SG-Mamba: Sparse Graph-Guided Mamba for Audio-Visual Speech Enhancement

Guo-Ruei Tseng*, Hung-Shin Lee‡, Hsin-Min Wang†, and Berlin Chen*
*Dept. of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
†Inst. of Information Science, Academia Sinica, Taiwan
‡Grad. Inst. of AI Interdisciplinary Applied Technology, National Taiwan Normal University, Taiwan

***Abstract*—Lightweight audio-visual speech enhancement (AVSE) models face a critical trade-off between computational efficiency and cross-modal alignment accuracy. While simple concatenation lacks relational expressiveness, dense cross-attention incurs computational overhead and is prone to unreliable cross-modal correspondence under strong acoustic interference. We propose Sparse Graph-Guided Mamba (SG-Mamba), a lightweight AVSE framework that integrates a sparse heterogeneous graph with a linear-complexity Mamba backbone. The graph explicitly models modality-specific relations through content-adaptive attention and cross-frame audio-visual connections, while Mamba captures long-range temporal context. We further introduce an audio skip connection to preserve spectral detail without sacrificing noise suppression. Evaluated on LRS3, SG-Mamba achieves competitive or superior performance against strong lightweight baselines and reaches 13.091 dB SI-SDR under noise-only condition. It also remains robust in cluttered multi-speaker conditions with a competitive cost of 3.45 G MACs (or 6.90 G FLOPs). Results on VoxCeleb2 further suggest that explicit structural priors improve robustness, generalizability, and computational efficiency in lightweight AVSE.**

***Index Terms*—audio-visual speech enhancement, graph-guided multimodal fusion, graph neural networks, state space models, Mamba**

I. Introduction

Speech enhancement (SE) is a core component in applications such as smartphones and hearing aids, where the objective is to recover clean speech from noisy observations [1], [2]. Compared with audio-only systems, audio-visual speech enhancement (AVSE) improves robustness by leveraging visual cues such as lip motion [3]–[5]. Recently, Real-Time Audio-Visual Speech Enhancement Using Pre-trained Visual Representations (RAVEN) [6] demonstrated strong performance by leveraging robust pre-trained visual front-ends [7]–[10]. To further improve temporal sequence modeling while maintaining computational efficiency, Mamba [11]–[14] provides linear-complexity sequence modeling while maintaining competitive long-range dependency modeling, making it an attractive backbone for lightweight AVSE.

Despite these advances, fusion design in AVSE still faces a trade-off between computational efficiency (e.g., latency and parameter count) and the expressive capacity required for effective cross-modal relational modeling [15]. Methods with strong enhancement performance, such as diffusion [16]–[18] and flow-matching models [19], typically require expensive sampling or parameter-heavy backbones. Transformer-based methods [20]–[22] achieve strong performance through expressive attention-based backbones, but their deep architectures remain costly for practical deployment. In contrast, efficient architectures often fall into a "Concatenation Trap," relying on simple concatenation, feature-wise modulation (FiLM) [23], or linear attention without explicit structural constraints. These designs fuse heterogeneous modalities in a shared feature space without explicitly modeling their structural correspondence, forcing the backbone to infer cross-modal relations implicitly. Alternatively, increasing the interaction capacity through dense global cross-attention [24] partially alleviates this limitation; however, without explicit structural constraints, these global interactions remain susceptible to unreliable cross-modal correspondence under severe acoustic interference [25]. Therefore, existing AVSE systems lack a structurally bounded fusion mechanism that simultaneously accommodates local audio-visual micro-asynchrony [26], [27] while preventing erroneous alignments from propagating globally.

Motivated by these limitations and inspired by the success of graph neural networks in multimodal fusion [28] and speech enhancement [29], we propose Sparse Graph-Guided Mamba (SG-Mamba) [1], a lightweight framework that employs explicit sparse graph-guided cross-modal interaction. Our main contributions are summarized as follows:

1) **Heterogeneous Graph for Fine-Grained Fusion:** We replace dense quadratic attention with a sparse heterogeneous graph processed via dynamic, content-adaptive attention. This explicitly models modality-specific interactions under varying acoustic conditions, striking a practical balance between the weak structural prior of concatenation and the high computational cost of full attention.
2) **Explicit Cross-Frame Modeling:** Instead of relying entirely on the backbone to implicitly infer temporal offsets, we explicitly introduce cross-frame edges connecting neighboring visual frames to the corresponding audio nodes. The resulting graph propagation enables flexible local alignment before temporal sequence modeling, improving robustness under temporal mismatch conditions.
3) **Synergistic SG-Mamba Backbone:** We combine a U-Net-structured dynamic graph attention network (GAT) [30] with a state-space architecture (Mamba). By dedi-

[1]The source code is available at: https://github.com/0u5gary/SG-Mamba

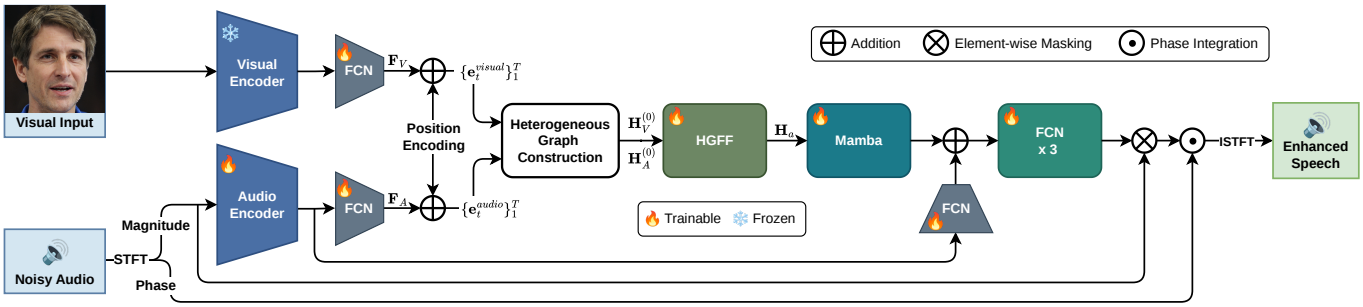


Fig. 1. Architecture of the proposed SG-Mamba. Modality-specific encoders extract audio and visual features ($\mathbf{F}_A, \mathbf{F}_V$), which are position-encoded and fed into the Heterogeneous Graph Construction module to yield node feature matrices $\mathbf{H}_A^{(0)}$ and $\mathbf{H}_V^{(0)}$. A HGFF module then performs content-adaptive cross-modal alignment, outputting $\mathbf{H}_a$ to a unidirectional Mamba block for global temporal modeling. Finally, 3 stacked FCN layers generate a mask for magnitude refinement. The refined magnitude is then integrated with the original noisy phase and reconstructed via ISTFT.

cating the graph-based U-Net to hierarchical cross-modal relational aggregation and the state-space model to long-range utterance-level temporal modeling, this targeted architectural division effectively mitigates severe acoustic interference without inflating computational costs.

II. Methodology

Our framework extends RAVEN [6] by introducing a structurally graph-guided fusion module and a Mamba-based temporal modeling backbone. As illustrated in Fig. 1, SG-Mamba adopts a modular four-stage design for lightweight offline AVSE, which consists of: (1) Feature Extraction and Node Embedding, (2) Heterogeneous Graph Construction, (3) Hierarchical Graph-Structured Feature Fusion, and (4) Mamba Modeling and Speech Reconstruction.

A. Feature Extraction and Node Embedding

Given a noisy audio stream and a synchronized visual stream, we first apply a Short-Time Fourier Transform (STFT) to obtain the magnitude and phase spectrograms, followed by power compression p of the magnitude. We feed the magnitude spectrogram into an Audio Encoder (five CNN layers, each followed by BatchNorm and ReLU, with a temporal receptive field of 5), followed by a Fully Connected Network (FCN) layer for dimensionality reduction, yielding audio features $\mathbf{F}_A \in \mathbb{R}^{T\times D}$, where T denotes the total number of frames and D denotes the feature dimension. We retain the phase component for final waveform reconstruction. Similarly, we pass visual frames through a pre-trained visual encoder, then temporally upsample them to match the audio frame rate, and apply an FCN to obtain lip-reading visual features $\mathbf{F}_V \in \mathbb{R}^{T\times D}$.

Because graph message passing does not inherently encode temporal order, we add Sinusoidal Positional Encoding (PE) [31] to both modalities to preserve temporal order. We obtain the final node embeddings as $\mathbf{e}_t^m = \mathbf{f}_t^m + PE_t, \quad m \in \{\text{audio}, \text{visual}\}$, where $\mathbf{e}_t^m$ denotes the embedding vector for modality m at frame t and serves as the input to Heterogeneous Graph Construction.

B. Heterogeneous Graph Construction

To explicitly capture inter-modal interactions and accommodate dynamic micro-asynchrony, we construct a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The node set $\mathcal{V}$ is partitioned into an audio subset $\mathcal{V}_a = \{A_1, \ldots, A_t, \ldots, A_T\}$ and a visual subset $\mathcal{V}_v = \{V_1, \ldots, V_t, \ldots, V_T\}$. Each node is initialized with embedding $\mathbf{e}_t^m$, yielding the audio and visual node feature matrices $\mathbf{H}_A^{(0)}$ and $\mathbf{H}_V^{(0)}$, which are fed separately into the Hierarchical Graph-Structured Feature Fusion (HGFF) module.

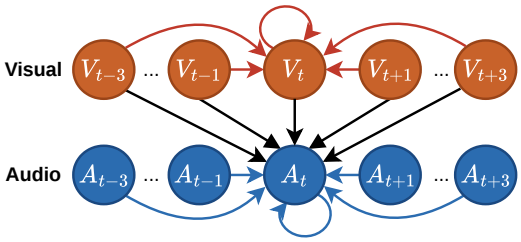


Fig. 2. The proposed heterogeneous graph centered at the current frame t. Red and blue arrows denote intra-modal temporal connections within each modality, while black arrows indicate unidirectional cross-modal interactions from the local visual window $V_{t\pm3}$ to the audio node A_t.

As shown in Fig. 2, all nodes retain self-loops ($A_t \to A_t$, $V_t \to V_t$) to preserve frame-level intrinsic features during message passing. To model intra-modal temporal continuity, we connect neighboring frames from $t-3$ to $t+3$ to the current node at frame t within each modality. For cross-modal interactions, audio node A_t establish directed edges from a local visual window $V_{t\pm3}$ (approximately ±30 ms). Crucially, these cross-modal edges are strictly unidirectional ($V_{t\pm3} \to A_t$), as allowing noisy acoustic features to propagate into the visual stream could destabilize its cleaner representations. Moreover, restricting aggregation to a local ±3-frame window reduces computational overhead. This design deliberately trades long-range dense alignment for local stability, ensuring the cross-modal graph remains lightweight while effectively capturing immediate phoneme-viseme correlations. Formally, we define the temporal neighbor window at frame t as $\mathcal{N}_t = \{t+k \mid k \in [-3,3],\ 0 \le t+k < T\}$, which is applied to audio-audio (AA), visual-audio (VA), and visual-visual (VV) interaction edge types. The resulting graph connectivity is denoted by the adjacency matrix $\mathbf{A}_\mathcal{G}$. In practice, $\mathbf{A}_\mathcal{G}$ is implemented as two sparse neighbor index tables of shape $T \times k$ (where $k = 2\times3+1 = 7$), one shared across AA and VV interactions and one for VA, avoiding materialization of the full $2T \times 2T$ matrix across the batch dimension.

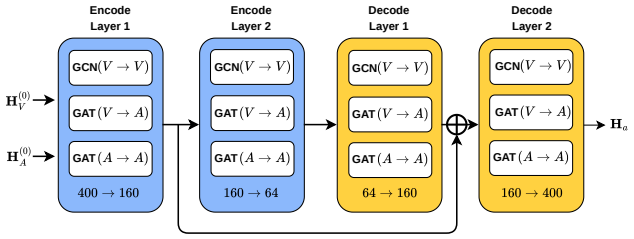


Fig. 3. Detailed architecture of the HGFF module, featuring modality-specific graph operations within a symmetrical U-Net backbone.

C. Hierarchical Graph-Structured Feature Fusion

As illustrated in Fig. 3, we employ the HGFF module to perform content-adaptive, fine-grained cross-modal structural alignment. Within each layer l in HGFF, the three edge types defined in $\mathbf{A}_{\mathcal{G}}$ are handled differently according to their respective characteristics to yield the layer's audio and visual features, denoted as $\mathbf{H}_A^{(l)}$ and $\mathbf{H}_V^{(l)}$. Let $[\mathbf{m}_{(\cdot)}]_t \in \mathbb{R}^D$ denote the aggregated message vector at frame t for interaction type $(\cdot) \in \{AA, VA, VV\}$. For interaction types AA and VA, the resulting message vector is aggregated via a GAT layer:

$$[\mathbf{m}_{(\cdot)}]_t = \sum_{j \in \mathcal{N}_t} \alpha_{tj} \mathbf{W}_{(\cdot)} \mathbf{h}_j^{(l)}, \tag{1}$$

where $\mathbf{h}_j^{(l)}$ denotes the features of neighbor j drawn from $\mathbf{H}_A^{(l)}$ for AA and $\mathbf{H}_V^{(l)}$ for VA, and $\mathbf{W}_{(\cdot)}$ is a linear projection matrix. Specifically, to compute the attention weight α_{tj}, which weights the contribution from neighbor j to t, we adopt the dynamic graph attention mechanism [30]. This approach avoids the static attention degradation present in standard GAT by applying a joint nonlinear transformation (LeakyReLU) to the query-key pair before projecting it into a scalar score via a learnable vector $\mathbf{a}$. Subsequently, a learnable temperature parameter τ is applied before softmax normalization to prevent degenerate attention distributions and stabilize training. All parameters and computations, including $\mathbf{a}$, τ, and the resulting α_{tj}, are maintained and performed independently for AA and VA to capture the distinct dynamics of each interaction type. For the remaining interaction type VV, a Graph Convolutional Network (GCN) mean aggregation is adopted:

$$[\mathbf{m}_{(\cdot)}]_t = \mathbf{W}_{(\cdot)} \left(\frac{1}{|\mathcal{N}_t|} \sum_{j \in \mathcal{N}_t} \mathbf{h}_j^{(l)} \right), \tag{2}$$

where $\mathbf{h}_j^{(l)}$ is drawn from $\mathbf{H}_V^{(l)}$. This uniform aggregation stabilizes visual anchor representations without attention overhead, reflecting the temporal redundancy in upsampled lip features, where highly correlated frames make averaging sufficient.

Stacking $[\mathbf{m}_{(\cdot)}]_t$ across all T frames yields the message matrix $\mathbf{M}_{(\cdot)} \in \mathbb{R}^{T \times D}$. The audio and visual node feature matrices are then updated via projected residual connections followed by layer normalization (LN):

$$\mathbf{H}_A^{(l+1)} \leftarrow \mathrm{LN}\Big(\mathbf{W}_A \, \mathbf{H}_A^{(l)} + \mathbf{M}_{AA} + \mathbf{M}_{VA} \Big), \tag{3}$$

$$\mathbf{H}_V^{(l+1)} \leftarrow \mathrm{LN}\Big(\mathbf{W}_V \, \mathbf{H}_V^{(l)} + \mathbf{M}_{VV} \Big), \tag{4}$$

where $\mathbf{W}_A$ and $\mathbf{W}_V$ are residual projections applied only when input and output dimensions differ. These HGFF layers are stacked into a 4-layer U-Net backbone, compressing dimensions at the encoder and symmetrically expanding at the decoder, with skip additions introduced before the final decoder block to preserve lower-level fine-grained cues, resulting in an effective receptive field of ± 12 frames (± 120 ms). Furthermore, this architecture provides inherent structural constraints and noise-robust multi-scale feature representations, ensuring stable multi-modal integration even under severe acoustic degradation. After the final decoder layer, we retain only the aggregated audio node representations $\mathbf{H}_a \in \mathbb{R}^{T \times D}$, which are passed to the subsequent Mamba block.

Unlike uniform 1D convolutions that mix modalities indiscriminately, or Transformer-based cross-attention that incurs quadratic $\mathcal{O}(T^2)$ complexity, our approach encodes explicit cross-modal structural priors through the sparse asymmetric $\mathbf{A}_{\mathcal{G}}$. By routing cross-modal information dynamically via GAT while using lightweight GCN mean pooling for highly redundant visual features, this sparse topology bridges the audio-visual semantic gap while preserving the flexibility of dynamic representation learning within a tightly controlled computational budget.

D. Mamba Modeling and Speech Reconstruction

To capture global temporal dependencies, we feed the aggregated audio features $\mathbf{H}_a$ into a unidirectional Mamba block. Since local bidirectional context has already been explicitly aggregated by the HGFF module, the subsequent Mamba is only responsible for modeling residual long-range temporal dependencies. Therefore, we adopt a unidirectional scan to avoid redundant temporal modeling while maintaining a lightweight backend. Specifically, Mamba performs selective scanning with input-dependent parameters $(\Delta_t, \mathbf{B}_t, \mathbf{C}_t)$. For the t-th frame $\mathbf{x}_t \in \mathbb{R}^D$ of $\mathbf{H}_a$, the discrete state update is:

$$\mathbf{h}_t = \bar{\mathbf{A}}_t \mathbf{h}_{t-1} + \bar{\mathbf{B}}_t \mathbf{x}_t, \quad \mathbf{y}_t = \mathbf{C}_t \mathbf{h}_t, \tag{5}$$

where the discretized parameters $\bar{\mathbf{A}}_t = \exp(\Delta_t \mathbf{A})$ and $\bar{\mathbf{B}}_t \approx \Delta_t \mathbf{B}_t$ (derived from the base transition $\mathbf{A}$) dynamically gate information flow and improve robustness.

To align latent dimensions and compensate for Mamba compression, we add the Mamba output with the skip connection from $\mathbf{F}_A$, and then apply LN to the summed representation. Three stacked FCN layers (each with ReLU activation) then serve as a non-linear fusion head, reconciling Mamba's structural consistency with fine spectral details. Subsequently, a sigmoid activation generates a multiplicative mask $\mathbf{M} \in [0, 1]^{T \times D}$. We apply $\mathbf{M}$ to the power-compressed magnitude, invert the compression $1/p$, combine with the noisy phase, and reconstruct the enhanced waveform via Inverse Short-Time Fourier Transform (ISTFT).

For training, following RAVEN [6], we adopt a composite spectral loss that combines complex-domain and magnitude reconstruction terms to jointly supervise phase consistency and spectral fidelity:

$$\mathcal{L}_{total}(\hat{\mathbf{S}}, \mathbf{S}) = ||\hat{\mathbf{S}} - \mathbf{S}||_2 + |||\hat{\mathbf{S}}| - |\mathbf{S}|||_2, \tag{6}$$

where $\hat{\mathbf{S}}$ and $\mathbf{S}$ denote the estimated and clean power-compressed spectrograms, respectively.

III. Experimental Setup

A. Dataset and Evaluation

We used 433 hours of audio-visual data from LRS3 [32] to build target speech, visual cues, and interfering speech samples, together with 180 hours of background noise and reverberation data from the DNS Challenge [33]. Following the official LRS3 split, the pretrain, trainval, and test sets were used for training, validation, and evaluation respectively, ensuring no speaker overlap across splits. To simulate complex acoustic environments, we dynamically mixed target speech with either LRS3 interferers or diverse DNS noise at SNRs uniformly sampled from -10 dB to 10 dB, covering both low- and high-interference conditions. All audio-visual segments were truncated or zero-padded to a fixed duration of 5 seconds for consistent model input length. To further evaluate cross-dataset generalizability, we applied the LRS3-trained models directly to the official VoxCeleb2 test set [34] without fine-tuning, using the same mixing protocol and SNR range. We evaluated performance using three standard metrics: PESQ [35] for perceptual quality, ESTOI [36] for intelligibility, and SI-SDR [37] for signal-level distortion reduction. Among these, SI-SDR is used as the sole metric in Secs. IV-C and IV-D to highlight performance differences.

B. Implementation Setup

The audio stream was resampled to 16 kHz and converted into 257-dimensional power-compressed magnitude spectrograms (25 ms Hann window, 10 ms hop size, 512-point FFT, 0.3 compression rate), while the video stream provided 96×96 lip ROIs. Following RAVEN [6], we extracted visual features using the frozen pre-trained visual front-end from AVHuBERT [7] and TalkNet [9], whose outputs were concatenated into 1,280-dimensional embeddings. Since the visual front-end is frozen and identical across all compared methods, the term "lightweight" refers to the trainable fusion and enhancement backend, whose complexity is reported for fair architectural comparison. Comparisons are limited to reproducible lightweight backbones under a unified experimental protocol. The backend consisted of the proposed HGFF backbone with single-head GAT (compressing from $D = 400$ to a bottleneck of 64 and symmetrically expanding to 400) followed by a unidirectional Mamba block ($d_{\text{state}} = 64$, $d_{\text{conv}} = 4$, expansion factor $= 2$). All compared models were trained with the Adam optimizer at an initial learning rate of 10^{-5}. To ensure fair comparison, no architecture-specific hyperparameter tuning was performed. A dropout rate of 0.1 was applied to all attention-based and graph aggregation layers. To stabilize training, a dynamic learning rate scheduling strategy was employed, which halved the learning rate when the validation loss plateaued. We used a micro-batch size of 16 with gradient accumulation, yielding an effective batch size of 128.

TABLE I
Performance comparison under noise-only scenario with mixed SNR condition from [-10 dB, 10 dB].

Method	PESQ	SI-SDR	ESTOI	Para. (↓)	MACs (↓)
Noisy	1.841	2.551	0.812	–	–
RAVEN (Base) [6]	2.300	12.648	0.837	<u>4.7 M</u>	<u>3.19 G</u>
Mamba	**2.305**	12.463	<u>0.840</u>	**4.5 M**	**3.16 G**
SG-RAVEN	2.286	<u>12.857</u>	0.838	5.4 M	3.47 G
SG-Mamba	<u>2.302</u>	**13.091**	**0.842**	5.3 M	3.45 G

C. Baseline Configuration

To assess the framework under lightweight constraints, all baselines share the same audio and visual encoders and operate at a hidden dimension of $D = 400$. RAVEN serves as the primary baseline, with two modifications applied for comparability. First, the concatenated audio-visual features are projected to $D = 400$ via an FCN before being fed into the LSTM, rather than passing the full high-dimensional representation directly as in the original implementation. Second, an audio skip connection with layer normalization is added before the output FCN layers to match the reconstruction pipeline of the proposed method. The Mamba baseline follows the same design principle as the modified RAVEN, replacing the LSTM with a unidirectional Mamba block under otherwise identical configurations. The Bi-Mamba ablation baseline extends the Mamba baseline with a bidirectional design, where forward and backward scans are performed independently and combined via element-wise addition prior to the residual connection. For the remaining ablation baselines, audio and visual features are independently projected to $D = 400$ before fusion: FiLM-Mamba applies feature-wise linear modulation to inject visual conditioning into the audio stream, Linear-Mamba employs a linear attention mechanism [38] where audio attends to visual features, and Cross-Mamba uses a 4-head cross-attention module where audio queries attend to visual keys and values. All four ablation baselines also incorporate the same audio skip connection as the RAVEN and Mamba baselines. SG-RAVEN adopts the same heterogeneous graph front-end as SG-Mamba but substitutes the Mamba block with LSTM, serving as a direct backbone comparison under identical fusion conditions.

IV. Results

A. Performance on Non-Speech Noise

As shown in Table I, SG-Mamba achieved the highest SI-SDR of 13.091 dB, while maintaining stable PESQ and ESTOI comparable to the strongest baselines, indicating that the graph-guided fusion improves signal-level separation without compromising perceptual quality or intelligibility. Compared with the strongest baselines, SG-Mamba improved SI-SDR by 0.443 dB over RAVEN and 0.628 dB over Mamba, confirming that explicit structural guidance at the fusion stage yields consistent signal-level gains under non-speech noise. Notably, SG-RAVEN also outperformed RAVEN by 0.209 dB in SI-SDR, suggesting that the heterogeneous graph fusion provides complementary structural cues that benefit both backbone types

TABLE II
PERFORMANCE COMPARISON UNDER **1-INTERFERER** SCENARIO ACROSS DIFFERENT SNR LEVELS. THE INTERFERING SPEAKER IS NOT VISIBLE.

Method	0 dB			5 dB			10 dB		
	PESQ	SI-SDR	ESTOI	PESQ	SI-SDR	ESTOI	PESQ	SI-SDR	ESTOI
Noisy	1.314	0.941	0.627	1.553	5.942	0.732	1.956	**10.942**	0.824
RAVEN (Base) [6]	1.493	2.176	0.636	1.763	5.871	0.738	2.138	8.773	0.822
Mamba	1.496	2.474	0.644	1.761	6.045	0.742	2.124	8.640	0.822
SG-RAVEN	1.485	2.388	0.637	1.750	6.124	0.739	2.127	9.054	0.823
SG-Mamba (Ours)	**1.509**	**2.752**	**0.647**	**1.775**	**6.229**	**0.745**	**2.143**	9.096	**0.827**

TABLE III
PERFORMANCE COMPARISON UNDER **3-INTERFERER** SCENARIO ACROSS DIFFERENT SNR LEVELS. THE INTERFERING SPEAKER IS NOT VISIBLE.

Method	0 dB			5 dB			10 dB		
	PESQ	SI-SDR	ESTOI	PESQ	SI-SDR	ESTOI	PESQ	SI-SDR	ESTOI
Noisy	1.200	1.064	0.526	1.406	**6.064**	0.665	1.794	**11.063**	0.786
RAVEN (Base) [6]	1.392	1.841	0.558	1.665	5.485	0.692	2.048	8.511	0.799
Mamba	1.401	2.293	0.565	1.672	5.716	0.696	2.044	8.414	0.800
SG-RAVEN	1.385	2.145	0.556	1.654	5.808	0.694	2.035	8.842	0.801
SG-Mamba (Ours)	**1.403**	**2.372**	**0.567**	**1.676**	6.011	**0.703**	**2.056**	9.190	**0.810**

under diverse noise conditions. Nevertheless, the performance gap between SG-Mamba and SG-RAVEN (0.234 dB) indicates that Mamba's selective state-space mechanism integrates the graph-aggregated features more effectively than LSTM-based gating, particularly in leveraging long-range temporal context for noise suppression. These improvements were achieved with 5.3 M parameters and 3.45 G MACs, demonstrating a favorable performance-complexity trade-off.

B. Performance on Interfering Speakers

To evaluate robustness against off-screen interfering voices—mimicking real-world scenarios where background chatter mismatches visible lip movements—we tested under 1- and 3-interferer conditions across 0, 5, and 10 dB SNRs. As expected, all methods showed reduced SI-SDR at 10 dB relative to the unprocessed noisy input, reflecting that mask-based enhancement on relatively clean speech can introduce reconstruction artifacts that outweigh the suppressed interference. Despite this, SG-Mamba consistently achieved the best or competitive performance across all SNR levels and interferer counts, demonstrating its strong robustness against various degrees of acoustic interference.

1) One Interfering Speaker: As shown in Table II, SG-Mamba achieved the best performance on all metrics across all SNR levels, confirming that the graph-guided fusion provides consistent gains over the RAVEN and Mamba baselines under single-interferer conditions. SG-Mamba outperformed Mamba in SI-SDR by 0.278 dB at 0 dB and 0.184 dB at 5 dB. SG-RAVEN similarly improved over RAVEN, suggesting that the heterogeneous graph front-end consistently provides structural benefits. Crucially, these gains became more pronounced under lower-SNR conditions: at 0 dB, SG-Mamba achieved an SI-SDR of 2.752 dB, outperforming RAVEN by 0.576 dB, whereas the margin narrowed at 5 dB (0.358 dB) and further at 10 dB, suggesting that explicit cross-modal structural priors are most effective when acoustic interference is most severe and visual anchoring is most critical. Moreover, SG-Mamba also attained the highest PESQ and ESTOI, indicating that the structural gains extended to perceptual quality and intelligibility. Such advantages are likely attributable to the graph's explicit cross-modal routing, which enables the audio stream to selectively attend to visual cues as a stable anchor when a single competing voice disrupts the acoustic scene.

2) Three Interfering Speakers: As shown in Table III, SG-Mamba maintained its advantage under the most challenging three-interferer scenario, achieving the best PESQ and ESTOI across all SNR levels and competitive SI-SDR throughout, demonstrating that the graph-guided fusion remains effective even under dense speech-to-speech mixtures. Compared to their non-graph counterparts, both SG-Mamba and SG-RAVEN yielded consistent SI-SDR improvements, suggesting that the heterogeneous graph front-end continues to provide structural benefits even as interferers increase. Notably, compared with the single-interferer condition, the absolute SI-SDR margins between SG-Mamba and the non-graph baselines narrowed across all SNR levels, which we attributed to the fixed graph topology: while content-adaptive attention weights remain dynamic, the pre-defined sparse local connectivity constrains the receptive field of cross-modal aggregation, making it less capable of suppressing interference patterns that span a broader or more irregular acoustic context under dense mixtures. Nevertheless, SG-Mamba exhibited no performance collapse across any condition, and its consistent lead in PESQ and ESTOI suggests that the visual anchor effectively preserved perceptual quality and intelligibility even when signal-level separation became more challenging.

C. Ablation Studies

1) Impact of Fusion Strategy: To evaluate our solution to the unstructured "Concatenation Trap," we compared multiple fusion mechanisms, as illustrated in Table IV. Despite carrying the highest parameter count (5.7 M), Bi-Mamba does not yield consistent gains under interferer conditions, possibly because the backward scan introduces conflicting future-

TABLE IV
PERFORMANCE COMPARISON UNDER DIFFERENT FUSION METHODS ON SI-SDR.

Method	Noise Only	1-Interferer			3-Interferer			Para. (↓)	MACs (↓)
	Mixed	0 dB	5 dB	10 dB	0 dB	5 dB	10 dB		
Bi-Mamba	12.939	2.607	5.517	7.652	2.210	5.306	7.769	5.7 M	3.27 G
Film-Mamba	12.578	2.641	6.047	8.543	2.305	5.548	8.105	**4.9 M**	**3.20 G**
Linear-Mamba	12.461	0.295	5.504	9.002	1.205	5.861	8.924	5.2 M	3.26 G
Cross-Mamba	12.583	0.198	5.466	9.023	1.055	5.868	9.035	5.2 M	3.27 G
SG-Mamba (Ours)	**13.091**	**2.752**	**6.229**	**9.096**	**2.372**	**6.011**	**9.190**	5.3 M	3.45 G

TABLE V
ABLATION ON GRAPH HYPERPARAMETERS (k, δ).

(K, δ)	Noise Only Mixed	1-Interferer Avg.	3-Interferer Avg.
$(1, +0)$	12.899	5.639	5.464
$(3, +0)$	**13.091**	**6.026**	**5.858**
$(5, +0)$	12.824	5.726	5.527
$(3, +2)$	12.984	5.910	5.643
$(3, -2)$	13.017	5.713	5.402

TABLE VI
CROSS-DATASET GENERALIZATION ON VOXCELEB2.

Method	Noise Only Mixed	Interferer Avg.	Overall Avg.
RAVEN (Base) [6]	7.712	2.061	4.887
Mamba	7.985	**2.940**	5.463
SG-RAVEN	7.945	2.514	5.230
SG-Mamba (Ours)	**8.222**	2.911	**5.567**

frame context that exacerbates cross-source contamination. While stable across all conditions, FiLM-Mamba's feature-wise modulation lacks the structural expressiveness needed for effective cross-modal integration. Attention-based methods achieved competitive SI-SDR in the noise-only and 10 dB conditions but degraded noticeably at 0 dB, accompanied by a PESQ degradation of approximately 0.2 across all interference conditions. We hypothesize that this degradation arises from negative transfer in unconstrained cross-modal fusion: without explicit structural boundaries, dense attention can erroneously align interfering acoustic features with the target's visual lip cues, causing the mask to suppress target speech. In contrast, SG-Mamba alleviates this by restricting cross-modal interactions to a sparse local window, bounding the propagation of such misalignment. Under a similar parameter budget, SG-Mamba achieved the best SI-SDR across all conditions while remaining stable under both interferer scenarios, suggesting a stronger robustness-fidelity balance than either feature modulation or dense attention under lightweight constraints.

2) Impact of Graph Hyperparameters: To examine sensitivity to graph hyperparameters, we vary the window size k and the cross-modal shift δ, with results shown in Table V. We first observe the impact of window size with the shift fixed at $\delta = 0$. Increasing k from 1 to 3 yields gains, as a narrow window lacks sufficient receptive field to capture cross-modal temporal variation. However, expanding to $k = 5$ degrades performance, as an oversized window introduces irrelevant distant frames and dilutes the graph's structural boundaries, risking negative transfer where acoustic interference erroneously aligns with distant visual cues. Regarding the cross-modal shift, both offsets ($\delta = \pm 2$) underperform the symmetry setting ($\delta = 0$). This aligns with the nature of speech production: since audio-visual asynchrony is dynamic and sentence-dependent, forcing a unidirectional offset misaligns phonemes that require an alternate context. The optimal performance of the symmetric, moderate window ($k = 3, \delta = 0$) suggests that our sparse graph topology functions as a localized tolerance buffer, encapsulating the dynamic range of asynchronies while bounding the propagation of cross-modal misalignment under adverse acoustic conditions.

D. Cross-Dataset Generalization

Table VI reports zero-shot transfer results on VoxCeleb2, where Interferer Avg. denotes the mean SI-SDR over 1- and 3-interferer conditions across 0, 5, and 10 dB SNRs, and Overall Avg. further includes the noise-only condition. SG-Mamba achieved the highest SI-SDR in the noise-only condition (8.222 dB) and the best Overall Avg. (5.567 dB). Under interferer conditions, Mamba achieved a marginally higher Interferer Avg. (2.940 dB vs. 2.911 dB), suggesting that the fixed graph topology becomes relatively less decisive under cross-dataset interference, consistent with the trend observed on LRS3. SG-RAVEN similarly improved over RAVEN across all conditions, suggesting that the structural benefits of the heterogeneous graph front-end transfer across datasets. These results demonstrate promising cross-dataset generalizability of the proposed framework without fine-tuning.

V. CONCLUSION

In this paper, we presented the SG-Mamba framework and showed that explicitly decoupling fine-grained cross-modal alignment from global temporal modeling helps address the "Concatenation Trap" in lightweight AVSE. Results on LRS3 and VoxCeleb2 show that explicit structural priors at the fusion stage can improve generalizability and robustness across diverse noise and multi-interferer conditions without substantially increasing computational cost. This graph-guided visual anchoring appears effective under severe acoustic interference. However, the current framework relies on a fixed graph topology and assumes uncorrupted visual inputs. Future work will investigate learnable sparse graph construction, robustness under audio-visual misalignment and visual degradations (e.g., occlusions), phase-aware reconstruction, and real-time streaming with end-to-end latency evaluation.

ACKNOWLEDGMENT

This work was supported in part by Realtek. The views expressed do not necessarily reflect those of the sponsor. We also used AI tools to assist with manuscript drafting and code development, with all outputs verified by the authors.

REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018.

[4] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[5] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[6] T. Ma, S. Yin, L.-C. Yang, and S. Zhang, "Real-time audio-visual speech enhancement using pre-trained visual representations," in *Proc. Interspeech*, 2025.

[7] B. Shi, W.-N. Hsu, and A. Mohamed, "Robust self-supervised audio-visual speech recognition," in *Proc. Interspeech*, 2022.

[8] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022.

[9] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection," in *Proc. ACM Multimedia*, 2021.

[10] X. Wang, F. Cheng, and G. Bertasius, "Loconet: Long-short context network for active speaker detection," in *Proc. CVPR*, 2024.

[11] X. Zhang, Q. Zhang, H. Liu, T. Xiao, X. Qian, B. Ahmed, E. Ambikairajah, H. Li, and J. Epps, "Mamba in speech: Towards an alternative to self-attention," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 1933–1948, 2025.

[12] X. Qian, J. Gao, Y. Zhang, Q. Zhang, H. Liu, L. P. Garcia, and H. Li, "Sav-se: Scene-aware audio-visual speech enhancement with selective state space model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 19, no. 4, pp. 623–634, 2025.

[13] R. Chao, W.-H. Cheng, M. La Quatra, S. M. Siniscalchi, C.-H. H. Yang, S.-W. Fu, and Y. Tsao, "An investigation of incorporating mamba for speech enhancement," in *Proc. SLT*, 2024.

[14] J. Wang, Z. Lin, T. Wang, M. Ge, L. Wang, and J. Dang, "Mamba-seunet: Mamba unet for monaural speech enhancement," in *Proc. ICASSP*, 2025.

[15] H. Xu, L. Wei, J. Zhang, J. Yang, Y. Wang, T. Gao, X. Fang, and L. Dai, "A multi-scale feature aggregation based lightweight network for audio-visual speech enhancement," in *Proc. ICASSP*, 2023.

[16] J. Richter, S. Frintrop, and T. Gerkmann, "Audio-visual speech enhancement with score-based generative models," in *Proc. ITG Conf. Speech Communication*, 2023.

[17] J.-E. Ayilo, M. Sadeghi, R. Serizel, and X. Alameda-Pineda, "Diffusion-based unsupervised audio-visual speech enhancement," in *Proc. ICASSP*, 2025.

[18] J.-C. Chou, C.-M. Chien, and K. Livescu, "Av2wav: Diffusion-based re-synthesis from continuous self-supervised features for audio-visual speech enhancement," in *Proc. ICASSP*, 2024.

[19] C. Jung, S. Lee, J.-H. Kim, and J. S. Chung, "Flowavse: Efficient audio-visual speech enhancement with conditional flow matching," in *Proc. Interspeech*, 2024.

[20] R. Mira, B. Xu, J. Donley, A. Kumar, S. Petridis, V. K. Ithapu, and M. Pantic, "La-voce: Low-snr audio-visual speech enhancement using neural vocoders," in *Proc. ICASSP*, 2023.

[21] F. E. Wahab, N. Saleem, A. Hussain, R. Ullah, and M. B. Hossen, "Multi-model dual-transformer network for audio-visual speech enhancement," in *Proc. AVSEC*, 2024.

[22] M. Sajid, D. Gupta, Y. Modi, S. Jain, H. J. S. Ganji, A. Rahaman, H. Choudhary, N. Saleem, A. Hussain, and M. Tanveer, "Aurexa-se: Audio-visual unified representation exchange architecture with cross-attention and squeezeformer for speech enhancement," in *Proc. AVSEC*, 2025.

[23] S. Ahmed, J.-C. Hou, and Y. Tsao, "Av-locofilm: Audio-visual speech enhancement using film-based fusion and hybrid local–global transformers," in *Proc. AVSEC*, 2025.

[24] N. Saleem, A. Hussain, K. Dashtipour, E. Sheikh, A. Sheikh, T. Arslan, and A. Hussain, "Viseme-gated multilayer cross-attentional feature fusion for cognitively-inspired multimodal speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 469–481, 2025.

[25] Z. Liu, X. Li, C. Chen, L. Guo, L. Li, and D. Wang, "Alignvsr: Audio-visual cross-modal alignment for visual speech recognition," in *Proc. ICCIP*, 2025.

[26] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, "The natural statistics of audiovisual speech," *PLoS computational biology*, vol. 5, no. 7, p. e1000436, 2009.

[27] J.-L. Schwartz and C. Savariaux, "No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag," *PLoS Computational Biology*, vol. 10, no. 7, p. e1003743, 2014.

[28] J. Chen and A. Zhang, "Hgmf: heterogeneous graph-based fusion for multimodal data with incompleteness," in *Proc. ACM SIGKDD*, 2020.

[29] H. N. Chau, T. D. Bui, H. B. Nguyen, T. T. H. Duong, and Q. C. Nguyen, "A novel approach to multi-channel speech enhancement based on graph neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1133–1144, 2024.

[30] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" in *Proc. ICLR*, 2022.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *ArXiv preprint arXiv:1809.00496*, 2018.

[33] C. K. A. Reddy *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, 2020.

[34] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018.

[35] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001.

[36] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[37] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *Proc. ICASSP*, 2019.

[38] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," in *Proc. ICML*, 2020.