AVQA-CoT: When CoT Meets Question Answering in Audio-Visual Scenarios

Guangyao Li, Henghui Du, Di Hu*
Gaoling School of Artificial Intelligence, Renmin University of China
{guangyaoli,cserdu,dihu}@ruc.edu.cn

Abstract

*The Audio Visual Question Answering task aims to answer questions derived from untrimmed audible videos. Existing research has focused on exploring multimodal representation learning for answering questions, there is often a lack of transparency in the perception and reasoning processes. It is crucial to enable machines to understand dynamic audio-visual scenes like human thinking and accurately infer answers to input questions. In this paper, we explores the reasoning process for answering complex questions using an **AVQA Chain-of-Thought (AVQA-CoT)**. Specifically, we decompose complex questions into multiple simpler sub-questions. Then employing the Large Language Model (LLM) to select relevant sub-questions through instructional learning, leveraging its reasoning capabilities to explore answering complex questions in a CoT manner. Extensive experiments on multiple benchmarks show that our framework excels addressing complex questions.*

1. Introduction

The real world revolves around sound and visual information, and their combination enhances our ability to perceive the world [12]. The past few years have observed a growing interest in tackling more pervasive and intricate audio-visual scenarios, such as localization [9], event localization [10], video parsing [11], dialog [8], question answering [4], *etc*., towards audio-visual scene understanding. Particularly, the Audio-Visual Question Answering (AVQA) task, requires comprehensive multimodal understanding and spatio-temporal reasoning over complex audio-visual scenes.

Previous AVQA research [4, 7] has primarily concentrated on creating joint feature representations of audio-visual inputs and questions to tackle complex inquiries but has often disregarded the reasoning process involved in answering. Several studies [2, 3] in Visual Question Answering (VQA) have addressed this issue by breaking down complex questions into simpler sub-questions. However, it is important to note that unlike in straightforward VQA scenarios, the distinctiveness of audio-visual scenes results in a dearth of specialized models to handle sub-questions. This limitation hinders the direct application of previous methods to reason complex questions in audio-visual scenarios. Hence, exploring the reasoning process behind complex questions in AVQA tasks is highly valuable.

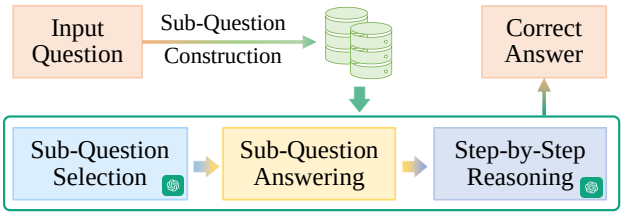

Figure 1. For input complex questions, allowing the LLM to select relevant sub-questions related to answering and utilizing perception models to answer them, gradually reasoning out the answers.

To achieve this goal, we propose an **AVQA Chain-of-Thought (AVQA-CoT)** to explore reasoning process of answering questions. Consider that answering questions process contributes to exploring the reasoning mechanism in audio-visual scenes. As shown in Fig. 1, we propose decompose-then-aggregate strategy in answering complex questions, utilizing the Large Language Model (LLM) to select relevant sub-questions through instructional learning, and leveraging its reasoning capabilities to explore answering complex questions in a CoT manner. Instead of simply decomposing complex questions, the designed sub-question is intended to be fine-grained perceptual units of the video, and utilizes the instruction learning to select the minimum set of sub-questions necessary for making precise inferences. For answering sub-questions, we opt for the latest AVQA perception model to minimize the impact of out-of-distribution (OOD) factors and enhance the accuracy of sub-question responses. Experiments on multiple benchmark has shown the significant potential of our method in audio-visual scenes reasoning.

2. Related Works

2.1. Audio Visual Question Answering

Audio-visual question answering, which requires comprehensive spatial-temporal reasoning of the audio-visual scenarios. The earliest works [4] emphasize the understanding

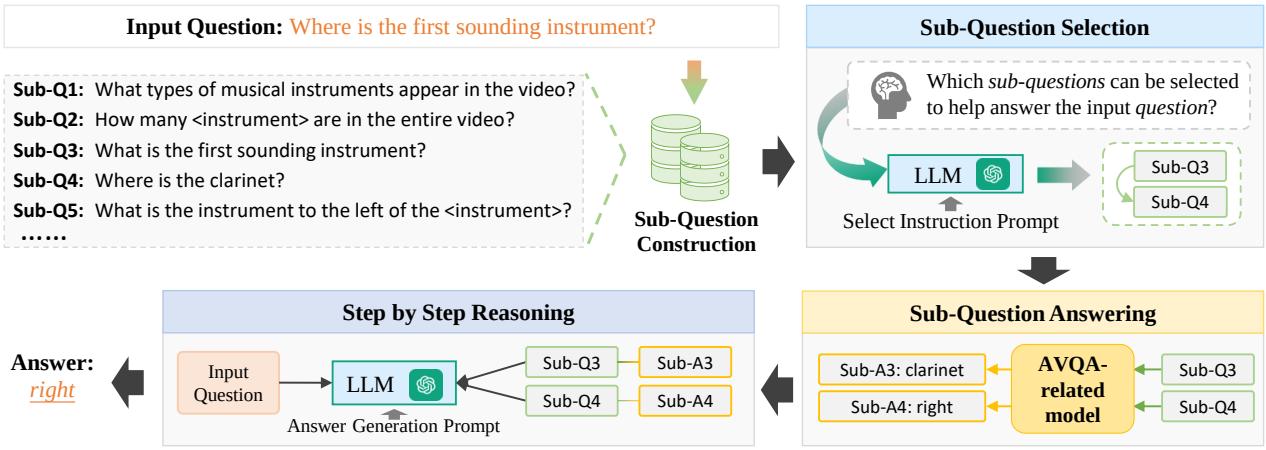

Figure 2. Framework of AVQA-CoT. We employ instruction learning, enabling LLM to select key sub-questions and apply its reasoning abilities in a CoT manner to explore the process of answering complex questions. It's achieved through the following steps: Sub-Question Construction, Selection, Answering and Step-by-step Reasoning.

of the whole video sequence and return a simple word that is relatively correct to the question. In recent studies, researchers [1, 5] consider the importance of the given question, which guides the feature extraction of both audio and visual signals. Unlike above existing works, our emphasis lies in investigating the reasoning process behind complex questions in AVQA tasks.

2.2. CoT Reasoning with LLM

As a prompt techniques, CoT has been widely used to elicit the multi-step reasoning abilities of LLM. Employing CoT to encourage LLM to generate intermediate reasoning chains proves beneficial in addressing complex questions. This inspired us to apply CoT in AVQA tasks to answer complex questions in audio-visual scenarios. Similarly, some VQA methods [2, 3] use CoT to answer complex questions by leveraging LLM's abilities. Chen *et al*. [2] proposed an iterative form of CoT, constantly updating the rationales and results until the correct answer is obtained. Hu *et al*. [3] leveraged external tools to solve a series of intermediate problems and obtains answers to complex questions. In this work, considering there is few available external foundation tools in effectively modeling audio-visual components as well as their interactions, we conducted another kind of initial exploration of the use of CoT in AVQA task. We propose to decompose complex questions into simpler sub-questions, utilizing AVQA-realted model to answer these sub-questions, and finally the LLM aggregates information from sub-questions to derive the final answer.

3. AVQA-CoT Method

Analyzing the question-answering process enhances our comprehension of the reasoning mechanisms within complex audio-visual scenarios. Previous works tend to on exploring joint multimodal representations to answer complex questions, often neglecting the underlying reasoning in the answering process. Given that LLM have shown impressive performance on complex reasoning by leveraging CoT prompting to generate intermediate reasoning chains as the rationale to infer the answer. We introduce an **AVQA Chian-of-Thought** (**AVQA-CoT**) that decompose complex questions into several simpler sub-questions. Then employing the LLM to select relevant sub-questions through instructional learning, leveraging its reasoning capabilities to explore answering complex questions in a CoT manner. As shown in Fig. 2, the proposed AVQA-CoT comprises four key steps: Sub-Question Construction, Sub-Question Selection, Sub-Question Answering and Step-by-step Reasoning.

Sub-Question Construction. We construct a sub-question set, which aims to build the basic ability to meet the requirements of answering complex questions, as well as encompassing rich video content. Subsequently, we aggregate these sub-questions into a collection, denoted as $\texttt{SubQ} = \{\texttt{Subq}_\gamma\}_{\gamma=1}^{\Gamma}$, where Γ is sub-question number.

Sub-Question Selection. For a given input complex question, humans can easily choose which sub-questions from SubQ to assist in answering the question. Hence, to empower machines with the human-like capability to select relevant sub-questions to answer question, the *Select Instruction Prompt*, denoted as SIPrompt, is designed to facilitate LLM in choosing ordered sub-questions through instructional learning. The selected sub-questions from SubQ can be denoted as $\texttt{SubQ}' = \{\texttt{Subq}_{\gamma'}\}_{\gamma'=1}^{\Gamma'}$, where $\Gamma' \leqslant \Gamma$.

Sub-Question Answering. Utilizing the model, which pretrained by the AVQA-related model, to answer each selected sub-question in SubQ′. Then the sub-question and its sub-answer pairs can be denoted as SubQA′.

Step-by-step Reasoning. To enable machines to answer complex questions posed in videos by aggregating simple sub-question-answering pair, the *Answer Generation Prompt*, referred to as AGPrompt is designed to facilitate

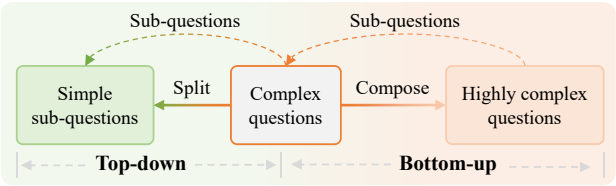

Figure 3. Top-down AVQA-CoT and Bottom-up AVQA-CoT.

LLM to progressively infer the final correct answers in a CoT manner for the given SubQA′ and input question.

Overall, it is expected that exploring the reasoning process of answering complex questions through the four steps of AVQA-CoT, providing a promising research paradigm for audio visual scene understanding and reasoning.

4. Experiments

4.1. Datasets

MUSIC-AVQA [4], it contains 9,288 videos covering 22 different instruments, with 45,867 question-answering pairs. The questions are designed under multi-modal scenes containing 33 question templates covering 9 types. The MUSIC-AVQA dataset is well suited for studying temporal-spatial reasoning for dynamic audio-visual scenes.

MUSIC-AVQA-Complex, is designed to exploration reasoning process of answering highly complex question in dynamic audio visual scenarios. It contains 100 videos, which from MUSIC-AVQA dataset, along with 500 complex QA pairs. For training and test, we randomly split the dataset into both sets with 350, and 150 QA pairs, respectively.

MUSIC-AVQA-Mini, given that the 33 question templates in MUSIC-AVQA cover both complex and simple questions, meaning that some complex questions can be answered by aggregating part of simple questions wintin MUSIC-AVQA. For these kinds of complex questions, we randomly selected 150 ones in the test set to explore AVQA-CoT's reasoning ability to answer complex questions.

4.2. Implementation Details

For audio and visual processing, we adopt a consistent approach with AVQA-realted models. The perception models involved in this paper are AVSD [8], PSTP-Net [5], and TSPM [6]. In CoT reasoning, we utilize the GPT-3.5 API to aggregate selected sub-question-answer pairs for answering complex questions.

4.3. AVQA-CoT Exploration

To explore efficacy of AVQA-CoT in answering complex questions, we designed two types of experiments: Bottom-up and Top-down AVQA-CoT, as shown in Fig. 3.

Bottom-up AVQA-CoT. Regarding the given questions in *MUSIC-AVQA-Complex*, QComplex, we consider the test set of MUSIC-AVQA as its sub-question set, denoted as

Method	MUSIC-AVQA Complex	MUSIC-AVQA Mini
AVSD [8]	60.17	42.31
w/. AVQA-CoT	66.45 (**+6.28**)	59.41 (**+17.28**)
PSTP-Net [5]	65.04	48.46
w/. AVQA-CoT	65.16 (**+0.12**)	50.00 (**+1.54**)
TSPM [6]	70.20	53.85
w/. AVQA-CoT	74.19 (**+3.99**)	65.33 (**+11.48**)

Table 1. The AVQA-CoT results on constructed datasets.

Method	MUSIC-AVQA *All*	MUSIC-AVQA *Localization + Temporal*
TSPM [6]	76.79	70.84
+ AVQA-CoT	74.55	65.61
+ AVQA-CoT *w/. Ft*	——	69.92

Table 2. Auxiliary AVQA-CoT results on the MUSIC-AVQA *All* and *Temporal + Localization* test set, and **Ft** means *fine-tuning*.

SubQComplex. For each question within QComplex, we employ the four-step incremental reasoning process from AVQA-CoT to provide answers. It's noteworthy that the forms of these sub-questions have appeared during the training process of AVQA-related model, thereby mitigating the out-of-distribution (OOD) issue, meaning that ensure that training and test data belong to the same distribution space. As shown in Tab. 1, it is observed that the performance of various methods improves when combined with AVQA-CoT, indicating the generalization of AVQA-CoT. And the proposed AVQA-CoT demonstrates satisfactory performance on MUSIC-AVQA-Complex, with a 3.99% improvement compared to TSPM. This indicates that Bottom-up AVQA-CoT can effectively answer highly complex questions and provide correct answers.

Top-down AVQA-CoT. For the questions in the *MUSIC-AVQA-Mini* dataset, denoted as QMini, we decompose them into sub-questions by considering the simpler questions from the MUSIC-AVQA test set, labeled as SubQMini. Both QMini and SubQMini originate from the same dataset, partially eliminating OOD concerns. Subsequently, we employ the three-step processes of AVQA-CoT to progressively reason through and answer the questions in QMini. Tab. 1 illustrates the outstanding performance of Top-Down AVQA-CoT on QMini, exhibiting a notable improvement of 11.48% compared to TSPM. The result further prove the proposed AVQA-CoT's effective reasoning and answering capabilities for complex questions.

To mitigate the impact of OOD issues in AVQA-CoT for improve answer accuracy, beyond exploring the MUSIC-AVQA-Complex and MUSIC-AVQA-Mini datasets, we also attempt alternative ways such as fine-tuning for sub-questions. Specifically, we initially select 200 sub-questions from SubQMA and annotate them. To enable the perception model to comprehend these sub-questions, we fine-tune the TSPM-trained model, denoted as *w/. Ft*. Con-

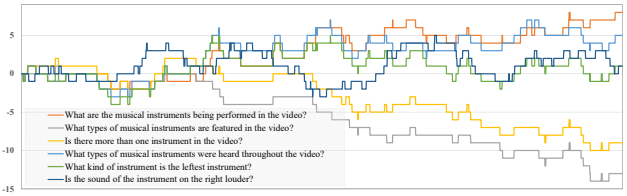

Figure 4. The evolution of different sub-question weights.

sidering the computational cost of using *GPT-3.5 API* for inference, we employ the AVQA-CoT maner to test on complex questions in *Localization* and *Temporal* types, where sub-questions are answered using *w/. Ft*. The results in Tab. 2 demonstrate that fine-tuning can, to a certain extent, address OOD issues, leading to a significant improvement in model performance (65.61% *vs.* 69.92%). Overall, these explorations have showcased the immense potential of AVQA-CoT in reasoning and answering complex questions in dynamic audio-visual scenarios.

4.4. Sub-question Selection and its generalization

The challenge in AVQA lies in the selection and answering of sub-questions, which is hindered by the absence of expert models. To solve this, we assess the efficacy of each sub-question during training, gradually augmenting the weights of valuable sub-questions and diminishing those less useful. Fig 4 shows the evolving weights of several sub-questions across training iterations (*horizontal axis*). To ensure the sub-questions possess generalizability, we construct them as fine-grained perception units for videos. Specifically, the basic video elements (*e.g.*, object categories list) are obtained through sub-questions, and then their spatial relationships are derived, progressively answering complex questions in a CoT manner.

4.5. Visualization Results

As shown in Fig. 5, the visualization results illustrate the proposed AVQA-CoT can adeptly engage in reasoning for answring complex questions in a CoT manner.

5. Conclusion

In this work, we propose an AVQA-CoT to explore the reasoning process for complex questions and provide answers. Leveraging the capabilities of the LLM, AVQA-CoT employs a CoT manner to investigate the reasoning process for answering complex questions, offering a promising paradigm research for audio-visual scene reasoning. Extensive experiments demonstrate that the proposed method effectively illustrates the reasoning process involved in answering questions. We believe that our work will serve as inspiration for researchers in AVQA reasoning.

Acknowledgement. This research was supported by National Natural Science Foundation of China (NO.62106272).

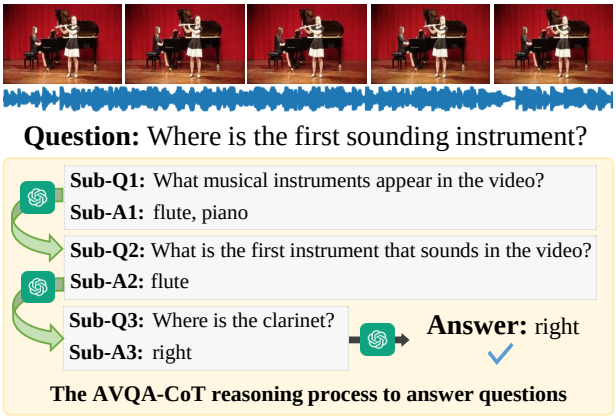

Figure 5. Visualized AVQA-CoT results. For a given question, it shows the AVQA-CoT slecting related sub-questions, answering them using TSPM, and then providing an overall response through the reasoning capabilities of LLM. Then AVQA-CoT accurately answered this complex question, thereby demonstrating the effectiveness of the AVQA-CoT.

References

[1] Zailong Chen, Lei Wang, Peng Wang, and Peng Gao. Question-aware global-local video understanding network for audio-visual question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[2] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *arXiv preprint arXiv:2301.05226*, 2023. 1, 2

[3] Ziniu Hu, Ahmet Iscen, Chen Sun, Kai-Wei Chang, Yizhou Sun, David A Ross, Cordelia Schmid, and Alireza Fathi. Avis: Autonomous visual information seeking with large language models. *arXiv preprint arXiv:2306.08129*, 2023. 1, 2

[4] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, pages 19108–19118, 2022. 1, 3

[5] Guangyao Li, Wenxuan Hou, and Di Hu. Progressive spatio-temporal perception for audio-visual question answering. page 7808–7816, 2023. 2, 3

[6] Guangyao Li, Henghui Du, and Di Hu. Boosting audio visual question answering via key semantic-aware cues. *arXiv preprint*, 2024. 3

[7] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *CVPR*, pages 2299–2309, 2023. 1

[8] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *CVPR*, pages 12548–12558, 2019. 1, 3

[9] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, pages 4358–4366, 2018. 1

[10] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018. 1

[11] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, pages 436–454. Springer, 2020. 1

[12] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022. 1