# Learning to Answer Questions in Dynamic Audio-Visual Scenarios

**Guangyao Li**[†]
Renmin University of China
guangyaoli@ruc.edu.cn

**Yake Wei**[†]
Renmin University of China
yakewei@ruc.edu.cn

**Yapeng Tian**[†]
University of Rochester
yapengtian@rochester.edu

**Di Hu**[*]
Renmin University of China
dihu@ruc.edu.cn

**Chengliang Xu**
University of Rochester
chenliang.xu@rochester.edu

**Ji-Rong Wen**
Renmin University of China
jrwen@ruc.edu.cn

## 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 AVQA dataset, which contains more than 45K question-answer pairs covering 33 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 testbed for evaluating and promoting progress in audio-visual scene understanding and spatio-temporal reasoning.*

## 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 [17]. For instance, imagining 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 [5, 18, 31, 43], audio scene analysis [8, 11, 42, 50], audio-visual scene parsing [35, 39], and content description [19, 34, 41] towards audio-visual scene understanding. Although these methods associate objects or sound events across audio and visual views, most of them remains 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) [7] 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-

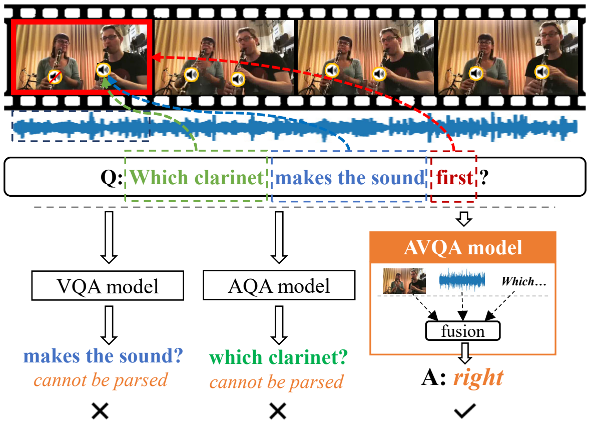


Figure 1. Audio-visual question answering requires auditory and visual modalities for multimodal scene understanding and spatio-temporal reasoning. For example, when we encounter a complex musical performance scene involving multiple sounding and non-sounding instruments above, it is difficult to analyze the *sound first* term in the question by VQA model that only considers visual modality. While if we only consider the AQA model with mono sound, the *left* or *right* position is also hard to be recognized. However, we can see that using both auditory and visual modalities can answer this question effortlessly.

visual scenario and focus on the "*first*" sounding "*clarinet*" in the timeline. To answer the question correctly, 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 AVQA* (ST-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 9,290 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 predefined 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 ST-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 [28, 31, 40] and separation [8, 11, 50], event localization [4, 36, 52], action recognition [12], video parsing [35, 39], captioning [19, 34, 41], and dialog [1, 54].

Regarding previous works on sound source localization and separation, the former mainly focuses on locating sounds in a visual context [28, 31], while the latter mainly centers around separating different sounds from corresponding visual objects [10, 50]. 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 [12, 36]. 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 [35], describe content [19], and leverage contextual cues for dialog [1, 54].

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 [29, 37], visual question answering [3, 20, 44, 48], audio question answering [7, 49], etc.

| Dataset | Origin | Main sound type | # Videos | Average video length | A Question | V Question | A-V Question | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Existential | Location | Counting | Comparative | Temporal |
| ActivityNet-QA [45] | ActivityNet | Background music | 5.8K | 180s | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| TVQA [24] | TV Show | Human speech | 21.8K | 60s/90s | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| AVSD [1] | Charades | Domestic sounds | 8.5K | 30s | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| Pano-AVQA [47] | Online | Visual object sound | 5.4k | 5s | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ |
| Our ST-AVQA | YouTube | Visual object sound | 9.3K | 60s | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1. **Comparison with other video QA datasets.** Our ST-AVQA dataset focuses on the interaction between visual objects and their produced sounds, offering QA pairs that covering audio questions, visual questions and audio-visual questions, which is more comprehensive than other datasets. The collected videos in ST-AVQA can facilitate audio-visual understanding in terms of spatial and temporal associations. Note that the Pano-AVQA dataset is not publicly available.

VQA [3, 15, 26] 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 between visual content, hence they are difficult to achieve effective visual reasoning in complex scene. To overcome this shortcoming, Johnson *et al.* [21] 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, 9, 27]. Recently, some methods proposed to improve the spatial-temporal reasoning ability of computational model further, by answering question in the video context [6, 22, 25, 45, 51]. Apart from the visual information, some other modality information in video, such as subtitles [24] or scripts [33], are used for advancing the understanding of video content. Similarly, some external knowledge are also utilized to achieve better content understanding [13, 38].

In addition to the visual modality-based QA, some researchers also proposed to answer questions in other modalities, such as audio [1, 7, 30, 47] and speech [49]. Pano-AVQA [47] 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 ST-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 ST-AVQA Dataset

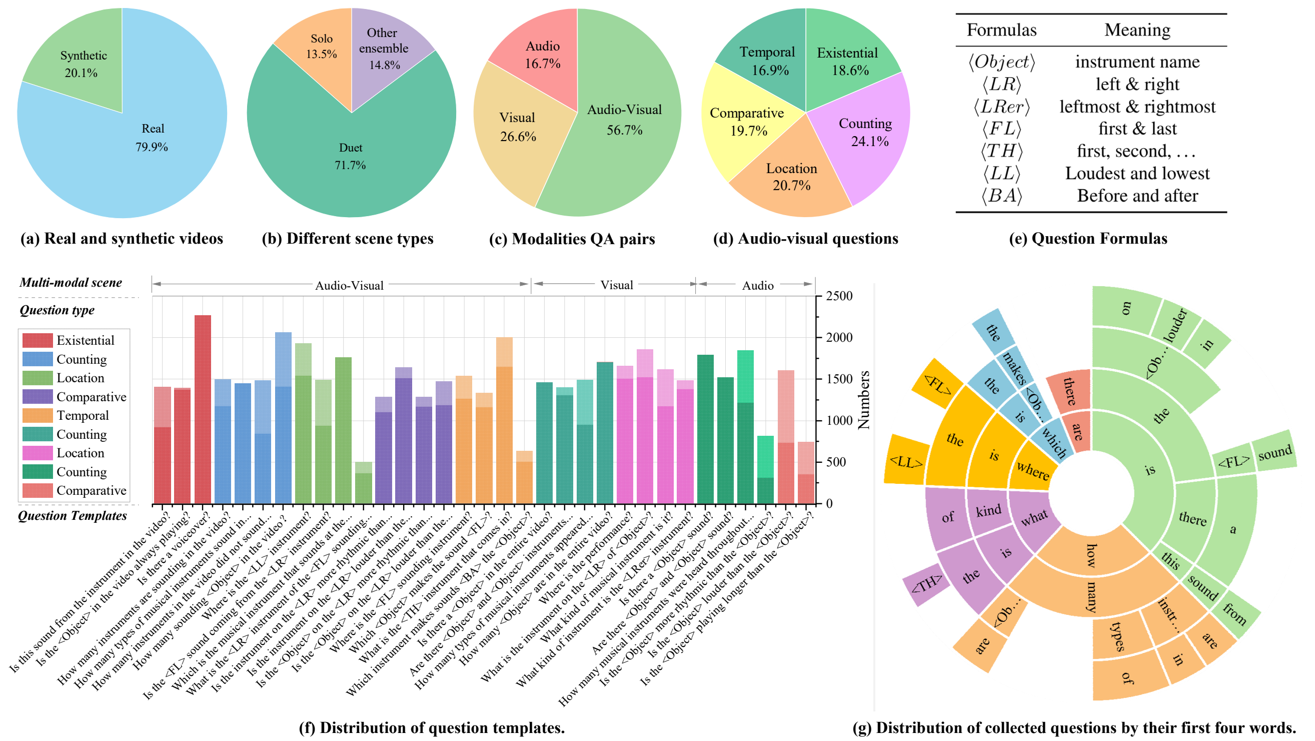
## 3.1. Overview

To explore scene understanding and spatio-temporal reasoning over audio and visual modalities, we build a large-scale audio-visual dataset, ST-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 ST-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 [45], TVQA [24], only contain visual question and provide limited possibility to explore audio-visual correlation. Although existing AVQA datasets, such as AVSD [1] and Pano-AVQA [47], 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. However, the audio information in most released datasets (*e.g.*, ActivityNet-QA [45] and AVSD [1]) is usually accompanied by severe noise (*e.g.*, background music), which makes them difficult to explore the association between different modalities. In addition, the TVQA [24] 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 ST-AVQA dataset.

## 3.2. Video Collection

**Real Videos.** We collect 7,423 real videos of musical performance from YouTube. Among these videos, three kinds of musical performance are covered to ensure the diversity, complexity and dynamic of audio-visual scenes: solo, ensemble of the same instrument (ESIT) and ensemble of different instruments (EDIT). In order to control the quantity



(a) Real and synthetic videos

(b) Different scene types

(c) Modalities QA pairs

(d) Audio-visual questions

| Formulas | Meaning |
|---|---|
| $\langle Object \rangle$ | instrument name |
| $\langle LR \rangle$ | left & right |
| $\langle LRer \rangle$ | leftmost & rightmost |
| $\langle FL \rangle$ | first & last |
| $\langle TH \rangle$ | first, second, ... |
| $\langle LL \rangle$ | Loudest and lowest |
| $\langle BA \rangle$ | Before and after |

(e) Question Formulas

(f) Distribution of question templates.

(g) Distribution of collected questions by their first four words.

Figure 2. **Illustrations of our ST-AVQA dataset statistics. (a-d)** statistical analysis of the videos and QA pairs. **(e)** Question formulas. **(f)** Distribution of question templates, where the dark color indicates the number of QA pairs generated from real videos while the light-colored area on the upper part of each bar means that from synthetic videos. **(g)** Distribution of first n-grams in questions. Our QA-pairs need fine-grained scene understanding and spatio-temporal reasoning over audio and visual modalities to be solved. For example, *existential* and *location* questions require spatial reasoning, and *temporal* questions require temporal reasoning. Best viewed in color.

balance of different instrument types, we design the following rules: **1) Solo**: about 50 solo videos are collected per instrument; **2) ESIT**: about 100 videos are collected per ESIT type; **3) EDIT**: each instrument is required to combine with every other instruments. For the collected 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 [21, 32], 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 ST-AVQA dataset contains 45,867 question-answer pairs, distributed in 9,290 videos for over 150 hours. Figure 2(a-d) provides the statistical analysis of our dataset. In this dataset, real videos and synthetic videos accounted for 79.9% and 20.1%, respectively. Real videos are composed of 14.8% solo videos, 71.7% duet videos and 13.5% other ensemble videos. Audio-visual questions makes up the majority of all QA pairs and consists of five types with a balanced share. Fig. 2(f) shows that all QA pairs types are divided into 3 modal scenarios, which contain 9 question types and 33 question templates. Finally, as an open-ended problem of our AVQA tasks, all 42 kinds of answers constitute a set for selection. For training and evaluation, we randomly split the dataset into training, validation, and testing sets with 32,087, 4,595, and 9,185 QA pairs, respectively. More details about the dataset construction and statistical analysis are in the *Supp. Materials*.

## 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 $\{V_t, A_t\}_{t=1}^{T}$, 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 $A_t$ into a feature vector $f_a^t$ using a pre-trained VGGish model [14], 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 [16] on video frames to extract visual feature map $f_{v,m}^t$ for each video segment $V_t$. The visual feature map is also extracted offline and the 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 $\{f_q^t\}_{n=1}^{N}$ 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 $V_t$, the visual feature map $f_{v,m}^t$ and the corresponding audio feature $f_a^t \in \mathcal{R}^C$ 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 $f_a^t$ . For each pair, we can compute the sound-related visual features, $f_{v,s}^t$, as:

$$f_{v,s}^t = f_{v,m}^t \cdot \sigma((f_a^t)^\intercal \cdot f_{v,m}^t), \tag{1}$$

where $\sigma$ is the softmax and $(\cdot)^\intercal$ represents the transpose operator. To prevent possible visual information loss, we averagely pool the visual feature map $f_{v,m}^t$, obtaining the global visual feature $f_{v,g}^t$. The two visual feature is fused as the visual representation:

$$f_v^t = \mathrm{fc}(\tanh[f_{v,g}^t, f_{v,s}^t]). \tag{2}$$

Then, the visual and the audio representation combines to predict the audio-visual pairs are matched or not:

$$\hat{y}^t = \sigma(\mathrm{fc}(\mathrm{concat}(f_a^t, f_v^t))), \tag{3}$$

$$\mathcal{L}_s = \mathcal{L}_{ce}(y^{match}, \hat{y}^t), \tag{4}$$

where $y^{match}$ indicates whether the audio and visual feature come from the matched pair, i.e., $y^{match} = 1$ when $f_v^t$ and $f_a^t$ 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., $f_v^t$ and $f_a^t$ 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 question embedding $f_q$ and audio-visual features $\{f_a^t, f_v^t\}_{t=1}^{T}$, the temporal grounding module will learn to aggregate question-aware audio and visual features. The grounded audio feature $\bar{f}_a$ and visual feature $\bar{f}_v$ can be computed as:

$$\bar{f}_a = \sum_{t=1}^{\intercal} w_t^a f_a^t = \sigma(\frac{f_q \boldsymbol{f}_a^\intercal}{\sqrt{d}}) \boldsymbol{f}_a \ , \tag{5}$$

$$\bar{f}_v = \sum_{t=1}^{\intercal} w_t^v f_v^t = \sigma(\frac{f_q \boldsymbol{f}_v^\intercal}{\sqrt{d}}) \boldsymbol{f}_v \ , \tag{6}$$

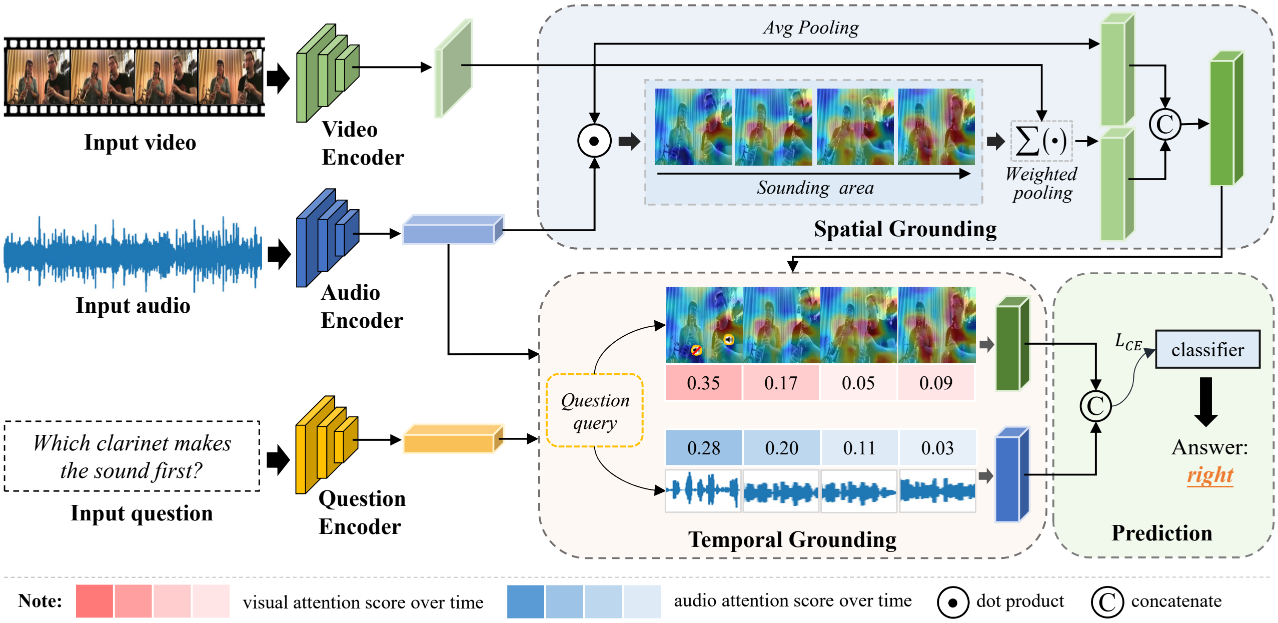


Figure 3. **The proposed audio-visual question answering model.** The model takes pre-trained CNNs to extract audio and visual features and uses a LSTM to obtain a question embedding. We associate specific visual locations with the input sounds to perform spatial grounding, based on which audio and visual features of key timestamps are further highlighted via question query for temporal grounding. Finally, multimodal fusion is exploited to integrate audio, visual, and question information for predicting the answer to the input question.

where $\boldsymbol{f}_a = [f_a^1; ...; f_a^T]$ and $\boldsymbol{f}_v = [f_v^1; ...; f_v^T]$; $d$ 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. Therefore, the question grounded audio and visual contextual embeddings are more capable of predicting correct answers.

### 4.4. Multimodal Fusion

Different modalities can contribute to correctly answer questions. To combine the features: $\bar{f}_a$, $\bar{f}_v$, 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 $f_{av} = \text{fc}(\tanh(\text{concat}(\bar{f}_a, \bar{f}_v)))$.

### 4.5. Answer Prediction

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 16000 $Hz$ and 1 $fps$, respectively. For each video, we divide it into non-overlapping segments of the same length with 1 frames and generate a 512-D feature vector for each visual segment. For each 1$s$-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 1$s$ every 6$s$. Batch size and number of epochs are 64 and 30, respectively. The initial learning rate is 1$e$-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, both $\mathcal{L}_{qa}$ and $\mathcal{L}_s$ are used as the loss function to train for AVQA task, $\mathcal{L} = \mathcal{L}_{qa} + \lambda \cdot \mathcal{L}_s$, where $\lambda$ is 0.5 in our experiment.

| Task | Method | Audio Question | | | Visual Question | | | Audio-Visual Question | | | | | | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Counting | Comparative | Avg. | Counting | Location | Avg. | Existential | Location | Counting | Comparative | Temporal | Avg. | Avg. |
| AudioQA | FCNLSTM [7] | 70.45 | 66.22 | 68.88 | 63.89 | 46.74 | 55.21 | 82.01 | 46.28 | 59.34 | 62.15 | 47.33 | 60.06 | 60.34 |
| | CONVLSTM [7] | 74.07 | **68.89** | 72.15 | 67.47 | 54.56 | 60.94 | **82.91** | 50.81 | 63.03 | 60.27 | 51.58 | 62.24 | 63.65 |
| VisualQA | GRU [3] | 72.21 | 66.89 | 70.24 | 67.72 | 70.11 | 68.93 | 81.71 | 59.44 | 62.64 | 61.88 | 60.07 | 65.18 | 67.07 |
| | BiLSTM Attn [53] | 70.35 | 47.92 | 62.05 | 64.64 | 64.33 | 64.48 | 78.39 | 45.85 | 56.91 | 53.09 | 49.76 | 57.10 | 59.92 |
| | HCAttn [26] | 70.25 | 54.91 | 64.57 | 64.05 | 66.37 | 65.22 | 79.10 | 49.51 | 59.97 | 55.25 | 56.43 | 60.19 | 62.30 |
| | MCAN [46] | 77.50 | 55.24 | 69.25 | 71.56 | 70.93 | 71.24 | 80.40 | 54.48 | 64.91 | 57.22 | 47.57 | 61.58 | 65.49 |
| VideoQA | PSAC [25] | 75.64 | 66.06 | 72.09 | 68.64 | 69.79 | 69.22 | 77.59 | 55.02 | 63.42 | 61.17 | 59.47 | 63.52 | 66.54 |
| | HME [6] | 74.76 | 63.56 | 70.61 | 67.97 | 69.46 | 68.76 | 80.30 | 53.18 | 63.19 | 62..69 | 59.83 | 64.05 | 66.45 |
| | HCRN [23] | 68.59 | 50.92 | 62.05 | 64.39 | 61.81 | 63.08 | 54.47 | 41.53 | 53.38 | 52.11 | 47.69 | 50.26 | 55.73 |
| AVQA | AVSD [30] | 72.41 | 61.90 | 68.52 | 67.39 | 74.19 | 70.83 | 81.61 | 58.79 | 63.89 | 61.52 | 61.41 | 65.49 | 67.44 |
| | Pano-AVQA [47] | 74.36 | 64.56 | 70.73 | 69.39 | 75.65 | 72.56 | 81.21 | 59.33 | 64.91 | 64.22 | 63.23 | 66.64 | 68.93 |
| | Our method | **78.18** | 67.05 | **74.06** | **71.56** | **76.38** | **74.00** | 81.81 | **64.51** | **70.80** | **66.01** | **63.23** | **69.54** | **71.52** |

Table 2. Audio-visual video question answering results of different methods on the test set of ST-AVQA. The top-2 results are highlighted.

| Method | A Question | V Question | A-V Question | All |
|---|---|---|---|---|
| Q | 65.19 | 44.42 | 55.15 | 54.09 |
| A+Q | 67.78 | 62.75 | 63.86 | 64.26 |
| V+Q | 68.76 | 67.28 | 63.23 | 65.28 |
| AV+Q | 70.67 | 69.72 | 65.84 | 67.72 |
| AV+Q+TG | 73.01 | 73.18 | 68.02 | 70.27 |
| AV+Q+TG+SG | 74.06 | 74.00 | 69.54 | 71.52 |

* TG: Temporal Grounding; SG: Spatial Grounding.

Table 3. Ablation study on input modalities and the proposed modules. We observe that leveraging audio, visual, and question information can boost AVQA task.

**Baselines.** To validate our method on the released ST-AVQA dataset, we compare it with recent audio QA methods: FCNLSTM [7] and CONVLSTM [7], visual QA methods: GRU [3], BiLSTM Attn [53], HCAttn [26] and MCAN [46], video QA methods: PSAC [25], HME [6] and HCRN [23], AVQA method: AVSD [30] and Pano-AVQA [47]. To investigate different modalities and modules, we compare several sub-models, as shown in Tab. 3.

**Evaluation.** To benchmark different models, we use answer prediction accuracy as the evaluation metric and evaluate performance of different models on answering different types of audio, visual, and audio-visual 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 1/42≈2.4%. Additionally, all models are trained on our AVQA dataset and the same audio/question features are used for fair comparison. Visual feature used in other method are the global visual feature in our method.

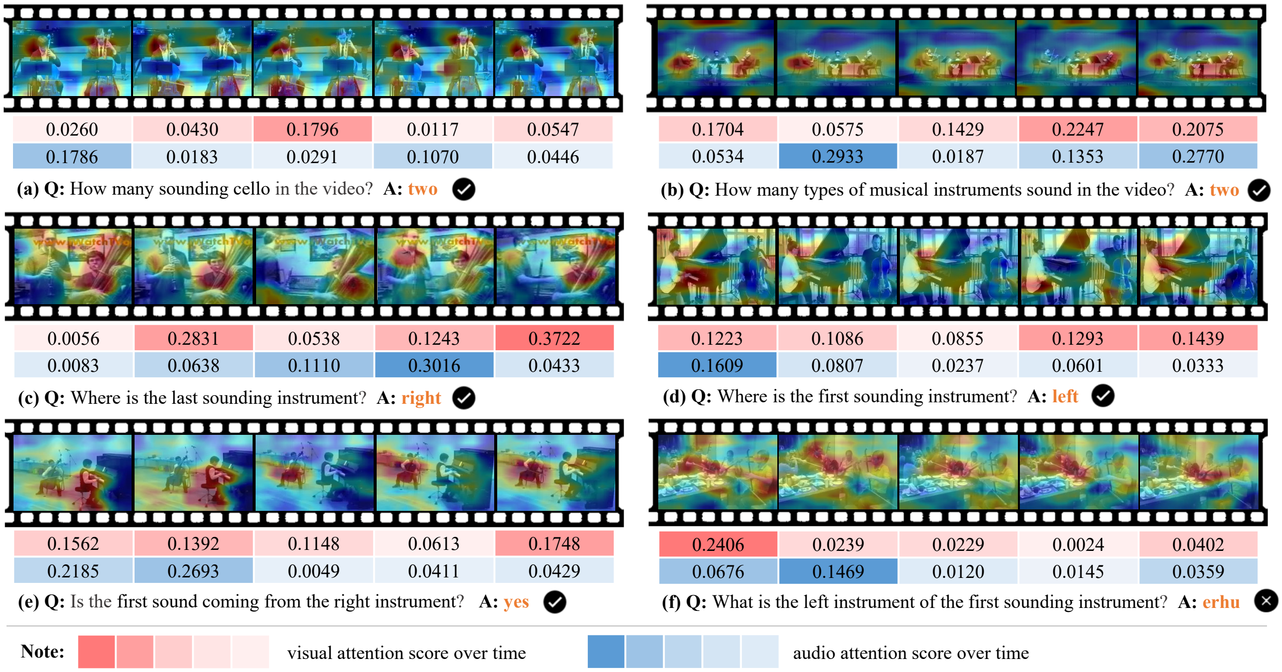
## 5.2. Results and analysis

To study different input modalities and validate the effectiveness of the proposed model, we conduct extensive ablations of our model (see Tab. 3) and compare to recent QA approaches (see Tab. 2).

**Question-only baseline.** Table 3 shows the results of the ablation study. The model Q, which only use questions as inputs, achieves accuracy of 54.90, since some type of questions can be answered fully based on common sense. This a common phenomenon that exists in the QA dataset [3, 47, 48]. For example, on Pano-AVQA dataset [47], the model Q even outperforms AVSD [30] method. However, the model Q is limited in handling complicate QA tasks (*e.g.*, *Location* and *Temporal*). After modeling the spatial and temporal association across modalities, the model performance gains a considerable improvement.

**Multisensory 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 multisensory 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 sounding source. 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



**(a) Q:** How many sounding cello in the video? **A:** two ✓

**(b) Q:** How many types of musical instruments sound in the video? **A:** two ✓

**(c) Q:** Where is the last sounding instrument? **A:** right ✓

**(d) Q:** Where is the first sounding instrument? **A:** left ✓

**(e) Q:** Is the first sound coming from the right instrument? **A:** yes ✓

**(f) Q:** What is the left instrument of the first sounding instrument? **A:** erhu ✗

Figure 4. **Visualized spatio-temporal grounding results.** Based on the grounding results of our method, the sounding area and key timestamps are accordingly highlighted in spatial and temporal perspectives (a-e), respectively, which indicates that our method can model the spatio-temporal association over different modalities well, facilitating the scene understanding and reasoning. Besides, the subfigure (f) shows one failure case predicted by our method, where the complex scenario with multiple sounding and silent objects makes it difficult to correlate individual objects with mixed sound, leading to a wrong answer for the given question.

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 ST-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 multisensory 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 [47] attempted to model audio-visual scenes, our methods explicitly constructs the association between audio and visual modalities and temporally aggregate audio and visual 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 multisensory content. To facilitate this task, we build a large-scale ST-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 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. 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.

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

## References

[1] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019. 2, 3

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016. 3

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2, 3, 7

[4] Mathilde Brousmiche, Jean Rouat, and Stéphane Dupont. Multi-level attention fusion network for audio-visual event recognition. *arXiv preprint arXiv:2106.06736*, 2021. 2

[5] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16867–16876, June 2021. 1

[6] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1999–2007, 2019. 3, 7

[7] Haytham M Fayek and Justin Johnson. Temporal reasoning via audio question answering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2283–2294, 2020. 1, 2, 3, 7

[8] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 1, 2

[9] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6639–6648, 2019. 3

[10] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 2

[11] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15495–15505, June 2021. 1, 2

[12] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020. 2

[13] Noa Garcia and Yuta Nakashima. Knowledge-based video question answering with unsupervised scene descriptions. *arXiv preprint arXiv:2007.08751*, 2020. 3

[14] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 5

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[17] Nicholas P Holmes and Charles Spence. Multisensory integration: space, time and superadditivity. *Current Biology*, 15(18):R762–R764, 2005. 1

[18] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *arXiv preprint arXiv:2010.05466*, 2020. 1

[19] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 958–959, 2020. 1, 2

[20] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766, 2017. 2

[21] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 3, 4

[22] Junyeong Kim, Minuk Ma, Trung Pham, Kyungsu Kim, and Chang D Yoo. Modality shifting attention network for multi-modal video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10115, 2020. 3

[23] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video

question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9972–9981, 2020. 7

[24] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 3

[25] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665, 2019. 3, 7

[26] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *arXiv preprint arXiv:1606.00061*, 2016. 3, 7

[27] Will Norcliffe-Brown, Efstathios Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. *arXiv preprint arXiv:1806.07243*, 2018. 3

[28] Rui Qian, Heinrich Dinkel Di Hu, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. *arXiv preprint arXiv:2007.06355*, 2020. 2

[29] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 2

[30] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12548–12558, 2019. 3, 7

[31] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 1, 2

[32] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*, 2021. 4

[33] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 3

[34] Yapeng Tian, Chenxiao Guan, Justin Goodman, Marc Moore, and Chenliang Xu. Audio-visual interpretable and controllable video captioning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops*, 2019. 1, 2

[35] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, 2020. 1, 2

[36] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2

[37] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015. 2

[38] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton Van Den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1367–1381, 2017. 3

[39] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1326–1335, 2021. 1, 2

[40] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6292–6300, 2019. 2

[41] Jun Xu, Ting Yao, Yongdong Zhang, and Tao Mei. Learning multimodal attention lstm networks for video captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 537–545, 2017. 1, 2

[42] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 882–891, 2019. 1

[43] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7177–7188, October 2021. 1

[44] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the ieee international conference on computer vision*, pages 2461–2469, 2015. 2

[45] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 3

[46] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019. 7

[47] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041, 2021. 3, 7, 8

[48] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022, 2016. 2, 7

[49] Ted Zhang, Dengxin Dai, Tinne Tuytelaars, Marie-Francine Moens, and Luc Van Gool. Speech-based visual question answering. *arXiv preprint arXiv:1705.00464*, 2017. 2, 3

[50] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 1, 2

[51] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *IJCAI*, pages 3518–3524, 2017. 3

[52] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021. 2

[53] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212, 2016. 7

[54] Ye Zhu, Yu Wu, Yi Yang, and Yan Yan. Describing unseen videos via multi-modal cooperative dialog agents. In *European Conference on Computer Vision*, pages 153–169. Springer, 2020. 2