Learning to Answer Questions in Dynamic Audio-Visual Scenarios

Guangyao Li\textsuperscript{1,†}, Yake Wei\textsuperscript{1,†}, Yapeng Tian\textsuperscript{3,†}, Chenliang Xu\textsuperscript{3}, Ji-Rong Wen\textsuperscript{1}, Di Hu\textsuperscript{1,2,*}

\textsuperscript{1}Gaoling School of Artificial Intelligence, Renmin University of China, Beijing
\textsuperscript{2}Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing
\textsuperscript{3}Department of Computer Science, University of Rochester, Rochester
\textsuperscript{1}\{guangyaoli, yakewei, jrwen, dihu\}@ruc.edu.cn, \textsuperscript{3}\{yapengtian, chenliang.xu\}@rochester.edu

Abstract

In this paper, we focus on the Audio-Visual Question Answering (AVQA) task, which aims to answer questions regarding different visual objects, sounds, and their associations in videos. The problem requires comprehensive multimodal understanding and spatio-temporal reasoning over audio-visual scenes. To benchmark this task and facilitate our study, we introduce a large-scale MUSIC-AVQA dataset, which contains more than 45K question-answer pairs covering different question templates spanning over different modalities and question types. We develop several baselines and introduce a spatio-temporal grounded audio-visual network for the AVQA problem. Our results demonstrate that AVQA benefits from multisensory perception and our model outperforms recent A-, V-, and AVQA approaches. We believe that our built dataset has the potential to serve as a testbed for evaluating and promoting progress in audio-visual scene understanding and spatio-temporal reasoning. Code and dataset: \url{http://gewulab.github.io/MUSIC-AVQA/}

1. Introduction

We are surrounded by audio and visual messages in daily life, and both modalities jointly improve our ability in scene perception and understanding [19]. For instance, imagine that we are in a concert, watching the performance and listening to the music at the same time contribute to better enjoyment of the show. Inspired by this, how to make machines integrate multimodal information, especially the natural modality such as the audio and visual ones, to achieve considerable scene perception and understanding ability as humans is an interesting and valuable topic.

In recent years, we have seen significant progress in sounding object perception [6, 22, 37, 52], audio scene analysis [7, 10, 13, 20, 21, 51, 59], audio-visual scene parsing [42, 47], and content description [24, 40, 50] towards audio-visual scene understanding. Although these methods associate objects or sound events across audio and visual views, most of them remain limited ability for cross-modal reasoning, under complex audio-visual scenarios. In contrast, humans are capable of performing multi-step spatial and temporal reasoning over multimodal contexts to solve complex tasks, such as answering an audio-visual question, but it is quite challenging for machines. Existing methods such as Visual Question Answering (VQA) [3] and Audio Question Answering (AQA) [9] only focus on single modality, which cannot reason well in a more natural scenario with both audio and visual modalities. For instance, as shown in Fig. 1, when answering the audio-visual question “Which clarinet makes the sound first?” for this instrumental ensemble, it requires to locate sounding objects “clarinet” in the audio-visual scenario and focus on the “first” sounding “clarinet” in the timeline. To answer the question cor-
rectly, both effective audio-visual scene understanding and spatio-temporal reasoning are essentially desired.

In this work, we focus on the Audio-Visual Question Answering (AVQA) task, which aims to answer questions regarding visual objects, sounds and their association. To this end, a computational model is essentially required to equip with effective multimodal understanding and reasoning ability on rich dynamic audio-visual scenes. To facilitate the aforementioned research, we built a large-scale Spatio-Temporal Music AVQA (MUSIC-AVQA) dataset. Considering that musical performance is a typical multimodal scene consisting of abundant audio and visual components as well as their interaction, it is appropriate to be utilized for the exploration of effective audio-visual scene understanding and reasoning. So we collected amounts of user-uploaded videos of musical performance from YouTube, and videos in the built dataset consist of solo, ensemble of the same instruments and ensemble of different instruments. It contains 9,288 videos covering 22 instruments, with a total duration of over 150 hours. 45,867 question-answer pairs are generated by human crowd-sourcing, with an average of about 5 QA pairs per video. The questions are derived from 33 templates and asked regarding content from different modalities at space and time, which are suitable to explore fine-grained scene understanding and spatio-temporal reasoning in the audio-visual context.

To solve the above AVQA task, we consider this problem from the spatial and temporal grounding perspective, respectively. Firstly, the sound and the location of its visual source is deemed to reflect the spatial association between audio and visual modality, which could help to decompose the complex scenario into concrete audio-visual association. Hence, we propose a spatial grounding module to model such cross-modal association through attention-based sound source localization. Secondly, since the audio-visual scene changes over time dynamically, it is critical to capture and highlight the key timestamps that are closely related to the question. Accordingly, the temporal grounding module that uses question features as queries is proposed to attend crucial temporal segments for encoding question-aware audio and visual embeddings effectively. Finally, the above spatial-aware and temporal-aware audio-visual features are fused to obtain a joint representation for Question Answering. As an open-ended problem, the correct answers to questions can be predicted by choosing words from a pre-defined answer vocabulary. Our results indicate that audio-visual QA benefits from effective audio-visual scene understanding and spatio-temporal reasoning, and our model outperforms recent A-, V-, and AVQA approaches.

To summarize, our contributions are threefold:

• We build the large-scale MUSIC-AVQA dataset of musical performance, which contains more than 9K videos annotated by over 45K QA pairs, spanning over different modal scenes.
• A spatio-temporal grounding model is proposed to solve the fine-grained scene understanding and reasoning over audio and visual modalities.
• Extensive experiments show that AVQA benefits from multisensory perception and our model is superior to recent QA approaches especially on the questions that measures spatio-temporal reasoning ability of models.

2. Related Work

2.1. Audio-Visual Learning

By integrating the audio and visual information in multimodal scenes, it is expected to explore more sufficient scene information and overcome the limited perception in single modality. Recently, there have been several works utilizing audio and visual modality to facilitate multimodal scene understanding in different perspectives, such as sound source localization [23, 31, 34, 37, 48] and separation [10, 13, 41, 59, 61, 63], audio inpainting [62], event localization [4, 43, 64], action recognition [14], video parsing [42, 47], captioning [24, 40, 50], and dialog [1, 66].

Regarding previous works on sound source localization and separation, the former mainly focuses on locating sounds in a visual context [34, 37], while the latter mainly centers around separating different sounds from corresponding visual objects [12, 59]. These works have made great progress for the interaction of audio and visual features, but they essentially focus on the perception of audio-visual objects. Further, some researchers propose to integrate audio and visual messages to explore semantic events and behaviors in multimodal scenes [14, 43]. As expected, these works have shown considerable performance by utilizing more sufficient information from audio and visual cues. Based on which, others took a step forward to parse the audio-visual scenes [42], describe content [24], and leverage contextual cues for dialog [1, 66].

Apart from the above methods that facilitate scene understanding by excavating and analyzing different modalities, a unified multimodal model should also be able to reason their spatio-temporal correlation. In this work, different from the previous methods, besides the fine-grained scene understanding, we further propose to explore spatio-temporal reasoning in the audio-visual context.

2.2. Question Answering

In the past years, several question answering tasks have been proposed but in different modalities, including text question answering [35, 44], visual question answering [3, 25, 53, 57], audio question answering [9, 58], etc.

VQA [3, 17, 32] aims to generate natural language answers about specific visual content. The early research in VQA focused on simple visual understanding in static images but ignored the spatial and semantic relationships be-
between visual content, hence they are difficult to achieve effective visual reasoning in complex scene. To overcome this shortcoming, Johnson et al. [26] released the simulated CLEVR dataset and expected the model to answer reasoning-oriented visual questions. Since then, more attentions are paid to the spatial and semantic relational reasoning of visual objects in VQA [2, 11, 33]. Recently, some methods proposed to improve the spatial-temporal reasoning ability of computational model further, by answering question in the video context [8, 27, 30, 49, 54, 60]. Apart from the visual information, some other modality information in video, such as subtitles [29] or scripts [39], are used for advancing the understanding of video content. Similarly, some external knowledge [15, 46] and situations [5, 45] are also utilized to achieve better content understanding.

In addition to the visual modality-based QA, some researchers also proposed to answer questions in other modalities, such as audio [1, 9, 36, 56] and speech [58]. Pano-AVQA [56] is a concurrent work to ours, also aiming at audio-visual question answering. But the QA-pairs within the dataset only covers relatively simple audio-visual association, such as existential or location questions. In contrast, our built MUSIC-AVQA dataset can facilitate study on spatio-temporal reasoning for dynamic and long-term audio-visual scenes. Meanwhile, the proposed method provides new perspectives in modeling such complex scenario and obtains noticeable results.

### 3. The MUSIC-AVQA Dataset

#### 3.1. Overview

To explore scene understanding and spatio-temporal reasoning over audio and visual modalities, we build a large-scale audio-visual dataset, MUSIC-AVQA, which focuses on question-answering task. As noted above, high-quality datasets are of considerable value for AVQA research. Hence, considering that musical performance is a typical multimodal scene consisting of abundant audio and visual components as well as their interaction, we choose to manually collect amounts of musical performance videos from YouTube. Specifically, 22 kinds of instruments, such as guitar, cello, and xylophone, are selected and 9 audio-visual question types are accordingly designed, which cover three different scenarios, i.e., audio, visual and audio-visual.

As shown in Tab. 1, compared to existing related datasets, our released MUSIC-AVQA dataset has the following advantages: 1) Our dataset offers QA pairs that covering audio question, visual question and audio-visual question, which is more comprehensive than other datasets. Most video QA datasets, like ActivityNet-QA [54], TVQA [29], only contain visual question and provide limited possibility to explore audio-visual correlation. Although existing AVQA datasets, such as AVSD [1] and Pano-AVQA [56], also offer audio-visual QA pairs, they focus on relatively simple audio-visual correlation that only needs spatial reasoning, such as existential or location questions. As a concurrent work of Pano-AVQA, our dataset is more comprehensive and much longer than it, which includes more spatial and temporal related question, such as existential, location, counting, comparative and temporal. 2) Our dataset consists of musical performance scenes that contains enriching audio-visual components, which contributes to better investigation of audio-visual interaction, and it can avoid the noise problem in the scene to some extent, where the visual objects and sounds are not related. The audio information in most released datasets (e.g., ActivityNet-QA [54] and AVSD [1]) is usually accompanied by severe noise that sound and visual objects in the video do not match (e.g. background music), which makes them difficult to explore the association between different modalities. In addition, the TVQA [29] dataset contains both visual and audio modality, but its sound mainly consists of human speech, and only the corresponding subtitle is used during QA pairs construction. In the followings, we provide detailed descriptions about the procedure of video collection, QA pairs annotation and collection, as well as the related statistical analysis about our MUSIC-AVQA dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Origin</th>
<th>Main sound type</th>
<th># Videos</th>
<th>Average video length</th>
<th>A Question</th>
<th>V Question</th>
<th>A-V Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>ActivityNet-QA</td>
<td>ActivityNet</td>
<td>Background music</td>
<td>5.8K</td>
<td>180s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TVQA [29]</td>
<td>TV Show</td>
<td>Human speech</td>
<td>21.8K</td>
<td>60s/960s</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>AVSD [1]</td>
<td>Charades</td>
<td>Domestic sounds</td>
<td>8.5K</td>
<td>30s</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Pano-AVQA [56]</td>
<td>Online</td>
<td>Visual object sound</td>
<td>5.4K</td>
<td>5s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUSIC-AVQA</td>
<td>YouTube</td>
<td>Visual object sound</td>
<td>9.3K</td>
<td>60s</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Comparison with other video QA datasets. Our MUSIC-AVQA dataset focuses on the interaction between visual objects and their produced sounds, offering QA pairs that cover audio, visual and audio-visual questions, which is more comprehensive than other datasets. The collected videos in MUSIC-AVQA can facilitate audio-visual understanding in terms of spatial and temporal associations.
untrimmed videos, we randomly cut them into one minute long for efficiency purpose. Moreover, human verification is performed to ensure whether the cut videos contain musical performance scenes.

**Synthetic Videos.** There are many solo and duet performance in real-world videos that contain limited visual objects and sounds. To further facilitate study on understanding and reasoning, we synthesize more challenging videos in which multiple visual objects and sounds are appeared with different associations.

### 3.3. QA Pairs Annotation and Collection

For the collected musical performance videos, the QA annotation is performed in three steps: question design, question collection and answer collection.

**Questions Design.** In order to better explore the contribution of the spatio-temporal correlation between visual and audio components to multimodal scene understanding, 33 question templates that cover 9 question types are proposed under different modality scenes. Concretely, to prevent from asking multiple simple questions and guarantee the diversity of questions, inspired by the mechanism of question templates in building VQA dataset [26, 38], we design several question templates before annotating the collected videos, as shown in Fig. 2(d).

**Questions Collection.** We design an audio-visual question answering labeling system to collect questions. To ensure the diversity and balance of different question templates, we set up the following rules for the labeling system: 1) the same question template in a video can only be annotated by the same annotator once; 2) each video needs to be watched for more than 30-seconds before it can be annotated; 3) the question templates that have been annotated will no longer be displayed to the subsequent annotators; 4) each video has to be annotated for 5 times. With these rules, we collect the questions for all the musical performance videos.

**Answers.** As each question template has certain answer, we ask annotators to directly choose the correct one from the answer vocabulary. And we also use the above labeling system to collect answers. In this process, we set up the following rules when answering questions: 1) when one answer that is selected for the same question twice, it will be considered as the correct answer; 2) when the answer to a question is confirmed, it will not be seen by the subsequent annotators. In addition, the unreasonable question is annotated as invalid, and the corresponding video will be asked one new question again.

### 3.4. Statistical Analysis

Our MUSIC-AVQA dataset contains 45,867 question-answer pairs, distributed in 9,288 videos for over 150 hours.
4. Method

To solve the AVQA problem, we propose a spatio-temporal grounding model to achieve scene understanding and reasoning over audio and visual modalities. An overview of the proposed framework is illustrated in Fig. 3.

4.1. Representations for Different Modalities

Given an input video sequence containing both visual and audio tracks, we first divide it into $T$ non-overlapping visual and audio segment pairs $\{V_i, A_i\}_{i=1}^{T}$, where each segment is 1s long. The question sentence $Q$ is tokenized into $N$ individual words $\{q_n\}_{n=1}^{N}$.

Audio Representation. We encode each audio segment $A_i$ into a feature vector $f_a^i$ using a pre-trained VGGish model [16], which is VGG-like 2D CNN network, employing over transformed audio spectrograms. The audio representation is extracted offline and the model is not fine-tuned.

Visual Representation. We sample a fixed number of frames for all video segments. We then apply pre-trained ResNet-18 [18] on video frames to extract visual feature map $f_v^{v,m}$ for each video segment $V_i$. The used pre-trained ResNet-18 model is not fine-tuned.

Question Representation. For an asked question $Q = \{q_n\}_{n=1}^{N}$, a LSTM is used to process projected word embeddings $\{f_q^i\}_{i=1}^{N}$ and encode the question into a feature vector $f_q$ using the last hidden state. The question encoder is trained from the scratch.

4.2. Spatial Grounding Module

We consider that the sound and the location of its visual source usually reflects the spatial association between audio and visual modality, the spatial grounding module, which performs attention-based sound source localization, is therefore introduced to decompose the complex scenarios into concrete audio-visual association. Specifically, for each video segment $V_i$, the visual feature map $f_v^{v,m}$ and the corresponding audio feature $f_a^i \in \mathcal{R}^C$ compose the matched pair. Then we randomly sample another visual segment and get its visual feature map, which composes the non-matched pair with the audio feature $f_a^i$. For each pair, we can compute the sound-related visual features, $f_{v,s}^t$, as:

$$f_{v,s}^t = f_v^{v,m} \cdot \sigma((f_a^i)^T \cdot f_v^{v,m}),$$

(1)
where $\sigma$ is the softmax and $(\cdot)^\top$ represents the transpose operator. To prevent possible visual information loss, we averagely pool the visual feature map $f^t_{v,m}$, obtaining the global visual feature $f^t_{v,g}$. The two visual feature is fused as the visual representation: $f^t_v = FC(\text{Tanh}(f^t_{v,g}, f^t_{v,u}))$, where FC represents fully-connected layers. Then, the visual and the audio representation combines to predict the audio-visual pairs are matched or not:

$$
\hat{y}^t = \sigma(FC(\text{Concat}(f^t_a, f^t_v)))
$$

$$
\mathcal{L}_s = \mathcal{L}_{ce}(y^{match}, \hat{y}^t)
$$

where $y^{match}$ indicates whether the audio and visual feature come from the matched pair, i.e., $y^{match} = 1$ when $f^t_a$ and $f^t_v$ is the matched pair, otherwise $y^{match} = 0$. $\mathcal{L}_{ce}$ is the cross-entropy loss. It should be noted that non-matched pairs are only used in the spatial grounding module, i.e., $f^t_a$ and $f^t_v$ is always the matched pair in other modules.

4.3. Temporal Grounding Module

To highlight the key timestamps that are closely associated to the question, we propose a temporal grounding module, which is designed for attending critical temporal segments among the changing audio-visual scenes and capturing question-aware audio and visual embeddings. Concretely, given a $t \in T$ and audio-visual features $\{f^t_{a}; f^t_{v}\}$, the temporal grounding module will learn to aggregate question-aware audio and visual features. The grounded audio feature $f_a$ and visual feature $f_v$ can be computed as:

$$
f_a = \sum_{t=1}^{T} w^t_a f^t_a = \sigma(\frac{f^T_a \sqrt{d}}{d}) f^t_a
$$

$$
f_v = \sum_{t=1}^{T} w^t_v f^t_v = \sigma(\frac{f^T_v \sqrt{d}}{d}) f^t_v
$$

where $f_a = [f^1_a; \ldots; f^T_a]$ and $f_v = [f^1_v; \ldots; f^T_v]$; $d$ is a scaling factor with the same size as the feature dimension. Obviously, the model will assign large weights to audio and visual segments, which are more relevant to the asked question. Hence, the question grounded audio/visual contextual embeddings are more capable of predicting correct answers.

4.4. Multimodal Fusion and Answer Prediction

Different modalities can contribute to correctly answer questions. To combine the features: $f_a$, $f_v$, and $f_q$, we introduce a simple multimodal fusion network. It firstly concatenates audio and visual features and then uses a linear layer with a tanh activation to generate an audio-visual embedding $f_{av}$. Finally, we integrate audio-visual and question features with employing an element-wise multiplication operation. Concretely, we can formulate the fusion function as: $e = f_{av} \circ f_q$, where $f_{av} = FC(\text{Tanh}(\text{Concat}(f_a, f_v)))$.

To achieve audio-visual video question answering, we predict the answer for a given question from the joint multimodal embedding $e$. It can be formulated as an open-ended task, which aims to choose one correct word as the answer from a pre-defined answer vocabulary. We utilize a linear layer and softmax function to output a probabilities $p \in \mathcal{R}^c$ for candidate answers. With the predicted probability vector and the corresponding ground-truth label $y$, we can optimize our network using a cross-entropy loss:

$$
\mathcal{L}_{qa} = - \sum_{c=1}^{C} y_c \log(p_c).
$$

During testing, we can select the predicted answer by $\hat{c} = \arg \max_c (p)$.

5. Experiments

5.1. Experiments Setting

Implementation Details. The sampling rates of sounds and video frames are 16 kHz and 1 fps, respectively. For each video, we divide it into non-overlapping segments of the same length with 1 frame and generate a 512-D feature vector for each visual segment. For each 1s-long audio segment, we use a linear layer to process the extracted 128-D VGGish feature into a 512-D feature vector. The dimension of the word embedding is set to 512. In experiments, due to the limitation of computing resources, we sampled the videos by taking 1s every 6s. Batch size and number of epochs are 64 and 30, respectively. The initial learning rate is $1e^{-4}$ and will drop by multiplying 0.1 every 10 epochs. Our networks is trained with the Adam optimizer.

Training Strategy. We use a two-stage training strategy, training the spatial grounding module first with $\mathcal{L}_s$. Later, based on stage one, using $\mathcal{L} = \mathcal{L}_{qa} + \lambda \cdot \mathcal{L}_s$ to train for AVQA task, where $\lambda$ is 0.5 in our experiment.

Baselines. To validate our method on the released MUSIC-AVQA dataset, we compare it with recent audio QA methods: FCNLSTM [9] and CONVLSTM [9], visual QA methods: GRU [3], BiLSTM Attn [65], HCAAttn [32] and MCAN [55], video QA methods: PSAC [30], HME [8] and HCRN [28], AVQA method: AVSD [36] and Pan-AVQA [56]. To investigate different modalities and modules, we compare several sub-models, as shown in Tab. 3.

Evaluation. We use answer prediction accuracy as the metric and evaluate model performance on answering different types of questions. The answer vocabulary consists of 42 possible answers (22 objects, 12 counting choices, 6 location types, and yes/no) to different types of questions in the dataset. For training, we use one single model to handle all questions without training separated models for each type. So the accuracy with random choice is 11/42≈2.4%. Additionally, all models are trained on our AVQA dataset using the same features for a fair comparison.

5.2. Results and analysis

To study different input modalities and validate the effectiveness of the proposed model, we conduct extensive
Table 2. AVQA results of different methods on the test set of MUSIC-AVQA. The top-2 results are highlighted.

<table>
<thead>
<tr>
<th>Method</th>
<th>Audio Question</th>
<th>Visual Question</th>
<th>Audio-Visual Question</th>
<th>All Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q</td>
<td>65.19</td>
<td>44.42</td>
<td>55.15</td>
<td>54.09</td>
</tr>
<tr>
<td>A+Q</td>
<td>67.78</td>
<td>62.75</td>
<td>63.86</td>
<td>64.26</td>
</tr>
<tr>
<td>V+Q</td>
<td>68.76</td>
<td>67.28</td>
<td>63.23</td>
<td>65.28</td>
</tr>
<tr>
<td>AV+Q</td>
<td>70.67</td>
<td>69.72</td>
<td>65.84</td>
<td>67.72</td>
</tr>
<tr>
<td>AV+Q+TG</td>
<td>73.01</td>
<td>73.18</td>
<td>68.02</td>
<td>70.27</td>
</tr>
<tr>
<td>AV+Q+TG+SG</td>
<td>74.06</td>
<td>74.00</td>
<td>69.54</td>
<td>71.52</td>
</tr>
</tbody>
</table>

A ablation study on input modalities and the proposed modules. We observe that leveraging audio, visual, and question information can boost AVQA task.

Multi-sensory perception boosts QA. As shown in Tab. 3, introducing A or V both facilitates the model performance. Also, the model V+Q adding visual features is overall better than the Q and the A+Q, which indicates that the visual modality is a strong signal for QA. It is not surprising to see that the V+Q is better than A+Q for visual question answering, but we also observe that V+Q outperforms A+Q for audio question answering. It is intuitive that recognizing sounds from complicated sound mixtures are very challenging, especially when two sounds are in the same category, while it is easy for visual modality since different sources are visually isolated. As shown in Fig. 4(a) shows, there are two sounding cellos in the video, which can be seen in visual effortlessly, while the sound of two trumpets is hard to recognized. What’s more, obviously, when combining audio and visual modalities, the AV+Q model performance is much better than the A+Q and V+Q models, indicating that multi-sensory perception helps to boost QA performance.

Spatio-temporal grounding analysis. With the spatio-temporal grounding module, our audio-visual model achieves the overall best performance among the compared methods. In Fig. 4, we provide several visualized spatial grounding results. The heatmap indicates the location of the question. Through the spatial grounding results, the sounding objects are visually captured, which can facilitate the spatial reasoning. For example, in the case of Fig. 4(c), the spatial grounding module offers the information that the sounding object in each timestamp. Also, the temporal grounding module aggregate the information of all timestamps based on the question. According to the keyword: last, the model can infer that at the last of the video, the instrument located on the right is playing. Combined with temporal grounding module, the model can capture the sounding objects in each timestamp and have a comprehensive understanding of the whole video.

Comparison to recent QA methods. Table 2 shows results of recent QA methods on our MUSIC-AVQA dataset. The results firstly demonstrate that all AVQA methods outperform A-, V- and VideoQA methods, which indicates that AVQA task can be boosted through multi-sensory perception. Secondly, our method achieves considerable improvement on most audio and visual questions. For the audio-visual question that desires spatial and temporal reasoning, our method is clearly superior over other methods on most question types, especially on answering the Counting and Location questions. Although the Pano-AVQA [56] attempted to model audio-visual scenes, our methods explicitly constructs the association between audio and visual modalities and temporally aggregate both features, solving the spatio-temporal reasoning problem more effectively. Moreover, the results confirm the potential of our dataset as a testbed for audio-visual scene understanding.

6. Discussion

In this work, we investigate the audio-visual question answering problem, which aims to answer questions regarding videos by fully exploiting multi-sensory content. To facilitate this task, we build a large-scale MUSIC-AVQA dataset, which consists of 45,867 question-answer pairs spanning over audio-visual modalities and different question types. We also propose a spatio-temporal grounding model to ex-
explore the fine-grained scene understanding and reasoning. Our results show that all of different modalities can contribute to addressing the AVQA task and our model outperforms recent QA approaches, especially when equipped with our proposed modules. We believe that our dataset can be a useful testbed for evaluating fine-grained audio-visual scene understanding and spatio-temporal reasoning, and has a potential to inspire more people to explore the field.

Limitation. Although we have achieved considerable improvement, the AVQA task still has a wide scope for exploration. Firstly, the scene of the current dataset is more limited to the musical scenario, while audio-visual interaction exists in more daily situations. We will explore audio-visual reasoning tasks in more general scenarios in the subsequent study. Our model simply decomposes the complex scenarios into concrete audio-visual association. However, some visual objects or sound sources, which are not relevant to the questions, are involved in the encoded unimodal embeddings, might introducing learning noises and make solving QA tasks challenging, as the shown failure example in Fig. 4(f). To alleviate the problem, we can parse each video into individual objects and isolated sounds and then adaptively leverage question-related audio and visual elements for more accurate question answering. Further, to facilitate temporal reasoning, we proposed to highlight the key timestamps that are close to the question. However, such module lacks explicit temporal modeling between audio and visual modality. More advanced model that could bridge the temporal association across modalities is expected to boost performance further. Though the scenarios are somewhat limited, we think this is the first step of audio-visual reasoning and we believe this paper will be a good start in this field.

Broader impacts. The released MUSIC-AVQA dataset is curated, which perhaps owns potential correlation between instrument and geographical area. This issue warrants further research and consideration.

Acknowledgement G. Li, Y. Wei, J-R. Wen and D. Hu were supported by Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the “Double-First Class” Initiative, Renmin University of China. They were also supported by Beijing Outstanding Young Scientist Program (NO.BJWZYJH012019100020098), the Research Funds of Renmin University of China (NO.21XNLG17), the National Natural Science Foundation of China (NO.62106272), the 2021 Tencent AI Lab Rhino-Bird Focused Research Program (No.JR202141), the Young Elite Scientists Sponsorship Program by CAST, the Large-Scale Pre-Training Program of Beijing Academy of Artificial Intelligence (BAAI) and the Public Computing Cloud, Renmin University of China. Y. Tian and C. Xu were supported by the National Science Foundation (NSF) under Grant 1741472. The article solely reflects the opinions and conclusions of its authors but not the funding agents.
References


