AV-BIMBA: A Cross-Modal State Space Model for Audio-Visual Question Answering

Varun Shanmugam, Abduljalil Radman[0000−0002−6317−9752], and Jorma Laaksonen[0000−0001−7218−3131]

Department of Computer Science, Aalto University, Finland
varunshan2001@gmail.com, {abduljalil.saif,jorma.laaksonen}@aalto.fi

Abstract. Audio-Visual Question Answering (AVQA) presents two core challenges: modeling long multimodal sequences and aligning heterogeneous audio and visual information. While state space models (SSMs) offer scalable sequence modeling, their potential for AVQA remains unexplored. In this work, we propose AV-BIMBA, an audio-visual state space model based on BIMBA. Its core contribution centers on the cross-modal selective-scan (CMSS) module, which aligns audio-visual tokens into compact spatiotemporal queries, distilling salient multimodal cues. CMSS injects temporally aware semantic audio features into visual tokens via element-wise addition, without increasing sequence length or computational cost. AV-BIMBA employs a LoRA-augmented decoder to bridge distilled multimodal representations and pre-trained linguistic knowledge with minimal additional parameters. Experiments on five AVQA benchmarks demonstrate that AV-BIMBA surpasses prior SSM-based models and achieves competitive performance against state-of-the-art AVQA approaches.

Keywords: Audio-visual question answering · State space models · Audio-visual learning · BIMBA.

1. INTRODUCTION

Audio-Visual Question Answering (AVQA) requires joint reasoning over synchronized temporal visual cues and audio events, demanding models that can capture long-range dependencies across high-resolution streams [17, 24]. However, dominant Transformer-based approaches [4, 20] face a critical bottleneck: the quadratic complexity ($O(L^2)$) of self-attention. When tackling longe videos, computational costs explode and redundant temporal tokens dilute attention, severely hindering scalability and reasoning efficiency in complex multimodal scenarios.

State Space Models (SSMs), particularly Mamba [8], offer a compelling alternative to achieve linear-time sequence modeling ($O(L)$) without sacrificing long-term dependency capture. VideoMamba [18] successfully established this paradigm for video understanding, proving that SSMs can master spatiotemporal representations efficiently. Building on this, bidirectional interleaved Mamba

for better answers (BIMBA) [14] further optimized the architecture for Visual Question Answering (VQA), employing bidirectional selective scanning to compress dense video inputs into information-rich representations. Yet, the extension of this efficiency to the audio-visual domain remains under-explored. While pioneering work such as structured hyperbolic Mamba (SHMamba) [30] has attempted this, it relies on complex hyperbolic geometries that introduce training instability and counteract the inherent simplicity of SSMs.

In this paper, we propose AV-BIMBA[1], an audio-visual state space model extending BIMBA's efficient bidirectional reasoning to multimodal sequences. It introduces the cross-modal selective-scan (CMSS) module, which encodes audio-visual tokens into compact spatiotemporal queries and refines them via bidirectional state-space scanning. This mechanism selectively retains salient multimodal cues, without increasing sequence length or computational cost. Our contributions are: **(1)** We propose AV-BIMBA, a novel SSM-based framework for AVQA that achieves robust audio-visual reasoning without increasing sequence length or computational complexity. **(2)** We propose the cross-modal selective-scan (CMSS) module, which compresses aligned audio-visual tokens into compact spatiotemporal queries and refines them via bidirectional scanning to capture high-priority multimodal cues. **(3)** Our extensive evaluations show that AV-BIMBA outperforms prior SSM-based AVQA models and remains competitive with state-of-the-art AVQA methods, offering an effective trade-off between accuracy and efficiency.

2. RELATED WORK

2.1 Audio-visual question answering

Early work on audio-visual question answering (AVQA) has focused on benchmark datasets such as MUSIC-AVQA [17] and AVQA [29], which introduce the challenge of jointly reasoning over audio and video streams with natural-language questions. In line with this, early models such as HCRN+HAVF [29] and the MUSIC-AVQA baseline [17] have been among the first attempts to address this challenge, adopting hierarchical reasoning and audio-visual fusion mechanisms. VAST [2] is an omni-modality foundational model that jointly represents vision, audio, subtitles, and text for question answering.

Recent advances in instruction tuning with vision and language [16, 28] have driven the development of multimodal large language models (MLLMs) for AVQA. Most of these models [19, 4, 21] are Transformer-based, extracting audio and visual features with pre-trained modality-specific encoders and fusing them, before projecting into an LLM's latent space for answer generation. Closed-source multimodal models such as GPT-4o [13] and Gemini-1.5 Pro [26] exhibit strong performance on multimodal tasks, including AVQA, often benefiting from massive training corpora and advanced fusion capabilities, though detailed performance on AVQA benchmarks remains proprietary.

[1] This work is based on the author's master's thesis [25]

Among open-source models, VideoLLaMA 2 [4] uses a spatial-temporal convolution connector and joint audio branch to better capture temporal and audio cues. Unified-IO 2 [21] processes images, text, audio, and actions in a shared encoder–decoder space, while Qwen2.5-Omni [27] unifies audio, image, video, and text understanding for real-time multimodal reasoning. Ola [19] employs an omni-modal design with dual audio branches and adapter-based connectors.

Crab [5] and PAVE [20] are adapter-based models that extend frozen LLMs to audio-visual tasks: Crab uses modality-specific adapters with an audio branch, while PAVE introduces lightweight "patches" for diverse downstream tasks including AVQA.

Our AV-BIMBA replaces Transformer-based quadratic self-attention with linear-scaling state space models (SSMs) for efficient long-range spatiotemporal modeling.

2.2 State space models (SSMs)

SSMs provide an efficient alternative for sequential modeling, capturing long-range dependencies with linear complexity. Mamba [8] advances this paradigm by introducing data-dependent dynamics and a selective parallel scan (S6), enabling scalable long-sequence processing. In vision, VideoMamba [18] demonstrates the effectiveness of purely SSM-based architectures for video understanding across short and long temporal ranges.

Building on these advances, BIMBA [14] extends SSM-based modeling to long-form video VQA by compressing spatiotemporal token sequences into a compact set of informative tokens using S6. Salient information is preserved through learned token selection with interleaved query tokens and bidirectional selective scanning, enabling efficient long-context reasoning.

SHMamba [30] is the only existing SSM-based model for AVQA, introducing structured hyperbolic state spaces to capture hierarchical audio-visual relationships. However, its reliance on hyperbolic geometry introduces training instability and compromises the simplicity typical of standard SSMs.

Unlike SHMamba, our AV-BIMBA uses simple addition to fuse spatiotemporal audio and visual information. Audio-visual query tokens integrate with this information, and bidirectional interleaved selective scanning captures short- and long-range dependencies efficiently.

3. THE PROPOSED AV-BIMBA

Audio-Visual Question Answering (AVQA) seeks to predict an answer $\mathbf{y}$ from a multimodal input triplet $\mathbf{X} = (\mathbf{x}_v, \mathbf{x}_a, \mathbf{x}_q)$, where $\mathbf{x}_v$ denotes the video frames, $\mathbf{x}_a$ the corresponding audio signals, and $\mathbf{x}_q$ the natural-language question. The proposed AV-BIMBA is a state-space-model-based framework for AVQA. As illustrated in Fig. 1, AV-BIMBA follows a modular design comprising modality-specific encoders, a cross-modal selective-scan (CMSS) module, and a multi-

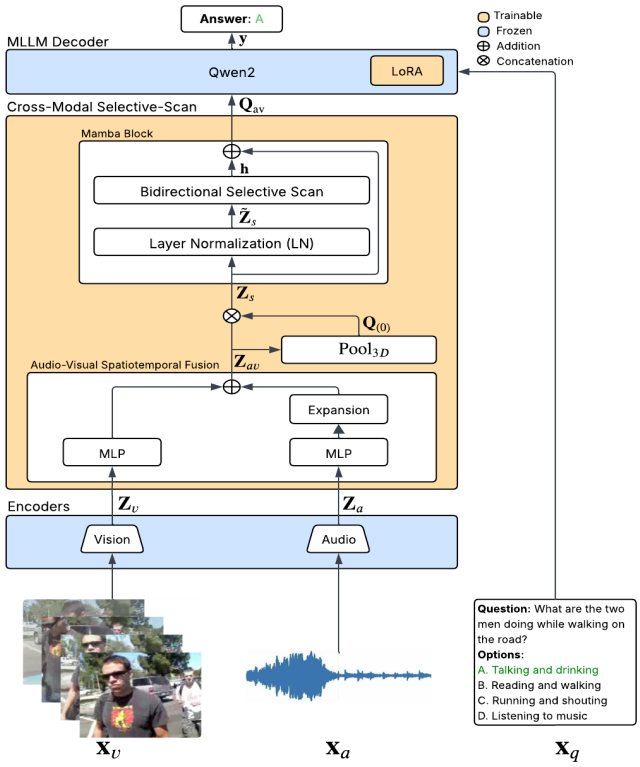

Fig. 1: Overview of the AV-BIMBA architecture.

modal LLM (MLLM) decoder. These components are described in detail in the following subsections.

3.1 Modality-Specific Encoders

The video input $\mathbf{x}_v \in \mathbb{R}^{T\times 3\times H\times W}$ consists of T frames uniformly sampled at 1 fps, where $3\times H\times W$ corresponds to RGB values at a spatial resolution of $H\times W$. The frames are processed by a SigLIP visual encoder [31], producing spatial token embeddings $\mathbf{Z}_v \in \mathbb{R}^{T\times n\times d_v}$, where d_v is the embedding dimensionality. Given $H = W = 384$, the spatial grid has $n = \lfloor H/14\rfloor \times \lfloor W/14\rfloor = 27\times 27 = 729$ patch tokens per frame. The audio input $\mathbf{x}_a$, temporally aligned with the video frames, is encoded by an audio encoder [7, 1, 23, 6] into embeddings $\mathbf{Z}_a \in \mathbb{R}^{T\times d_a}$, where d_a denotes the audio embedding dimensionality. The question $\mathbf{x}_q$ is tokenized and fed into the MLLM [28] along with the audio-visual inputs.

3.2 Cross-Modal Selective-Scan Module

As shown in Fig. 1, the cross-modal selective-scan (CMSS) module includes three sequential stages: (1) audio-visual spatiotemporal fusion, (2) audio-visual query

initialization via spatiotemporal pooling, and (3) Mamba bidirectional selective scanning.

Audio-visual spatiotemporal fusion To ensure temporally consistent audio-visual alignment, both modalities are projected into a shared embedding space of dimension d, matching the dimensionality of the MLLM decoder. The audio features are then broadcast along the spatial dimension to match the visual token layout and integrated via element-wise addition:

$$\mathbf{Z}_{av} = \phi_v(\mathbf{Z}_v) + \mathrm{Expand}\big(\phi_a(\mathbf{Z}_a)\big), \quad \mathbf{Z}_{av} \in \mathbb{R}^{T\times n\times d}, \tag{1}$$

where $\phi_v(\cdot)$ and $\phi_a(\cdot)$ are multilayer perceptron (MLP) projection layers, and $\mathrm{Expand}(\cdot)$ replicates audio features across spatial patches per frame, yielding audio tokens in $\mathbb{R}^{T\times n\times d}$. This design enforces fine-grained spatiotemporal correspondence between audio and visual signals, enabling audio semantics to modulate visual tokens at both frame and patch levels without increasing sequence length.

Audio-visual query initialization The CMSS module compresses the long spatiotemporal sequence $\mathbf{Z}_{av}$ by initializing a fixed set of audio–visual queries $\mathbf{Q}_{(0)}$ using adaptive 3D pooling over the temporal and spatial dimensions:

$$\mathbf{Q}_{(0)} = \mathrm{Pool}_{3D}(\mathbf{Z}_{av}), \quad \mathbf{Q}_{(0)} \in \mathbb{R}^{N_q\times d}, \tag{2}$$

where N_q denotes the number of audio-visual queries. These pooled queries $\mathbf{Q}_{(0)}$ serve as compact initial states that summarize coarse spatiotemporal structure while retaining global audio-visual context. These queries are then interleaved with the flattened audio-visual token sequence $\mathbf{Z}_{av} \in \mathbb{R}^{N_{av}\times d}$, where $N_{av} = T\cdot n$, to form the input sequence for Mamba bidirectional selective scanning stage:

$$\mathbf{Z}_s = \mathbf{Q}_{(0)} \otimes \mathbf{Z}_{av}, \quad \mathbf{Z}_s \in \mathbb{R}^{N_s\times d}, \tag{3}$$

where $N_s = N_{av} + N_q$, and $\otimes$ denotes interleaved concatenation distributing the queries $\mathbf{Q}_{(0)}$ uniformly along the sequence $\mathbf{Z}_s$. Since $N_q \ll N_{av}$, corresponds to a 16× compression ratio, this interleaved positioning allows each query to repeatedly scan the entire spatiotemporal sequence rather than attending locally.

Mamba bidirectional selective scanning The CMSS module processes the interleaved sequence $\mathbf{Z}_s$ using a bidirectional Mamba block (Fig. 1), built on Mamba's selective state-space scanning (S6) framework [8] and adopted in BIMBA [14], which applies layer normalization (LN), a bidirectional selective state-space scan, and a residual connection. The audio-visual queries $\mathbf{Q}_{\mathrm{av}}$ are extracted from

the Mamba block output $\mathbf{Z}_{\text{Mamba}}$ at the initialized query positions:

$$
\begin{aligned}
& z_t = \tilde{\mathbf{Z}}_s\big[t\big], \qquad \text{and} \quad \tilde{\mathbf{Z}}_s = \mathrm{LN}\big(\mathbf{Z}_s\big), \quad t = 1, \ldots, N_s, \\
& \overrightarrow{\mathbf{h}}_t = \mathbf{A}_f\, \overrightarrow{\mathbf{h}}_{t-1} + \mathbf{B}_t z_t, \qquad t = 1, \ldots, N_s, \\
& \overleftarrow{\mathbf{h}}_t = \mathbf{A}_b\, \overleftarrow{\mathbf{h}}_{t+1} + \mathbf{B}_t z_t, \qquad t = N_s, \ldots, 1, \\
& \mathbf{h} = \big[\mathbf{h}_t\big]_{t=1}^{N_s}, \qquad \text{and} \qquad \mathbf{h}_t = \overrightarrow{\mathbf{h}}_t + \overleftarrow{\mathbf{h}}_t, \\
& \mathbf{Z}_{\text{Mamba}} = \mathbf{Z}_s + \mathbf{h}, \\
& \mathbf{Q}_{\text{av}} = \mathbf{Z}_{\text{Mamba}}\big[\mathcal{Q}\big], \quad \text{and} \quad \mathcal{Q} \subset \{1, \ldots, N_s\}, \quad |\mathcal{Q}| = N_q,
\end{aligned}
\tag{4}
$$

where $\overrightarrow{\mathbf{h}}_t$ and $\overleftarrow{\mathbf{h}}_t$ denote the forward and backward state updates at time step t, $\mathbf{h}_t \in \mathbb{R}^d$ denotes the sequence of hidden states, $\mathbf{A}_f, \mathbf{A}_b \in \mathbb{R}^{d \times d}$ are learned input-independent state transition matrices, and $\mathbf{B}_t = \mathrm{Linear}_d(z_t)$ is an input-dependent parameter produced by a linear layer. This bidirectional selective scan enables $\mathbf{Q}_{\text{av}}$ to efficiently aggregate global spatiotemporal audio-visual context.

3.3 Multimodal LLM (MLLM) Decoder

The audio-visual queries $\mathbf{Q}_{\text{av}}$ are prepended to the tokenized question embeddings $\mathbf{x}_q$ and fed into the AV-BIMBA MLLM decoder (Qwen-2 [28]). The decoder is augmented with a LoRA adapter [12] for parameter-efficient fine-tuning while preserving visual question knowledge from BIMBA [14]. AV-BIMBA is trained with the autoregressive cross-entropy loss.

4. EXPERIMENTS AND RESULTS

4.1 Exprimental Settings

AV-BIMBA is initialized from BIMBA [14] weights, with the MLLM decoder Qwen2-7B [28] ($d = 3584$). The visual encoder, SigLIP (siglip-so400m-patch14-384, $d_v = 1152$) [31], and the audio encoder are kept frozen during training; only the MLLM decoder, fine-tuned with LoRA [12] initialized from BIMBA's LoRA weights, and the proposed CMSS module are trained (Fig. 1). We evaluate four pretrained audio encoders, Audio Spectrogram Transformer (AST) [7], ImageBind (IB) [6], Wav2Vec2 (W2V) [1], and Whisper (WSP) [23], with $d_a = 768$ for AST/W2V/WSP and $d_a = 1024$ for IB. Training is conducted on NVIDIA A100 GPUs (80 GB) for one epoch with batch size 1, gradient accumulation 2, and AdamW with cosine scheduling, a base learning rate of $1 \cdot 10^{-6}$, warmup ratio 0.03, and no weight decay. BF16 mixed precision, gradient checkpointing, and a maximum sequence length of 32,768 tokens are used.

For fine-tuning, videos are sampled at 1 fps, with dataset-specific frame limits. Evaluation is also conducted at 1 fps with up to 64 frames per video. Model predictions are mapped to candidate options for multiple-choice or evaluated via exact match for generative task, with accuracy (%) as the metric.

Method	SSM	Decoder MLLM	Encoder		M-AVQA	AVQA	V-AVQA
			Visual	Audio			
MUSIC-AVQA [17]	–	–	ResNet-18 [9]	VGGish [10]	71.52	–	–
HCRN+HAVF [29]	–	–	ResNet-101 [9]	PANNs [15]	–	89.00	–
VAST [2]	–	–	CLIP-ViT-L/14 [22]	BEATs [3]	80.70	–	69.49
Crab [5]	–	Vicuna-7B [32]	CLIP-ViT-L/14 [22]	BEATs [3]	–	–	39.23
Ola [19]	–	Qwen2.5-7B	SigLIP	WSP+BEATs [3]	–	–	72.26
VideoLLaMA 2 [4]	–	Qwen2-7B	CLIP-ViT-L/14 [22]	BEATs [3]	79.20	–	66.80
PAVE [20]	–	Qwen2-7B	SigLIP	IB	82.30	93.80	–
SHMamba [30]	✓	–	CLIP-ViT-L/14 [22]	VGGish [10]	74.12	90.80	–
AV-BIMBA (M-AVQA)	✓	Qwen2-7B	SigLIP	AST	75.12	84.04	65.08
				IB	76.13	84.70	66.94
				W2V	**76.56**	84.78	65.92
				WSP	75.67	83.81	64.72
AV-BIMBA (AVQA)	✓	Qwen2-7B	SigLIP	AST	66.74	91.98	65.45
				IB	66.29	91.98	65.40
				W2V	58.83	91.95	65.30
				WSP	64.20	**92.03**	65.12
AV-BIMBA (V-AVQA)	✓	Qwen2-7B	SigLIP	AST	57.94	78.80	**76.97**
				IB	60.10	79.94	76.51
				W2V	57.00	78.73	76.35
				WSP	61.23	80.68	76.51
AV-BIMBA (All)	✓	Qwen2-7B	SigLIP	AST	75.10	<u>91.93</u>	<u>76.10</u>
				IB	76.13	91.23	75.93
				W2V	<u>76.52</u>	91.08	75.42
				WSP	75.67	91.30	75.87

Table 1: Fine-tuning comparison of AV-BIMBA variants across evaluated datasets. Reported results for compared methods are taken from the corresponding dataset or model publications. Accuracy (%) is used as the evaluation metric.

4.2 Fine-Tuning-Results

We fine-tune four AV-BIMBA variants separately on MUSIC-AVQA (M-AVQA) [17], AVQA [29], Valor32k-AVQA v2.0 (V-AVQA) [24], and on a combined dataset of all three (All), using four different audio encoders. Table 1 reports the fine-tuning results and compares AV-BIMBA with state-of-the-art (SoTA) methods also fine-tuned on these datasets.

From Table 1, methods without an MLLM decoder generally underperform MLLM-based approaches across datasets. A key limitation is that they formulate AVQA as a classification problem rather than generative reasoning, which restricts flexibility and generalization to diverse and open-ended scenarios. Among MLLM-based but non-SSM methods, VideoLLaMA 2 [4] and PAVE [20] achieve slightly better performance than the AV-BIMBA (M-AVQA) and AV-BIMBA (AVQA) variants fine-tuned on their respective datasets.

In contrast, the AV-BIMBA (V-AVQA) variant, fine-tuned on V-AVQA, achieves SoTA performance, surpassing the strongest non-SSM method, Ola [19], by 4.71%. Notably, the V-AVQA dataset is more challenging, with real-world scenarios unlike AVQA and the music-focused M-AVQA.

All AV-BIMBA variants fine-tuned on their respective datasets outperform SHMamba [30], the only SSM-based model, making AV-BIMBA the SoTA SSM-based model for AVQA task, without increasing sequence length or computa-

Method		Daily-Omni	WorldSense
Gemini 2.0 Flash		67.84	–
Gemini 1.5 Pro [26]		–	48.00
Unified-IO-2 XXL [21]		28.24	25.90
VideoLLaMA 2 [4]		35.17	25.40
Qwen2.5-Omni [27]		47.45	**45.40**
Ola [19]		50.71	–
Daily-Omni Agent [33]		**61.82**	–
AV-BIMBA (M-AVQA)	AST	40.35	32.02
	IB	43.94	35.88
	W2V	44.27	34.74
	WSP	39.77	33.07
AV-BIMBA (AVQA)	AST	43.36	36.38
	IB	43.27	38.15
	W2V	41.94	36.92
	WSP	39.68	36.06
AV-BIMBA (V-AVQA)	AST	51.46	39.37
	IB	<u>52.46</u>	<u>40.80</u>
	W2V	<u>52.46</u>	40.67
	WSP	50.54	38.21
AV-BIMBA (All)	AST	46.45	39.31
	IB	49.29	38.65
	W2V	47.03	38.02
	WSP	49.96	39.22

Table 2: Zero-shot evaluation of AV-BIMBA variants with different audio encoders on the Daily-Omni and WorldSense (% accuracy).

tional complexity, thanks to its simple addition-based audio-visual spatiotemporal fusion method (Eq. 1). The AV-BIMBA (All) variant shows consistent performance across datasets and audio encoders, though it remains slightly below AV-BIMBA variants fine-tuned on their respective datasets.

Although no single audio encoder consistently outperforms the others, each AV-BIMBA variant excels on its respective fine-tuned dataset across all audio encoders.

4.3 Zero-Shot Evaluation

Zero-shot performance of the AV-BIMBA variants is summarized in Table 2. Evaluation is performed on the more recent Daily-Omni [33] and WorldSense [11] benchmarks, which feature longer videos than the 10-second clips used for fine-tuning and require integrated audio-visual reasoning. Published results from several MLLM-based systems are included for comparison.

The results in Table 2 show that closed models like Gemini 2.0 Flash and Gemini 1.5 Pro [26] achieve the best performance across both datasets, likely due to extensive pretraining on large-scale multimodal data. Among open-source MLLMs, Daily-Omni Agent [33], the base model for the Daily-Omni dataset, achieves SoTA performance, but relies on very large audio (7B), visual (7B), and language (14B) models, limiting its scalability for AVQA tasks.

Apart from Daily-Omni Agent [33], AV-BIMBA (V-AVQA) outperforms all other open-source models on the Daily-Omni dataset. On WorldSense, with much

Method		$\text{LoRA}_{weights}$	V-AVQA			Daily-Omni	WorldSense
$\text{AV-BIMBA}_{wo/audio}$ (V-AVQA)		–		69.64		45.87	37.21
AV-BIMBA (V-AVQA)	AST	–		71.00		47.95	38.49
	IB			70.96		48.96	38.36
	W2V			70.84		48.45	38.55
	WSP			70.93		49.62	37.98
Fusion method			†		‡		
AV-BIMBA (V-AVQA)	AST	BIMBA	62.20	**76.97**	62.81	51.46	39.37
	IB		62.78	76.51	61.48	**52.46**	**40.80**
	W2V		62.70	76.35	61.04	**52.46**	40.67
	WSP		62.70	76.51	61.85	50.54	38.21

Table 3: Impact of audio integration, MLLM LoRA initialization, and concatenation† and cross-attention‡ fusion methods in AV-BIMBA.

longer videos (average 140 seconds) compared to the 10-second V-AVQA clips used for fine-tuning, AV-BIMBA (V-AVQA) performs slightly below Qwen2.5-Omni [28]. This stems from AV-BIMBA's audio-visual queries (Eq. 4), which efficiently capture cross-modal temporal and semantic correlations.

4.4 Ablation Study

We conduct an ablation study to assess the impact of audio integration, MLLM decoder adaptation, and audio–visual fusion strategies in AV-BIMBA. Table 3 indicates that audio integration consistently improves AV-BIMBA's performance, while the audio-agnostic variant ($\text{AV-BIMBA}_{wo/audio}$) underperforms, highlighting the importance of audio cues and the effectiveness of simple addition–based integration (Eq. 1) for SSM-based AVQA. Table 3 also demonstrates the critical role of MLLM decoder adaptation. Fine-tuning the MLLM decoder without LoRA yields similar performance with and without audio, indicating limited learning of audio cue integration. In contrast, initializing LoRA from BIMBA's pretrained weights significantly improves fine-tuning and zero-shot performance by retaining pretrained visual knowledge and incorporating audio context.

Table 3 further reveals the effect of replacing AV-BIMBA's addition-based audio–visual fusion (Eq. 1) with concatenation† or cross-attention‡. Both alternatives degrade performance compared to simple addition, suggesting heavier fusion is less suited to SSM-based architectures that favor lightweight, order-preserving interactions for robust sequential modeling.

4.5 Qualitative Evaluation

Fig. 2a demonstrates that AV-BIMBA (V-AVQA) correctly answers a multiple-choice AVQA question by associating a spoken utterance with the corresponding visual speaker. Fig. 2b illustrates the model's generative capability, where

Question: Who says, "We look forward to seeing you and your brother square off", in the video?

A. A man wearing a red headband
B. A man wearing yellow gloves
C. A man wearing a red jacket
D. A man wearing a blue jacket

Ground Truth: **D** Predicted Answer: **D**

(a)

Question: How many sounding ukulele in the video?

Ground Truth: **Three** Predicted Answer: **Three**

(b)

Fig. 2: Successful AVQA examples by AV-BIMBA (V-AVQA) on (a) WorldSense [11] and (b) M-AVQA [17].

temporally distributed acoustic signals are aligned with visual evidence to enable coherent reasoning about sound-producing objects. Both examples show that AV-BIMBA (V-AVQA) effectively aligns audio events with visual objects, thanks to its CMSS module that enables efficient spatiotemporal cross-modal fusion.

5. CONCLUSION

We presented AV-BIMBA, an SSM-based audio-visual model leveraging an MLLM decoder and fusing audio and visual information spatiotemporally via a cross-modal selective state-space (CMSS) module. Learned audio-visual queries within CMSS capture temporal and semantic correlations, enabling effective performance on AVQA tasks. AV-BIMBA consistently outperforms existing SSM-based methods and shows competitive performance compared to larger open-source MLLMs in both fine-tuning and zero-shot evaluations, without increasing sequence length or computational complexity. Future work includes extending AV-BIMBA to longer, real-world videos and exploring more efficient MLLM decoders.

Acknowledgements We acknowledge the computational resources provided by the Aalto Science-IT project and the support of the Research Council of Finland through the FCAI Flagship (grant 345552).

Disclosure of Interests. The authors declare no conflict of interest.

References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. NeurIPS **33**, 12449–12460 (2020)
2. Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., Liu, J.: VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. NeurIPS **36**, 72842–72866 (2023)
3. Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X., Wei, F.: BEATs: audio pre-training with acoustic tokenizers. In: ICML (2023)
4. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs. arXiv preprint arXiv:2406.07476 (2024)
5. Du, H., Li, G., Zhou, C., Zhang, C., Zhao, A., Hu, D.: Crab: A unified audio-visual scene understanding model with explicit cooperation. In: CVPR. pp. 18804–18814 (2025)
6. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K., Joulin, A., Misra, I.: ImageBind: One embedding space to bind them all. In: CVPR. pp. 15180–15190 (2023)
7. Gong, Y., Chung, Y., Glass, J.: AST: Audio spectrogram transformer. In: Interspeech. pp. 571–575 (2021)
8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First Conference on Language Modeling (2024)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
10. Hershey, S., Chaudhuri, S., Ellis, D., Gemmeke, J., Jansen, A., Moore, R., Plakal, M., Platt, D., Saurous, R., Seybold, B., et al.: CNN architectures for large-scale audio classification. In: ICASSP. pp. 131–135 (2017)
11. Hong, J., Yan, S., Cai, J., Jiang, X., Hu, Y., Xie, W.: WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. arXiv preprint arXiv:2502.04326 (2025)
12. Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
13. Hurst, A., Lerer, A., Goucher, A., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
14. Islam, M., Nagarajan, T., Wang, H., Bertasius, G., Torresani, L.: BIMBA: selective-scan compression for long-range video question answering. In: CVPR. pp. 29096–29107 (2025)
15. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.: PANNS: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE Trans. ASLPRO. **28**, 2880–2894 (2020)
16. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: LLaVA-OneVision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
17. Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios. In: CVPR. pp. 19108–19118 (2022)
18. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: VideoMamba: State space model for efficient video understanding. In: ECCV. pp. 237–255 (2024)
19. Liu, Z., Dong, Y., Wang, J., Liu, Z., Hu, W., Lu, J., Rao, Y.: Ola: Pushing the frontiers of omni-modal language model. arXiv preprint arXiv:2502.04328 (2025)

20. Liu, Z., Li, Y., Nguyen, K., Zhong, Y., Li, Y.: PAVE: Patching and adapting video large language models. In: CVPR. pp. 3306–3317 (2025)
21. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-IO 2: Scaling autoregressive multimodal models with vision language audio and action. In: CVPR. pp. 26439–26455 (2024)
22. Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
23. Radford, A., Kim, J., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: ICML. pp. 28492–28518 (2023)
24. Riahi, I., Radman, A., Guo, Z., Hedjam, R., Laaksonen, J.: Valor32k-AVQA v2.0: Open-ended audio-visual question answering dataset and benchmark. In: ACM MM. pp. 13097–13103 (2025)
25. Shanmugam, V.: Audio-visual state space model for question answering with bimba (2026), https://aaltodoc.aalto.fi/items/19bbbcf9-41d9-47ed-b4af-83626a2a1489, accessed: 2026-06-22
26. Team, G., Georgiev, P., Lei, V., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
27. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., et al.: Qwen2.5-omni technical report. arXiv preprint arXiv:2503.20215 (2025)
28. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., et al.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)
29. Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., Zhu, W.: AVQA: A dataset for audio-visual question answering on videos. In: ACM MM. pp. 3480–3491 (2022)
30. Yang, Z., Li, W., Cheng, G.: SHMamba: Structured hyperbolic state space model for audio-visual question answering. IEEE Trans. ASLPRO. **33**, 3582–3593 (2025)
31. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV. pp. 11975–11986 (2023)
32. Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging LLM-as-a-judge with mt-bench and chatbot arena. NeurIPS **36**, 46595–46623 (2023)
33. Zhou, Z., Wang, R., Wu, Z.: Daily-Omni: Towards audio-visual reasoning with temporal alignment across modalities. arXiv preprint arXiv:2505.17862 (2025)