Audio-Visual Adaptive Fusion Network for Question Answering Based on Contrastive Learning

Xujian Zhao[1*†], **Yixin Wang**[1†], **Peiquan Jin**[2]

[1]School of Computer Science and Technology, Southwest University of Science and Technology
[2]School of Computer Science and Technology, University of Science and Technology of China
jasonzhaoxj@swust.edu.cn, yixin1202@mails.swust.edu.cn, jpq@ustc.edu.cn

Abstract

The Audio-Visual Question Answering (AVQA) task involves extracting question-related audio-visual clues from both temporal and spatial perspectives to answer questions accurately. Despite the promising performance of existing multi-modal AVQA models, thanks to large-scale pre-trained models, challenges remain in the field. Firstly, aligning audio-visual information across temporal and spatial dimensions is difficult. Secondly, the fusion of audio-visual information is often weighted inadequately, limiting model performance. To address the above issues, we design the Audio-Visual Adaptive Fusion Network (AVAF-Net), which uses contrastive learning to align audio-visual information temporally and spatially and adaptively adjusts fusion weights based on the question. Specifically, we initially align visual and audio information temporally through a temporal-alignment contrastive loss. This is followed by an audio-visual clue-mining module that highlights question-related cues, aligning them with the vocal region spatially using spatial alignment contrastive loss. Additionally, a question-oriented adaptive fusion module assigns different weights to audio and visual modalities based on the question content and then fuses them. The fused audio-visual cues are finally used to predict the answer. Extensive experiments on the MUSIC-AVQA dataset show that AVAF-Net surpasses all baseline models, with a maximum improvement of 15.90% in average accuracy and an average improvement of 9.80%.

Introduction

Audio and visual clues greatly help us perceive the world and transmit information daily. By combining the two, our ability to perceive and understand dynamic audio-visual scenes is greatly improved (Wei et al. 2022). In recent years, the research on audio-visual scene understanding has received widespread attention and made significant progress, such as sound source localization (Hu et al. 2023; Senocak et al. 2018) and separation (Zhou et al. 2022a; Gan et al. 2020), event localization (Zhou et al. 2021, 2024a; Zhou, Guo, and Wang 2023; Zhou et al. 2024b), video parsing (Tian, Li, and Xu 2020; Mo and Tian 2022), segmentation

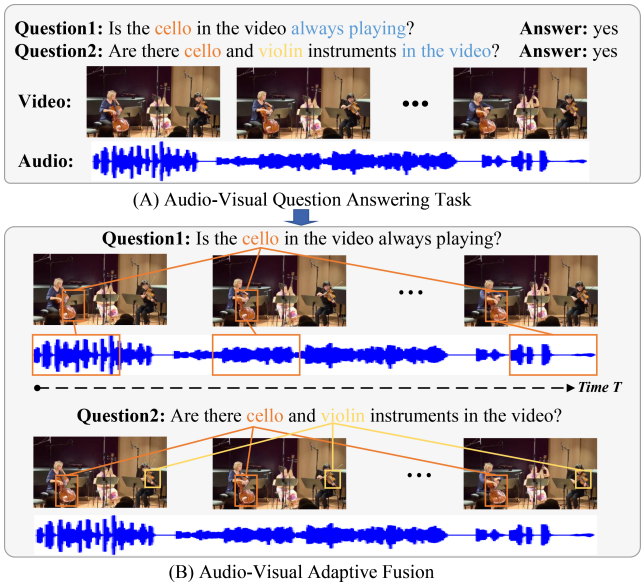

Figure 1: Illustration of the AVQA task and our audio-visual adaptive fusion method. (A) The AVQA task requires mining audio-visual clues related to the question. (B) Different questions require different modalities to be focused on. For question 1, the audio modality needs to be given higher attention, while for question 2, the visual modality needs to be focused on.

(Zhou et al. 2022b, 2024c; Liu et al. 2024), and question answering (Li et al. 2022; Yang et al. 2022; Li, Hou, and Hu 2023). Different from other audio-visual tasks, the audio-visual question answering (AVQA) task not only requires exploring the temporal and spatial relations in audio-visual scenes, but also requires comprehensive spatio-temporal reasoning based on the question content (Li et al. 2024). As shown in Fig.1, to obtain the correct answer, the AVQA model should identify the "cello" and "violin" in both temporal and spatial dimensions. For question one, "Is the cello in the video always playing?", the AVQA model should pay high attention to the audio modality. And for question two, "Are there violin and cello instruments in the video?", the visual modality should be given high attention.

To explore the AVQA task under long-term videos, Li et al. (2022) constructed a large-scale dynamic audio-visual

*Corresponding authors.
†These authors contributed equally.

scene dataset MUSIC-AVQA and a spatio-temporal grounding model to enhance the comprehension and reasoning of dynamic audio-visual scenes. Jiang and Yin (2023) proposed to introduce the question topic into the visual modality and construct a single-stream audio-visual question network. Moreover, Chen et al. (2024) also considered taking the question as a guiding factor to achieve accurate queries of question-related information. Li, Hou, and Hu (2023) designed the PSTP-Net to perform spatio-temporal reasoning by gradually perceiving the temporal segments and spatial regions related to the question. In addition, Lin et al. (2023) introduced a parameter-efficient framework to improve the performance of the model by efficiently fine-tuning the existing pre-trained models. Although these efforts have significantly promoted the progress of AVQA research, several challenges still need to be addressed:

Challenge 1: *Incomplete Alignment of Vision and Audio in Time and Space.* The AVQA task requires understanding the visual content in the video and associating the audio with the vocalization area, which makes it necessary to align the visual and audio in the temporal dimension and align the audio and vocalization areas in the spatial dimension. Some methods use pre-trained object detection models or attention mechanisms to locate the vocalization areas. Such as Li et al. (2024) used the object detection model DETR as the visual feature extractor to extract object-level visual features, and Li, Hou, and Hu (2023) located key temporal segments and spatial regions based on the attention map between questions and audio-visual clues. However, due to the complex visual objects and audio-visual correspondences in dynamic audio-visual scenes, it is still unable to effectively associate the audio and vocalization areas in time and space, which limits the performance of the model.

Challenge 2: *Inappropriate Fusion Weights between Audio and Vision.* For AVQA tasks, it is necessary not only to mine clues related to the question from the visual and audio modalities but also to select the required modal information based on the question, that is, the fusion weight value of the visual and audio modalities is assigned according to the question. As illustrated in Fig. 1 (B), for the question "Is the cello in the video always playing?", the network should pay more attention to the audio modality. Conversely, for "Are there cello and violin instruments in the video?", the visual model should be given more attention. Simply assigning the same weight to visual and audio modalities and then merging them may affect answer prediction.

To address the challenges mentioned, we propose the **Audio-Visual Adaptive Fusion Network** (**AVAF-Net**), which can perceive question-related audio-visual cues in complex dynamic audio-visual scenes. For the first challenge, we adopt two different contrastive loss functions, namely temporal alignment contrastive loss and spatial alignment contrastive loss, to optimize our network. The temporal alignment contrastive loss is used to align the visual information and the audio information in the temporal dimension, while the spatial alignment contrastive loss aligns the visual areas with the audio in the spatial dimension to find the effective vocalization areas. For the second challenge, we consider assigning weights to the mined visual and audio cues according to the given question. Specifically, we propose a question-oriented adaptive fusion module that allows the network to dynamically assign different weights to visual and audio clues according to questions, making AVAF-Net dynamically focus on different modalities and improving the accuracy of answer prediction. Finally, the mined audio-visual clues are used to predict the answers. Extensive experiments demonstrate that our AVAF-Net can effectively mine the spatio-temporal relationship in videos, highlighting its great potential in audio-visual task. Briefly, the contributions in this paper are as follows:

- We propose an audio-visual adaptive fusion network based on contrastive learning. This network adaptively aligns visual and audio information across temporal and spatial dimensions using temporal and spatial alignment contrast losses, thereby optimizing and enhancing network performance.
- We consider adaptively assigning weights to visual and audio cues based on the question's content. To achieve this, we design a question-oriented adaptive fusion module that enables the network to dynamically focus on and fuse different modalities according to the question, thereby improving answer prediction accuracy.
- The proposed AVAF-Net is extensively experimented on the public dataset MUSIC-AVQA. The results show that AVAF-Net outperforms all baseline models on audio-related, visual-related, and audio-visual-related question-answering tasks, achieving an average accuracy of 75.90%. Particularly, AVAF-Net achieves up to 15.90% improvements in the average accuracy compared with all baselines.

Related Work

Audio-Visual Representation Learning

Inspired by how humans use multi-sensory perception to understand and analyze the world, research on multimodal audio-visual scene understanding has received increasing attention (Wei et al. 2022). Compared with other modalities, vision and hearing are the main sources of human perception of the world. Audio-visual representation learning includes various interesting tasks, such as action recognition (Gao et al. 2020), sound source localization (Hu et al. 2023; Senocak et al. 2018), event localization (Zhou et al. 2021, 2024a; Zhou, Guo, and Wang 2023; Zhou et al. 2024b), audio-visual segmentation (Zhou et al. 2022b, 2024c; Liu et al. 2024), etc. Sun et al. (2023) proposed a learning strategy called FNAC based on contrastive learning to alleviate the poor performance problem caused by false negative samples misleading the training process of the source localization models. Wang et al. (2024) proposed a new encoder-prompt-decoder paradigm that helps the visual foundation model focus on sounding objects by constructing a semantic-aware audio prompt, improving the accuracy of audio-visual segmentation. To address the data scarcity problem in the field of Audio-Visual Emotion Recognition, Sun et al. (2024) proposed the HiCMAE, a three-pronged approach to promote

hierarchical audio-visual feature learning inspired by self-supervised learning.

These studies overcame the perceptual limitations of single-modal models by integrating the rich audio-visual information in multi-modal scenes, thereby leveraging visual and audio modalities for a more fine-grained exploration of multi-modal scenes.

Audio-Visual Question Answering

As an audio-visual pattern recognition and spatio-temporal reasoning task, audio-visual question answering is attracting increasing attention from researchers (Li et al. 2022; Yang et al. 2022; Li, Hou, and Hu 2023; Li et al. 2024). It necessitates a comprehensive understanding and analysis of the audio-visual content in the video based on the given question, resulting in accurate answers. To explore the AVQA task, Li et al. (2022) constructed a MUSIC-AVQA dataset consisting of dynamic, long-term music performance videos and introduced a spatio-temporal grounding model to enhance the comprehension and reasoning of dynamic audio-visual scenes. Jiang and Yin (2023) considered introducing the topic words of the question into the visual modality and proposed a single-stream audio-visual question-answering network. Furthermore, Li, Hou, and Hu (2023) proposed a progressive spatio-temporal perception model PSTP-Net, which located the temporal segments and spatial regions related to the question by using questions as guiding factors. Li et al. (2024) considered mining question-related visual objects from object-level visual features. They applied the pre-trained target detection model to extract object-level features and identified key objects through the contrastive learning strategy. Lin et al. (2023) focused on enhancing the connections between visual and audio modality and optimizing training efficiency. They fine-tuned existing extraction models by adding efficient parameters to improve the adaptability of extracted features to downstream tasks.

Although promising progress has been made in the field of AVQA, some challenges still exist. Our work aligns audio-visual cues in temporal and spatial dimensions through contrastive learning, and adaptively assigns fusion weights to visual and audio modalities according to the question, to effectively encode the multi-modal relationships among visual, audio, and question.

Method

To address the challenges mentioned, we introduce an effective Audio-Visual Adaptive Fusion Network (AVAF-Net), which aligns audio-visual information in temporal and spatial dimensions and dynamically adjusts attention to visual and auditory modalities. An overview of AVAF-Net is shown in Fig. 2.

Data Representation

Firstly, we first down-sample videos to T non-overlapping, one-second-long visual and auditory segments $\{v_t, a_t\}_{t=1}^{T}$.

Visual Representation. For each one-second-long visual segment v_t, we sample a fixed number of frames from it. Then, we apply the pre-trained CLIP's vision encoder (Radford et al. 2021) with frozen parameters to extract token-level features f_p^t on each frame, where $f_p^t \in \mathbb{R}^{N\times D}$, N is the token numbers of one frame and D is the feature dimension. Finally, all T seconds of visual features can be denoted as $F_p = \{f_p^1, f_p^2, ..., f_p^T\}$, where $F_p \in \mathbb{R}^{T\times N\times D}$.

Audio Representation. For audio segment a_t, the pre-trained VGGish (Gemmeke et al. 2017) with frozen parameters is used as extractor to extract audio features as $f_a^t \in \mathbb{R}^D$, which is trained on the large-scale AudioSet (Gemmeke et al. 2017) dataset. Finally, all T seconds of audio features can be denoted as $F_a = \{f_a^1, f_a^2, ..., f_a^T\}$, where $F_a \in \mathbb{R}^{T\times D}$.

Question Representation. Given a question Q, we divide it into M individual words q_m through the Tokenizer and then embed each word into a fixed-length vector. Finally, it is fed back to the pre-trained CLIP's text encoder (Radford et al. 2021) to generate sentence-level features F_q and word-level features F_w, where $F_q \in \mathbb{R}^{1\times D}$ and $F_w \in \mathbb{R}^{M\times D}$. We take the [CLS] token as the sentence-level feature.

Temporal Alignment Contrastive Learning

Aligning visual and audio streams in the temporal dimension helps the network better understand contextual information and learn the temporal correlation between visual and audio in dynamic audio-visual scenes. We design the Temporal Alignment Contrastive Learning (TACL) method to encourage positive cross-modal pairs to be mapped nearby, while negative pairs are as far away as possible in the shared semantic space. We first adopt two linear functions to project the patch-level visual features $F_p \in \mathbb{R}^{T\times N\times D}$ and audio features $F_a \in \mathbb{R}^{T\times D}$ into the semantic space of the same dimension, and obtain the frame-level visual features $F_v = \{f_v^1, f_v^2, ..., f_v^T\} \in \mathbb{R}^{T\times D}$ by average pooling. Then, we calculate the interaction matrix of F_p and F_a and average it to obtain the vocalization visual features $F_s = \{f_s^1, f_s^2, ..., f_s^T\} \in \mathbb{R}^{T\times D}$. For the sample at time i, the temporal alignment contrastive loss is calculated as follows:

$$\mathcal{L}_{\text{Temp1}} = \frac{1}{T}\sum_{j}^{T}(\mathcal{L}_d(sim(f_a^i, f_v^j), sim(f_a^i, f_a^j)) + \mathcal{L}_d(sim(f_a^i, f_v^j), sim(f_v^i, f_v^j))) \quad (1)$$

$$\mathcal{L}_{\text{Temp2}} = \frac{1}{T}\sum_{j}^{T}\mathcal{L}_d(sim(f_a^i, f_a^j), sim(f_s^i, f_s^j)) \quad (2)$$

$$\mathcal{L}_{\text{Temp}} = \mathcal{L}_{\text{Temp1}} + \mathcal{L}_{\text{Temp2}} \quad (3)$$

where $sim(\cdot)$ is the similarity function, $\mathcal{L}_d(\cdot)$ represents the $L1$ distance, and T denotes the length of the audio-visual sequence.

Audio-Visual Clue Mining

Dynamic audio-visual scenes contain much audio-visual information, so question-related key clues should be mined in the visual and audio modalities to answer questions accurately. For example, for the question "Is the cello in the video

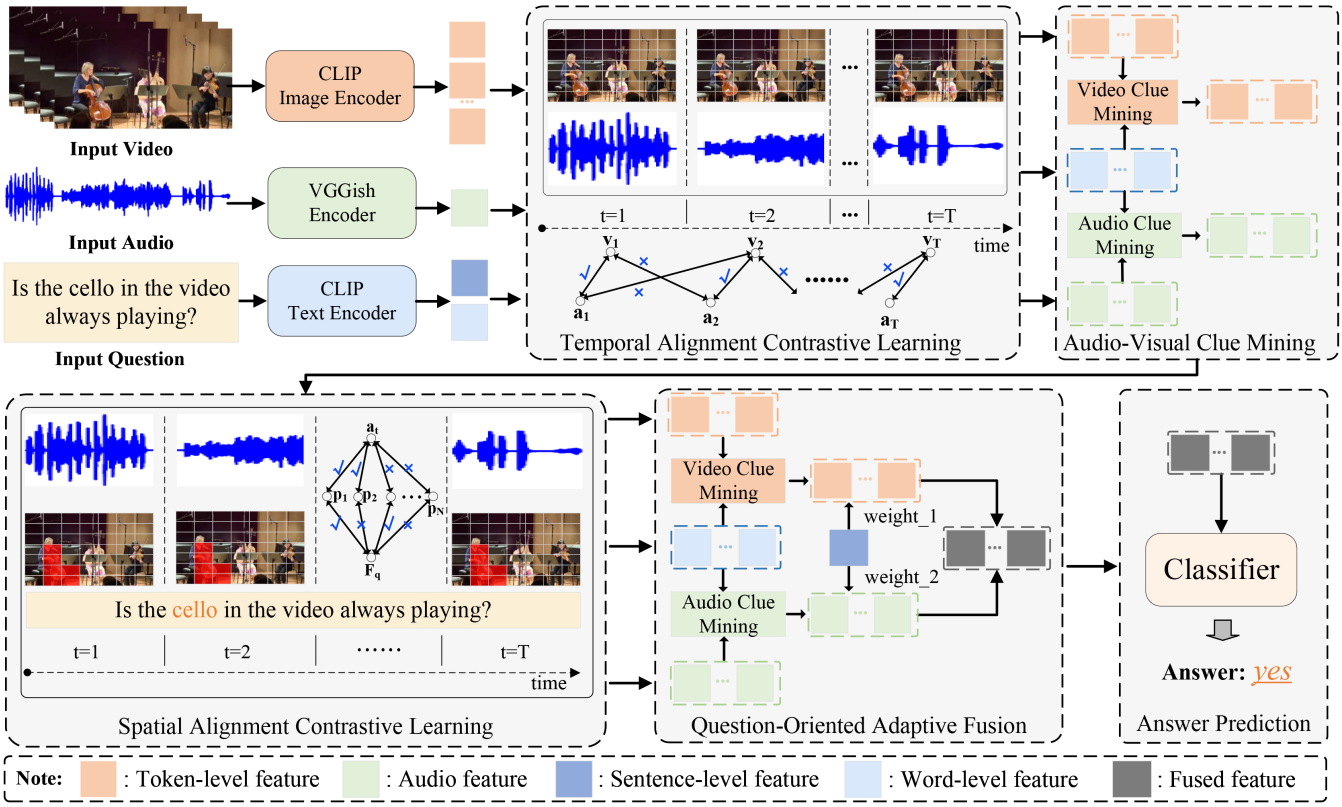

Figure 2: An overview of our proposed Audio-Visual Adaptive Fusion Network (AVAF-Net). First, we use the pre-trained models to extract multi-modality features. Then, we align the visual and audio features in the temporal dimension through TACL. Subsequently, we highlight the question-related audio-visual cues through the AVCM module. After that, the SACL is used to make the network focus on the matched question-visual region pairs and audio-visual region pairs in the spatial dimension. Finally, we mine the question-related audio-visual cues and adaptively fuse the audio-visual features based on the question content.

always playing?". The visual modality should mine clues related to "cell", while the corresponding audio modality should mine clues related to "always playing". Specifically, we use the Audio-Visual Clue Mining (AVCM) module to obtain question-related audio-visual clues.

The AVCM mainly comprises two multi-head attention (MultiHead) and two feed-forward neural networks (FFN), which model the cross-modal relations between the visual or audio features and the question, thereby mining key audio-visual clues. Overall, the calculation process is as follows:

$$F'_p = \mathrm{FFN}(\mathrm{MultiHead}(F_p, F_w, F_w)) \tag{4}$$

$$F'_a = \mathrm{FFN}(\mathrm{MultiHead}(F_a, F_w, F_w)) \tag{5}$$

Here, F_p, F_a, and F_w denote token-level visual features, audio features, and word-level features, respectively. Consequently, we obtain the updated audio feature $F'_a \in \mathbb{R}^{T\times D}$ and token-level visual feature $F'_p \in \mathbb{R}^{T\cdot N\times D}$.

Spatial Alignment Contrastive Learning

To better answer questions, the network should learn positive and question-related semantic information in multi-modalities. Based on this, we consider the learning optimization of the network from two aspects. On the one hand, the network should concentrate on the visual areas related to the question. On the other hand, it should also associate audio signals with corresponding visual areas. We exploit the contrastive loss proposed by Li et al. (2024) to design the Spatial Alignment Contrastive Learning (SACL) method, which optimizes the network by highlighting the matched question-visual area pairs and audio-visual area pairs.

Taking the question-visual area contrast learning as an example, for the audio-visual sequence at the i-th moment, the question-visual area contrast loss formula is calculated as follows:

$$s_i^{qp} = softmax(\frac{F_q}{\|F_q\|} \otimes (\frac{f_p^{i\prime}}{\|f_p^{i\prime}\|})^\top) \tag{6}$$

$$\begin{cases} \mathbf{A}_i^{qp} = \{s_i^{qp} | s_i^{qp} > \varphi\}, \\ \mathbf{B}_i^{qp} = \{s_i^{qp} | s_i^{qp} \le \varphi\}, \end{cases} \tag{7}$$

$$\begin{cases} l_i^{qp} = -\log \frac{\sum_{k=1}^{K} \exp(s_{i,k}^{qp,+}/\tau)}{\sum_{m=1}^{M} \exp(s_{i,m}^{qp,+}/\tau) + \sum_{j=1}^{N-M} \exp(s_{i,j}^{qp,-}/\tau)} \\ \mathcal{L}_{\mathrm{Spat1}} = \frac{1}{T} \sum_{i=1}^{T} l_i^{qp} \end{cases} \tag{8}$$

Here, $\otimes$ and φ denote the matrix multiplication and threshold. $\mathbf{A}_i^{qp}$ and $\mathbf{B}_i^{qp}$ represent the question-related visual regions and question-irrelated visual regions respectively, and $s_{i,m}^{qp,+} \in \mathbf{A}_i^{qp}, s_{i,j}^{po,-} \in \mathbf{B}_i^{qp}$. M denotes the number of question-related visual areas and N denotes the total number of visual areas. τ is a hyper-parameter that balances the computation of the loss.

Similarly, we can also calculate the audio-visual region contrast loss $\mathcal{L}_{\text{Spat2}}$. The final spatial alignment contrast loss is calculated as follows:

$$\mathcal{L}_{\text{Spat}} = \mathcal{L}_{\text{Spat1}} + \mathcal{L}_{\text{Spat2}} \tag{9}$$

Question-Oriented Adaptive Fusion

To reasonably fuse visual and audio features, as well as further mine question-related audio-visual clues to answer the question more accurately, we propose the Question-Oriented Adaptive Fusion (QOAF) module.

Firstly, we use the question information as the $query$ of the multi-head attention mechanism to further mine key audio-visual clues from the visual and audio modalities. The calculation formula is as follows:

$$F_p'' = \text{FFN}(\text{MultiHead}(F_w, F_p', F_p')) \tag{10}$$

$$F_a'' = \text{FFN}(\text{MultiHead}(F_w, F_a', F_a')) \tag{11}$$

Here, $F_p'' \in \mathbb{R}^{M \times D}$ and $F_a'' \in \mathbb{R}^{M \times D}$ denote the mined question-related visual clues and question-related audio clues, respectively. Next, to more reasonably integrate the visual and audio modalities, we expect to dynamically assign different weights to them. We generate modality-weights δ_1 and δ_2 according to the sentence-level question feature F_q. Then they are multiplied with the corresponding modal features. The calculation formula is as follows:

$$\delta_1, \delta_2 = \text{Weight}(F_q) \tag{12}$$

$$F_{fuse} = \text{FC}(\text{Concat}(\delta_1 \cdot F_p'', \delta_2 \cdot F_a'')) \tag{13}$$

Here, δ_1 and δ_2 represent the generated modality weight values. $F_{fuse} \in \mathbb{R}^{M \times D}$ denotes the fused audio-visual feature for answer prediction.

Answer Prediction

Following previous works (Li et al. 2022; Li, Hou, and Hu 2023; Li et al. 2024), the process of answer prediction is conducted in an open-ended manner, identifying the right answer from a given set of candidate answers. Specifically, we feed F_{fuse} into a linear classification layer and softmax layer to obtain the probability $p \in \mathbb{R}^C$ of the candidate answers, where C denotes the size of the candidate vocabulary. During training, we use the following loss function to optimize the network:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \cdot \mathcal{L}_{\text{Temp}} + \beta \cdot \mathcal{L}_{\text{Spat}} \tag{14}$$

Here, $\mathcal{L}_{CE}$ represents the cross entropy loss. Meanwhile, α and β denote the scaling factors used to balance the weights among the three losses. During inference, we use $\hat{c} = \arg\max_c(p)$ to predict the answer.

Experiments

In this section, we first introduce the used datasets, evaluation metrics, and experimental settings. Then, we demonstrate the effectiveness of our approach through comparative experiments, ablation experiments, and hyper-parameter exploration.

Dataset

In this paper, we validate our proposal on two datasets: MUSIC-AVQA (Li et al. 2022) and MUSIC-AVQA-Real. The detailed descriptions of these datasets are as follows:

MUSIC-AVQA. This dataset contains 9,288 music performer videos, with a total of 45,867 question-answering pairs. And 7,422 real videos were collected from YouTube and 1,867 human-made synthetic videos. The questions are designed based on 33 question templates, covering 9 types, including audio questions (*Counting* and *Comparative*), visual questions (*Counting* and *Localization*), audio-visual questions (*Existential, Localization, Counting, Comparative*, and *Temporal*).

MUSIC-AVQA-Real. Our goal with this dataset is to evaluate the performance of our network in a noise-free environment. It only consists of the real videos and the corresponding question-answer pairs, i.e., a real music performance scenario, extracted from the MUSIC-AVQA. We also compare it with that in a noisy environment (i.e., MUSIC-AVQA) to verify the robustness of the network.

Implementation Details

About the visual features, we initially split the video into one-second-long segments and sample frames at $1fps$. Then, we use the pre-trained CLIP-ViT model with frozen parameters to extract 512-D features for each visual segment. And the token-level visual features are regarded as visual input. For the audio input, we first sample it at 16kHz and then convert it to the corresponding spectrogram. Then, the VGGish network pre-trained on AudioSet is used to generate 128-D audio features. For question features, we use the CLIP-BERT model to obtain sentence-level and word-level features. In all experiments, we employ the Adam optimizer, initializing the learning rate at $1e$-4 and reducing it by a factor of 0.1 every 8 epochs. The batch size is set to 32, and running for 30 epochs. All our experiments are implemented on an NVIDIA GeForce RTX A6000 and implemented in PyTorch.

Evaluation Metrics

In this paper, following the definition of previous work, we verify the network performance on each type of question and take the average answer prediction accuracy as the main indicator to measure network performance, which is calculated by Eq. 15:

$$\textbf{Accuracy} = \frac{N_p}{N_{all}}, \tag{15}$$

Here, N_p represents the count of correctly predicted questions, while N_{all} donates the total number of questions.

Method	Audio			Visual			Audio-Visual						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
GRU (Agrawal et al. 2017)	74.55	66.80	72.89	68.32	73.46	70.83	84.25	65.75	62.52	62.12	65.83	68.43	69.75
HME (Fan et al. 2019)	77.47	64.75	74.74	66.45	67.97	67.19	81.66	66.70	56.44	63.27	60.03	66.34	67.88
MCAN (Yu et al. 2019)	80.49	50.82	74.12	71.87	71.69	71.78	82.90	69.35	55.59	57.13	60.50	65.71	68.73
PASC (Li et al. 2019)	77.58	63.93	74.65	68.41	69.25	68.82	78.85	68.66	55.87	61.91	60.82	65.66	67.94
AVSD (Schwartz, Schwing, and Hazan 2019)	74.10	64.75	72.10	65.79	75.51	70.54	82.11	66.07	62.80	63.27	65.83	68.27	69.50
Pano-AVQA (Yun et al. 2021)	76.91	65.57	74.47	69.25	77.18	73.12	82.56	67.02	62.09	64.93	67.55	69.12	71.08
ST-AVQA (Li et al. 2022)	80.49	63.11	76.76	72.71	76.69	74.65	82.45	65.77	71.16	64.72	70.38	71.05	72.95
PSTP-Net (Li, Hou, and Hu 2023)	75.32	**76.67**	75.42	77.29	62.04	75.03	73.77	72.61	**75.96**	**75.05**	**72.91**	**73.86**	74.07
QAGL (Chen et al. 2024)	84.75	67.62	81.07	77.38	78.55	77.95	**85.83**	75.19	69.02	64.20	68.97	72.91	75.60
APL (Li et al. 2024)	84.75	67.62	**81.07**	74.77	81.39	78.00	84.03	76.99	69.73	65.04	68.03	73.10	75.72
AVAF-Net	**85.65**	62.70	80.70	**80.84**	**85.11**	**82.93**	83.46	**77.31**	73.97	63.16	70.85	73.78	**77.45**

Table 1: Accuracy (%) of baselines and our AVAF-Net on the test set of MUSIC-AVQA-Real. The best and second-best results of each question type are highlighted in bold and underlined form, respectively.

Method	Audio			Visual			Audio-Visual						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
GRU (Agrawal et al. 2017)	72.21	66.89	70.24	67.72	70.11	68.93	81.71	62.64	59.44	61.88	60.07	65.18	67.07
HME (Fan et al. 2019)	74.76	63.56	70.61	67.97	69.46	68.76	80.30	63.19	53.18	62.69	59.83	64.05	66.45
MCAN (Yu et al. 2019)	77.50	55.24	69.25	71.56	70.93	71.24	80.40	64.91	54.48	57.22	47.57	61.58	65.49
PASC (Li et al. 2019)	75.64	66.06	72.09	68.64	69.79	69.22	77.59	63.42	55.02	61.17	59.47	63.52	66.54
AVSD (Schwartz, Schwing, and Hazan 2019)	72.41	61.90	68.52	67.39	74.19	70.83	81.61	63.89	58.79	61.52	61.41	65.49	67.44
Pano-AVQA (Yun et al. 2021)	74.36	64.56	70.73	69.39	75.65	72.56	81.21	64.91	59.33	64.22	63.23	66.64	68.93
ST-AVQA (Li et al. 2022)	76.89	67.00	73.25	72.60	75.76	74.19	81.98	68.54	64.35	**63.94**	65.57	68.92	71.08
PSTP-Net (Li, Hou, and Hu 2023)	80.24	58.92	72.38	75.19	79.27	77.25	83.40	70.83	64.02	61.67	67.40	69.51	72.07
QAGL (Chen et al. 2024)	82.99	**71.04**	78.58	80.12	77.88	78.89	82.29	72.73	62.83	63.40	64.36	69.43	73.58
APL (Li et al. 2024)	81.15	69.19	76.97	77.11	82.29	79.73	81.78	73.75	66.52	62.40	64.72	70.09	73.86
AVAF-Net	**83.09**	69.70	**78.15**	**80.20**	**84.49**	**82.37**	**84.51**	**75.05**	**68.37**	61.94	**70.07**	**72.12**	**75.90**

Table 2: Accuracy (%) of baselines and our AVAF-Net on the test set of MUSIC-AVQA.

Results

To verify the effectiveness of AVAF-Net, we test its performance on MUSIC-AVQA-Real and MUSIC-AVQA datasets and compare it with ten state-of-the-art models. Please note that the results of the comparison methods are all re-implemented based on the official codes. From Table 1 and Table 2, it can observed that the performance of AVAF-Net surpasses all the comparison models. In addition, we can find that on the dataset MUSIC-AVQA, which contains synthetic data, the performance of AVAF-Net decreases by 1.55%, while the performance of ST-AVQA, PSTP-Net, QAGL, and APL decreases by 1.87%, 2.00%, 2.02%, and 1.86%, respectively. This proves that our method is more robust when faced with noisy information, that is, one of the modalities is interfering information.

Moreover, we further explore and analyze the experimental results on the MUSIC-AVQA dataset. As shown in Table 2, AVAF-Net shows significant improvements in subtask types of audio, visual, and audio-visual, compared to the recent model APL (Li et al. 2024). Meanwhile, AVAF-Net achieves remarkable improvements of 1.18%, 2.64%, and 2.03% in the above-mentioned subtask types. Specifically, compared with APL, AVAF-Net obtains remarkable improvements of 1.94% and 0.51% in the *Counting* and *Comparative* subtasks of the audio modality. Additionally, 3.09% and 2.20% performance boosting are obtained in the *Counting* and *Localization* subtasks of the visual modality, respectively. Furthermore, the performance of our proposal has been improved by 2.73%, 1.30%, 1.85%, and 5.35% in *Existential*, *Localization*, *Counting*, and *Temporal* subtask types of audio-visual modality. The remarkable improvements indicate that our AVFA-Net is effective in accurately identifying crucial audio-visual clues.

Overall, our AVAF-Net achieves significant performance improvements compared to existing methods and contributes to the advancement of dynamic audio-visual question answering scene tasks.

	Method	Average Accuracy (%)			
		Audio	Visual	Audio-Visual	All
1	AVAF-Net w/o. all	46.87	16.60	25.12	26.70
2	AVAF-Net w/o. TACL	78.27	82.20	71.17	75.35
3	AVAF-Net w/o. AVCM	75.11	77.58	67.78	71.67
4	AVAF-Net w/o. SACL	77.28	82.62	71.55	75.50
5	AVAF-Net w/o. QOAF	75.36	77.50	69.41	72.78
6	AVAF-Net	78.15	82.37	72.12	75.90

Table 3: Ablation study on the different modules of AVFA-Net on the test set of MUSIC-AVQA.

Ablation Studies

In this section, we will mainly explore the effects of different hyper-parameters and modules on the performance of AVAF-Net based on the MUSIC-AVQA dataset.

To explore the effectiveness of each module in AVAF-Net, we remove TACL, AVCM, SACL, and QOAF respectively, and re-evaluate their performance. As shown in Table 3, after removing different components, AVAF-Net shows a different performance drop. The analysis of the ablation experiments is as follows:

- **AVFA-Net w/o. all.** To verify that the performance improvement is the result of the combined effect of the dif-

ferent components we proposed, we remove all the designed modules and only keep the input audio, video, and question features. As shown in Table 3, the performance dropped significantly (75.90% vs 26.70%). This significant deterioration is convincing evidence that the complexly designed multiple modules in AVAF-Net contribute to enhancing the overall effectiveness.

- **AVFA-Net w/o. TACL.** The goal of designing TACL is to enable the network to align visual and audio features in the temporal dimension, thereby enhancing the ability to understand audio-visual scenes. To verify the necessity of TACL, we removed it from AVAF-Net and re-evaluated the performance of the network. As shown in Table 3, when TACL is removed, the performance drops to 75.35%, a decrease of 0.55%.
- **AVFA-Net w/o. AVCM.** The purpose of AVCM is to discover question-related audio-visual clues in the audio-visual scenes, making the network focus on the required information and reduce the interference of irrelevant information. To illustrate the importance of AVCM, we re-test the model without the AVCM module. The results are shown in Table 3, the performance dropped by 4.23%, compared with AVAF-Net. Furthermore, remarkable performance declines are observed in the three subtask types, demonstrating that the AVCM module contributes to enhancing performance.
- **AVFA-Net w/o. SACL.** SACL is designed to encourage the network to focus on semantically matched question-visual region pairs and audio-visual region pairs in the spatial dimension. Removing this module means that the network will lack attention to the matched question-visual region pairs and audio-visual region pairs in the spatial dimension, which may lead to a decrease in the performance of the AVAF-Net. As shown in the table 3, without SACL, the performance drops to 75.50%.
- **AVFA-Net w/o. QOAF.** We design the QOAF to further explore the key clues in the audio-visual scenes and fuse the visual and audio modalities more reasonably according to the content of the question. To prove the effectiveness of the QOAF, we remove it and retest the performance of the network. Specifically, we directly fuse the visual and audio modalities without assigning weights. As shown in table 3, when QOAF is removed, the performance drops from 75.90% to 72.78%. This significant drop proves that QOAF has made a great contribution to performance improvement.

In summary, each module in AVAF-Net plays a role in performance improvement. The optimal results can be achieved when all modules are presented.

Meanwhile, we also explored the effects of different hyper-parameter configurations on the performance of AVAF-Net.

- **Effects of Threshold φ in Eqn. 9.** Here, φ is the threshold used to select highly matched question-visual region pairs and audio-visual region pairs in Equ. 9. As shown in table 4, we explore the impact of different φ values on the performance according to the number of patches N. When the value of φ is set to 0.019, the overall average performance reaches the highest. If the value of φ is large, the network may not be able to obtain enough clues. On the contrary, if the value of φ is small, the network's attention may be distracted.

	Threshold φ	Average Accuracy (%)			
		Audio	Visual	Audio-Visual	All
1	φ = 0.000	77.28	82.62	71.55	75.50
2	φ = 0.017	77.65	82.37	71.53	75.48
3	φ **= 0.019**	**78.15**	82.37	**72.12**	**75.90**
4	φ = 0.021	77.59	**82.82**	71.88	75.79
5	φ = 0.023	77.34	82.29	71.76	75.54
6	φ = 0.025	78.09	82.41	71.80	75.73
7	φ = 0.027	77.72	82.37	71.68	75.58

Table 4: Impact of the threshold φ in spatial alignment contrastive learning. In this experiment, each frame is divided into N patches, and we explore φ around $\frac{1}{N}$.

	Scaling factors β	Average Accuracy (%)			
		Audio	Visual	Audio-Visual	All
1	β = 0.1	**78.27**	82.54	71.74	75.76
2	β = 0.2	78.15	82.37	72.12	**75.90**
3	β = 0.3	77.78	**82.82**	71.76	75.76
4	β = 0.4	77.34	82.37	71.98	75.68
5	β = 0.5	76.29	82.29	71.82	75.39
6	β = 0.6	76.29	82.25	70.92	74.87
7	β = 0.7	73.36	81.63	71.09	74.64
8	β = 0.8	74.43	81.79	**72.29**	74.07

Table 5: Impact of the scaling factor β in the loss function.

- **Effects of β in Eqn. 14.** Here, β is a scaling factor used to balance the weights of different loss functions. As shown in table 5, we explore the impact of different β values on network performance from 0.1 to 0.8. It can be observed that when the value of β is 0.2, the best performance is achieved. Then the performance gradually decreases as the value of β increases.

Conclusion

In this work, we propose a novel audio-visual adaptive fusion network based on contrastive learning for audio-visual question-answering tasks. The proposed network aligns visual and audio information in temporal and spatial dimensions through the contrastive learning strategy, and also adaptively assigns different weights to the visual and audio modalities according to the type of question, thereby improving the accuracy of answer prediction.

We conduct extensive evaluation experiments and compare our method with 10 existing models. The experimental results indicate that AVAF-Net outperforms all the comparison models and obtains an accuracy of 75.90%, proving the effectiveness of the proposed network. To verify the impact of different modules and hyper-parameters of AVAF-Net, we also conduct a large number of ablation experiments. The results of the ablation experiments demonstrate the effectiveness of AVAF-Net.

Acknowledgments

This paper is supported by the Sichuan Science and Technology Program (2024YFFK0120), the Major Project of Sichuan Provincial Natural Science Foundation (25ZNSFSC0022), and the National Science Foundation of China (62072419).

References

Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Parikh, D.; and Batra, D. 2017. VQA: Visual Question Answering. *IJCV*, 123: 4–31.

Chen, Z.; Wang, L.; Wang, P.; and Gao, P. 2024. Question-Aware Global-Local Video Understanding Network for Audio-Visual Question Answering. *IEEE Trans. Circuits Syst. Video Technol.*, 34(5): 4109–4119.

Fan, C.; Zhang, X.; Zhang, S.; Wang, W.; Zhang, C.; and Huang, H. 2019. Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering. In *CVPR*, 1999–2007.

Gan, C.; Huang, D.; Zhao, H.; Tenenbaum, J. B.; and Torralba, A. 2020. Music Gesture for Visual Sound Separation. In *CVPR*, 10475–10484.

Gao, R.; Oh, T.; Grauman, K.; and Torresani, L. 2020. Listen to Look: Action Recognition by Previewing Audio. In *CVPR*, 10454–10464.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 776–780. IEEE.

Hu, D.; Wang, Z.; Nie, F.; Wang, R.; and Li, X. 2023. Self-Supervised Learning for Heterogeneous Audiovisual Scene Analysis. *IEEE Trans. Multim.*, 25: 3534–3545.

Jiang, Y.; and Yin, J. 2023. Target-Aware Spatio-Temporal Reasoning via Answering Questions in Dynamic Audio-Visual Scenarios. In *EMNLP*, 9399–9409.

Li, G.; Hou, W.; and Hu, D. 2023. Progressive Spatio-temporal Perception for Audio-Visual Question Answering. In *ACM MM*, 7808–7816.

Li, G.; Wei, Y.; Tian, Y.; Xu, C.; Wen, J.; and Hu, D. 2022. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. In *CVPR*, 19086–19096.

Li, X.; Song, J.; Gao, L.; Liu, X.; Huang, W.; He, X.; and Gan, C. 2019. Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering. In *AAAI*, 8658–8665.

Li, Z.; Guo, D.; Zhou, J.; Zhang, J.; and Wang, M. 2024. Object-Aware Adaptive-Positivity Learning for Audio-Visual Question Answering. In *AAAI*, 3306–3314.

Lin, Y.; Sung, Y.; Lei, J.; Bansal, M.; and Bertasius, G. 2023. Vision Transformers are Parameter-Efficient Audio-Visual Learners. In *CVPR*, 2299–2309.

Liu, J.; Wang, Y.; Ju, C.; Ma, C.; Zhang, Y.; and Xie, W. 2024. Annotation-free Audio-Visual Segmentation. In *WACV*, 5592–5602.

Mo, S.; and Tian, Y. 2022. Multi-modal Grouping Network for Weakly-Supervised Audio-Visual Video Parsing. In *NeurIPS*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Schwartz, I.; Schwing, A. G.; and Hazan, T. 2019. A Simple Baseline for Audio-Visual Scene-Aware Dialog. In *CVPR*, 12548–12558.

Senocak, A.; Oh, T.; Kim, J.; Yang, M.; and Kweon, I. S. 2018. Learning to Localize Sound Source in Visual Scenes. In *CVPR*, 4358–4366.

Sun, L.; Lian, Z.; Liu, B.; and Tao, J. 2024. HiCMAE: Hierarchical Contrastive Masked Autoencoder for self-supervised Audio-Visual Emotion Recognition. *Inf. Fusion*, 108: 102382.

Sun, W.; Zhang, J.; Wang, J.; Liu, Z.; Zhong, Y.; Feng, T.; Guo, Y.; Zhang, Y.; and Barnes, N. 2023. Learning Audio-Visual Source Localization via False Negative Aware Contrastive Learning. In *CVPR*, 6420–6429.

Tian, Y.; Li, D.; and Xu, C. 2020. Unified Multisensory Perception: Weakly-Supervised Audio-Visual Video Parsing. In *ECCV*, volume 12348, 436–454.

Wang, Y.; Liu, W.; Li, G.; Ding, J.; Hu, D.; and Li, X. 2024. Prompting Segmentation with Sound Is Generalizable Audio-Visual Source Localizer. In *AAAI*, 5669–5677.

Wei, Y.; Hu, D.; Tian, Y.; and Li, X. 2022. Learning in Audio-visual Context: A Review, Analysis, and New Perspective. *CoRR*, abs/2208.09579.

Yang, P.; Wang, X.; Duan, X.; Chen, H.; Hou, R.; Jin, C.; and Zhu, W. 2022. AVQA: A Dataset for Audio-Visual Question Answering on Videos. In *ACM MM*, 3480–3491.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep Modular Co-Attention Networks for Visual Question Answering. In *CVPR*, 6281–6290.

Yun, H.; Yu, Y.; Yang, W.; Lee, K.; and Kim, G. 2021. Pano-AVQA: Grounded Audio-Visual Question Answering on 360° Videos. In *ICCV*, 2011–2021.

Zhou, D.; Zhou, X.; Hu, D.; Zhou, H.; Bai, L.; Liu, Z.; and Ouyang, W. 2022a. SepFusion: Finding Optimal Fusion Structures for Visual Sound Separation. In *AAAI*, 3544–3552.

Zhou, J.; Guo, D.; Mao, Y.; Zhong, Y.; Chang, X.; and Wang, M. 2024a. Label-anticipated Event Disentanglement for Audio-Visual Video Parsing. In *ECCV*, 1–22.

Zhou, J.; Guo, D.; and Wang, M. 2023. Contrastive positive sample propagation along the audio-visual event line. *TPAMI*, 7239–7257.

Zhou, J.; Guo, D.; Zhong, Y.; and Wang, M. 2024b. Advancing Weakly-Supervised Audio-Visual Video Parsing via Segment-wise Pseudo Labeling. *IJCV*, 1–22.

Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y.

2024c. Audio-visual segmentation with semantics. *IJCV*, 1–21.

Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2022b. Audio–visual segmentation. In *ECCV*, 386–403.

Zhou, J.; Zheng, L.; Zhong, Y.; Hao, S.; and Wang, M. 2021. Positive sample propagation along the audio-visual event line. In *CVPR*, 8436–8444.