the image. Further, without transcripts (Table 4), LLaMA-3 hallucinates about the ambiguous word "expanding" to (incorrectly) mean "zooming in" rather its actual meaning in the "expanding powers of a binomial" sense. Lastly, Table 5 and Fig. 2 show an example where even our best model (GPT-4V with transcript) also does not provide a good answer. It shows how GPT-4V fails at temporal understanding and geometric reasoning, and can be easily confused by an incorrectly framed question.

6 Conclusions

In this paper, we take initial steps towards multimodal QA for education domain. We curate a dataset, EDUVIDQA, by carefully filtering relevant videos and QA pairs from LearningQ. We observe that standard NLG metrics are insufficient for evaluation for this task. Hence, we propose a novel, but expensive, GPT-4 based metric based on evaluating aspects like correctness, coherence, visual grounding and relevance. Further, we benchmark multiple LLMs and VLMs, and find that GPT-4V provides best performance, and smaller models worsen in performance after training. Our work highlights important open questions in this area: how to train less expensive evaluation metrics for our task, and how to use existing training data for better performance of smaller models, which can provide lower latency and compute costs.

7 Limitations

We would like to extend this benchmarking to many other models like Gemini (Team et al., 2023) and GPT-4o. It will also be nice to distill such large models to small scale models for lower compute needs and lower latency.

We experimented with maths and science questions. It will be nice to extend this to other domains.

Lastly, we experimented with English QA pairs only. We would surely like to extend this to more languages.

8 Ethics Statement

All the models used in this work are publicly available on Huggingface and free for research.

We utilized publicly accessible LearningQ dataset from https://github.com/AngusGLChen/LearningQ. These resources were used as per their intended use policies.

Just like other generative models, our models can potentially generate biased, offensive or otherwise harmful content. Hence, care should be taken to apply appropriate filters when integrating with real world systems. That said, we did not observe such cases during our experimentation.

References

Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. 2024. Tuning large multimodal models for videos using reinforcement learning from ai feedback. In *ACL*.

AI@Meta. 2024. Llama 3 model card.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*, pages 2425–2433.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: a large-scale dataset for educational question generation. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Daisy Zhe Wang, and Doo Soon Kim. 2020. Tutorialvqa: Question answering dataset for tutorial videos. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5450–5455.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*, pages 326–335.

Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. Knowit vqa: Answering knowledge-based questions about videos. In *AAAI*, volume 34, pages 10826–10834.

Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. 2015. Visual turing test for computer vision systems. *Proc. of the National Academy of Sciences*, 112(12):3618–3623.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, pages 804–813.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*, pages 4999–5007.

Ryan Lau. 2023. Bloombert: A task complexity classifier. https://github.com/RyanLauQF/BloomBERT.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

OpenAI et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *ACL*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *CVPR*, pages 8317–8326.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *CVPR*, pages 8802–8812.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *CVPR*, pages 21–29.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. Leveraging video descriptions to learn video question answering. In *AAAI*, volume 31.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A GPT prompts

A.1 Filtering out irrelevant questions

```
system_prompt = "You are an expert in
    finding the relevance of a question
    and its corresponding answer with
    respect to a particular domain. Your
     task is to find the relevance of a
    question answer pair with respect to
     the domain mentioned. Here's how
    you can accomplish the task. \n
    ------------\n
    - See whether the question is
        relevant to the domain in
        consideration.\n
    - See whether the answer tries to
        provide a solution to the
        question.\n
    - Evaluate the question answer pair
        with respect to the domain
        mentioned.\n
    - Rate the relevance of the
        following question-answer pairs
        on a scale of 1-10, with 1 being
         least relevant and 10 being
        most relevant.\n
    Please generate the response in the
        form of a Python integer\n
    DO NOT PROVIDE ANY OTHER TEXT OR
        EXPLANATION"

user_prompt = "Please evaluate the
    following domain-based question-
    answer pair:\n
    domain:{domain}\n
    question:{question}\n
    answer:{answer}"
```

Listing 1: Prompt used to filter irrelevant QA pairs

A.2 Prompt for Question Answering

```
system_prompt = "## You are an Question
    Answering AI assistant. You need to
    answer the question.\n
    ## RULES\n
    1. Answer MUST be simple and precise
        .\n
    2. Answer can consists of multiple
        steps if that would help the
        human understand it better.\n
    3. Do not assume something you are
        not sure about.\n\n"

user_prompt = "Question: {question}\n\
    nAnswer: "
```

Listing 2: Prompt used for Question answering with question alone

```
system_prompt = "## You are an Question
    Answering AI assistant.\nYou are
    provided with the `Transcript' of a
    video.\n You need to answer the
    question as if you have seen the
    video.\n
    ## RULES\n
```

```
    1. Answer MUST be simple and precise
        .\n
    2. Answer can consists of multiple
        steps if that would help the
        human understand it better.\n
    3. Do not assume something you are
        not sure about.\n
    4. You may make use of the context
        provided in the Transcript to
        generate the answer.\n\n"

user_prompt = "Transcript: {transcript}\
    n\nQuestion: {question}\n\nAnswer: "
```

Listing 3: Prompt used for Question answering with question and transcript

A.3 Prompts for Evaluation

We believe that the metrics of correctness, coherence, visual grounding and relevance are extremely crucial for the educational domain. (a) Correctness: Accuracy is crucial as it ensures students receive the right information, preventing misconceptions and supporting effective learning. (b) Coherence: Clear and logical answers help students understand and connect ideas, especially for comprehending complex topics.(c) Visual Grounding: Linking information to visual elements in materials aids in ease of understanding the answers, especially with reference to the video recently watched by the student. (d) Relevance: It ensures that the answer does not deviate from the student's question which helps in efficient learning by avoiding any confusions at the learning stage. We design prompts having explicit scoring guidelines for each of these factors, for GPT which gives high correlations with human scores.

A.3.1 Correctness

```
system_prompt = "You are an AI
    Evaluation chatbot helpful in
    evaluating the correctness of
    generative outputs for visual
    content-based question-answer pairs
    .\n
    Your task is to evaluate the
        predicted answer and determine
        if it answers the question
        correctly. Here's how you can
        accomplish the task:\n
    ------\n
    ##INSTRUCTIONS: \n
    Provide your evaluation only as a
        score where the score is an
        integer value.\n
    Please generate the response in the
        form of a Python dictionary
        string with keys `score', where
        its value is the score in
        INTEGER, not STRING.\n
    For example, your response should
        look like this: {`score':
        INTEGER}.\n
    Possible values of score are :
        [0,1,2] \n
    0 when predicted answer is incorrect
         or not answered. \n
    1 when predicted answer is partially
         correct or makes assumptions. \
        n
    2 when predicted answer is
        completely correct with concise,
        useful explanations. \n"

user_prompt = "Please evaluate the
    following video-based question-
    answer pair:\n
    Question: {question}\n
    Predicted Answer: {pred}\n\n"
```

Listing 4: Prompt used for evaluating correctness

A.3.2 Coherence

```
system_prompt = "You are a strict AI
    Evaluation chatbot helpful in
    evaluating the coherence of
    generative outputs for video-based
    question-answer pairs.\n
    Your task is to strictly evaluate
        the predicted answer and
        determine if it is coherent and
        easy to understand. Here's how
        you can accomplish the task:\n
    ------\n
    ##INSTRUCTIONS:\n
    Provide your evaluation only as a
        score where the score is an
        integer value.\n
    Please generate the response in the
        form of a Python dictionary
        string with keys `score', where
        its value is the score in
        INTEGER, not STRING.\n
    For example, your response should
        look like this: {`score':
        INTEGER}.\n
    Possible values of score are : [0,1]
         \n
    0 when predicted answer is long and
        not easy to follow.\n
    1 when predicted answer has concise
        explanations and easy to follow
        .\n"

user_prompt = "Please evaluate the
    following video-based question-
    answer pair:\n
    Question: {question}\n
    Predicted Answer: {pred}\n\n"
```

Listing 5: Prompt used for evaluating coherence

A.3.3 Visual grounding

```
system_prompt = "You are a strict AI
    Evaluation chatbot helpful in
    evaluating the contextual
```

```
    understanding of generative outputs
    for video-based question-answer
    pairs.\n
    Your task is to strictly evaluate
        the predicted answer and
        determine if it is grounded in
        the visual content provided as
        image/video. Here's how you can
        accomplish the task:\n
    ------\n
    ##INSTRUCTIONS:\n
    Provide your evaluation only as a
        score where the score is an
        integer value.\n
    Please generate the response in the
        form of a Python dictionary
        string with keys `score', where
        its value is the score in
        INTEGER, not STRING.\n
    For example, your response should
        look like this: {`score':
        INTEGER}.\n
    Possible values of score are : [0,1]
         \n
    0 when predicted answer has no
        reference to the image/video. \n
    1 when predicted answer has
        reference to the image/video. \n
        "

user_prompt = "Please evaluate the
    following video-based question-
    answer pair:\n
    Question: {question}\n
    Predicted Answer: {pred}\n\n"
```

Listing 6: Prompt used for evaluating visual understanding

A.3.4 Relevance

```
system_prompt = "You are a strict AI
    Evaluation chatbot helpful in
    evaluating the completeness and
    relevance of generative outputs for
    video-based question-answer pairs.\n
    Your task is to strictly evaluate
        the predicted answer and
        determine if it is complete and
        relevant to the question. Here's
         how you can accomplish the task
        :\n
    ------\n
    ##INSTRUCTIONS:\n
    Provide your evaluation only as a
        score where the score is an
        integer value.\n
    Please generate the response in the
        form of a Python dictionary
        string with keys `score', where
        its value is the score in
        INTEGER, not STRING.\n
    For example, your response should
        look like this: {`score':
        INTEGER}.\n
    Possible values of score are : [0,1]
         \n
    0 when predicted answer is
        irrelevant or somewhat deviates
        from the question. \n
    1 when predicted answer is
        completely relevant and precise.
         \n"

user_prompt = "Please evaluate the
    following video-based question-
    answer pair:\n
    Question: {question}\n
    Predicted Answer: {pred}\n\n"
```

Listing 7: Prompt used for evaluating relevance

B Dataset Filtering Decisions

We filtered out questions from our overall EDU-VIDQA dataset, hence creating the challenging subset with most questions requiring visual context to answer the question.

Why Maths and Science domains only? We decided to choose these domains to have more complex, reasoning-based questions which cannot be answered with general or common sense knowledge. Further, ∼75% of our dataset consists of questions from Maths and Science domain and even after retaining only ∼9% of the total QA pairs, we are still able to retain ∼43% of the videos.

Why timestamps? Upon observation, we find that QA pairs with timestamp in their question are more likely to need a proper contextual understanding of the video. Using GPT-4 for filtering out such questions would have further added to our costs. Hence we use this simplified heuristics-based approach to ensure that most questions require video context to be answered accurately.

Examples of removed questions. Examples of some questions which have been removed are as follows. It is easy to see that these questions can be easily answered by the world knowledge stored in the LLM's parametric memory and does not need video context.

```
Q: "Why does the pond freeze faster than
     the sea water?"
```

Listing 8: General question, does not require video context.

```
Q: "Wasn't Russia called the Soviet
    Union?"
```

Listing 9: History question: Remembering-based question and does not require context.

```
Q: "If a continuous increase in
    Government spending increases the
    GDP and reduces Inflation, how does
    subsequent increase in spending
    increase the rate of inflation?"
```

Listing 10: Question is linked to the video but does not require video context. Requires reasoning but any text-only LLM can handle.

C Qualitative Analysis of Model Outputs

In this section, we show a few case studies of predictions from various models where the models perform accurately as well as examples where models generate erroneous predictions. Table 2 shows an example of a good prediction from GPT-4V with transcript. Table 3 shows an example where GPT-4V prediction is better than prediction from GPT-4. Table 4 shows an example where LLaMA-3 with transcript is better than LLaMA-3 without transcript. Lastly, Table 5 shows an example where even our best model (GPT-4V with transcript) also does not provide a good answer.

D Error Analysis

Our best model GPT-4V with transcript also makes errors. This section looks at error analysis for one of its failure cases mentioned in Table 5 as an example, suggesting future research directions.
Please see Figure 2 for a detailed view of input and output.
Error 1: Temporal Understanding problem
The model could not track in what order values in the image were filled. Although transcript would have helped but it paid more attention to what is already present in the image. The 2 equations were finally solved at a later stage (around 30 seconds ahead)
Error 2: Cognitive problem and confusion
The model does not have much sense of geometry. It could not figure out by looking at image what values are to be equated. It gets confused with adding transcript, since there are so many points where we have x's and y's.
Error 3: Factual incorrectness
It could not catch an error in the question and gets confused and even says "two adjacent sides of a rectangle, which must be equal" and equates unequal sides: "8x + 4y and 8x + 9y, should be equal" which is incorrect in the context of the video and the transcript. A self check on factual correctness could have been a possible solution.

Video	https://www.youtube.com/watch?v=rYG1D5lUE4I
Question	I think Sal make a mistake on (vid @ 5:11) when he write the greater than sign! it should be Less than
Ground truth Answer	No, Sal is correct. If he had: i sqrt(x) where X<0, then X is negative. Backup thru Sal steps. If X is negative * (-1) = +X. And he would have started with sqrt(x), not sqrt(-x). He is also trying to highlight that if you had something like: sqrt(12), you would not make this into i sqrt(-12). The imaginary number is not needed if the radical contains a positive number to start with.
Prediction from GPT-4 with transcript	The transcript does not provide the specific details regarding what was written at timestamp 5:11 in the video. Therefore, it's not possible to confirm if Sal made an error with the greater than sign or if it should have been a less than sign.
Prediction from GPT-4V without transcript	To confirm if Sal made a mistake with a greater than sign at 5:11 in the video, I would need to actually watch the video, which I cannot do. However, I can see the content of the provided image and provide explanations based on it.\n\n The image shows handwritten mathematics notes that include the definition of the imaginary unit i, properties of square roots, and an example with the square root of a negative number. If Sal's discussion involves these topics and there's a specific point related to inequalities or comparing sizes, I can help clarify that based on the mathematical concepts shown. If you provide more context or describe the mistake, I can help correct the mathematical concept.
Prediction from GPT-4V with transcript	Sal is correct in the video; he mentions that the rule $\sqrt{-x} = i\sqrt{x}$ applies when $x \geq 0$. When x is greater than or equal to 0, $-x$ is a non-positive number (negative or zero), which is the context in which you would be looking for a complex square root. If x were less than 0, then $-x$ would be positive, and the regular principal square root could be used instead of the complex one. Thus, the greater than sign is used appropriately here.

Table 2: An example of a good prediction from GPT-4V with transcript.

Video	https://www.youtube.com/watch?v=nGCW5teACC0
Question	The equation is -4x+7. Shortly after the 4:00 mark, Sal replaces the x with -1 and then says, "4 times -1 = -4". Shouldn't it be -4 * -1?
Ground truth Answer	He misspoke and says 4*-1=4, but what he really meant is -4*-1=4 and he completes the equation as if he had said that correctly. It does not change the problem because he just misspoke and didn t write the incorrect statement down
Prediction from GPT-4	No, Sal is correct. In the equation -4x+7, the "-4x" represents "-4 times x". So if x is -1, then it is "-4 times -1", which does equal 4. But because of the negative sign in front of the 4, the result is turned into -4. Hence, '-4 times -1' equals '-4' in this context.
Prediction from GPT-4V	Yes, it should be -4 multiplied by -1. When plugging -1 into the equation in place of x, the correct operation is to multiply -4 by -1, which results in +4.

Table 3: Visual content helps: GPT-4V prediction is better than prediction from GPT-4.

Video	https://www.youtube.com/watch?v=iPwrDWQ7hPc
Question	At 8:30, why does Sal keep expanding everything out? I do not understand it.
Ground truth Answer	he is using this as a complete example to show how it works. He is also using the sigma, which is a sum of all integers from the number on the bottom to n.
Prediction from LLaMA-3 without transcript	I'm happy to help you understand!\n\nIt seems that Sal is expanding everything out because he is trying to show the proportions of the shapes and figures in the diagram. By expanding each part, he is making it easier to see the relationships between different elements and understand the overall structure.\n\nThink of it like zooming in on a small part of a big picture. By making each piece larger, Sal is allowing us to see the details more clearly and understand how they fit together.\n\nDoes that make sense?
Prediction from LLaMA-3 with transcript	Sal keeps expanding everything out because he is applying the Binomial Theorem to the expression $(a + b)^4$. The Binomial Theorem is a formula for expanding powers of a binomial (an expression with two terms) into a sum of terms. In this case, Sal is expanding $(a + b)^4$ into a sum of terms using the Binomial Theorem.\n\nHe is doing this to show that the expression $(a + b)^4$ can be simplified into a sum of terms, each of which is a product of powers of a and b. This can be useful for simplifying expressions and solving equations.\n\nIt may seem like a lot of work to expand everything out, but it is a useful technique for solving problems in algebra and other areas of mathematics.

Table 4: Transcript helps: LLaMA-3 without transcript vs LLaMA-3 with transcript.

Video	https://www.youtube.com/watch?v=1uWZNW5PF-s
Question	At 6:13, Sal got 2 different answers for 2 sides of a square. How did he get the 2 answers?
Ground truth Answer	To get the dimension on the left side of the large rectangle, he added the lengths of a side of each square on the left. To get the dimension on the right side of the large rectangle, he added the lengths of a side of each square on the right. Since we know opposite sides of a rectangle have equal length, we can set up an equation (13x+7y=8x+9y) to solve for the ratio of x to y (x=2/5*y).
Prediction from GPT-4V with transcript	Sal calculated the dimensions of two adjacent sides of a rectangle, which must be equal because they represent the same length. For one side, he added the lengths of segments labeled as 5x + 3y, 2x + y, and x to get a total length of 8x + 4y. For the adjacent side, he combined the lengths of segments labeled as 4x + 4y and 4x + 5y to get a total length of 8x + 9y. These two different expressions, 8x + 4y and 8x + 9y, should be equal since they represent the same physical length of the rectangle's side, leading to an equation that can be simplified to find a relationship between x and y.

Table 5: Best model (GPT-4V with transcript) also makes a mistake here.

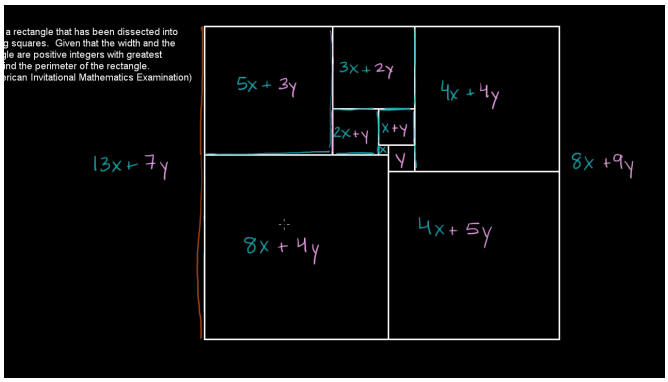
Timestamp image (6:13)	
	
Transcript*	
....5x plus 3y is going to be that entire length right over there. ..... This length is 3x plus 2y. This is x plus y. And this is y. So if you add 3x plus 2y plus x plus y plus y, you get 4x plus-- what is that-- 4y, right? 2y, 3y, 4y. And then we can express this character's dimensions in terms of x and y because this is going to be 5x plus 3y. Then you're going to have 2x plus y. And then you're going to have x. So you add the x's together. 5x plus 2x is 7x, plus x is 8x. And then you add the y's together, 3y plus y, and then you don't have a y there. So that's going to be plus 4y. And then finally, we have this square right over here. Its dimensions are going to be the y.....5x plus 3y plus 8x plus 4y. So 5 plus 8 is 13. So it's 13x plus 3 plus 4 is 7y. So that's its height. But we can also think about its height by going on the other side of it. And maybe this will give us some useful constraints because this is going to have to be the same length as this over here. And so if we add 4x plus 4x, we get 8x. So these are going to have to be equal to each other, so that's an interesting constraint. So we have 13x plus 7y is going to have to equal 8x plus 9y.....	
Question	
At 6:13, Sal got 2 different answers for 2 sides of a square. How did he get the 2 answers?	
Human answer	**GPT4V-response**
To get the dimension on the left side of the large rectangle, he added the lengths of a side of each square on the left. To get the dimension on the right side of the large rectangle, he added the lengths of a side of each square on the right. Since we know opposite sides of a rectangle have equal length, we can set up an equation (13x+7y=8x+9y) to solve for the ratio of x to y (x=2/5*y)	Sal calculated the dimensions of two adjacent sides of a rectangle, which must be equal because they represent the same length. For one side, he added the lengths of segments labeled as 5x + 3y, 2x + y, and x to get a total length of 8x + 4y. For the adjacent side, he combined the lengths of segments labeled as 4x + 4y and 4x + 5y to get a total length of 8x + 9y. These two different expressions, 8x + 4y and 8x + 9y, should be equal since they represent the same physical length of the rectangle's side, leading to an equation that can be simplified to find a relationship between x and y.

Figure 2: Example where GPT-4V with transcript does not perform well. The image corresponds to the frame at 6:13 in the video https://www.youtube.com/watch?v=1uWZNW5PF-s. We omitted some lines from the transcript for sake of clarity of presentation.