AV-Master: Dual-Path Comprehensive Perception Makes Better Audio-Visual Question Answering

Jiayu Zhang, Qilang Ye, Shuo Ye, Xun Lin, Zihan Song, Zitong Yu, *Senior Member, IEEE*

Abstract—Audio-Visual Question Answering (AVQA) requires models to effectively utilize both visual and auditory modalities to answer complex and diverse questions about audio-visual scenes. However, existing methods lack sufficient flexibility and dynamic adaptability in temporal sampling and modality preference awareness, making it difficult to focus on key information based on the question. This limits their reasoning capability in complex scenarios. To address these challenges, we propose a novel framework named AV-Master. It enhances the model's ability to extract key information from complex audio-visual scenes with substantial redundant content by dynamically modeling both temporal and modality dimensions. In the temporal dimension, we introduce a dynamic adaptive focus sampling mechanism that progressively focuses on audio-visual segments most relevant to the question, effectively mitigating redundancy and segment fragmentation in traditional sampling methods. In the modality dimension, we propose a preference-aware strategy that models each modality's contribution independently, enabling selective activation of critical features. Furthermore, we introduce a dual-path contrastive loss to reinforce consistency and complementarity across temporal and modality dimensions, guiding the model to learn question-specific cross-modal collaborative representations. Experiments on four large-scale benchmarks show that AV-Master significantly outperforms existing methods, especially in complex reasoning tasks.

Index Terms—Audio-visual question answering, multimodal fusion, collaborative learning.

I. INTRODUCTION

Humans perceive the world through various modalities, such as vision, hearing, and touch. Inspired by such multisensory experiences, researchers have increasingly focused on a range of multimodal understanding tasks [1], [2], [3]. Among these tasks, Audio-Visual Question Answering (AVQA) [4], [5], [6] is a practical task with broad application prospects. AVQA utilizes both visual and auditory modalities and requires the model to discover the associations between them to answer various questions. This process involves dynamically understanding audio-visual segments and addressing question-specific modality preferences, significantly increasing the complexity of the task.

For AVQA, we argue it is essential to focus on the following: *(i) How can the model identify the most relevant visual and audio segments of the video related to the given*

Jiayu Zhang, Zihan Song and Zitong Yu are with Great Bay University, and Dongguan Key Laboratory for Intelligence and Information Technology.
Qilang Ye is with Nankai University.
Shuo Ye is with Great Bay University. He is also with the Tsinghua Shenzhen International Graduate School, Tsinghua University
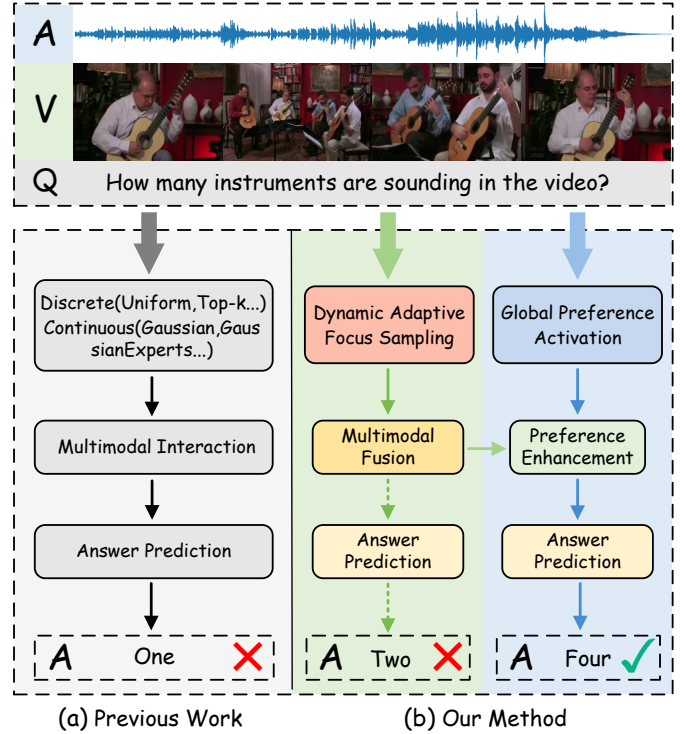Xun Lin is with The Chinese University of Hong Kong.

Fig. 1. Illustration of the AVQA task and the comparison of our method with previous work. Our method employs dynamic adaptive focus sampling to capture key audio-visual segments and predicts modal preferences through global preference activation to enhance the model, ultimately generating the correct answer.

question? (ii) How can the model extract cues from the most relevant modality information related to the given question? For (i), most current models employ discrete [7], [5] or continuous [6] sampling methods. Although these methods have achieved promising results, they still have notable limitations. The former compromises the inherent temporal cues in audio-visual segments, resulting in hallucinations during fine-grained understanding. The latter partially alleviates this issue but introduces significant redundancy, which hinders further enhancement of model performance. For (ii), most mainstream methods are not specifically designed to differentiate between various input modalities but instead treat them merely as supplementary information. A small number of researchers [8] have recognized these challenges and opted to design fusion weights for features from different modalities during the multimodal interaction process, thereby biasing the model towards a particular modality. However, this method overlooks the challenges in learning significant differences among fine-grained features and requires the model to reinterpret modality

preferences in complex multimodal semantics, without explicitly enhancing the learning process, thus providing limited guidance for improving the model's performance.

To solve these issues, we propose AV-Master, which can find the important fine-grained areas related to the current question within complex dynamic audio-visual scenes, and generate the optimal answer by integrating global preference information. As shown in Fig. 1, unlike previous work [5], [6], AV-Master employs dynamic adaptive focus sampling to extract the fine-grained focus features from the audio-visual segments relevant to the current questioning scenario. The focus sampling combines the advantages of discrete (*e.g.*, uniform [7], top-k [5]) and continuous (*e.g.*, gaussian, gaussian-experts [6]) sampling methods, allowing the model to capture all continuous time steps while significantly reducing redundant information obtained from sampling. Furthermore, we propose a global preference activation strategy focused on model modality preferences. This strategy determines the preference distribution of the model in different modalities through independent perception and improves the decision-making capabilities of the model. Compared to using dynamic weights at the multimodal fusion stage, our enhancement strategy focuses on the global information in the initial features. This avoids biases caused by decoding from complex multimodal features, while also complementing the fine-grained information obtained through temporal dynamic perception, achieving comprehensive audio-visual understanding. Our contributions are summarized as follows:

- We propose a dual-path audio-visual learning model named AV-Master that enhances the understanding of audio-visual scenes by perceiving fine-grained details and global modality preferences related to questions, achieving cross-modal mapping from audio-visual signals to textual answers.
- We introduce a dynamic adaptive focus sampling method that progressively performs adaptive learning from the input audio-visual segments during the encoding, enabling precise capture of the focal areas within a large number of redundant segments.
- Extensive evaluation on four benchmark AVQA datasets demonstrates that our proposed AV-Master is superior and achieves new state-of-the-art performance compared to existing AVQA methods.

The remainder of this paper is organized as follows. In Section II, we review the related work relevant to the research direction of this study. Section III provides a detailed description of the architecture of the proposed model. Section IV presents experimental results that validate the effectiveness of our proposed methods. Finally, in Section V, we conclude by summarizing the main contributions of this paper.

II. RELATED WORK

A. Audio-Visual Scene Understanding

In recent years, audio-visual scene understanding has emerged as a significant research area, garnering ever-increasing attention from the academic community. The core idea of this field originates from the human instinct to perceive the world through the synergy of multiple senses. Specifically, visual and auditory information are not only complementary but also often inseparable in understanding complex, dynamic environments. These two modalities are tightly linked through semantic consistency (e.g., the image of a dog matching the sound of its bark) and spatio-temporal correlation (a sound synchronizing with and emanating from its visual source). This interplay provides the key elements for achieving a sophisticated understanding of scenes that surpasses the capabilities of single-modality perception. Building on this foundation, researchers have explored numerous specific subtasks such as sound source localization [9], [10], [11], action recognition [12], [13], event detection [14], [15], [16], video parsing [17], [18], [19], audio-visual source separation [20], [21], etc. These tasks aim to fully leverage the interaction and fusion of audio and visual information to overcome the limitations of unimodal perception, thereby enhancing the fine-grained understanding of dynamic audio-visual scenes.

Within this broad research landscape, our work focuses on the more advanced cognitive task of audio-visual question answering. We achieve precise scene understanding and reasoning based on multimodal information by specifically emphasizing the temporal dynamic perception of audio-visual segments and employing a global preference activation strategy to dynamically capture the dependency of different questions on specific modalities.

B. Audio-Visual Question Answering

The audio-visual question answering (AVQA) aims at achieving fine-grained comprehension and reasoning of complex audio-visual scenes. This task involves a comprehensive analysis of audio, visual, and their fused information, allowing the model to provide accurate answers to different questions. Existing research focuses mainly on unimodal audio question answering (AQA) and visual question answering (VQA), but a single modality is insufficient to fully capture the rich semantic information contained in natural videos. To this end, recent works [5], [8], [6] have begun exploring the field of AVQA. Among these works, Li et al. constructed declarative sentence prompts based on question templates to help the temporal awareness module better identify key segments relevant to the question, and designed a novel spatial awareness module to facilitate efficient fusion of visual tokens. Zhao et al. aligned audio-visual cues across spatial and temporal dimensions through contrastive learning, and adaptively assigned fusion weights to the visual and audio modalities according to the question. Kim et al. proposed the QA-TIGER framework, which adaptively focuses on both continuous and discontinuous frames according to the question, explicitly injects question information, and applies progressive optimization.

However, existing works only adopt simple audio-visual segment selection methods, which introduce a large amount of redundant information, making it difficult for the subsequent decoder to extract key clues from the coarse-grained multimodal features. Moreover, they overlook the fact that different questions may have varying demands on visual and audio modalities. As a result, the models often fail to dynamically
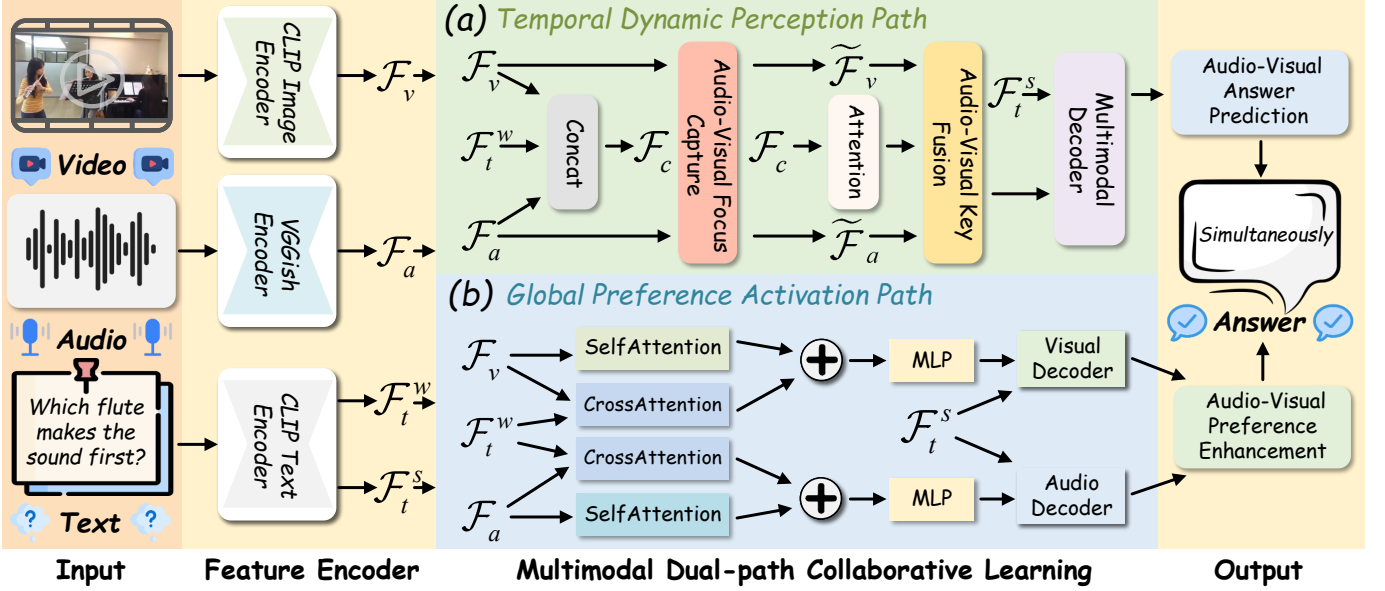
Fig. 2. Overview of AV-Master. We utilize three separate pre-trained encoders to extract features from video, audio, and question inputs. The encoded features are then fed into the temporal dynamic perception path and the global preference activation path, respectively. Finally, the model predicts the correct answer based on the outputs of these two paths.

leverage the relevant dominant modality for different questions and instead perform a coarse multimodal fusion, which, to some extent, undermines the overall performance of the model. In contrast, our work enables fine-grained key audio-visual learning through dynamic adaptive sampling based on the question, and enhances the model from a global perspective by leveraging modality preference, ultimately leading to more accurate answers.

III. METHODOLOGY

In this section, we will introduce our proposed AV-Master in detail. Specifically, we first introduce the feature information used and the feature extraction settings in Section III-A. Subsequently, in Section III-B and Section III-C, we respectively explain the proposed temporal dynamic perception path and global preference activation path. Finally, in Section III-D, we describe in detail the learning objectives involved in the training phase, which include the dual-path prediction loss and the dual-path contrastive loss. The overall architecture is shown in Fig. 2, and the specific implementation flow of the temporal dynamic perception path proposed in Fig. 2 is illustrated in Fig. 3.

A. Input Representation

(a) Visual representation: For a given video, we split it into T non-overlapping $1s$ segments, each with paired audio and visual elements. Each visual segment is processed by a pre-trained vision-language model CLIP [22]. In this process, a special token is added at the beginning of each segment and is used as the visual feature. The visual features can be represented as $\mathcal{F}_v = \left\{ \mathcal{F}_v^1, \mathcal{F}_v^2, \cdots, \mathcal{F}_v^T \right\} \in \mathbb{R}^{T \times D}$, where D denotes the feature dimension.

(b) Audio representation: For each audio segment, we follow previous works [5], [4], [6] that use the pre-trained

VGGish model [23] to extract audio features. The audio features can be represented as $\mathcal{F}_a = \left\{ \mathcal{F}_a^1, \mathcal{F}_a^2, \cdots, \mathcal{F}_a^T \right\} \in \mathbb{R}^{T \times D}$. The parameters of the CLIP and VGGish models are frozen during training.

(c) Question representation: For the input question, we use the CLIP text encoder to obtain word-level features $\mathcal{F}_t^w = \left\{ \mathcal{F}_t^2, \mathcal{F}_t^3, \cdots, \mathcal{F}_t^L \right\} \in \mathbb{R}^{L \times D}$, and extract the sentence-level feature $\mathcal{F}_t^s \in \mathbb{R}^{1 \times D}$ by taking the embedding of the first token. L denotes the number of question tokens.

B. Temporal Dynamic Perception Path

To extract key information from complex initial audio-visual features, we propose a temporal dynamic perception path, which consists of two modules: an audio-visual focus capture module and an audio-visual key fusion module. The former focuses more on the relationships within each modality, aiming to discover key hidden clues in the current modality, while the latter concentrates on establishing connections between different modalities. Additionally, the inputs to the perception path are the outputs from the previous feature encoding stage, including visual features $\mathcal{F}_v$, audio features $\mathcal{F}_a$, word-level features $\mathcal{F}_t^w$, and sentence-level features $\mathcal{F}_t^s$.

(a) Audio-visual focus capture: As shown in Fig. 3 (a), to achieve fine-grained perception, the visual features $\mathcal{F}_v$ and the audio features $\mathcal{F}_a$ are fed into the audio-visual focus capture module for dynamic adaptive focus sampling at each time step from time 0 to time n, where $n = T - 1$. During the perception process, we utilize predefined templates ($\mathcal{A}_{cls}$ and $\mathcal{V}_{cls}$) to focus on the audio-visual features at each moment, thereby extracting global effective information from the complex initial features. Moreover, we introduce learnable biases to highlight key regions with significant variations between different moments for improving sampling accuracy. Over time, the predefined templates progressively extract the critical
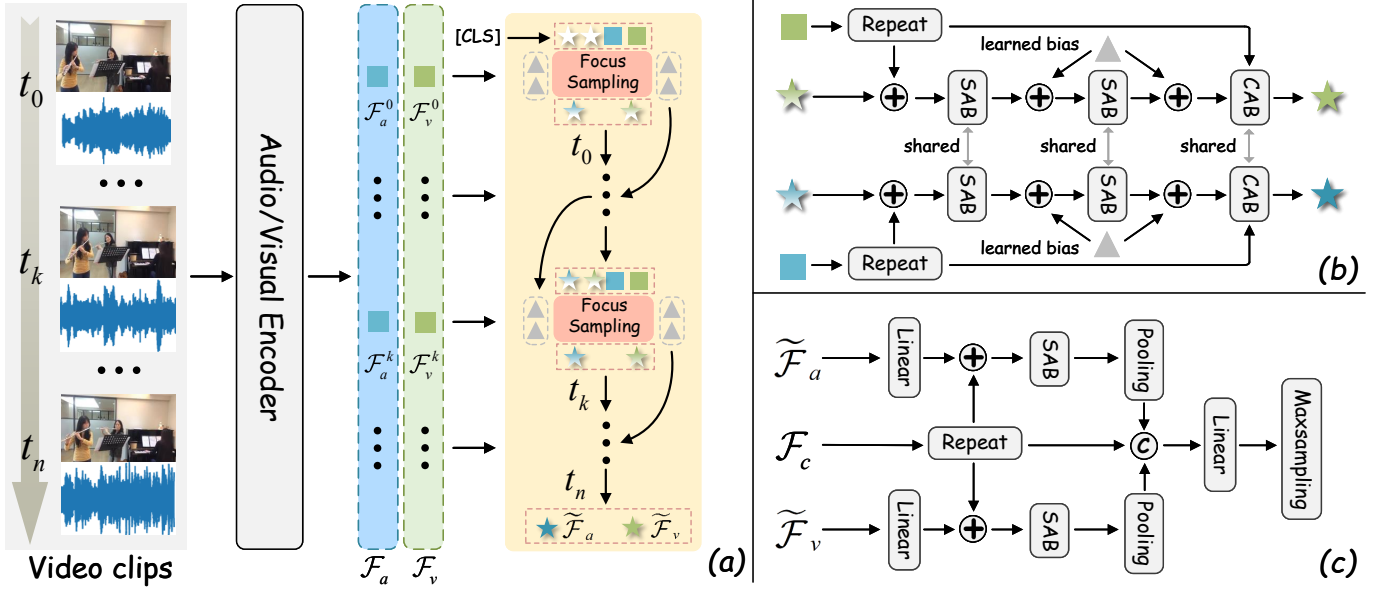
Fig. 3. The pipeline of *(a) audio-visual focus capture* and *(c) audio-visual key fusion* in the **temporal dynamic perception path**, where *(b)* represents the specific implementation process of focus sampling in *(a) audio-visual focus capture*. SAB and CAB represent the self-attention block and the cross-attention block, respectively. ★ ★ represent the input predefined CLS tokens (serve as audio-visual templates), ■ ■ represent the audio-visual features at a certain time step, and ▲ is a learned bias.

information hidden at different moments throughout the entire audio-visual segment. Finally, we take the templates from the n-th time step as the output and obtain the fine-grained focus features $\widetilde{\mathcal{F}}_v$ and $\widetilde{\mathcal{F}}_a$.

(b) Dynamic adaptive focus sampling: For a specific time step k, the detailed focus sampling process is shown in Fig. 3 (b). First, the k-th time step visual feature $\mathcal{F}_v^k$ and audio feature $\mathcal{F}_a^k$ together with the $(k-1)$-th time step template $\mathcal{A}_{cls}^{k-1}$ and $\mathcal{V}_{cls}^{k-1}$ serve as inputs for the entire sampling process. Features $\mathcal{F}_v^k$ and $\mathcal{F}_a^k$ are repeated to match the length of their corresponding templates and then added to the templates. The resulting sums are fed into a self-attention block (SAB) for attention enhancement. These enhanced features are then added to a learnable bias and go through the above operation once more. Afterward, these features, along with the repeated feature, are passed into a cross-attention block (CAB) for modality interaction. Finally, we can obtain the updated templates $\mathcal{A}_{cls}^k$ and $\mathcal{V}_{cls}^k$ for the current time step. This process is formulated as follows:

$$\widetilde{\mathcal{F}}_v = \left(\mathcal{V}_{cls}^k | k = T - 1\right)$$
$$\mathcal{V}_{cls}^k = Focus\,Sampling\left(\mathcal{F}_v^k, \mathcal{V}_{cls}^{k-1} | bias\right), \quad (1)$$

where $k \in [0, T-1]$ and T represents the total number of video 1s segments. The focus sampling is formulated as:

$$\mathcal{V}_{tp1}^{k-1} = \mathcal{F}_v^k + \mathcal{V}_{cls}^{k-1}$$
$$\mathcal{V}_{tp2}^{k-1} = \text{SAB}\left(\text{SAB}\left(\mathcal{V}_{tp1}^{k-1}\right) + bias\right) + bias, \quad (2)$$
$$\mathcal{V}_{cls}^k = \text{CAB}\left(\mathcal{V}_{tp2}^{k-1}, \mathcal{F}_v^k\right)$$

where $\mathcal{V}_{tp2}^{k-1}$ serves as $query$ and $\mathcal{F}_v^k$ serves as $key, value$ in CAB. For the k-th time step audio feature $\mathcal{F}_a^k$, the same applies following the above Eqs. (1-2).

(c) Audio-visual key fusion: After obtaining the fine-grained audio-visual focal features $\widetilde{\mathcal{F}}_a$ and $\widetilde{\mathcal{F}}_v$, we feed them into the audio-visual key fusion module, where multimodal fusion is performed under the guidance of the word-level question feature $\mathcal{F}_t^w$ to provide a refined semantic anchor for the subsequent decoding. The detailed audio-visual key fusion process is shown in Fig. 3 (c), $\widetilde{\mathcal{F}}_a$ and $\widetilde{\mathcal{F}}_v$ are transformed by linear layers and then added to the multimodal feature $\mathcal{F}_c$. Subsequently, each of these is fed into an SAB for enhancement, followed by a pooling operation, and then concatenated with $\mathcal{F}_c$ along the feature dimension. Finally, the result is passed through a linear layer and max sampling to obtain the final fused feature $\mathcal{F}_{fu}$. Where, $\mathcal{F}_c$ is formed by concatenating the visual feature $\mathcal{F}_a$, audio feature $\mathcal{F}_v$, and question feature $\mathcal{F}_t^w$ along the sequence length. The fusion process is formulated as follows:

$$\mathcal{F}_c = \text{SAB}\left(Concat\left(\mathcal{F}_a, \mathcal{F}_v, \mathcal{F}_t^w\right)\right)$$
$$\mathcal{O}_a^l = Pooling\left(\text{SAB}\left(Linear\left(\widetilde{\mathcal{F}}_a\right) + \mathcal{F}_c\right)\right)$$
$$\mathcal{O}_v^l = Pooling\left(\text{SAB}\left(Linear\left(\widetilde{\mathcal{F}}_v\right) + \mathcal{F}_c\right)\right), \quad (3)$$
$$\mathcal{F}_{fu} = Max\left(Linear\left(Concat\left(\mathcal{O}_a, \mathcal{O}_v, \mathcal{F}_c\right)\right)\right)$$

where the $Pooling\,(\cdot)$ performs summation along the sequence dimension, while the $Max\,(\cdot)$ takes the maximum value along the feature dimension. The fused feature $\mathcal{F}_{fu} \in \mathbb{R}^{1 \times D}$ and $\mathcal{O}$ represents the intermediate outputs.

C. Global Preference Activation Path

Considering the potential mismatch between audio-visual segments and the possibility that the current question may be more biased toward a specific modality or scene in audio-visual question answering, we propose a global preference

activation path. The preference activation path is designed to independently perceive and decouple auditory and visual inputs, activating the global contextual information within them. It serves to complement the fine-grained features from the temporal dynamic perception path, providing the model with an additional auxiliary perspective to achieve a more comprehensive understanding of the audio-visual scene.

As shown in Fig 2 (b), the visual feature $\mathcal{F}_v$ and the word-level question feature $\mathcal{F}_t^w$ are first fed into a cross-attention block. Meanwhile, the $\mathcal{F}_v$ is also separately enhanced via a self-attention block. The outputs are then summed and passed through a multi-layer perceptron (MLP) to obtain the visual preference feature $\mathcal{F}_v^p$. Similarly, the audio feature $\mathcal{F}_a$ goes through the same procedure to obtain the audio preference feature $\mathcal{F}_a^p$. The calculation process is as follows:

$$
\begin{aligned}
\mathcal{O}_v^g &= \mathrm{SAB}\left(\mathcal{F}_v\right) + \mathrm{CAB}\left(\mathcal{F}_v + \mathcal{F}_t^w\right) \\
\mathcal{F}_v^p &= \mathrm{MLP}\left(\mathcal{O}_v^g\right) \\
\mathcal{O}_a^g &= \mathrm{SAB}\left(\mathcal{F}_a\right) + \mathrm{CAB}\left(\mathcal{F}_a + \mathcal{F}_t^w\right) \\
\mathcal{F}_a^p &= \mathrm{MLP}\left(\mathcal{O}_a^g\right)
\end{aligned}
\tag{4}
$$

where visual feature $\mathcal{F}_v$ or audio feature $\mathcal{F}_a$ serves as $query$ and $\mathcal{F}_t^w$ serves as $key, value$ in CAB. The activated preference features $\mathcal{F}_v^p$ and $\mathcal{F}_a^p \in \mathbb{R}^{T \times D}$.

D. Optimization and Answer Prediction

During training, the fused feature $\mathcal{F}_{fu}$ and the sentence-level question feature $\mathcal{F}_t^s$ are jointly fed into the multimodal decoder to generate the answer prediction distribution, which is then used together with the ground-truth labels to compute the answer loss (denoted as $\mathcal{L}_{qa}$). In the preference activation path, features $\mathcal{F}_v^p$ and $\mathcal{F}_a^p$ are separately input into two independent audio/visual decoders, which also generate prediction distributions under the guidance of $\mathcal{F}_t^s$, resulting in two preference losses (denoted as $\mathcal{L}_v^p$ and $\mathcal{L}_a^p$). Moreover, a contrastive loss (denoted as $\mathcal{L}_c$) is applied between the dynamic perception and preference activation paths, aiming to enhance the stability of the dual-path paradigm and improve the model's discriminative ability by leveraging hard negative samples. During inference, we sum all the predicted distributions from the model and use the argmax function to obtain the final answer.

(a) Dual-path prediction loss: For the answer loss $\mathcal{L}_{qa}$, we follow the standard training procedure for the AVQA. The goal of the model is to minimize the negative log-likelihood of the probabilities over multiple answer choices generated by the multimodal decoder. For the preference losses $\mathcal{L}_v^p$ and $\mathcal{L}_a^p$, we also adopt a similar computation method to calculate them based on the probability distributions generated by the two audio/visual decoders. The formula for the above loss calculation can be represented as:

$$
\begin{aligned}
\mathcal{L}_{qa} &= -\sum_{ans=1}^{C} y_{ans} \log\left(P_{ans}|\mathcal{F}_{fu}, \theta_l\right) \\
\mathcal{L}_v^p, \mathcal{L}_a^p &= -\sum_{pef=1}^{C} y_{pef} \log\left(P_{pef}|\mathcal{F}_v^p, \mathcal{F}_a^p, \theta_g\right)
\end{aligned}
\tag{5}
$$

TABLE I
DETAILED DESCRIPTION OF AVQA, MUSIC-AVQA, MUSIC-AVQA-R, AND MUSIC-AVQA-V2.0 DATASETS.

Dataset	# Videos	# Train QA	# Valid QA	# Test QA
AVQA	57,015	40,425	-	16,910
MUSIC-AVQA	9,288	31,904	4,568	9,129
MUSIC-AVQA-R	9,288	-	-	211,572
MUSIC-AVQA-v2.0	10,492	37,408	5,346	10,819

where C is the total number of answer choices and θ is the set of learnable parameters of the decoders. Both the multimodal decoder and the audio/visual decoders include transformer blocks and linear layers.

(b) Dual-path contrastive loss: For the contrastive loss $\mathcal{L}_c$, we enhance the stability of the dual-path architecture and improve the accuracy of joint prediction by increasing the similarity between the positive sample feature $\mathcal{F}_{fu}$ from the dynamic perception path and $\mathcal{F}_g^j$ from the preference activation path, $j \in \{v, a\}$. Meanwhile, we compare the feature $\mathcal{F}_{fu}$ from the positive sample with the feature $\mathcal{F}_g^j$ from the negative sample and reduce their similarity to enhance the model's ability to distinguish positive samples. The contrastive $\mathcal{L}_c$ is expressed as follows:

$$
\begin{aligned}
p_1^{j(i)} &= exp\left(cos\left(\mathcal{F}_{fu}, \overline{\mathcal{F}}_g^{j(i)}\right)/\mathcal{T}\right) \\
p_1^{(i)} &= p_1^{v(i)} + p_1^{a(i)} \\
\mathcal{L}_c &= -\frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{p_1^{(i)}}{\sum_{k \neq i}^{neg} p_1^{(k)} + p_1^{(i)}}\right]
\end{aligned}
\tag{6}
$$

where neg is the number of negative pairs and $cos\left(\cdot,\cdot\right)$ is the cosine function used to compute similarity. $\overline{\mathcal{F}}$ represents the operation of averaging the feature $\mathcal{F}$ along the sequence length dimension, $\mathcal{T}$ denotes the temperature. The overall loss is the weighted sum of the above losses:

$$
\mathcal{L} = \lambda_{qa}\mathcal{L}_{qa} + \lambda_v^p\mathcal{L}_v^p + \lambda_a^p\mathcal{L}_a^p + \lambda_c\mathcal{L}_c
\tag{7}
$$

where λ_{qa}, λ_v^p, λ_a^p, and λ_c are hyperparameters used to trade-off each loss functions.

IV. EXPERIMENTS

A. Experimental Setting

(a) Dataset and evaluation metric: In this paper, we validate our proposal on four datasets: MUSIC-AVQA [7], MUSIC-AVQA-R [24], MUSIC-AVQA-v2.0 [25], and AVQA [26]. The details are summarized in Table I. Consistent with previous work, the ablation and analysis experiments are conducted on the MUSIC-AVQA dataset by default.

MUSIC-AVQA [7] is a large-scale AVQA dataset focused on multimodal understanding and reasoning in music performance scenarios. It contains 45,601 question-answer pairs distributed across 9,288 videos, with a total duration of over 150 hours. The videos are primarily collected from YouTube and cover 22 types of musical instruments (such as guitar, cello, marimba, etc.). Annotations are manually created by human annotators. The question-answer pairs are divided into

TABLE II
EXPERIMENTAL RESULTS ON THE MUSIC-AVQA TEST SET. THE BEST AND SECOND BEST PERFORMANCE OF EACH TASK ARE HIGHLIGHTED IN **BOLD** AND <u>UNDERLINE</u> RESPECTIVELY. FOR A FAIR COMPARISON, WE REPORT THE PERFORMANCE OF THE VERSION USING THE SAME AUDIO ENCODER AS OUR METHOD, DENOTED AS †. COMPARISONS WITH OTHER VERSIONS ARE PRESENTED IN SUBSEQUENT EXPERIMENTS.

Methods	Audio QA			Visual QA			Audio-Visual QA						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
MCAN *[CVPR'19]*	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
PSAC *[AAAI'19]*	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
HME *[CVPR'19]*	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
AVSD *[CVPR'19]*	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
HCRN *[CVPR'20]*	68.59	50.92	62.05	64.39	61.81	63.08	54.47	53.38	41.53	52.11	47.69	50.26	55.73
Pano-AVQA *[ICCV'21]*	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 *[CVPR'22]*	78.18	67.05	74.06	71.56	76.38	74.00	81.81	70.80	64.51	**66.01**	63.23	69.54	71.52
LAVISH *[CVPR'23]*	82.09	65.56	75.97	78.98	81.43	80.22	81.71	75.51	66.13	63.77	67.96	71.26	74.46
QAGL *[TCSVT'24]*	82.99	**71.04**	<u>78.58</u>	80.12	77.88	78.89	82.29	72.73	62.83	63.40	64.36	69.43	73.58
TSPM *[ACMMM'24]*	84.07	64.65	76.91	82.29	84.90	83.61	82.19	76.21	71.85	<u>65.76</u>	**71.17**	73.51	76.79
APL *[AAAI'24]*	82.40	<u>70.71</u>	78.09	76.52	82.74	79.69	82.99	73.29	66.68	64.76	65.95	70.96	74.53
PSOT† *[AAAI'25]*	–	–	78.22	–	–	80.07	–	–	–	–	–	72.61	75.29
AVAF-Net *[AAAI'25]*	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
SHMamba *[TASLP'25]*	82.30	63.64	75.42	78.53	81.31	79.93	82.89	72.65	67.93	61.31	68.37	70.64	74.12
CoQo† *[IJCV'25]*	–	–	78.90	–	–	83.70	–	–	–	–	–	73.92	77.40
QA-TIGER *[CVPR'25]*	<u>84.86</u>	67.85	78.58	<u>83.96</u>	<u>86.29</u>	<u>85.14</u>	83.10	<u>78.58</u>	**72.50**	63.94	69.59	73.74	<u>77.62</u>
AV-Master (Ours)	**87.02**	67.85	**79.95**	**86.55**	**86.61**	**86.58**	<u>83.60</u>	**79.13**	<u>72.39</u>	64.21	<u>70.80</u>	**74.22**	**78.51**

three modality scenarios (audio, visual, and audio-visual), encompassing nine question types (such as existential, location, counting, etc.) and 33 question templates. The answer set includes 42 different answers. The questions require fine-grained scene understanding and spatiotemporal reasoning of both audio and visual content, for example: "Is a certain instrument present?" or "Where is the source of the sound located?". The MUSIC-AVQA dataset emphasizes the interaction between audio and visual modalities in music performance scenes. Compared to other video question answering datasets (such as general video datasets), it places a stronger focus on audio-visual correlation.

MUSIC-AVQA-R [24] is an extended version of the MUSIC-AVQA, designed to evaluate the robustness of AVQA models and to address the issues of limited question expression variety and potential biases in the original dataset. It is constructed by rephrasing and splitting the questions in the MUSIC-AVQA test set. Specifically, researchers used AI tools to rephrase the 9,129 questions in the test set 25 times, generating semantically equivalent but more diverse expressions. Then, three annotators independently voted to retain most of the questions. As a result, the number of test questions expanded from 9,129 to 211,572, and the vocabulary size increased from 93 to 465. All questions are divided into two subsets: frequent (head) and rare (tail), classified based on the frequency of answer occurrences. Compared to MUSIC-AVQA, MUSIC-AVQA-R enhances question diversity and is better suited for testing AVQA models under conditions of diverse questions and imbalanced answer distributions.

MUSIC-AVQA-v2.0 [25] is an improved version of the MUSIC-AVQA, aimed at addressing the data bias issues in the original dataset and constructing a more balanced and challenging dataset. Researchers manually collected 1,230 new instrument performance videos (sourced from YouTube) and created 8,100 new question-answer pairs to supplement the original MUSIC-AVQA dataset. They ensured a more even answer distribution across each question category and sub-category, especially for binary questions, where the answers are nearly evenly distributed to avoid significant bias. Some of the videos were horizontally flipped to generate symmetric question-answer pairs, enhancing data diversity. MUSIC-AVQA-v2.0 resolves the answer bias problems present in the MUSIC-AVQA dataset, making it a more reliable benchmark suitable for testing model performance in unbiased AVQA tasks.

To evaluate the performance of AV-Master in scenarios beyond musical performance, we additionally introduce the AVQA dataset. AVQA [26] is a large-scale benchmark specifically designed for audio-visual question answering in real-world settings. The dataset contains 57,015 real-life videos and 57,335 question-answer pairs, with a total duration of over 158 hours. The videos are sourced from the VGG-Sound dataset [27] and cover 165 categories of daily activities and natural sounds. The questions are manually designed to ensure reliance on both modalities for reasoning, involving various types of relations such as existence, location, temporal, causal, and intentional. With its large scale, diversity, and high-quality human annotations, the dataset has become a widely used benchmark for evaluating multimodal fusion methods in complex real-world scenarios.

(b) Implementation details: For a fair comparison, we follow previous work [6] and adopt a similar setup: videos are uniformly sampled at a rate of 1 frame per second. Audio representations are extracted using the pre-trained VGGish model [23], while visual inputs and corresponding questions are encoded through the CLIP-ViT-L/14 model [22]. All extracted features are projected into a 512-dimensional space via a linear transformation. The predefined templates, $\mathcal{A}_{cls}$ and $\mathcal{V}_{cls}$, are both composed of a set of learnable embeddings, each with a length of 8 and randomly initialized. The model is optimized using Adam [28] and starts with a learning rate of 1e-4, which is reduced by a factor of 0.1 every 8 epochs.

7

TABLE III
EXPERIMENTAL RESULTS ON THE MUSIC-AVQA-R TEST SET, WITH H AND T REPRESENTING PERFORMANCE ON HEAD (FREQUENT) AND TAIL (RARE) ANSWER CATEGORIES, RESPECTIVELY. ALL RESULTS ARE OBTAINED FROM OFFICIAL REPORTS OR REPRODUCED FROM OTHER WORKS.

Methods	Audio QA				Visual QA				Audio-Visual QA												Avg
	Count		Comp		Count		Local		Exist		Count		Local		Comp		Temp				
	H	T	H	T	H	T	H	T	H	T	H	T	H	T	H	T	H	T			
HCAttn [NeurIPS'16]	61.67	41.63	59.09	47.14	56.52	9.20	67.01	53.16	66.57	61.13	59.53	12.48	37.05	42.48	48.81	60.12	33.82	39.26			51.90
MCAN [CVPR'19]	75.02	60.16	58.89	50.09	64.58	26.69	66.48	62.25	51.29	67.29	64.76	25.28	46.11	61.61	50.57	52.40	34.64	58.05			57.27
PSAC [AAAI'19]	53.01	56.68	57.41	48.12	49.55	26.43	72.96	60.69	50.56	55.54	56.70	19.58	41.98	52.30	38.13	58.92	26.68	46.24			50.45
HME [CVPR'19]	62.60	53.95	54.97	58.29	50.95	16.46	73.25	58.60	65.74	66.49	63.18	17.18	33.79	46.03	53.20	69.57	33.95	41.57			53.66
AVSD [CVPR'19]	54.00	47.84	60.61	47.79	60.34	10.07	74.78	61.43	66.28	61.98	46.21	8.06	33.00	40.35	51.98	66.00	40.14	41.52			52.33
FCNLSTM [TASLP'20]	66.23	36.48	64.78	51.24	61.75	5.31	54.86	51.06	64.76	78.52	62.69	7.23	46.66	57.30	43.13	71.67	37.02	30.78			54.12
HCRN [CVPR'20]	55.53	53.31	47.17	32.44	41.87	23.55	39.40	51.27	41.81	65.45	54.58	19.57	36.62	42.72	33.33	36.87	40.47	44.13			43.92
Pano-AVQA [ICCV'21]	50.57	43.45	50.78	44.93	47.28	15.50	67.19	65.51	52.37	22.04	52.21	21.52	44.35	61.69	45.61	40.49	35.00	49.33			47.40
ST-AVQA [CVPR'22]	56.40	41.48	62.28	57.59	59.86	12.94	63.31	54.00	73.35	77.26	48.31	8.41	35.35	40.49	53.30	62.44	40.25	38.15			52.80
LAVISH [CVPR'23]	61.73	43.99	65.06	60.38	65.53	11.13	70.21	64.73	77.83	79.46	49.88	14.87	41.76	41.20	59.26	65.10	41.84	46.26			57.63
TSPM [ACMMM'24]	81.65	71.80	67.66	49.56	78.29	47.56	80.58	73.18	69.15	82.79	77.09	38.64	42.24	57.37	52.07	68.86	39.23	49.36			66.30
QA-TIGER [CVPR'25]	82.67	75.82	71.75	43.11	81.30	54.59	84.76	75.59	72.84	78.56	76.70	33.55	48.22	64.65	37.55	80.47	36.85	62.96			67.99
AV-Master (Ours)	84.90	72.61	70.67	49.22	83.48	57.40	87.39	79.33	75.55	78.41	80.18	35.28	55.39	77.41	47.76	70.80	46.49	69.99			71.19

TABLE IV
EXPERIMENTAL RESULTS ON THE MUSIC-AVQA-v2.0 FOR (A) BIAS AND (B) BALANCED TEST SETS.

Test	Training	Methods	A-QA	V-QA	AV-QA	Avg
(a) Bias	Bias	ST-AVQA	76.86	77.70	69.59	73.07
		LAVISH	76.73	80.96	70.80	74.59
		QA-TIGER	79.13	84.83	72.37	76.93
		AV-Master	79.31	86.54	74.12	78.39
	Balance	ST-AVQA	76.18	77.20	67.96	71.92
		LAVISH	75.56	80.83	69.27	73.51
		LAST	77.10	82.99	70.86	75.24
		LAST-Att	77.29	83.47	71.05	75.45
		QA-TIGER	77.07	85.93	71.20	76.57
		AV-Master	79.25	86.87	71.52	77.03
(b) Balance	Bias	ST-AVQA	73.34	76.82	64.51	69.40
		LAVISH	73.14	79.70	65.01	70.39
		QA-TIGER	77.57	84.84	67.43	73.91
		AV-Master	78.22	86.42	69.11	75.37
	Balance	ST-AVQA	75.50	77.67	66.32	71.02
		LAVISH	76.15	81.32	68.28	73.18
		LAST	78.08	83.29	69.72	74.85
		LAST-Att	78.56	84.07	70.30	75.44
		QA-TIGER	79.90	86.95	70.22	76.43
		AV-Master	80.84	87.37	70.29	76.75

TABLE V
EXPERIMENTAL RESULTS ON THE TEST SET OF AVQA DATASET.

Methods	Ensemble	Total Accuracy (%)
HME	HAVF	85.0
PSAC	HAVF	87.4
LADNet	HAVF	84.1
ACRTransformer	HAVF	87.8
HGA	HAVF	87.7
HCRN	HAVF	89.0
SaSR-Net	–	89.9
PSTP-Net	–	90.2
TSPM	–	90.8
MCD	–	90.8
AV-Master (Ours)	–	**91.4**

TABLE VI
COMPARISON WITH PRETRAINING-BASED METHODS. Z-S INDICATES WHETHER THE ZERO-SHOT SETTING IS USED AND PT REPRESENTS THE AMOUNT OF DATA USED FOR MODEL PRE-TRAINING.

Methods	V-Enc.	A-Enc.	Z-S	PT	Params	ACC
OneLLM	CLIP$_L$	CLIP$_L$	✓	1008.5M	7B	47.6
ChatBridge	ViT$_G$	BEAT	✓	130.0M	13B	43.0
CAT	ImageBind	Imagebind	✓	3.1M	7B	48.6
CAT+	ImageBind	Imagebind	✓	0.2M	7B	50.1
VideoLLaMa	EVACLIP$_G$	Imagebind	✓	2.8M	7B	36.6
AVLLM	CLIP$_L$	CLAP	✓	1.6M	13B	45.2
AVicuna	CLIP$_L$	CLAP	✓	1.1M	7B	49.6
CAD	ViT	PANNs	✗	100.0M	- -	78.3
VAST	EVACLIP$_G$	BEATs	✗	42.0M	1.3B	80.7
VALOR	CLIP$_L$	AST	✗	33.5M	593M	78.9
AV-Master	CLIP$_L$	VGGish	✗	N/A	61M	78.5

The batch size is set to 32, and the model is trained for 30 epochs. Our proposed model is trained on NVIDIA GeForce RTX 4090 and implemented in PyTorch.

B. Quantitative Results

(a) Compared methods: To evaluate the effectiveness and superiority of our model, we compare AV-Master with existing state-of-the-art AVQA methods across multiple datasets. These methods include: QA-TIGER [6], CoQo [29], SHMamba [30], AVAF-Net [8], PSOT [31], SaSR-Net [32], PSTP-Net [4], TSPM [5], APL [33], LAST [25], MCD [34], QAGL [35], LAVISH [36], ST-AVQA [7], Pano-AVQA [37], ACRTrans-former [38], HGA [39], HCRN [40], FCNLSTM [41], LAD-Net [42], AVSD [43], HME [44], PSAC [45], MCAN [46] and HCAttn [47].

(b) MUSIC-AVQA: As shown in Table II, AV-Master achieves an overall accuracy of 78.51%, outperforming all existing models, including the state-of-the-art method QA-TIGER (77.62%). Notably, our model demonstrates strong

TABLE VII
ABLATION ON THE DIFFERENT COMPONENTS OF AV-MASTER.

#	Methods	Average Accuracy (%)			
		A-QA	V-QA	AV-QA	Avg
1	w/o. ALL	73.37	79.23	69.82	72.94
2	w/o. AVFC	79.83	85.55	74.04	78.11
3	w/o. DPCL	78.83	85.71	73.78	77.84
4	w/o. GPAP	77.90	85.05	73.10	77.12
5	w/o. TDPP	78.77	83.73	72.94	76.83
6	**AV-Master**	**79.95**	**86.58**	**74.22**	**78.51**

TABLE IX
IMPACT OF LENGTHS OF AUDIO-VISUAL TEMPLATES.

#	Lengths	Average Accuracy (%)			
		A-QA	V-QA	AV-QA	Avg
1	16	80.51	85.96	74.04	78.34
2	12	80.26	85.88	74.08	78.30
3	8	79.95	**86.58**	**74.22**	**78.51**
4	4	**80.63**	85.22	73.92	78.10
5	2	80.38	85.84	73.98	78.26

TABLE VIII
ABLATION STUDY OF DIFFERENT LOSS FUNCTIONS IN AV-MASTER
TRAINING.

#	$\mathcal{L}_{qa}$	$\mathcal{L}_c$	$\mathcal{L}_a^p$	$\mathcal{L}_v^p$	Average Accuracy (%)			
					A-QA	V-QA	AV-QA	Avg
1	✓				77.84	84.31	71.11	75.80
2	✓	✓			78.15	84.43	70.76	75.69
3	✓	✓	✓		79.83	82.74	72.49	76.50
4	✓	✓		✓	78.83	86.33	72.63	77.36
5	✓		✓	✓	78.83	85.71	73.78	77.84
6	✓	✓	✓	✓	**79.95**	**86.58**	**74.22**	**78.51**

TABLE X
IMPACT OF WEIGHT SHARING STRATEGIES. A-S AND B-S REPRESENT
ATTENTION BLOCK WEIGHT SHARING AND LEARNED BIAS WEIGHT
SHARING, RESPECTIVELY.

#	A-S	B-S	Average Accuracy (%)			
			A-QA	V-QA	AV-QA	Avg
1			79.70	85.80	73.92	78.09
2	✓		79.95	**86.58**	**74.22**	**78.51**
3		✓	**80.26**	85.88	73.59	78.03
4	✓	✓	80.07	86.29	73.88	78.27

performance on complex reasoning tasks such as counting, significantly outperforming the second-best method (i.e., A-Counting: 87.02% vs. 84.86%, V-Counting: 86.55% vs. 83.96%, AV-Counting: 79.13% vs. 78.58%).

(c) MUSIC-AVQA-R: As shown in Table III, compared to the MUSIC-AVQA dataset, AV-Master demonstrates more significant improvements (+3.20%) on the MUSIC-AVQA-R dataset, achieving an overall accuracy of 71.19% and setting a new state-of-the-art performance. The notable performance gain can be attributed to AV-Master's dual-path learning paradigm, which provides the model with powerful generalization capabilities.

(d) MUSIC-AVQA-v2.0: As shown in Table IV, AV-Master outperforms existing models across all types. Notably, when trained on the biased dataset, AV-Master still achieves significant improvements over the second-best method on both the balanced and biased test sets. These results highlight the robustness and adaptability of AV-Master in handling various training environments, demonstrating its strong generalization capability even under distributional bias.

(e) AVQA: To further validate the generalization capability of our model in real-world scenarios, we conducted experiments on the AVQA dataset. As shown in Tab. V, AV-Master achieved an overall accuracy of 91.4%, surpassing all previous methods both with and without the HAVF [26] module. These results strongly demonstrate that AV-Master maintains its exceptional performance in complex, real-world audio-visual question answering tasks. It is worth noting that although the performance improvement of AV-Master on the AVQA dataset may seem limited compared to its performance on the MUSIC-AVQA series of datasets, this is primarily due to the shorter duration and simpler audio-visual content of the AVQA dataset.

(f) Comparison with pretraining-based methods: In this work, our approach focuses on designing efficient AVQA expert models that achieve competitive performance under limited training data and hardware conditions by developing powerful modules. This is also the mainstream direction in current AVQA research [6], [29], [32]. Additionally, there are some methods [48], [49], [50] that rely on large-scale pretraining, which attempt to apply large language models to audio-visual scenarios to handle downstream AVQA tasks. However, pretraining-based AVQA methods require substantial computational resources and multimodal data, and their generalization ability in specific audio-visual scenarios falls short of expectations. We present relevant experiments to compare AV-Master with pretraining-based models, including OneLLM [51], Chatbridge [52], CAT [53], CAT+ [54], VideoLLaMa [55], AVLLM [56], AVicuna [57], VideoLLaMa2 [48], CAD [58], VAST [49] and VALOR [50].

As shown in Table VI, AV-Master achieves competitive performance (78.5% accuracy) compared to various pretraining-based models, despite having significantly fewer parameters (61M) and not relying on large-scale pretraining. In contrast, methods such as CAT+ and AVicuna, which are based on large language models, achieve notably lower performance (50.1% and 49.6% accuracy) under zero-shot conditions. Although these approaches leverage abundant resources and powerful backbone models, they often exhibit weaker generalization capabilities in domain-specific audio-visual tasks. Other pretraining-based models achieve impressive performance (80.7% accuracy) after fine-tuning on the MUSIC-AVQA dataset. However, our method achieves comparable results using significantly less training data, fewer trainable parameters, and a more lightweight audio-visual feature encoder. This highlights the practicality of AV-Master and its suitability for resource-constrained audio-visual scenarios.
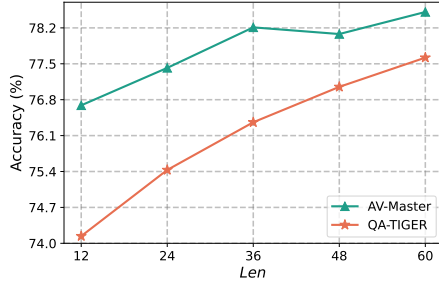
Fig. 4. The ablation study on the lengths of visual and audio segments and comparison with QA-TIGER.

Fig. 5. The ablation study on input modalities and comparison with other popular models (AVTS and QA-TIGER).

TABLE XI
DIFFERENT VISUAL AND AUDIO FEATURE EXTRACTORS.

Visual Encoder	Audio Encoder	Methods	Average Accuracy (%)			
			A-QA	V-QA	AV-QA	Avg
Resnet-18	VGGish	ST-AVQA	74.06	74.00	69.54	71.52
		AV-Master	**78.21**	**79.23**	**70.25**	**74.04**
CLIP$_B$	VGGish	PSTP-Net	70.91	77.26	72.57	73.52
		AV-Master	**78.40**	**82.25**	**72.82**	**76.31**
CLIP$_L$	VGGish	TSPM	76.91	83.61	73.51	76.79
		QA-TIGER	78.58	85.14	73.74	77.62
		AV-Master	**79.95**	**86.58**	**74.22**	**78.51**
	CLAP	PSOT	79.08	87.12	74.07	78.42
		AV-Master	**80.63**	**87.78**	**74.92**	**79.34**
Internvideo2	Internvideo2	CoQo	79.27	87.90	75.80	79.60
		AV-Master	**81.50**	**88.32**	**75.86**	**80.15**

C. Ablation Studies

(a) Ablation study on main components: To explore the effectiveness of each component, we removed them individually and re-evaluated performance. As shown in Tab. VII, removing different components leads to varying degrees of performance degradation for AV-Master. Specifically, when the TDPP (Temporal Dynamic Perception Path) is removed, the performance drops to 76.83%; removing the GPAP (Global Preference Activation Path) results in a drop to 77.12%; removing the DPCL (Dual-path Contrastive Loss) lowers performance to 77.84%, and removing the AVFC (Audio-Visual Focus Capture) decreases it to 78.11%. When all components are removed simultaneously, the performance drops significantly from 78.51% to 72.94%. These results show that each component in AV-Master contributes to performance improvement, and the best results are achieved only when all components are present.

(b) Impact of template lengths: We explored the impact of the lengths of predefined templates $\mathcal{A}_{cls}$ and $\mathcal{V}_{cls}$ on model performance. As shown in Tab. IX, the average accuracy of the model reaches its highest when the template length is set to 8. If the template length is too short, important information may be missed during sampling, leading to reduced performance. Conversely, if the template length is too long, redundant information may be introduced, which can impair the model's performance.

(c) Impact of weight sharing strategies: We also explored the impact of different weight-sharing strategies in focus sampling. As shown in Tab. X, the model achieves the best overall performance when attention blocks share weights while biases remain unshared — this configuration is also adopted as our default setting. Interestingly, when only biases are shared, the model attains the lowest overall accuracy but achieves
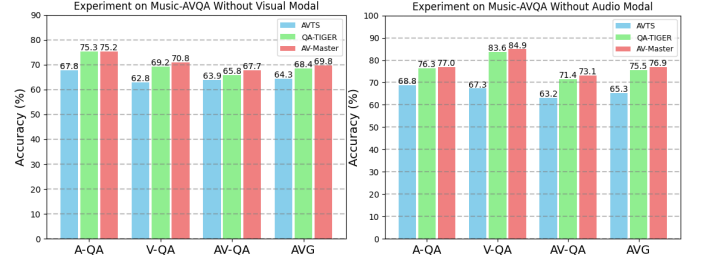
the highest accuracy on the A-QA task. This suggests that bias sharing may benefit certain subtasks while potentially hindering overall model performance.

(d) Ablation study on loss functions: To analyze the contribution of each loss function, we conducted an ablation study with the results presented in the Tab. VIII. This study systematically evaluates the model's performance by incrementally adding different loss components. The baseline model, trained only with the answer loss $\mathcal{L}_{qa}$, achieves an average accuracy of 75.80%. Consecutively adding the contrastive loss $\mathcal{L}_c$ and the preference losses $\mathcal{L}_a^p$ and $\mathcal{L}_v^p$ provides progressive gains. The final model, which integrates all four loss functions, achieves the highest average accuracy of 78.51%. This demonstrates that each loss component plays a vital role, and their combined effect is crucial for optimizing the model's overall performance. It is worth noting that, as can be seen from the results in the third, fourth, and fifth rows of the table, visual preference training provides the most significant overall improvement to the model compared to other training objectives (aside from basic $\mathcal{L}_{qa}$). This also indicates that in the vast majority of scenarios, the model relies more heavily on visual information for question answering.

(e) Impact of different feature extractors: The results in Table XI demonstrate that different visual encoders have a significant impact on the performance of AV-Master. When using ResNet-18 [59] as the visual encoder, AV-Master achieves an average accuracy of 74.04%, showing a considerable improvement compared to ST-AVQA (71.52%) with the same visual encoder. When a more powerful visual encoder, CLIP$_B$ [22], is adopted, the average accuracy of AV-Master further increases to 76.31%, surpassing PSTP-Net (73.52%). Moreover, when using CLIP$_L$ [22] and Internvideo2 [60] as visual encoders, AV-Master achieved the best results across all subtasks (A-QA, V-QA, AV-QA), significantly outperforming comparable methods. In summary, as the visual encoder was progressively upgraded from ResNet-18 to Internvideo2, the performance of AV-Master on all tasks steadily improved.

Furthermore, the choice of audio encoders also plays a crucial role. When fixing the visual encoder to CLIP$_L$, switching from VGGish [23] to a more advanced audio encoder like CLAP [61] boosts the average accuracy from 78.51% to 79.34%, outperforming its competitor PSOT (78.42%). This trend continues with the most powerful backbones; using Internvideo2 for both visual and audio encoding, AV-Master achieves the highest overall average accuracy of 80.15%, again

TABLE XII
DIFFERENT MODALITY PREFERENCE ENHANCEMENTS.

#	Methods	Average Accuracy (%)			
		A-QA	V-QA	AV-QA	Avg
1	w/o. APE	79.21	86.50	73.31	77.85
2	w/o. VPE	79.83	84.02	73.51	77.41
3	w/o. AVPE	77.65	85.59	72.98	77.15
4	**AV-Master**	**79.95**	**86.58**	**74.22**	**78.51**

Fig. 6. Visualization of the audio-visual focus capturing process, including the accuracy at different time steps and the overall trend curve.

surpassing the competing method CoQo (79.60%). These results collectively demonstrate that the performance of AV-Master is consistently enhanced by leveraging stronger feature extractors for both the visual and audio modalities, underscoring the importance of high-quality unimodal representations for complex audio-visual reasoning tasks.

(f) Impact of different audio-visual segment lengths: To investigate the effect of audio-visual segment lengths on model performance and the overall robustness of AV-Master in scenarios with limited audio-visual input, we conducted an ablation study on the segment length. As shown in Fig. 4, when the amount of audio–visual input decreases, the overall performance of all models shows a downward trend, indicating that more complete audio-visual information can provide richer cues and is beneficial for improving model performance. Meanwhile, compared with QA-TIGER, AV-Master maintains a relatively stable performance decline when facing reduced audio-visual content, suggesting that it is capable of extracting and integrating key features from limited audio-visual information. This can be attributed to the proposed audio–visual focus capture module, which progressively refines coarse-grained audio–visual features into fine-grained cues. In addition, the involvement of the preference activation strategy and the dual-path model architecture further strengthen the model's robustness. Note that when the segment length is 48, AV-Master's performance shows a slight drop. This may reflect the model reaching near saturation during the middle-to-late period or being less sensitive to redundant tail segments.

(g) Ablation study on input modalities: To investigate the contribution of different input modalities and to verify the stability of the AVQA model, we conduct an ablation study using two input settings: without visual modal (A+Q) and without audio modal (V+Q). As shown in Fig. 5, all models demonstrate improved performance when the visual modality is present. Specifically, AV-Master achieves the highest average accuracy across both input settings, reaching 69.8%

TABLE XIII
DIFFERENT ENHANCEMENT METHOD.

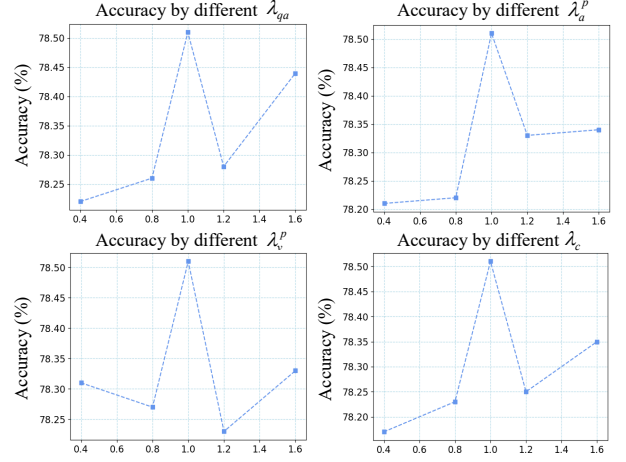| #	Methods	Average Accuracy (%)			
		A-QA	V-QA	AV-QA	Avg
1	W-ADD	78.71	81.79	71.86	75.70
2	MUL	79.70	85.30	73.70	77.84
3	ADD	79.95	86.58	74.22	78.51

Fig. 7. The effects of four trade-off parameters on the MUSIC-AVQA dataset including λ_{qa}, λ_a^p, λ_v^p and λ_c.

Fig. 8. The attention visualization for video-question (upper) and audio-question (lower), with attention intensity indicated by the color scale on the right.

with A+Q and 76.9% with V+Q. Compared to the baseline model (QA-TIGER), AV-Master shows consistent improvements across all sub-tasks (A-QA, V-QA, AV-QA), suggesting its superior ability to extract and fuse relevant features from the input. Notably, the performance gap between A+Q and V+Q settings is larger for QA-TIGER and AV-Master than for AVST, indicating that advanced models are more effective at leveraging visual cues. These results highlight the dominant role of the visual modality in multimodal question answering, while also confirming the robustness of AV-Master under varying modality conditions. The results for QA-TIGER in *(f)* and *(g)* were reproduced using the official code.

(h) Ablation study on modality preference enhancement: To further investigate the effectiveness of our proposed modality preference enhancement strategy and to identify which modality has the greatest impact on overall model performance, we conduct an ablation study by selectively disabling different
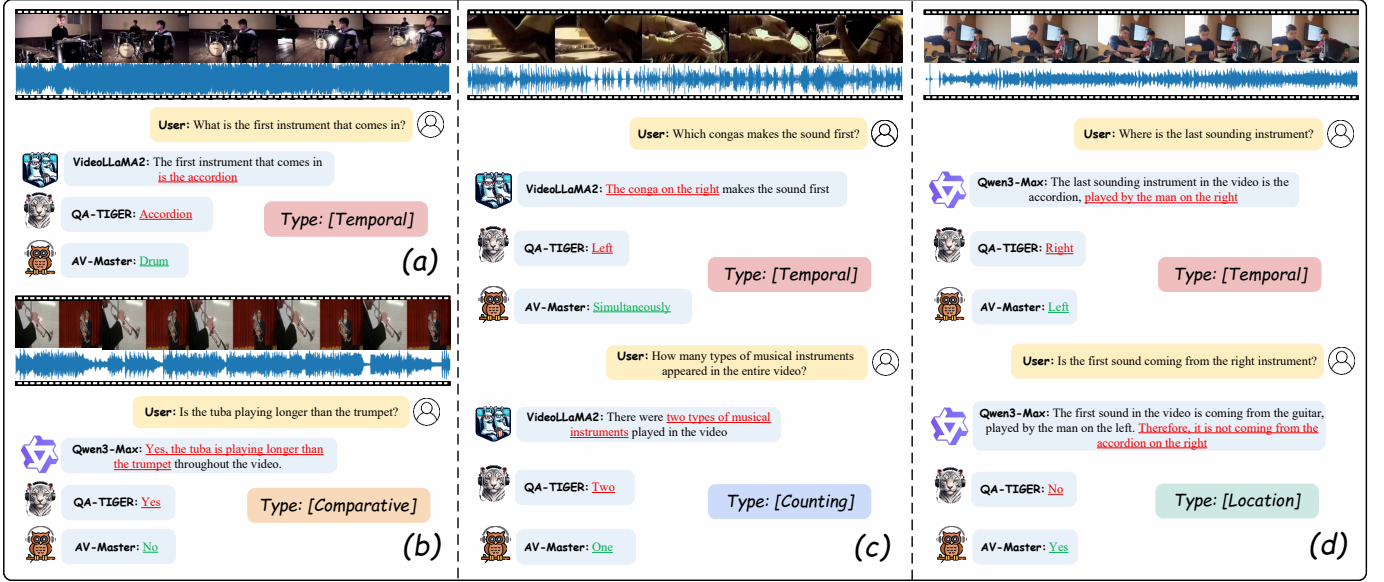
Fig. 9. Qualitative demonstration of our proposed AV-Master and comparison with MLLM (VideoLLaMa2-7B and Qwen3-Max) and AVQA expert model (QA-TIGER). Correct parts are highlighted in green while incorrect parts are highlighted in red.

components during the inference stage. As shown in Table XII, removing the audio-visual preference enhancement (AVPE) leads to the most significant performance drop, reducing the average accuracy to 77.15%, which highlights the importance of jointly modeling audio-visual scenarios. Additionally, disabling visual preference enhancement (VPE) results in lower V-QA accuracy (84.02% vs. 86.58%) and overall average accuracy (77.41%), indicating that visual preference plays a critical role in AVQA. Similarly, removing the audio preference enhancement (APE) leads to degraded performance on A-QA and AV-QA, showing the necessity of modeling audio preference. Overall, the full model achieves the best performance across all metrics, confirming the complementary benefits of modality-specific preference enhancement and their joint contribution to robust AVQA performance.

(i) Impact of different enhancement methods: To explore the impact of different preference distribution fusion methods on model performance, we conduct a comparative study using three representative methods: (1) W-ADD, which first computes a weighted ratio based on the initial overall scores of the visual and auditory preference distributions before performing a weighted summation; (2) MUL, which applies element-wise multiplication between modality preferences; and (3) ADD, a simple summation of the two distributions. As shown in Table XIII, the ADD strategy achieves the best overall performance, with an average accuracy of 78.51%, outperforming both MUL (77.84%) and W-ADD (75.70%). This result suggests that a direct and unweighted summation of visual and auditory preference distributions can better preserve cross-modal complementary information and prevent overfitting to any single modality. Interestingly, although the W-ADD method introduces an adaptive weighting mechanism, its performance lags behind due to potential imbalances introduced by dynamic scaling. When viewed together with the previous ablation results in Table XII, it is evident that both

the design of modality-specific preference enhancements and the fusion mechanism of these preferences play a crucial role in optimizing overall AVQA performance.

(j) Impact of different trade-off hyperparameters: As shown in Fig. 7, we investigate the impact of four trade-off hyperparameters λ_{qa}, λ_a^p, λ_v^p and λ_c, which are used to balance different loss functions. To assess the importance of each loss function in AV-Master, we vary one hyperparameter at a time while keeping the others fixed at a default value of 1.0. The experimental results show that increasing λ_{qa} from 0.4 to 1.0 significantly improves accuracy. However, when the value increases beyond a certain threshold (e.g., 1.6), accuracy declines slightly, suggesting that overemphasizing a single objective may diminish the contributions of others. Similar patterns are observed for λ_a^p, λ_v^p and λ_c, underscoring the independent yet equally important roles of each objective in the AV-Master. Based on these findings, we set all trade-off parameters to a default value of 1.0 to ensure a balanced contribution from each loss component. This configuration achieves optimal overall performance.

D. Qualitative Analysis

(a) Focus effects at different time steps: As shown in Fig. 6, we present the variation in accuracy over time during the audio-visual focus capture process. The blue dots in the figure represent the original accuracy data at different moments, while the red curve indicates the overall trend. Although the original data exhibits some fluctuations, the trend line shows that the accuracy remains relatively stable during the first 40 steps. However, starting at step 45, the accuracy begins to rise significantly, reaching its peak at step 60. This suggests that the audio-visual focusing mechanism becomes more effective in the later stages, and the model progressively improves its ability to extract and fuse relevant information during the focusing process.

(b) Attention visualization of different modalities: As shown in Fig. 8, we illustrate which parts of the text the visual and audio modalities attend to within the global preference activation path. For the input question "Where is the loudest instrument?", the visual modality primarily attends to the word "instrument" while paying minimal attention to the audio-related word "loudest." In contrast, the audio modality places significant focus on "loudest." These results indicate that AV-Master can effectively distinguish between visual and auditory cues and accurately align them with the corresponding textual elements, which further demonstrates AV-Master's capability in cross-modal understanding.

(c) Qualitative results comparison of different models: Fig. 9 presents a qualitative comparison among our proposed AV-Master, MLLM (VideoLLaMA2 [48] and Qwen3-Max [62]), and another AVQA expert model (QA-TIGER [6]) across four distinct audio-visual scenarios. In each case, AV-Master exhibits a superior ability to comprehend audio-visual content and provides accurate answers that align closely with both visual and auditory cues. In contrast, the other models fail to achieve such consistency. For example, in the first scenario (a), both VideoLLaMA2 and QA-TIGER incorrectly identify the instrument as an "accordion", whereas AV-Master correctly identifies it as a "drum", owing to its accurate interpretation of the audio signal. Similarly, in the other scenarios (c-d), AV-Master accurately infers the spatial and temporal characteristics of sound events, while the competing models fail to effectively integrate these multimodal cues. These qualitative results reinforce our quantitative findings, showing that AV-Master is more robust in fine-grained audio-visual reasoning, particularly in scenarios that require synchronized and cross-modal understanding. This further validates the effectiveness of our carefully designed modules in fully leveraging both audio and visual modality information. The inference results for VideoLLaMA2 and Qwen3-Max were obtained from its official demo, simulating application scenarios in low-resource environments.

V. CONCLUSION

In this paper, we propose AV-Master, a novel dual-path audio-visual question answering expert model designed to address the challenges faced by existing models in processing complex audio-visual scenes. Current methods often struggle to flexibly focus on the most question-relevant spatiotemporal segments when confronted with a large amount of redundant information, and they lack the ability to dynamically perceive the importance of different modalities for different questions. This limits their comprehensive understanding of audio-visual scenes. To tackle these difficulties, AV-Master introduces a dynamic adaptive focus sampling mechanism and a global modality preference activation strategy. These enable it to effectively capture question-relevant audio-visual segments and modality preferences, thereby enhancing its decision-making capabilities. Extensive experiments on multiple large-scale AVQA benchmark demonstrate that by meticulously capturing key audio-visual details and integrating a global understanding of modality preferences, AV-Master provides an efficient and powerful solution for AVQA, particularly excelling in complex reasoning tasks and under challenging data distributions.

REFERENCES

[1] J. Zhang, P. Tang, Y. Tan, and H. Wang, "Mgtr-miss: More ground truth retrieving based multimodal interaction and semantic supervision for video description," *Neural Networks*, p. 107817, 2025.

[2] P. Tang, J. Zhang, H. Wang, Y. Tan, and Y. Yi, "Srvc-la: Sparse regularization of visual context and latent attention based model for video description," *Neurocomputing*, vol. 630, p. 129639, 2025.

[3] X. Lin, X. Guo, T. Wang, Y. Ma, J. Huang, J. Zhang, J. Cao, and Z. Yu, "Svc 2025: the first multimodal deception detection challenge," *arXiv preprint arXiv:2508.04129*, 2025.

[4] G. Li, W. Hou, and D. Hu, "Progressive spatio-temporal perception for audio-visual question answering," in *Proceedings of the 31st ACM international conference on multimedia*, pp. 7808–7816, 2023.

[5] G. Li, H. Du, and D. Hu, "Boosting audio visual question answering via key semantic-aware cues," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5997–6005, 2024.

[6] H. Kim, I. Jung, D. Suh, Y. Zhang, S. Lee, and S. Hong, "Question-aware gaussian experts for audio-visual question answering," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13681–13690, 2025.

[7] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19108–19118, 2022.

[8] X. Zhao, Y. Wang, and P. Jin, "Audio-visual adaptive fusion network for question answering based on contrastive learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 10483–10491, 2025.

[9] M. Risoud, J.-N. Hanson, F. Gauvrit, C. Renard, P.-E. Lemesre, N.-X. Bonne, and C. Vincent, "Sound source localization," *European annals of otorhinolaryngology, head and neck diseases*, vol. 135, no. 4, pp. 259–264, 2018.

[10] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba, "Music gesture for visual sound separation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10478–10487, 2020.

[11] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15490–15500, IEEE, 2021.

[12] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, "Listen to look: Action recognition by previewing audio," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10457–10467, 2020.

[13] M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar, "Multimodal fusion for audio-image and video action recognition," *Neural Computing and Applications*, vol. 36, no. 10, pp. 5499–5513, 2024.

[14] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 247–263, 2018.

[15] H. Xuan, Z. Zhang, S. Chen, J. Yang, and Y. Yan, "Cross-modal attention network for temporal inconsistent audio-visual event localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 279–286, 2020.

[16] J. Zhou, D. Guo, and M. Wang, "Contrastive positive sample propagation along the audio-visual event line," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7239–7257, 2022.

[17] Y. Wu and Y. Yang, "Exploring heterogeneous clues for weakly-supervised audio-visual video parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1326–1335, 2021.

[18] Y.-H. Lai, Y.-C. Chen, and F. Wang, "Modality-independent teachers meet weakly-supervised audio-visual event parser," *Advances in Neural Information Processing systems*, vol. 36, pp. 73633–73651, 2023.

[19] J. Zhou, D. Guo, Y. Mao, Y. Zhong, X. Chang, and M. Wang, "Label-anticipated event disentanglement for audio-visual video parsing," in *European Conference on Computer Vision*, pp. 35–51, Springer, 2024.

[20] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 35–53, 2018.

[21] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3879–3888, 2019.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021.

[23] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135, IEEE, 2017.

[24] J. Ma, M. Hu, P. Wang, W. Sun, L. Song, H. Pei, J. Liu, and Y. Du, "Look, listen, and answer: Overcoming biases for audio-visual question answering," *arXiv preprint arXiv:2404.12020*, 2024.

[25] X. Liu, Z. Dong, and P. Zhang, "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*, pp. 4478–4487, 2024.

[26] P. Yang, X. Wang, X. Duan, H. Chen, R. Hou, C. Jin, and W. Zhu, "Avqa: A dataset for audio-visual question answering on videos," in *Proceedings of the 30th ACM international conference on multimedia*, pp. 3480–3491, 2022.

[27] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, IEEE, 2020.

[28] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] B. Pei, Y. Huang, G. Chen, J. Xu, Y. Wang, L. Wang, T. Lu, Y. Qiao, and F. Wu, "Guiding audio-visual question answering with collective question reasoning," *International Journal of Computer Vision*, pp. 1–18, 2025.

[30] Z. Yang, W. Li, and G. Cheng, "Shmamba: Structured hyperbolic state space model for audio-visual question answering," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[31] Z. Li, J. Zhou, J. Zhang, S. Tang, K. Li, and D. Guo, "Patch-level sounding object tracking for audio-visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 5075–5083, 2025.

[32] T. Yang, Y. Nan, L. Dai, Z. Liang, Y. Tian, and X. Zhang, "Sasrnet: Source-aware semantic representation network for enhancing audio-visual question answering," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15894–15904, 2024.

[33] Z. Li, D. Guo, J. Zhou, J. Zhang, and M. Wang, "Object-aware adaptive-positivity learning for audio-visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 3306–3314, 2024.

[34] Q. Ye, Z. Yu, and X. Liu, "Answering diverse questions via text attached with key audio-visual clues," *arXiv preprint arXiv:2403.06679*, 2024.

[35] Z. Chen, L. Wang, P. Wang, and P. Gao, "Question-aware global-local video understanding network for audio-visual question answering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 4109–4119, 2023.

[36] Y.-B. Lin, Y.-L. Sung, J. Lei, M. Bansal, and G. Bertasius, "Vision transformers are parameter-efficient audio-visual learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2299–2309, 2023.

[37] H. Yun, Y. Yu, W. Yang, K. Lee, and G. Kim, "Pano-avqa: Grounded audio-visual question answering on 360deg videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2031–2041, 2021.

[38] J. Zhang, J. Shao, R. Cao, L. Gao, X. Xu, and H. T. Shen, "Action-centric relation transformer network for video question answering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 63–74, 2020.

[39] P. Jiang and Y. Han, "Reasoning with heterogeneous graph alignment for video question answering," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 11109–11116, 2020.

[40] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9972–9981, 2020.

[41] H. M. Fayek and J. Johnson, "Temporal reasoning via audio question answering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2283–2294, 2020.

[42] X. Li, L. Gao, X. Wang, W. Liu, X. Xu, H. T. Shen, and J. Song, "Learnable aggregating net with diversity learning for video question answering," in *Proceedings of the 27th ACM international conference on multimedia*, pp. 1166–1174, 2019.

[43] I. Schwartz, A. G. Schwing, and T. Hazan, "A simple baseline for audio-visual scene-aware dialog," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12548–12558, 2019.

[44] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1999–2007, 2019.

[45] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan, "Beyond rnns: Positional self-attention with co-attention for video question answering," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 8658–8665, 2019.

[46] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6281–6290, 2019.

[47] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," *Advances in neural information processing systems*, vol. 29, 2016.

[48] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, and L. Bing, "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms," *arXiv preprint arXiv:2406.07476*, 2024.

[49] S. Chen, H. Li, Q. Wang, Z. Zhao, M. Sun, X. Zhu, and J. Liu, "Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset," *Advances in Neural Information Processing Systems*, vol. 36, pp. 72842–72866, 2023.

[50] S. Chen, X. He, L. Guo, X. Zhu, W. Wang, J. Tang, and J. Liu, "Valor: Vision-audio-language omni-perception pretraining model and dataset," *arXiv preprint arXiv:2304.08345*, 2023.

[51] J. Han, K. Gong, Y. Zhang, J. Wang, K. Zhang, D. Lin, Y. Qiao, P. Gao, and X. Yue, "Onellm: One framework to align all modalities with language," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

[52] Z. Zhao, L. Guo, T. Yue, S. Chen, S. Shao, X. Zhu, Z. Yuan, and J. Liu, "Chatbridge: Bridging modalities with large language model as a language catalyst," *arXiv preprint arXiv:2305.16103*, 2023.

[53] Q. Ye, Z. Yu, R. Shao, X. Xie, P. Torr, and X. Cao, "Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios," in *European Conference on Computer Vision*, pp. 146–164, Springer, 2024.

[54] Q. Ye, Z. Yu, R. Shao, Y. Cui, X. Kang, X. Liu, P. Torr, and X. Cao, "Cat+: investigating and enhancing audio-visual understanding in large language models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[55] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[56] F. Shu, L. Zhang, H. Jiang, and C. Xie, "Audio-visual llm for video understanding," *arXiv preprint arXiv:2312.06720*, 2023.

[57] Y. Tang, D. Shimada, J. Bi, and C. Xu, "Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue," *arXiv preprint arXiv:2403.16276*, vol. 2, 2024.

[58] A. Nadeem, A. Hilton, R. Dawes, G. Thomas, and A. Mustafa, "Cad-contextual multi-modal alignment for dynamic avqa," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7251–7263, 2024.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[60] Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, Z. Wang, Y. Shi, *et al.*, "Internvideo2: Scaling foundation models for multimodal video understanding," in *European Conference on Computer Vision*, pp. 396–416, Springer, 2024.

[61] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[62] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.