AV-Mamba: Cross-Modality Selective State Space Models for Audio-Visual Question Answering

Ziru Huang[1*] Jia Li[2] Wenjie Zhao[2] Yunhui Guo[2] Yapeng Tian[2]
[1]IIIS, Tsinghua University [2] The University of Texas at Dallas
[1]huangzr21@mails.tsinghua.edu.cn, [2]{jxl220096, wxz220013, yunhui.guo, yapeng.tian}@utdallas.edu

1. Introduction

Audio-visual question answering (AVQA) aims to answer questions relevant to visual objects, sounds, and their relationships within videos [6, 10, 11]. This task delves into the complexities of multimodal scenes, which are both diverse and dynamic. The key challenges of AVQA lie in accurately identifying the audio and video segments that directly relate to the question and establishing whether the identified visual regions produce sounds relevant to the question. Prior research primarily leverages attention mechanisms to tackle these challenges. For instance, employing audio-guided visual attention helps localize sounding visual regions, while question-guided temporal attention aggregates relevant audio and visual segments [4–6]. Nonetheless, audio and visual segments don't always correlate, and multimodal video segments can vary dynamically over time. Furthermore, when faced with lengthy sequences of audio-visual data accompanied by textual inputs, attention mechanisms may struggle to accurately discern the relationships across different modalities over extended durations. Consequently, the selection mechnism, driven by attention weights, may falter with long sequences, inadvertently aggregating irrelevant segments despite minimal weights.

The Mamba model [1] has proven its strength in modeling long sequences across diverse tasks. It dynamically adjusts the parameters of State Space Models (SSMs) [2] guided by the input, allowing for context-aware reasoning. Mamba's unique capability to select and retain information indefinitely inspired us to expand its success story into the realm of multimodal video modeling. In this paper, we introduce CM-Mamba, a new extension that leverages a cross-modality selection mechanism within Mamba models. The model is designed to efficiently leverage information across audio, visual, and textual modalities, in which the parameters of SSMs will dynamically adjust in response to inputs from an alternative modality. To leverage the power of CM-Mamba for AVQA, we introduce AV-Mamba (see Fig. 1), which consists of four components: feature extraction, audio-guided spatial grounding, question-guided cross-modal temporal grounding, and prediction. CM-Mamba significantly enhances the spatial and temporal grounding components, enabling more precise selection of relevant information. To evaluate the efficacy of the proposed model, we conducted extensive experiments on the MUSIC-AVQA dataset [6]. The experimental results demonstrate that our AV-Mamba can effectively answer audio, visual, and audio-visual questions relevant to videos and outperform several recent state-of-the-art approaches. Our main contributions include: 1) a new CM-Mamba model that can leverage cross-modal information for selecting useful contexts for multimodal sequence modeling; 2) a novel AV-Mamba framework for AVQA that leverages CM-Mamba for effective audio-guided spatial grounding and question-guided temporal grounding. Extensive experiments on the Music-AVQA dataset can validate the effectiveness of the proposed approach.

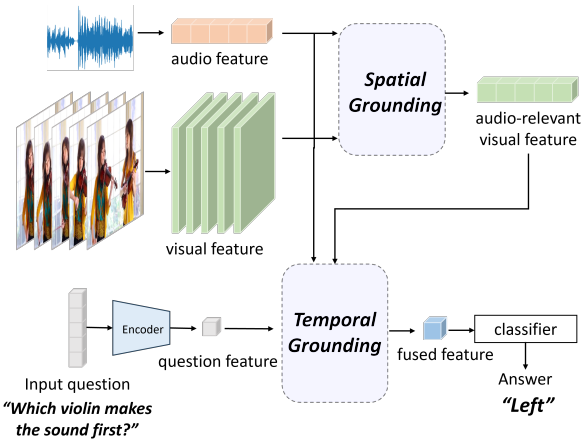

Figure 1: Our AVQA framework.

2. Method

2.1. Overview

One of the primary challenges in addressing the AVQA task is the effective fusion of multimodal information. It is essential to capitalize on both spatial and temporal cues to enhance the fusion process. To achieve this, our approach draws inspiration from the architecture of the Audio-Visual Spatial-Temporal (AVST) model [6]. As illustrated in Figure 1, our first step involves feature extraction from text, audio, and visual inputs. This is followed by using an audio-

*Work done while Ziru Huang was a visiting student at UT Dallas.

guided spatial grounding module, which extracts visual information relevant to the sound sources. Then, a question-guided cross-modal temporal grounding module processes these audio and visual features along the time dimension, integrating them with features derived from the question. Finally, we use the fused feature to obtain the final answer.

2.2. Cross-Modality Selective State Models

In this section, we first introduce the key concepts of the State Space Models and Mamba. Then, we present CM-Mamba, a novel framework designed to seamlessly integrate cross-modality information.

2.2.1 Preliminaries: State Space Models and Mamba

A standard form of State Space Models is as follows.

$$\dot{h}(t) = Ah(t) + Bx(t)$$
$$y(t) = Ch(t) + Dx(t)$$

where $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$, $C \in \mathbb{R}^{1\times N}$, and N is the dimension of the hidden states. The input function or sequence $x(t) \in \mathbb{R}$ is mapped to the output function or sequence $y(t) \in \mathbb{R}$ through the hidden states $h(t) \in \mathbb{R}^N$. After applying zero-order hold discretization:

$$\begin{aligned} h_t &= \bar{A}h_{t-1} + \bar{B}x_t \\ y_t &= Ch_t + Dx_t \\ \bar{A} &= \exp(\Delta A) \\ \bar{B} &= (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B. \end{aligned}$$

S4 [2] is characterized by four parameters (Δ, A, B, C) and possesses a significant property known as linear time invariance (LTI), indicating that the model's dynamics remain constant over time. While this property enhances computational efficiency, it also constrains the model's capacity for context-aware reasoning.

The Mamba model [1] facilitates the context-aware reasoning capability of SSMs by dynamically adjusting its parameters based on input relevance:

$$\begin{aligned} B &= s_B(x) \\ C &= s_C(x) \\ \Delta &= \tau_\Delta(\text{Parameter} + s_\Delta(x)). \end{aligned}$$

Here, $s_B(x) = \text{Linear}_N(x)$, $s_C(x) = \text{Linear}_N(x)$, $s_\Delta(x) = \text{Broadcast}_D(\text{Linear}_1(x))$ and $\tau_\Delta = \text{softplus}$.

2.2.2 CM-Mamba with Two-Tier Selective Scan

In audio-visual modeling, the relevance of audio and visual information may vary at different timesteps. Therefore, we aim to devise a more effective selective mechanism for aggregating pertinent and valuable information. We seek to enhance our approach by leveraging the long-range and selective modeling capabilities of Mamba. To enable cross-modal information selection, a natural approach is to utilize other modalities to determine the Mamba parameters.

From the results of the Mamba [1], we observed that parameter A exerts the most significant influence on the system as it governs the evolution of the entire system. Following this, parameter B demonstrates considerable influence, with parameter C holding lesser significance, as it can be incorporated in the post-processing phase. Inspired by this observation, we enable cross-modal selectivity in our CM-Mamba by leveraging cross-modal information to determine parameters A and B.

The overall architecture of CM-Mamba is shown in Fig. 2, given a query q, we employ a two-tier selective scan module based on Mamba to extract relevant information from a sequence s. We apply bidirectional mamba in each cross-modality selective scan block. In the first tier, the Mamba parameters B_1, B_2, C_1, C_2 and Δ_1, Δ_2 are determined solely by the query q. Subsequently, the parameters of the second-tier Mamba B_3, B_4, C_3, C_4 and Δ_3, Δ_4 are determined by the output of the first tier s', resulting in the final output denoted as O. It is noteworthy that for a fixed query, the first tier operates as a linear time invariant system for the sequence. However, in the second tier, parameters are dynamically adapted to the sequence. In this way, the query's cross-modal information directly guides parameter selection in CM-Mamba, enabling powerful cross-modal context-aware reasoning.

The CM-Mamba module can be formulated as:

$$\begin{aligned} P_1 &= (B_1, C_1, \Delta_1) \leftarrow q, \\ P_2 &= (B_2, C_2, \Delta_2) \leftarrow q, \\ s' &= \text{Mamba}_1(s, P_1) + \text{Mamba}_2(\text{reverse}(s), P_2) \end{aligned}$$

$$\begin{aligned} P_3 &= (B_3, C_3, \Delta_3) \leftarrow s', \\ P_4 &= (B_4, C_4, \Delta_4) \leftarrow \text{reverse}(s'), \\ O &= \text{Mamba}_3(s, P_3) + \text{Mamba}_4(\text{reverse}(s), P_4) \end{aligned}$$

where $B, C, \Delta \leftarrow r$ represents the following operations:

$$\begin{aligned} B &= s_B(r) \\ C &= s_C(r) \\ \Delta &= \tau_\Delta(\text{Parameter} + s_\Delta(r)). \end{aligned}$$

CM-Mamba-Based Spatial Grounding. In this stage, our objective is to extract audio-relevant visual features at each time step. It aligns perfectly with the purpose of our CM-Mamba, which is designed to prioritize and select cross-modality relevant information. Leveraging CM-Mamba's long-range and multimodal selective modeling, we utilize information from the audio modality to guide the spatial aggregation of visual features. As illustrated in Fig 3, in our

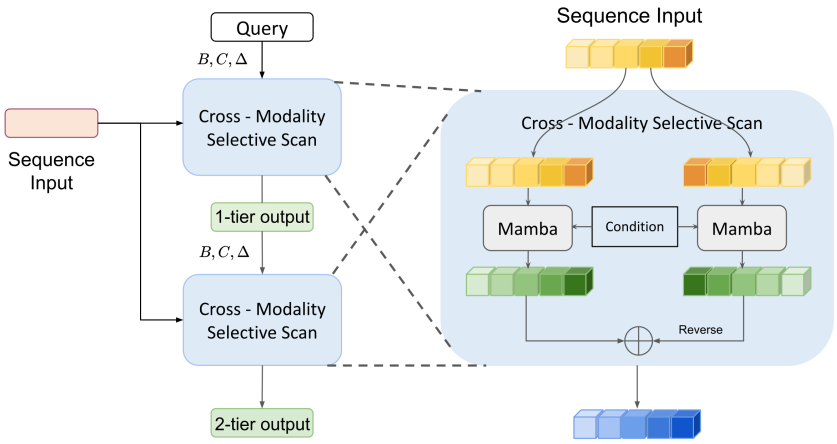

Figure 2: Our proposed CM-Mamba. Within each Cross-Modality Selective Scan block, bidirectional Mamba parameters are dynamically generated from the conditional information. The query serves as the condition for the first tier, while the output from the first tier serves as the condition for the second tier.

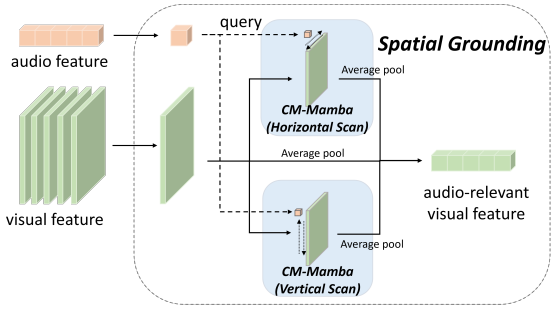

Figure 3: Audio-Guided Spatial Grounding.

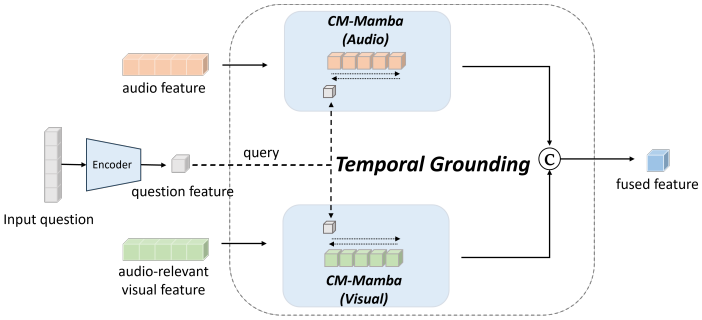

Figure 4: Question-Guided Cross-Modal Temporal Grounding.

spatial grounding module, we employ our CM-Mamba to extract audio-relevant visual information. Here, the query consists of audio features: $f_a^t \in \mathbb{R}^{D_a}$ at each timestep, while the sequence consists of the corresponding visual features $f_v^t \in \mathbb{R}^{(H\times W)\times D_v}$. Drawing inspiration from [7], we employ two CM-Mamba modules, one along the horizontal direction and the other along the vertical direction, to obtain the spatially grounded visual feature. Specifically, the visual features are flattened in horizontal direction and vertical direction, respectively.

CM-Mamba-Based Temporal Grounding. In this stage, we aim to utilize the question feature as a query to temporally aggregate audio and visual features for question answering. In Fig. 4, the temporal grounding module employs our CM-Mamba approach, with the question feature $f_q \in \mathbb{R}^{D_q}$ serving as the query and the audio feature $\mathbf{f}_a \in \mathbb{R}^{L\times D_a}$ and visual feature $\mathbf{f}_v \in \mathbb{R}^{L\times D_v}$ sequences as the candidates. Note that here we made a slight modification by integrating a direct combination of audio and visual features, employing two linear projection layers for each modality, as the first-tier input sequence for both modules. We empirically found that this approach can achieve improved modality fusion.

3. Experiments

3.1. Experimental Setup

We conduct our experiments on the MUSIC-AVQA dataset [6], which contains more than 9K videos annotated with over 45K QA pairs. We compare our proposed approach with AVSD [9], LAVIT [11], AVST [6], and AVST blended with COCA [4]. We evaluate the performance of AVQA models by measuring their accuracy in predicting answers to audio, visual, and audio-visual questions. In implementation, we use an LSTM for text encoding and utilize CLIP visual features [8] along with VGGish audio features [3]. For a given 60s video segment, our model samples a subset of 10 frames, along with their corresponding audio.

3.2. Experimental Comparison

The comparison results are shown in Tab. 1. Our proposed AV-Mamba Model demonstrates superior performance compared to other models. Notably, our model excels in addressing visual questions and temporal questions. This suggests that our proposed model effectively enhances the extraction of both spatial and temporal information through its cross-modality selective scan mechanism,

Method	Audio Question			Visual Question			Audio-Visual Question						All
	Counting	Comparative	Avg.	Counting	Location	Avg.	Existential	Location	Counting	Comparative	Temporal	Avg.	Avg.
AVSD [9]	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
LAViT [11]	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 [6]	79.74	63.47	73.74	77.61	75.43	76.51	80.97	62.83	73.44	62.49	63.14	68.96	71.80
AVST+COCA [4]	79.94	**67.68**	**75.42**	75.10	75.43	75.23	**83.50**	**66.63**	69.72	64.12	65.57	69.96	72.33
Ours	**81.61**	62.79	74.67	**78.86**	**79.67**	**79.27**	82.49	65.33	**74.94**	61.31	**68.13**	**70.62**	**73.63**

Table 1: Evaluation Results on MUSIC-AVQA. The top-2 results are highlighted.

while also implicitly incorporating positional information.

SG	TG	A Question	V Question	AV Question	All
AP	CA	74.74	76.30	68.56	71.71
WP	CA	73.74	76.51	68.96	71.80
CM	CA	74.24	77.83	69.09	72.32
CM	CM I	75.36	77.99	70.51	73.35
CM	CM II	**75.61**	78.74	70.19	73.41
CM	CM III	74.67	**79.27**	**70.62**	**73.63**

Table 2: Ablation Study. SG: spatial grounding, TG: temporal grounding; AP: average pool, WP: weighted pool, CM: CM-Mamba, CA: cross-attention.

3.3. Ablation Study

As shown in Tab. 2, we conduct an ablation study to evaluate the effectiveness of our proposed modules. We test three settings for our CM-Mamba in the temporal grounding step. They differ from each other with different inputs in each tier. Denote the audio feature sequence and visual feature sequence as s_a, s_v, respectively. And let $s_{av} = \text{Linear}(s_a) + \text{Linear}(s_v)$. Let $i \in \{a, v\}$ and we consider the input sequence for modality i: CM I 1-tier input sequence: s_i; 2-tier input sequence: s_i. CM II 1-tier input sequence: s_i; 2-tier input sequence: s_{av}. CM III 1-tier input sequence: s_{av}; 2-tier input sequence: s_i.

Our proposed CM-Mamba within the spatial grounding module has significantly enhanced the visual question answering ability, showcasing its efficacy in extracting visual features. Additionally, the application of CM-Mamba in the temporal grounding module, where we integrate question, audio, and visual features, has resulted in a higher total score, further substantiating the effectiveness of our cross-modality selective scan method.

The results across three distinct settings of the CM-Mamba are close, with a slightly higher score observed in the third setting. This suggests that incorporating straightforward fusion techniques at the intermediate stage could potentially enhance performance.

4. Conclusion

In this paper, we investigate the efficacy of leveraging the selectivity of the Mamba module to capture relevant cross-modality information. Our experimental results on the MUSIC-AVQA dataset highlight the superiority of our AV-Mamba framework for AVQA and showcase the promising cross-modality capabilities of our CM-Mamba.

Acknowledgments. This work was supported in part by a Cisco Faculty Research Award, an Amazon Research Award, and a research gift from Adobe. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 2

[2] A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 1, 2

[3] 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 *Icassp*, pages 131–135. IEEE, 2017. 3

[4] M. Lao, N. Pu, Y. Liu, K. He, E. M. Bakker, and M. S. Lew. Coca: Collaborative causal regularization for audio-visual question answering. *AAAI*, 37(11), 2023. 1, 3, 4

[5] G. Li, W. Hou, and D. Hu. Progressive spatio-temporal perception for audio-visual question answering, 2023. 1

[6] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, 2022. 1, 3, 4

[7] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 3

[8] 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 *ICML*, 2021. 3

[9] I. Schwartz, A. G. Schwing, and T. Hazan. A simple baseline for audio-visual scene-aware dialog. In *CVPR*, 2019. 3, 4

[10] 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 *ACM MM*, 2022. 1

[11] H. Yun, Y. Yu, W. Yang, K. Lee, and G. Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *ICCV*, 2021. 1, 3, 4