SaSR-Net: Source-Aware Semantic Representation Network for Enhancing Audio-Visual Question Answering

Anonymous ACL submission

Abstract

Audio-Visual Question Answering (AVQA) is a challenging task that involves answering questions based on both auditory and visual information in videos. A significant challenge is interpreting complex multi-modal scenes, which include both visual objects and sound sources, and connecting them to the given question. In this paper, we introduce the Source-aware Semantic Representation Network (SaSR-Net), a novel model designed for AVQA. SaSR-Net utilizes *source-wise learnable tokens* to efficiently capture and align audio-visual elements with the corresponding question. It streamlines the fusion of audio and visual information using spatial and temporal attention mechanisms to identify answers in multi-modal scenes. Extensive experiments on the Music-AVQA and AVQA-Yang datasets show that SaSR-Net outperforms state-of-the-art AVQA methods. We will release our source code and pre-trained models.

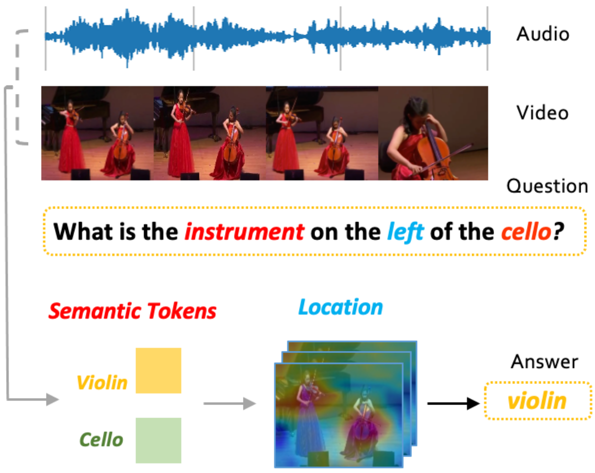

Figure 1: Leveraging semantic representation for AVQA involves: (1) Extracting features of various instrument types based on semantic tokens, (2) Identifying the location of the relevant sounding instruments, and (3) Establishing connections between the extracted semantic features, identified instrument locations, and the crucial parts of the question, guiding the model to answer the question accurately.

1 Introduction

Recent contributions to the field of audio-visual question answering (AVQA) include the creation of diverse datasets and sophisticated models (Yun et al., 2021; Yang et al., 2022; Li et al., 2022, 2023; Jiang and Yin, 2023). For example, the Pano-AVQA dataset (Yun et al., 2021) contains 360-degree videos paired with corresponding QA sets, while the AVQA-Yang dataset (Yang et al., 2022) is designed for answering audio-visual questions in real-world scenarios. The MUSIC-AVQA dataset (Li et al., 2022) further broadened the research scope by focusing on spatio-temporal understanding in audio-visual scenes. This dataset uses a dual attention mechanism, identifying sound-producing areas visually first and then applying attention for spatio-temporal reasoning. More recently, PSTP-Net (Li et al., 2023) was introduced, which progressively identifies key regions relevant to audio-visual questions using refined attention mechanisms.

Existing AVQA methods typically employ general audio and visual encoders to extract features from videos. However, this strategy often fails to link certain sound-producing objects in the video with the responses. Consider questions like ***What is the instrument on the left of the cello?*** which necessitates specific type and location awareness, as shown in Fig. 1. Current models often find it difficult to associate the *cello* mentioned in the question with its actual representation in the video scene.

To address these challenges, we propose the Source-aware Semantic Representation Network (SaSR-Net). This model enhances the understanding and integration of individual sound sources and visual objects in AVQA by two strategies: (1) **Source-wise Learnable Tokens:** Embedded within the Source-aware Semantic Representation Block, these tokens capture essential semantic features from both audio and visual data. This facilitates precise alignment and enhances semantic richness, enabling the model to accurately associate auditory and visual elements based on the query context. (2) **Attention Mechanisms:** The model utilizes spatial and temporal attention mechanisms to identify and synchronize relevant visual and audio regions with the query. This not only enhances the accuracy of localization but also strengthens cross-modal associations, crucial for forming a coherent understanding of the scene.

The efficacy of SaSR-Net is demonstrated by its performance on the Music-AVQA (Li et al., 2022) and AVQA-Yang (Yang et al., 2022) datasets, where it surpasses state-of-the-art AVQA approaches. The results highlight the effectiveness of the model's source-aware and semantically driven approach in managing complex audio-visual data. Our key contributions are as follows:

1. We introduce SaSR-Net, a novel framework that enriches the understanding of sound and visual information, leveraging Source-wise Learnable Tokens to extract semantic-aware audio and visual representations for AVQA.
2. SaSR-Net integrates multi-modal spatial and temporal attention mechanisms to adaptively leverage visual and audio information in videos for accurate scene understanding.
3. Our extensive experiments and ablation studies can validate the effectiveness of our proposed method.

2 Related Works

Audio-Visual Scene Understanding: Audio-visual learning focuses on understanding and correlating information from both modalities, aiming to mimic the human's multi-modal perception. This field has been extensively researched in various directions, showing remarkable progress in tasks, *e.g.,* sound source localization (Hu et al., 2021; Liu et al., 2022; Qian et al., 2020; Mo and Tian, 2023), action recognition (Gao et al., 2020), event localization (Mahmud and Marculescu, 2023; Brousmiche et al., 2021; Tian et al., 2018; Zhou et al., 2021), video parsing (Wu and Yang, 2021; Tian et al., 2020; Rachavarapu et al., 2023), captioning (Iashin and Rahtu, 2020; Tian et al., 2019), separation (Gao and Grauman, 2021; Tian et al., 2021; Zhao et al., 2018; Chen et al., 2023), and dialog (Zhu et al., 2020; Alamri et al., 2019; Hori et al., 2019). Despite this progress, these models still face challenges in integrating the audio modality with visual scene understanding. Effectively leveraging both audio and visual inputs for comprehensive video understanding remains concern. It is essential to consider both audio and visual signals holistically for effective video comprehension. In this work, we propose using Source-wise Learnable Tokens to leverage semantically-aware representations for audio-visual scene understanding.

Audio-Visual Question Answering: Audio-Visual Question Answering (AVQA) integrates both modalities, offering a more holistic understanding of scenes. Recent efforts in AVQA include the introduction of datasets such as the Pano-AVQA dataset (Yun et al., 2021), which features 360-degree videos (Yun et al., 2021), the real-life AVQA-Yang dataset (Yang et al., 2022), and the MUSIC-AVQA dataset (Li et al., 2022), which focuses on various musical performances (Li et al., 2022). The MUSIC-AVQA *v2.0* dataset was recently introduced to further reduce dataset bias (Liu et al., 2024). Innovations like PSTP-Net (Li et al., 2023), which identifies key regions relevant to audio-visual questions through refined attention mechanisms, have been instrumental. Additionally, LAVISH (Lin et al., 2023) introduced a novel parameter-efficient framework for encoding audios and videos, enhancing the potential for practical applications. Despite these advancements, challenges remain in accurately learning video semantics, which can limit the effectiveness of AVQA. Our approach aims to enhance video understanding by modeling semantic entities and strengthening the connections between questions and video content, thereby achieving competitive accuracy.

3 The Proposed SaSR-Net

Given a video with both visual and audio tracks, along with a question related to the content within the video, the objective of the AVQA task is to predict an accurate answer response. To achieve this, we propose a novel SaSR-Net architecture. This model is designed to generate compact, semantic-aware embeddings by identifying salient sounding objects present in the audio-visual input that are relevant to the given query. The overview of our proposed framework is illustrated in Figure 2.

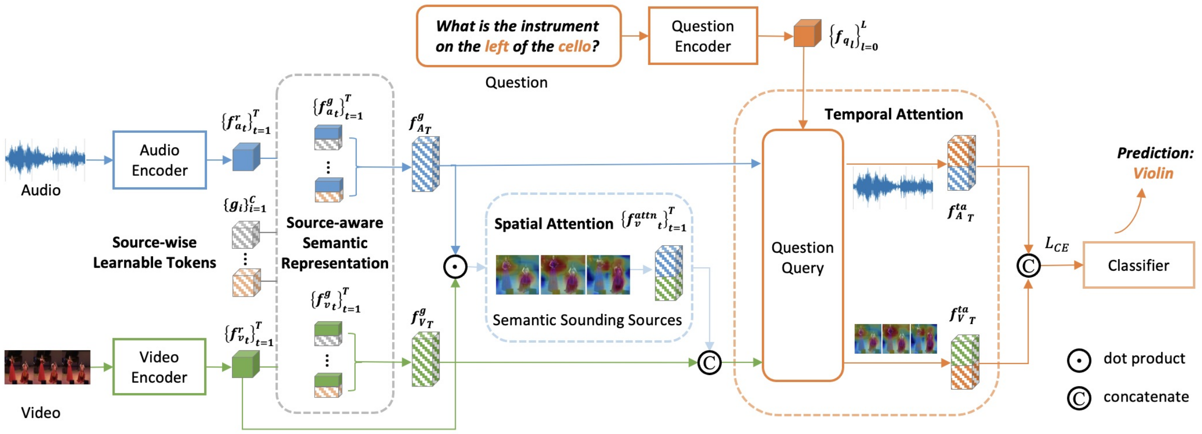

Figure 2: The architecture of the proposed SaSR-Net.

3.1 Representations for Different Modalities

Given a video with both visual and audio tracks, V_T and A_T, we split it into 1-second non-overlapping segment pairs $\{(v_t, a_t)\}_{t=1}^{T}$, where v_t and a_t are the video and audio clips during time $[t-1, t)$. Besides, each sample has a related question $Q_L = \{q_l\}_{l=1}^{L}$ and answer $\mathbf{y}$, *i.e.*, $(\{(v_t, a_t)\}_{t=1}^{T}, \{q_l\}_{l=1}^{L}, \mathbf{y})$, where q_l is a word and $\mathbf{y}$ is a one-hot encoding representing the correct answer.

Audio Feature: Each audio segment a_t is converted into a raw feature vector $\mathbf{f}_{a_t}^{r}$ using the pre-trained VGGish (Gemmeke et al., 2017) model, which works on transformed audio spectrograms. In all, the full audio will be transformed to a set of raw feature vectors $\mathbf{f}_{A_T}^{r} = \{\mathbf{f}_{a_t}^{r}\}_{t=1}^{T}$.

Visual Feature: Using ResNet-18 (He et al., 2016), we process the initial frames from V_T into raw vectors $\mathbf{f}_{V_T}^{r} = \{\mathbf{f}_{v_t}^{r}\}_{t=1}^{T}$ and feature maps $\mathbf{X}_{PT}^{r} = \{\mathbf{X}_{Pt}^{r}\}_{t=1}^{T} = \{\{\mathbf{x}_{pt}^{r}\}_{p=1}^{P}\}_{t=1}^{T}$, where p denotes positions on the feature maps, up to P positions.

Question Feature: For a question $Q_L = \{q_l\}_{l=1}^{L}$, word embeddings are passed through an LSTM. The resulting feature vectors $\mathbf{f}_{Q_L} = \{f_{q_l}\}_{l=1}^{L}$ are derived from the LSTM's final hidden state. Here, L is the max sequence length. The encoder is trained from scratch along with the entire model.

3.2 Source-wise Learnable Tokens

Distinguishing between audio sources and visual objects in videos fundamentally requires the association of these two modalities. A video may contain several visual objects and sound sources. To accurately respond to questions related to these video scenes, it is essential that our model effectively aligns and associates audio and visual content that are semantically synchronized. To achieve this, we introduce a series of Source-wise Learnable Tokens (SLT). Each token represents a distinct semantic category, such as a *guitar* or *piano*. These tokens will be utilized to align the two modalities and aggregate multimodal source-aware contexts for QA.

We denote Source-wise Learnable Tokens as $\mathbf{G}_C = \{\mathbf{g}_i\}_{i=1}^{C}$. Here, C represents the total number of distinct categories of sounding objects within our dataset.

Initially, we align the Source-wise Learnable Tokens with features from both video and audio by concatenating them. This computation will help ensure each token matches one of our intended categories, such as guitar or piano. To achieve this, we prepare category annotations in the labels and guide the model by applying penalties to the tokens during training. This will be elaborated in the following sections.

Subsequently, we apply self-attention $\mathrm{SelfAttn}$ to aggregate the auditory features $\mathbf{f}_{a_t}^{r}$ and visual features $\mathbf{f}_{v_t}^{r}$ separately. Here, we use the notation $[\mathbf{a}; \mathbf{b}]$ to represent the concatenation operation between tensor $\mathbf{a}$ and tensor $\mathbf{b}$, or the split operation between tensor $\mathbf{a}$ and tensor $\mathbf{b}$

$$[\mathbf{f}_{a_t}^{s}; \mathbf{G}_C^{a}] = \mathrm{SelfAttn}([\mathbf{f}_{a_t}^{r}; \mathbf{G}_C])$$

$$[\mathbf{f}_{v_t}^{s}; \mathbf{G}_C^{v}] = \mathrm{SelfAttn}([\mathbf{f}_{v_t}^{r}; \mathbf{G}_C])$$

After applying self-attention and splitting, we obtain source-aware audio embedding $\mathbf{f}_{a_t}^{s}$, source-aware visual embedding $\mathbf{f}_{v_t}^{s}$, and tokens $\mathbf{G}_C^{a}$ and $\mathbf{G}_C^{v}$. In detail, if we assume D is the dimension for each single feature embedding above, the self-attention $\mathbf{S}$ can be represented as ($\mathbf{f}$ is an input feature),

$$\mathbf{S}(\mathbf{f}) = \sigma\left(\frac{\mathbf{f} \cdot \mathbf{f}^{\intercal}}{\sqrt{D}}\right) \cdot \mathbf{f}$$

where σ is representing Softmax function.

The obtained representation $\mathbf{f}^s_{a_t}$, $\mathbf{f}^s_{v_t}$, $\mathbf{G}^a_C$ and $\mathbf{G}^v_C$ will be used next to compute the source-aware semantic representation.

3.3 Source-aware Semantic Representation

In this section, we assign semantic attention more directly and introduce training penalties to ensure that all learnable tokens accurately represent specific semantic categories. This design aims to improve our model's capability to precisely represent multi-modal scenes in videos and generate source-aware audio and visual semantic embeddings.

We introduce a source-aware semantic representation block. In the previous section, we have already got both semantically enriched audio and visual embeddings which enhanced with token information. Instead of treating the embeddings and Source-wise Learnable Tokens within the same modality as a single entity as we did in Sec. 3.2, we hope the model to learn specific information fusion / weighting relationships between the Source-wise Learnable Tokens and the embeddings. As a result, as for the audio/video features that are contained in the embedding and we are also interested in, the model will finally enhance them by properly-learned tokens. To achieve it, we will use our Source-aware Semantic Representation Block to perform cross attention from learnable tokens $\mathbf{G}^a_C$ and $\mathbf{G}^v_C$ to the semantically enriched audio and visual embeddings.

The resulting semantically-enriched audio embedding $\mathbf{f}^g_{A_T} = \{\mathbf{f}^g_{a_t}\}^T_{t=1}$ and video embedding $\mathbf{f}^g_{V_T} = \{\mathbf{f}^g_{v_t}\}^T_{t=1}$ are computed as the following equations performing cross-attention:

$$\begin{aligned}
\mathbf{G}^{a\prime}_C &= \mathbf{G}^a_C + \mathrm{FC}((\mathrm{CrossAttn}(\mathbf{G}^a_C, \mathbf{f}^s_{A_T})) \\
\mathbf{G}^{v\prime}_C &= \mathbf{G}^v_C + \mathrm{FC}((\mathrm{CrossAttn}(\mathbf{G}^v_C, \mathbf{f}^s_{V_T})) \\
\mathbf{f}^g_{A_T} &= \mathrm{FC}((\mathrm{CrossAttn}(\mathbf{f}^s_{A_T}, \mathbf{G}^{a\prime}_C)) \\
\mathbf{f}^g_{V_T} &= \mathrm{FC}((\mathrm{CrossAttn}(\mathbf{f}^s_{V_T}, \mathbf{G}^{v\prime}_C,))
\end{aligned}$$

where $\mathbf{f}^s_{A_T} = \{\mathbf{f}^s_{a_t}\}^T_{t=1}$, $\mathbf{f}^s_{V_T} = \{\mathbf{f}^s_{v_t}\}^T_{t=1}$, $\mathbf{G}^{a\prime}_C$ and $\mathbf{G}^{v\prime}_C$ are source-aware represented tokens, FC represents a fully-connected layer, LN is layer normalization, and the cross-attention works as:

$$\mathrm{CrossAttn}(\mathbf{a}, \mathbf{b}) = \sigma(\frac{\mathrm{FC}(\mathbf{a}) \cdot \mathrm{FC}(\mathbf{b})}{\sqrt{D}}) \cdot \mathrm{FC}(\mathbf{b})$$

The calculation of cross-attention for $\mathrm{CrossAttn}(\mathbf{G}^{a\prime}_C, \mathbf{f}^s_{A_T})$ and $\mathrm{CrossAttn}(\mathbf{G}^{v\prime}_C, \mathbf{f}^s_{V_T})$ follows the equations above. The fully-connected layer FC is used to align the dimensions of features from different latent spaces.

While the entire set of trainable parameters in SaSR-Net is optimized for minimizing the AVQA loss function that we will define later, it is also important to incorporate auxiliary loss functions specifically targeting the Source-wise Learnable Tokens. These additional loss functions are basically utilizing the prior knowledge to force the Source-wise Learnable Tokens to become the centroids in the hidden space. It will highlight the task-specific significance of these tokens, ensuring that they capture the characteristics of sound sources present in the audio and video. At last, they facilitate the extraction of more meaningful, source-aware representations, which are essential for the AVQA task.

The first auxiliary loss function is the binary cross-entropy (BCE) loss, which focuses on identifying individual sound sources' presence in the input audio and video channel,

$$\begin{aligned}
\mathcal{L}_{\text{source}} = \mathrm{BCE}(\sigma(\mathrm{FC}(\mathbf{G}^{a\prime}_C)), \mathbf{p}_C) + \\
\mathrm{BCE}(\sigma(\mathrm{FC}(\mathbf{G}^{v\prime}_C)), \mathbf{p}_C)
\end{aligned}$$

where $\mathbf{p}_C$ is the ground truth label for the source class. This label is compared against the predicted labels generated by applying the sigmoid activation function σ to a fully connected layer, operating on the semantically enriched audio embedding $\mathbf{f}^g_{A_T}$ and video embedding $\mathbf{f}^g_{V_T}$.

The second auxiliary loss function serves as a regularization term to ensure that each learned token uniquely represents a distinct type of sound source. Specifically, we aim for each token vector $\mathbf{g}_i$ to exclusively represent a single type of sound source. To achieve this, we define the loss using cross-entropy (CE) for sound source classification:

$$\mathcal{L}_{\text{reg}} = \mathrm{CE}(\mathrm{FC}(\mathbf{g}_i), \{c\}^C_{c=1})$$

3.4 Multi-modal Spacial Attention

One significant challenge involves localizing visual areas relevant to the given question in the AVQA task. This entails two tasks: firstly, identifying areas with key items by allocating reasonable spatial attention on the visual feature map, and secondly, establishing a temporal connection between the weighted feature map and the question.

Fortunately, the sections from 3.1 to 3.3 have already provided us with semantic-aware audio and visual embeddings. The semantic information in these embeddings proves beneficial in creating a meaningful association between the two modalities through shared semantic tokens.

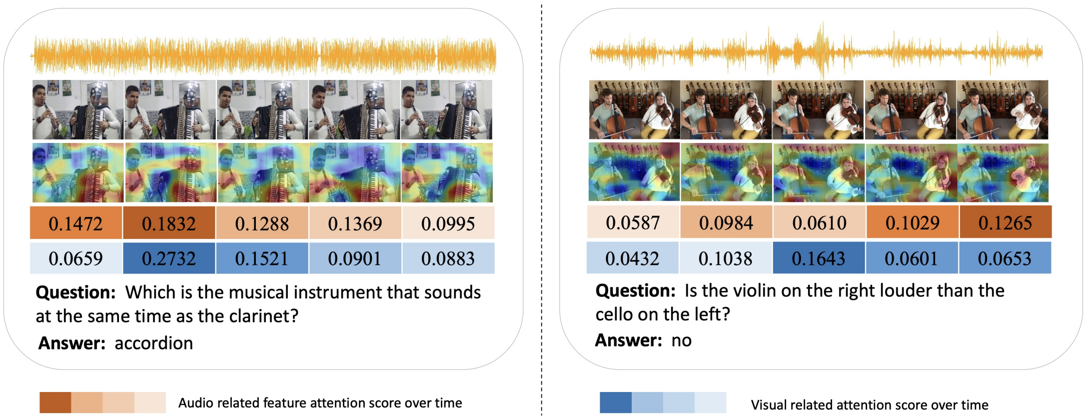

Figure 3: Visualization of Spatial Attention (SA) and Temporal Attention (TA) Blocks. The SA Block heatmaps pinpoint sounding object locations, and the TA Block displays audio-visual feature scores. SA localizes critical visual areas, while TA synchronizes video moments with questions, boosting overall audio-visual comprehension.

To address the first task, in our model, visual features differentiate semantic items from the background spatially based on their associated sounds. This involves applying a multi-modal spatial attention between the source-aware audio embedding $\mathbf{f}^g_{a_t}$ and the initial video encoding feature maps $\mathbf{X}^r_P$. By incorporating the source-aware video embedding $\mathbf{f}^g_{v_t}$, we derive the spatially-attended video representation $\mathbf{f}^{\text{sa}}_{v_t}$:

$$\mathbf{f}^{\text{attn}}_{v_t} = \sigma(\mathbf{X}^{r\,\intercal}_{Pt} \circledast \mathbf{f}^g_{a_t})) \cdot \mathbf{X}^r_{Pt}$$

$$\mathbf{f}^{\text{sa}}_{v_t} = \text{FC}(\tanh([\mathbf{f}^g_{v_t}; \mathbf{f}^{\text{attn}}_{v_t}]))$$

where $\circledast$ represents the convolution operation, which means this incorporating is broadcasting to all locations on the feature map.

In practice, based on the computations above, we also observed the presence of contrastive information, allowing the model to better learn how to accurately extract semantic object embeddings spacially on the feature maps. Essentially, it is crucial not only allow the model to learn how to successfully align visual and audio information but also to penalize those errors in cases where visual and audio inputs do not belong to the same scene at all. This will ultimately enhance SaSR-Net's spatial attention capabilities.

To achieve this, during training, we supplement both a matched (positive) audio-video pair $\{(v_t, a_t)\}_{t=1}^T$ along with a mismatched (negative) pair, $\{(v'_t, a_t)\}_{t=1}^T$, where v'_t is from a 1-second random video clip that belongs to a different video than a_t. Let $\mathbf{f}^{\text{sa}}_{v_t}$ be the spatially-attended representation for a matched sample, and $\mathbf{f}^{\text{sa}}_{v'_t}$ be that for a mismatched sample. For the optimization of the traning process, we employ a loss function to distinguish between matched and mismatched samples using a binary classifier:

$$\mathcal{L}_{\text{match}} = \text{CE}(\mathbf{f}^{\text{sa}}_{v_t}, 1) + \text{CE}(\mathbf{f}^{\text{sa}}_{v_t}, 0)$$

This optimization will make the learned representations more discriminative.

3.5 Multi-modal Temporal Attention

In this section, we address the second task outlined in Sec. 3.4.

Traditional QA methods treat questions as single entities, as in (Alamri et al., 2019). Our AVQA approach, however, utilizes the temporal sequences of data, such as frames and audio, to align questions with specific content moments. For example, a *violin* query directs the focus to relevant video segments. This alignment leads to contextually accurate responses by linking question tokens to the correct temporal embeddings.

To achieve this, we introduce multi-modal temporal attention block that employs cross-attention through $t = 0$ to $T - 1$ for updated audio embedding $\mathbf{f}^{\text{ta}}_{A_T}$ and visual embedding $\mathbf{f}^{\text{ta}}_{V_T}$ based on the question's embedding $\mathbf{f}_{Q_L}$. The cross attention is calculated as follows,

$$\mathbf{f}^{\text{ta}}_{A_T} = \sigma\left(\frac{\mathbf{f}_{Q_L} \mathbf{f}^{g\,\intercal}_{A_T}}{\sqrt{D}}\right)\mathbf{f}^g_{A_T}, \quad \mathbf{f}^g_{A_T} = \{\mathbf{f}^g_{a_t}\}_{t=1}^T$$

$$\mathbf{f}^{\text{ta}}_{V_T} = \sigma\left(\frac{\mathbf{f}_{Q_L} \mathbf{f}^{\text{sa}\,\intercal}_{V_T}}{\sqrt{D}}\right)\mathbf{f}^{\text{sa}}_{V_T}, \quad \mathbf{f}^{\text{sa}}_{V_T} = \{\mathbf{f}^{\text{sa}}_{v_t}\}_{t=1}^T$$

3.6 Answer Prediciton

To predict the final answer to the question, we utilize the multi-modal temporal embeddings and semantically-enriched embeddings, as they have already been proven to contain competent high-dimensional values after attention masks. The implementation includes a shortcut connection structure and a necessary fusion network.

For the shortcut connection structure, we (averagely) reduce the semantically-enriched embeddings across their time dimension and aggregate them with the multi-modal temporal embeddings, modality by modality. This operation is expected to help maintain global information and facilitate gradient back-propagation.

We hope the fusion network could integrate both the audio-text modal and visual-text modal into a final mixed modal that could be directly taken advantage of by its classifier and output predictions. Hence, we concatenate the two embeddings after the shortcut connection structure and employ a fully-connected layer as a classifier to predict the answer. The full operation is formulated as follows,

$$\mathbf{f}_{av} = \mathrm{FC}(\tanh([\mathbf{f}^{\mathrm{ta}}_{A_T} + \mathbf{f}^{g}_{A_T}; \mathbf{f}^{\mathrm{ta}}_{V_T} + \mathbf{f}^{g}_{V_T}]))$$

$$\hat{\mathbf{y}} = \sigma(\mathrm{FC}(\tanh(\mathbf{f}_{av} \cdot \mathbf{f}_{Q_L})))$$

Here $\mathbf{y}$ denotes the right answer id encoded by an one-hot vector.$\hat{\mathbf{y}}$ represents the probabilities of selection among all the answers, to match $\mathbf{y}$ closely. Therefore, we use cross-entropy loss for AVQA to penalize incorrect predictions,

$$\mathcal{L}_{\mathrm{avqa}} = \mathrm{CE}(\mathbf{y}, \hat{\mathbf{y}})$$

At last, the overall training loss is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{avqa}} + \lambda_1 \mathcal{L}_{\mathrm{source}} + \lambda_2 \mathcal{L}_{\mathrm{reg}} + \lambda_3 \mathcal{L}_{\mathrm{match}}$$

4 Experiment

4.1 Experiments Setting

Datasets: The MUSIC-AVQA dataset (Li et al., 2022) includes 9,290 videos, featuring 7,423 real and 1,867 synthetic examples, and 45,867 question-answer pairs. This dataset spans 9 audio-visual question types and 33 templates, showcasing 22 instruments categorized into Strings, Winds, Percussion, and Keyboards. Each video is annotated with instrument category labels. The dataset, designed for answering questions about the appearance, sounds, and associations of different objects in videos, is published under the Creative Commons Attribution-NonCommercial 4.0 International License and is public for research use. The question type primarily involves estimating answers.

The AVQA-Yang dataset (Yang et al., 2022) contains 57,015 videos paired with 57,335 questions that require understanding both audio and visual clues. The question type in this dataset is multiple-choice.

Implementation: The audio data has a sampling rate of 16 Hz, and video data has 1 fps. Videos are segmented into non-overlapping 1-frame segments, each yielding a $512D$ feature vector. We sample 1-second video segments every 6 seconds. Audio segments, also 1-second long, are processed using a linear layer, converting them from $128D$ VGGish features to $512D$ feature vectors. Word embeddings are set to 512 dimensions. Our batch size is 16, and we train for 80 epochs using the Adam optimizer with an initial learning rate of $1e-4$, which decreases by a factor of 0.3 every 16 epochs. Also, we set $\lambda_1 = \lambda_2 = \lambda_3 = 0.5$. Our model and related utility codes are based on PyTorch. We use *torchinfo* to summary our model's configuration. Our model contains 65,117,283 parameters (approximately 205.24 MB storage). We put our model trained as well as evaluated on an NVIDIA GeForce GTX 1080 Ti.

Evaluation: Following (Li et al., 2022), we use answer prediction accuracy as our evaluation metric.

4.2 Comparison to Prior Work

In this study, we introduced SaSR-Net, a novel multi-modal AVQA framework, and compared it with established unimodal and cross-modal question answering systems in Tab. 1 to demonstrate its effectiveness. The baselines include: (1) Audio Question Answering: FCNLSTM (Fayek and Johnson, 2020), CONVLSTM (Fayek and Johnson, 2020). (2) Visual Question Answering: HCAttn (Lu et al., 2016), MCAN (Yu et al., 2019) (3) Video Question Answering: PSAC (Li et al., 2019b), HME (Fan et al., 2019), HCRN (Le et al., 2020). (4) Audio-Visual Question Answering: AVSD (Schwartz et al., 2019), Pano-AVQA (Yun et al., 2021), AVST (Li et al., 2022). PSTP-Net (Li et al., 2023) and TJSTG (Jiang and Yin, 2023).

These baselines primarily use general encoders to extract video features, which are then processed through attention mechanisms for question answering. In contrast, our SaSR-Net uses Source-wise Learnable Tokens to extract semantically compact features from videos and employs Source-aware Semantic Representation to align these with visual and audio features. This enhances the model's capability to integrate and understand individual sound sources and visual objects in AVQA queries, enriching the features semantically.

SaSR-Net not only delivers robust performance in audio and visual QA but also showcases exceptional results in audio-visual QA, a domain where

Task	Method	Audio Question			Visual Question			Audio-Visual Question						All
		Count	Comp	Avg.	Count	Local	Avg.	Exist	Local	Count	Comp	Temp	Avg.	Avg.
AudioQA	FCNLSTM	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	73.55	67.17	71.20	67.17	55.84	61.44	82.49	63.08	51.85	62.13	50.36	62.56	63.79
VisualQA	HCAttn	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	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	HME	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	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	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	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
	AVST	77.78	67.17	73.87	73.52	75.27	74.40	82.49	69.88	64.24	64.67	65.82	69.53	71.59
	PSTP-Net	73.97	65.59	70.91	**77.15**	77.36	**77.26**	76.18	73.23	71.80	71.79	69.00	72.57	73.52
	TJSTG	**80.38**	**69.87**	**76.47**	76.19	**77.55**	76.88	**82.59**	71.54	64.24	66.21	64.84	70.13	73.04
	SaSR-Net(ours)	73.95	69.81	73.56	73.76	71.84	73.28	69.76	**73.43**	**73.64**	**79.15**	**77.46**	**74.66**	**74.21**

Table 1: Different methods on Music-AVQA dataset. The top-2 results are highlighted.

SLT	SaSR	Accuracy	Improvement
✗	✗	70.31%	-
✗	✓	71.78%	↑ 1.47%
✓	✗	72.16%	↑ 1.85%
✓	✓	74.21%	↑ 3.90%

Table 2: Ablation on Source-wise Learnable Tokens (SLT) and Source-aware Semantic Representation (SaSR)

TA	SA	Accuracy	Improvement
✗	✗	70.17%	–
✓	✗	72.03%	↑ 1.03%
✓	✓	74.21%	↑ 2.18%

Table 3: Ablation studies on Multi-modal Special Attention (SA), Multi-modal Temporal Attention (TA) blocks

previous AVQA methods have been less effective. We have made substantial improvements in this area. SaSR-Net excels particularly in Audio-Visual Questions, significantly outperforming AVST (Li et al., 2022) with notable improvements in **Counting (3.55%)**, **Localization (9.4%)**, **Comparative (14.48%)**, and **Temporal (11.64%)** questions. Moreover, our method surpasses AVSD by 9.22%, Pano-AVQA by 7.9%, AVST by 5.13%, PSTP-Net by 2.09%, and TJSTG by 4.53% in average accuracy,indicating a strong advancement in AVQA. In Audio QA, SaSR-Net achieves an average accuracy of 73.56%, exceeding specialized models like FCNLSTM and CONVLSTM.

These exceptional results provide strong evidence of the effectiveness of our proposed Source-wise Learnable Tokens and Source-aware Semantic Representation. By embedding audio and visual features with semantic context relevant to the queries, these innovations significantly enhance the representational capabilities of the framework. The effective use of Source-wise Learnable Tokensfacilitates a deeper integration of audio and visual modalities, allowing SaSR-Net to accurately identify and address complex multimodal interactions inherent in AVQA tasks.

4.3 Ablation Studies

In this section, we conducted ablation studies on Music-AVQA dataset to quantitatively evaluate the Source-wise Learnable Tokens (SLT) and the Source-aware Semantic Representation (SaSR) block, as presented in Table 2. Additionally, we performed ablation studies to quantitatively assess the Multi-modal Spacial Attention (SA) and Multi-modal Temporal Attention (TA) blocks, as presented in Table 3.

Effectiveness of SLT and SaSR: The inclusion and removal of the SLT (Source-wise Learnable Tokens) and SaSR (Source-aware Semantic Representation) blocks impact the performance of the AVQA model. Removing both blocks leads to a considerable accuracy drop to 70.31%. This decline occurs primarily because the model struggles to extract distinct semantic visual and auditory features without the SLT and fails to integrate these features without the SaSR, highlighting the critical roles these components play in comprehending complex audio-visual content. Conversely, introducing the SLT block in the baseline model increases the AVQA accuracy by 1.85%, demonstrating its effectiveness in enhancing video comprehension by extracting more semantic information from diverse sources. Additionally, retaining the SaSR block while eliminating the SLT block results in a 1.47% increase in accuracy, emphasizing the SaSR's crucial role in integrating diverse audio and visual features. More importantly, incorporating both SLT and SaSR into the model leads to a substantial improvement in accuracy by 3.90%. These findings underscore the importance of both SLT and SaSR in aligning auditory elements with their corresponding visual cues and enhancing the model's question-answering capabilities.

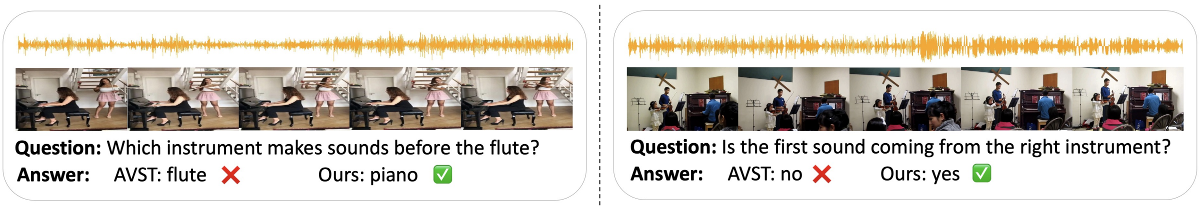

Figure 4: Comparison of our SaSR-Net and AVST (Li et al., 2022). Our SaSR-Net provides more precise answers to complex questions by effectively integrating semantic information into audio and visual features.

Method	Avg(%)
HME (Fan et al., 2019)+HAVF (Yang et al., 2022)	85.0
PSAC (Li et al., 2019b)+HAVF (Yang et al., 2022)	87.4
LADNet (Li et al., 2019a)+HAVF (Yang et al., 2022)	84.1
HGA (Jiang and Han, 2020)+HAVF (Yang et al., 2022)	87.7
HCRN (Le et al., 2020)+HAVF (Yang et al., 2022)	89.0
SaSR-Net(ours)	**89.9**

Table 4: Results of different methods on AVQA-Yang dataset.

Effectiveness of SA and TA: Removing the TA (Multi-modal Temporal Attention) and SA (Multi-modal Spatial Attention) blocks significantly reduces accuracy to 70.17%, underscoring their importance. Without SA, the model cannot accurately locate sounding instruments in videos, and without TA, it struggles to understand temporal dynamics, severely impairing its ability to identify key frames and localize sound sources. Introducing SA enhances the model's ability to link sounding objects with their sounds in complex scenes, improving spatial precision. Adding TA helps align temporal sequences, pinpointing key video frames relevant to the query. Together, SA and TA increase AVQA accuracy by 1.03%, highlighting their synergistic effect in boosting the model's comprehension of audio-visual content.

4.4 Visualization

Visualization of SA and TA: In Fig. 3, we visualize the results of the Spatial Attention and Temporal Attention Blocks.

Comparative Results: In Fig. 4, we present the results of our SaSR-Net method, compared with the results of AVST (Li et al., 2022). Our approach more accurately answers complex questions with specific semantic information due to our SLT and SaSR blocks. The SLT extracts and aggregates semantic category information from various sources, while the SaSR effectively integrates these semantic-aware features into both audio and visual features. These aggregated features outperform the original features, leading to superior performance.

Previous AVQA methods often fail to accurately associate visual objects with corresponding sounds in complex scenes, leading to incorrect answers. In contrast, our SaSR-Net, with its SLT and SaSR blocks, effectively connects sounding objects with mixed audio sources and accurately pinpoints their locations using spatial attention. It also employs temporal attention to identify key timestamps related to the posed question. This enhances the model's ability to map sound sources accurately, significantly improving audio-visual analysis in dynamic multi-modal environments.

4.5 Experiments on AVQA Dataset

While most existing methods are tested on the MUSIC-AVQA dataset (Li et al., 2022), we extend the validation of our method to the AVQA-Yang dataset (Yang et al., 2022) to further demonstrate its effectiveness. This confirms its applicability across different question formats and more complex scenarios. Following the approach in (Yang et al., 2022), we integrate various strategies (Fan et al., 2019; Li et al., 2019b,a; Jiang and Han, 2020; Le et al., 2020) with HAVF (Yang et al., 2022) as our evaluation metric. The comparative results in Table 4 show that our method outperforms others on the AVQA dataset. This underscores the robustness of our proposed SaSR-Net in diverse audio-visual question answering environments.

5 Conclusion

In this paper, we present SaSR-Net, a novel AVQA approach that introduces source-aware learnable tokens to effectively capture and integrate semantic-aware audio-visual representations. This enhances alignment between audio elements and visual cues, crucial for identifying relevant scene regions and their association with questions. By excelling at extracting and understanding single-source information within complex scenes, SaSR-Net significantly improves performance on AVQA tasks.

Limitation: SaSR-Net marks a transformative milestone in AVQA research. However, it may still face challenges in handling extremely noisy audio-visual data or scenarios with highly complex and overlapping audio sources, which could be areas for future improvement and research.

References

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *CVPR*.

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

Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. 2023. iquery: Instruments as queries for audio-visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14675–14686.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*.

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

Ruohan Gao and Kristen Grauman. 2021. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE.

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

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 ICASSP*, pages 776–780. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. 2019. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356. IEEE.

Di Hu, Yake Wei, Rui Qian, Weiyao Lin, Ruihua Song, and Ji-Rong Wen. 2021. Class-aware sounding objects localization via audiovisual correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9844–9859.

Vladimir Iashin and Esa Rahtu. 2020. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959.

Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11109–11116.

Yuanyuan Jiang and Jianqin Yin. 2023. Target-aware spatio-temporal reasoning via answering questions in dynamics audio-visual scenarios. *arXiv preprint arXiv:2305.12397*.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. In *CVPR*.

Guangyao Li, Wenxuan Hou, and Di Hu. 2023. Progressive spatio-temporal perception for audio-visual question answering. *arXiv preprint arXiv:2308.05421*.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*.

Xiangpeng Li, Lianli Gao, Xuanhan Wang, Wu Liu, Xing Xu, Heng Tao Shen, and Jingkuan Song. 2019a. Learnable aggregating net with diversity learning for video question answering. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1166–1174.

Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019b. Beyond rnns: Positional self-attention with co-attention for video question answering. In *AAAI*.

Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. 2023. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2309.

Xian Liu, Rui Qian, Hang Zhou, Di Hu, Weiyao Lin, Ziwei Liu, Bolei Zhou, and Xiaowei Zhou. 2022. Visual sound localization in the wild by cross-modal interference erasing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1801–1809.

Xiulong Liu, Zhikang Dong, and Peng Zhang. 2024. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4478–4487.

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

Tanvir Mahmud and Diana Marculescu. 2023. Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization. In *WACV*, pages 5158–5167.

Shentong Mo and Yapeng Tian. 2023. Audio-visual grouping network for sound localization from mixtures. In *CVPR*, pages 10565–10574.

Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. 2020. Multiple sound sources localization from coarse to fine. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 292–308. Springer.

Kranthi Kumar Rachavarapu et al. 2023. Boosting positive segments for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10192–10202.

Idan Schwartz, Alexander G Schwing, and Tamir Hazan. 2019. A simple baseline for audio-visual scene-aware dialog. In *CVPR*.

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

Yapeng Tian, Di Hu, and Chenliang Xu. 2021. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2745–2754.

Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 436–454. Springer.

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

Yu Wu and Yi Yang. 2021. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *CVPR*, pages 1326–1335.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3480–3491.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *CVPR*, pages 6281–6290.

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. 2021. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *CVPR*.

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

Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. 2021. Positive sample propagation along the audio-visual event line. In *CVPR*, pages 8436–8444.

Ye Zhu, Yu Wu, Yi Yang, and Yan Yan. 2020. Describing unseen videos via multi-modal cooperative dialog agents. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 153–169. Springer.