CogCM: Cognition-Inspired Contextual Modeling for Audio-Visual Speech Enhancement

Supplementary Material

A. Display Page

We provide a website to showcase the actual enhancement performance of our CogCM and its comparison with other AVSE methods, available at: https://SoarCld.github.io/CogCM/.

B. Implementation Details

B.a. Data Preprocessing

In all experiments, audio samples are resampled to a sampling rate of 16 kHz, and the frame rate for videos is set to 25 fps. The length of speech segments is consistently set at 2 seconds. The STFT / iSTFT is performed using a Hamming window of 400 units in length and a hop size of 100 units. The frame rate of videos is set to 25fps. For facial inputs of SeCM_{V}, the dimensions are set to $H = W = 112$; for lip inputs of SeCM_{PV} and SeCM_{PAV}, $H = W = 88$. During training, random cropping and horizontal flipping are introduced. In testing, only center cropping is utilized.

B.b. Hyperparameter Settings

The initial learning rate for our model is established at 0.001, with a batch size of 48. Training is conducted throughout 20 epochs. The StepLR strategy is utilized with step size of 5 and a gamma value of 0.5. The AdamW optimizer is used for training, configured with a weight decay parameter of 0.01 and beta coefficients of (0.9, 0.999). The loss weights are empirically established as $\alpha = 0.9$, $\beta = 0.1$, and $\gamma = 0.05$. The number of Time-Frequency Blocks is set to $N = 4$.

B.c. Model Details

Signal Context Modeling Module (SiCM): As illustrated in Figure 1, the SiCM is composed of a bidirectional MAMBA module and a convolutional module. The forward and reversed copies of the input features are each processed by a separate MAMBA module. Then, the output from the reverse branch is reversed again and added to the forward output. After a convolution layer extracts local features, a residual connection from the input is added to produce the final output.

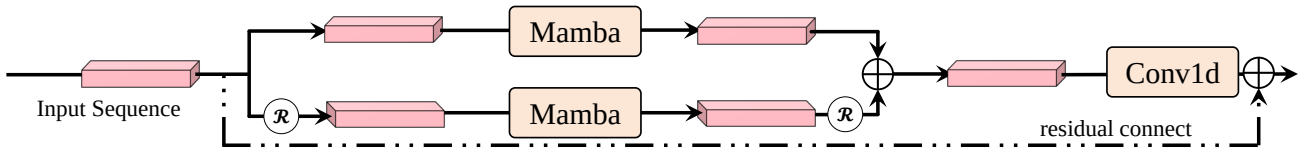

Figure 1. Structure of our SiCM. It consists of two mamba and a convolution layer. $\mathcal{R}$ denotes the reverse operation.

Table 1 shows the architectural details of our modules that are not described in detail in the main text.

C. Evaluation Metrics

We employ a comprehensive set of evaluation metrics to assess our system's performance. These metrics are grouped into Absolute Metrics and Improvement Metrics.

Absolute Metrics:

Signal-to-Distortion Ratio (SDR) [22]: Measures speech quality by comparing the signal power to the distortion between the enhanced output and the original clean speech.

Module	Configuration
Audio Encoder	• (0) ConvBlock Encoder – Conv2d(3, 64, k=(1,1), s=(1,1)) – InstanceNorm2d(64, eps=1e-05, momentum=0.1) – PReLU(64) • (1) BasicBlock(c,k,s,p) – Conv2d(64, 64, k=(3,3), s=(1,1), p=(1,1)) – BatchNorm2D(64) – PReLU(64) • (2) BasicBlock(c,k,s,p) – Conv2d(64, 64, k=(3,3), s=(1,1), p=(1,1)) – BatchNorm2D(64) – PReLU(64) • (3) Downsample ConvBlock – Conv2d(64, 64, k=(1,3), s=(1,2), p=(0,1)) – BatchNorm2D(64) – PReLU(64)
TF-Upsampler	• Block 0 – ConvTranspose2d(C_{in}, C_{mid}, k=(4,12), s=(2,10), p=(1,1) – BatchNorm2D(C_{mid}) – PReLU(C_{mid}) • Block 1 – ConvTranspose2d(C_{mid}, 64, k=(4,12), s=(2,10), p=(1,1)) – BatchNorm2D(64) – PReLU(64) • Block 2 – Linear Interpolate(1.6, along time dim)
Magnitude Decoder	• (0) BasicBlock(c=64,k=(3,3),s=(1,1),p=(1,1)) • (1) BasicBlock(c=64,k=(3,3),s=(1,1),p=(1,1)) • (2) Upsample ConvBlock – TransConv2d(64, 64, k=(1,3), s=(1,2), p=(0,1)) • (3) ConvBlock_Decoder(outchannel=1) – Conv2d(64, outchannel, k=(1,2), s=(1,1)) – InstanceNorm2d(outchannel) – PReLU(outchannel) – Conv2d(outchannel, outchannel, k=(1,1), s=(1,1)) – PReLU(outchannel)
Complex Decoder	• (0) BasicBlock(c=64,k=(3,3),s=(1,1),p=(1,1)) • (1) BasicBlock(c=64,k=(3,3),s=(1,1),p=(1,1)) • (2) Upsample ConvBlock • (3) ConvBlock_Decoder(outchannel=2)

Table 1. Simplified summary of the our modules.

Scale-invariant Signal-to-Noise Ratio (SI-SNR): Measures the quality of speech by comparing the signal power to the noise power.

Short-Time Objective Intelligibility (STOI) [29]: Quantifies speech intelligibility on a scale from 0 to 1, with higher values

indicating better intelligibility.

Extended Short-Time Objective Intelligibility (ESTOI): Provides an enhanced measure of speech intelligibility by capturing extended temporal dynamics.

Perceptual Evaluation of Speech Quality (PESQ) [25]: Assesses overall perceptual quality on a scale from 0.5 to 4.5, where higher scores denote superior quality.

DNSMOS [24]: A neural network–based metric that estimates the perceptual quality of speech.

Word Error Rate (WER): Evaluates speech recognition accuracy by comparing the recognized transcript to the reference, with lower percentages indicating better performance.

Mel Cepstral Distortion (MCD): Measures spectral distortion in the mel-cepstral domain between the enhanced and clean speech.

CSIG: Estimates the degree of signal distortion.

CBAK: Evaluates the quality of the background noise.

COVL: Assesses the overall quality of the enhanced speech.

Note that, except for MCD and WER, all other metrics are interpreted such that higher values indicate better performance.

Improvement Metrics:

Improvement is computed as the difference between the metric value for the enhanced speech and that for the noisy speech. This approach is adopted because some methods report these metric gains [19], while others are not open source—meaning that our constructed noisy test set may not perfectly match the noisy speech used by those methods. Therefore, for a fair comparison, we report relative improvements rather than absolute metric values:

PESQi: Improvement in PESQ.

MCDi: Improvement in MCD.

ESTOIi: Improvement in ESTOI.

STOIi: Improvement in STOI.

SI-SDRi: Improvement in Scale-Invariant SDR.

D. Dataset Details

We conducted our AVSE experiments on LRS3+DNS, GRID+CHiME3, TCD-TIMIT+NTCD-TIMIT, MEAD+DEMAND, AVSpeech+DNS, (GRID, TCD-TIMIT) + PNL Nonspeech, and AVSEC3 datasets. In each dataset group, the first dataset comprises clean audio-visual data, while the second dataset consists of noise samples collected from a variety of settings. For all visual inputs, SeCM_{V} processes a 112×112 grayscale image of the face following [32]. Meanwhile, SeCM_{PV} and SeCM_{PAV} receive an 88×88 grayscale image of the mouth Region of Interest (ROI), in accordance with the configurations of [18, 28, 36]. In the training phase, the 112×112 facial images and the 88x88 mouth ROI images are randomly cropped from larger 128×128 facial images and 96×96 mouth images, respectively. In the testing phase, a center cropping technique is employed.

LRS3 + DNS: LRS3 contains 438 hours of talking videos from TED and TEDX clips downloaded from YouTube. We evaluate our method on the pretrain subset which contains 407 hours of video. We partitioned this subset into training, validation, and testing sets with a ratio of 8:1:1. We follow Défossez et al. [4] to obtain the noise signal from the noise subset of the DNS dataset. The subset contains approximately 181 hours of noise audio collected from a wide variety of events. During training and evaluation, we utilized these samples as background noise to add noise to the clean speech and construct synthetic noisy audio inputs.

GRID + CHiME3: GRID consists of 33 speakers. For our experiments, we follow the general setting Balasubramanian et al. [1] to designate speakers s2 and s22 as the validation set, speakers s1 and s12 as the unseen unheard test set, and the remaining 29 speakers as the training set. We sample noise from CHiME to corrupt the clean speech. The noise in CHiME is categorized into 4 types: Cafe, Street, Bus, and Pedestrian. The CHiME dataset is divided into training, validation and testing sets with an 8:1:1 ratio.

TCD-TIMIT + NTCD-TIMIT: TCD-TIMIT consists of AV speech data from 56 English speakers with an Irish accent. Each utterance is approximately 5 seconds long and sampled at 16kHz. As recommended in Harte and Gillen [12], we split the dataset into training, validation, and testing sets, with 39 speakers for training, 8 for validation, and 9 for testing. The noisy speech input is derived from the NTCD-TIMIT dataset. This dataset is created by adding six different types of noise to the original speech data from the TCD-TIMIT corpus. The noise types include Living Room, White, Cafe, Car, Bable, and Street, and each noise type is associated with a specific SNR. Similar to the approach in Golmakani et al. [10], we selected 5 utterances per noise level and noise type for each test speaker to create a test set of 1350 utterances.

MEAD + DEMAND: The MEAD dataset consists of recordings from 46 participants, who uttered sentences expressing eight different emotions at three intensity levels under seven camera viewpoints. This dataset is extensively employed in research across various fields, including affective computing, human-computer interaction, and robust AVSE. Following Kang et al. [13], choose videos that captured frontal views and the highest level (level 3) of emotion intensity for experiment. For training, we utilized approximately 5 hours of videos from the MEAD dataset. Additionally, 0.7 hours were reserved for validation, and another 0.7 hours were allocated for testing purposes. The DEMAND dataset comprises noise recordings from multiple real-world environments and is extensively used in fields such as speech enhancement and speech recognition.

AVSpeech + DNS: AVSpeech is a large-scale audio-visual dataset designed for speech enhancement and related tasks. Each clip lasts between 3 and 10 seconds and features a single visible speaker whose voice is captured in the audio track. In total, the dataset contains roughly 4700 hours of video segments, sourced from about 290,000 YouTube videos, and includes nearly 150,000 unique speakers. This extensive diversity in speakers, languages, and facial poses makes AVSpeech an invaluable resource for advancing audio-visual processing research.

(GRID, TCD-TIMIT) + PNL Nonspeech: PNL Nonspeech is a collection containing 100 types of non-speech noise recordings. Following EANet [37], clean speeches from GRID and TCD-TIMIT are divided into 8:1:1 for training, validating, and testing. During training and validation, noises are randomly selected from the PNL Nonspeech dataset. While testing, noises from NoiseX-92 [31] are used to synthesize The noisy input.

AVSEC3: The 3rd Audio-Visual Speech Enhancement Challenge (AVSEC3) provide a benchmark for the evaluation of AVSE systems. The paired clean data are obtained from the LRS3 dataset. There are two types of interferers: 1) Speech: competing speakers are derived from the LRS3 dataset. 2) Noise: the noise dataset is built based on CEC1 [11], DEMAND, and DNS datasets.

E. More Comparison Results

E.a. Comparison Results on GRID + CHiME3

We compare the proposed CogCM with the SOTA AVSE approaches on GRID datasets. Following Wang et al. [33], we utilize the noises from the CHiME3 dataset to synthesize the noisy input audios and perform an evaluation with the test signal-to-noise ratio (SNR) levels of both -5dB and 0dB. As shown in Tabel 2, CogCM achieves the best performance in both the PESQ improvement (PESQi) and STOI improvement (STOIi) metrics with different test SNR levels.

Method	-5dB		0dB	
	STOIi(%)	PESQi	STOIi(%)	PESQi
L2L [5]	11.14	0.54	8.86	0.62
VSE [6]	-	0.45	-	0.60
OVA [33]	-	0.40	-	0.66
VSET [23]	-	0.50	-	0.75
MHCA-AVCRN [35]	13.51	0.76	11.25	0.88
M3Net [34]	13.42	0.75	11.31	0.89
DualAVSE [32]	15.79	0.76	13.56	0.92
CogCM (Ours)	**26.25**	**1.01**	**20.50**	**1.21**

Table 2. Comparison results on GRID + CHiME3 datasets. '-' denotes that the results are not reported in the original paper.

E.b. Comparison Results on TCD-TIMIT + NTCD-TIMIT

We further evaluate our CogCM model on the TCD-TIMIT dataset, comparing it with SOTA AVSE methods. The reporting metrics in Golmakani et al. [10] contains SI-SDR [14], PESQ, and STOI. We report the score improvement as a means of comparison. As illustrated in Table 3, the proposed CogCM achieves the best performance across all metrics at all SNR levels.

Method	SI-SDRi (dB)↑					PESQi↑					STOIi↑				
	-5dB	0dB	5dB	10dB	15dB	-5dB	0dB	5dB	10dB	15dB	-5dB	0dB	5dB	10dB	15dB
A-VAE [26]	4.34	5.12	5.93	6.07	5.76	0.16	0.19	0.20	0.21	0.05	0.02	0.02	0.04	0.04	0.04
AV-VAE [26]	6.15	6.86	7.38	7.22	6.52	0.24	0.27	0.29	0.28	0.08	0.02	0.03	0.04	0.05	0.04
A-DKF [10]	5.78	6.80	7.67	8.35	7.71	0.27	0.32	0.36	0.38	0.18	0.02	0.05	0.07	0.09	0.08
AV-DKF [10]	9.02	9.50	10.10	9.62	8.56	0.43	0.48	0.49	0.43	0.20	0.05	0.08	0.09	0.10	0.08
DualAVSE [32]	18.50	17.18	15.35	12.93	10.71	0.45	0.67	0.88	1.06	1.16	0.15	0.15	0.13	0.10	0.06
CogCM (Ours)	**20.89**	**19.95**	**18.39**	**16.27**	**13.86**	**1.21**	**1.50**	**1.77**	**1.97**	**2.02**	**0.27**	**0.26**	**0.22**	**0.16**	**0.11**

Table 3. Comparison results on TCD-TIMIT + NTCD-TIMIT datasets.

E.c. Comparison Results on MEAD + Demand

We conduct a comparison between CogCM and the AVSE methods on the MEAD dataset. The results presented in Table 4 demonstrate that our proposed CogCM model outperforms all other methods in terms of all evaluated metrics across various SNR conditions.

Method	SI-SDRi (dB)↑					PESQi↑					STOIi↑				
	-10dB	-5dB	0dB	5dB	10dB	-10dB	-5dB	0dB	5dB	10dB	-10dB	-5dB	0dB	5dB	10dB
A-VAE [15]	8.91	10.33	10.52	9.81	8.14	0.03	0.27	0.35	0.38	0.31	0.01	0.03	0.04	0.01	-0.01
AV-CVAE [27]	8.96	10.58	10.45	9.46	7.65	0.12	0.32	0.39	0.37	0.31	0.02	0.04	0.03	0.01	-0.02
AV-CVAE-WithHM [13]	8.08	10.02	10.12	9.21	7.70	0.12	0.29	0.32	0.30	0.28	0.01	0.02	0.01	-0.01	-0.03
AV-CVAE-RFF [13]	9.62	10.72	10.68	9.70	8.00	0.22	0.45	0.46	0.43	0.35	0.03	0.05	0.05	0.01	-0.01
DualAVSE [32]	16.06	15.21	14.09	12.98	11.27	0.35	0.54	0.74	0.92	1.01	0.10	0.10	0.08	0.05	**0.03**
CogCM (Ours)	**22.02**	**21.07**	**19.22**	**16.90**	**14.13**	**1.26**	**1.57**	**1.68**	**1.60**	**1.33**	**0.17**	**0.14**	**0.09**	**0.06**	**0.03**

Table 4. Comparison resutls on MEAD + DEMAND datasets.

E.d. Comparison Results on AVSpeech + DNS

Following LA-VocE [19], we evaluate our method on the AVSpeech + DNS dataset under various noise conditions and compare it with highly influential methods such as LA-VocE and RT-LA-VocE. Due to the inconsistent quality of AVSpeech data and to reduce training costs, we employed DNSMOS scoring to assess speech quality and selected a subset of high-quality (i.e., low-background-noise) samples for our training and validation sets. For testing, in line with LA-VocE, we randomly selected 1% (approximately 1500 clips) from the test set. It is noteworthy that our training set is smaller than that used by LA-VocE.

Following LA-VocE, we conducted comparative experiments under three noise conditions. In Condition 1, there is one background noise at 0 dB SNR and one interfering speaker at 0 dB SIR. Condition 2 comprises three background noises at -5 dB SNR and two interfering speakers at -5 dB SIR. Finally, Condition 3 includes five background noises at -10 dB SNR and three interfering speakers at -10 dB SIR.

Table 5 presents a detailed comparison of multiple methods using several evaluation metrics, including MCDi, PESQi, STOIi, and ESTOIi. Due to the failure to install ViSQOL, we did not report the ViSQOL results. The results demonstrate that our proposed CogCM consistently outperforms competing approaches under challenging acoustic scenarios, highlighting its robustness and effectiveness in enhancing speech quality.

Method	Input	MCDi↓	PESQi↑	STOIi↑	ESTOIi↑
Noise condition 1 (1 background noise at 0 dB SNR + 1 interfering speaker at 0 dB SIR)					
GCRN [30]	A	0.41	0.044	-0.052	-0.038
AV-GCRN [30]	AV	-1.193	0.394	0.220	0.235
AV-Demucs [4]	AV	-5.581	0.738	0.270	0.298
MuSE [20]	AV	-5.528	0.787	0.276	0.299
VisualVoice [7]	AV	-3.781	0.606	0.249	0.270
LA-VocE [19]	AV	-6.653	0.931	0.294	0.333
RT-LA-VocE [3]	AV	-6.157	0.653	0.255	0.282
CogCM (Ours)	AV	**-13.155**	**1.466**	**0.307**	**0.445**
Noise condition 2 (3 background noises at -5 dB SNR + 2 interfering speakers at -5 dB SIR)					
GCRN [30]	A	0.416	-0.010	-0.015	-0.015
AV-GCRN [30]	AV	-1.354	0.096	0.234	0.214
AV-Demucs [4]	AV	-5.548	0.274	0.308	0.300
MuSE [20]	AV	-5.314	0.297	0.308	0.289
VisualVoice [7]	AV	-3.388	0.164	0.253	0.237
LA-VocE [19]	AV	-6.863	0.511	0.379	0.397
RT-LA-VocE [3]	AV	-6.313	0.297	0.315	0.321
CogCM (Ours)	AV	**-15.391**	**1.056**	**0.343**	**0.446**
Noise condition 3 (5 background noises at -10 dB SNR + 3 interfering speakers at -10 dB SIR)					
GCRN [30]	A	-0.015	0.045	-0.020	-0.005
AV-GCRN [30]	AV	-1.263	0.043	0.171	0.139
AV-Demucs [4]	AV	-5.170	0.013	0.262	0.230
MuSE [20]	AV	-4.418	0.011	0.231	0.218
VisualVoice [7]	AV	-3.125	0.045	0.181	0.166
LA-VocE [19]	AV	-6.170	0.159	0.371	0.206
RT-LA-VocE [3]	AV	-5.671	0.039	0.287	0.272
CogCM (Ours)	AV	**-17.705**	**0.629**	**0.342**	**0.373**

Table 5. Comparison results on AVSpeech + DNS datasets under different noise conditions.

E.e. Comparison Results on (GRID, TCD-TIMIT) + PNL Nonspeech

We further evaluate our method on the (GRID, TCD-TIMIT) + PNL Nonspeech dataset, which combines clean audio-visual pairs with PNL Nonspeech noise. Table 6 presents performance metrics—namely PESQi and STOIi—across a wide range of SNR levels. Our proposed CogCM achieves the best average performance, demonstrating its effectiveness in leveraging visual and semantic cues to enhance speech quality in diverse noisy conditions.

Method	SNR								AVG
	-12dB	-9dB	-6dB	-3dB	0dB	6dB	9dB	12dB	
PESQi↑									
DCCRN+ [17]	0.43	0.42	0.39	0.33	0.22	0.44	0.33	0.46	0.38
PerceptNet+ [8]	0.46	0.47	0.34	0.61	0.49	0.29	0.24	0.35	0.41
DNSIP [16]	0.11	0.06	0.06	0.01	-0.03	0.12	0.54	0.78	0.22
CASE [9]	0.57	0.53	0.47	0.52	0.81	0.57	0.60	0.48	0.57
SE+AVVAD+SI [2]	0.75	0.72	0.84	0.87	0.83	0.66	0.64	0.59	0.74
AVMCNN [2]	0.86	0.87	0.87	0.86	0.84	0.75	0.76	0.70	0.81
EANet [37]	1.14	1.21	1.21	1.24	1.14	1.01	0.96	0.88	1.10
CogCM (Ours)	**1.15**	**1.23**	**1.34**	**1.38**	**1.42**	**1.43**	**1.40**	**1.31**	**1.33**
STOIi↑									
DCCRN+ [17]	0.11	0.10	0.07	0.05	0.09	0.04	0.07	0.07	0.08
PerceptNet+ [8]	0.15	0.15	0.14	0.13	0.14	0.09	0.08	0.07	0.12
DNSIP [16]	0.11	0.11	0.09	0.07	0.12	0.09	0.09	0.07	0.09
CASE [9]	0.12	0.12	0.11	0.08	0.10	0.05	0.06	0.04	0.09
SE+AVVAD+SI [2]	0.13	0.12	0.11	0.10	0.12	0.08	0.07	0.05	0.10
AVMCNN [2]	0.11	0.11	0.10	0.11	0.13	0.09	0.09	0.07	0.10
EANet [37]	0.20	0.18	0.16	0.14	0.14	0.10	0.09	0.08	0.14
CogCM (Ours)	**0.26**	**0.25**	**0.24**	**0.21**	**0.19**	**0.13**	**0.11**	**0.09**	**0.18**

Table 6. Comparison resutls on (GRID, TCD-TIMIT)+PNL Nonspeech datasets. STOIi and PESQi of various methods under different SNR conditions.

F. More Ablation Studies

F.a. Structure of SeCM

We further explore the impact of employing different pre-trained models as SeCM. Specifically, we substitute SeCM_{PAV} in CogCM with various pre-trained models known for their strong performance in AVSR tasks, including AVHuBERT, VATLm, and Auto-AVSR. Given that these models were pre-trained exclusively on clean audio-visual paired data, they have learned the shared information from audio-visual data. While the audio inputs of AVSE are interfered by noise, the shared information would be destroyed thus resulting in performance degradation. To ensure a fair comparison, we also investigate a modified variant of SeCM_{PAV}, which similarly uses only the visual modality.

The experimental results, as shown in Table 7, reveal that (1) all pre-trained models used as SeCM significantly improve enhancement performance, validating the importance of semantic contextual information in AVSE; (2) robust AVHuBERT (SeCM_{PAV}) shows marked superiority over other pre-trained models, irrespective of whether the input is visual-only or audio-visual. This advantage is attributed to the inclusion of noise during the training process, which enables Robust AVHuBERT to learn more robust audio-visual representations under noisy conditions.

F.b. Different-Layer Features of AVHuBERT

To conduct a more detailed assessment of how various degrees of semantic information affect the AVSE task, we utilize features from different encoder layers of robust AV-HuBERT as examples for evaluation. As indicated in Table 8, features from mid-to-high layers generally outperform those from lower layers. Typically, performance improves with higher-layer levels. The results supports again our motivation of introducing semantic context for AVSE because of the consensus that features extracted from higher layers are more closely associated with semantic-level cognition. Nevertheless, there are also exceptions; for example, under -5dB and 0dB SNR ratios, features from the 12th layer yield a higher SDR than those from the 24th layer. Given that different metrics assess different aspects of speech quality, these findings reveal that semantic contexts from different layers prioritize distinct aspects of speech. This suggests that integrating features across various layers could

Method	Input	-15dB			-10dB			-5dB			0dB		
		SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
SeCM_{V}													
VSR	V	5.814	1.988	0.782	8.986	2.321	0.847	11.768	2.673	0.892	14.354	3.031	0.925
SeCM_{PV}													
AVHuBERT	V	6.877	2.233	0.836	9.885	2.557	0.879	12.504	2.885	0.912	14.842	3.208	0.936
VATLM	V	6.878	2.245	0.836	10.074	2.543	0.879	12.725	2.858	0.911	15.030	3.179	0.935
VATLM	AV	6.305	2.084	0.799	9.799	2.466	0.868	12.561	2.831	0.909	15.008	3.170	0.935
Auto-AVSR	V	6.384	2.126	0.813	9.588	2.455	0.866	12.299	2.798	0.904	14.692	3.131	0.930
SeCM_{PAV}													
AVHuBERT	V	6.789	2.231	0.836	9.869	2.543	0.878	12.511	2.869	0.911	14.916	3.196	0.935
AVHuBERT	AV	**7.284**	**2.269**	**0.842**	**10.385**	**2.608**	**0.886**	**12.998**	**2.922**	**0.917**	**15.339**	**3.235**	**0.939**

Table 7. Comparisons of different pre-trained models for SeCM. In our evaluation, "V" indicates that SeCM receives visual inputs only, while "AV" indicates audio-visual inputs. Models under SeCM_{PAV} are pre-trained on paired video and noisy audio, whereas those under SeCM_{PV} are pre-trained on paired video and clean audio.

potentially enhance the overall performance for AVSE.

Layer	-15dB			-10dB			-5dB			0dB		
	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
24^{th} Layer	**7.284**	**2.269**	**0.842**	10.385	**2.608**	**0.886**	12.998	**2.922**	**0.917**	15.339	**3.235**	**0.939**
23^{rd} Layer	7.257	2.246	0.840	10.339	2.589	0.885	12.838	2.897	0.916	15.003	3.207	0.938
12^{th} Layer	7.140	2.164	0.826	**10.351**	2.511	0.878	**13.049**	2.858	0.913	**15.439**	3.189	0.937
1^{st} Layer	5.814	2.096	0.805	9.475	2.405	0.863	12.377	2.751	0.903	14.858	3.085	0.930
w/o SeC	5.171	1.904	0.752	8.509	2.241	0.831	11.393	2.616	0.884	14.005	2.992	0.921

Table 8. Evaluation of semantic contexts extracted from different encoder layers of robust AV-HuBERT. 'w/o SeC' denotes that no semantic contexts are utilized (AOSE Baseline).

F.c. Structure of SSGM

As illustrated in Figure 3 in the main text, the integrated output includes the residual connection for semantic context, denoted as E'_{se}. To further explore the role of semantic context in the fusion process, we developed a variant of SSGM that omits E'_{se}, simulating conditions devoid of semantic context. This variant differs from the AOSE Baseline presented in Table 8 in that it still employs the AVSE approach but excludes the direct influence of semantic context. Table 9 demonstrates that the residual connections consistently enhance performance. This indicates that semantic context plays a role not only during the initial stages of fusion at shallower network layers but also continues to guide performance improvements in deeper network layers where contextual information is fully integrated.

E'_{se}	-15dB			-10dB			-5dB			0dB		
	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
✓	**7.284**	**2.269**	**0.842**	**10.385**	**2.608**	**0.886**	**12.998**	**2.922**	**0.917**	**15.339**	**3.235**	**0.939**
✗	7.033	2.220	0.835	10.036	2.560	0.880	12.697	2.900	0.914	14.947	3.203	0.936

Table 9. Comparison results of different SSGM configurations, E'_{se} indicates whether to add semantic context residual connections

G. Complete Ablation Study Results of Main Text

Due to space limitations in the main text, we were unable to present the full range of SNR results for our ablation experiments. Here, we provide the complete ablation study results covering all SNR values (Table 10- 13). Detailed analysis of these results can be found in the main text.

SeC	-15dB			-10dB			-5dB			0dB		
	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
w/o SeC	5.171	1.904	0.752	8.509	2.241	0.831	11.393	2.616	0.884	14.005	2.992	0.921
SeCM$_V$	5.414	1.973	0.779	8.637	2.307	0.845	11.526	2.663	0.891	14.210	3.022	0.924
SeCM$_{PV}$	6.112	2.183	0.826	9.344	2.503	0.872	12.178	2.829	0.907	14.725	3.143	0.932
SeCM$_{PAV}$	**6.680**	**2.214**	**0.833**	**9.838**	**2.548**	**0.879**	**12.588**	**2.865**	**0.912**	**15.054**	**3.174**	**0.935**

Table 10. Evaluation of SeC. 'w/o SeC' denotes that no semantic contexts are utilized (AOSE Baseline).

SiC	-15dB			-10dB			-5dB			0dB		
	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
Conformer	4.311	1.780	0.730	7.797	2.138	0.817	10.742	2.496	0.875	13.467	2.875	0.914
SiCM (ours)	**5.171**	**1.904**	**0.752**	**8.509**	**2.241**	**0.831**	**11.393**	**2.616**	**0.884**	**14.005**	**2.992**	**0.921**

(a) Evaluation of SiCM and Conformer Modules in AOSE.

SiC	-15dB			-10dB			-5dB			0dB		
	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
Conformer	4.656	1.845	0.755	7.952	2.161	0.829	10.831	2.521	0.88	13.492	2.897	0.917
SiCM (ours)	**5.414**	**1.973**	**0.779**	**8.637**	**2.307**	**0.845**	**11.526**	**2.663**	**0.891**	**14.210**	**3.022**	**0.924**

(b) Evaluation of SiCM and Conformer Modules in AVSE.

Table 11. Evaluation of SiC. For the experiments in Table(b), the visual inputs are processed by SeCM$_V$. The context fusion strategy is a simple additional fusion.

Fusion Strategies	-15dB			-10dB			-5dB			0dB		
	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
Add (naive)	5.414	1.973	0.779	8.637	2.307	0.845	11.526	2.663	0.891	14.210	3.022	0.924
SSGM (ours)	**5.814**	**1.988**	**0.782**	**8.986**	**2.321**	**0.847**	**11.768**	**2.673**	**0.892**	**14.354**	**3.031**	**0.925**

(a) Evaluation of different contextual information fusion strategies, with SeCM$_V$ to obtain semantic context.

Fusion Strategies	-15dB			-10dB			-5dB			0dB		
	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
Add (naive)	6.680	2.214	0.833	9.838	2.548	0.879	12.588	2.865	0.912	15.012	3.189	0.937
SSGM (ours)	**7.284**	**2.269**	**0.842**	**10.385**	**2.608**	**0.886**	**12.998**	**2.922**	**0.917**	**15.339**	**3.235**	**0.939**

(b) Evaluation of different contextual information fusion strategies, with SeCM$_{PAV}$ to obtain semantic context.

Table 12. Evaluation of SSGM.

US	FM	-15dB			-10dB			-5dB			0dB		
		SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI	SDR	PESQ	STOI
✓	✓	**7.284**	**2.269**	**0.842**	**10.385**	**2.608**	**0.886**	**12.998**	**2.922**	**0.917**	**15.339**	**3.235**	**0.939**
✗	✓	6.992	2.252	0.839	10.098	2.593	0.884	12.737	2.918	0.915	15.064	3.229	0.937
✓	✗	7.025	2.214	0.835	10.075	2.560	0.882	12.681	2.876	0.914	15.012	3.189	0.937
✗	✗	6.803	2.199	0.831	9.866	2.535	0.879	12.559	2.869	0.913	14.950	3.190	0.936

Table 13. Ablation study of visual frequency for AVSE. US indicates frequency upsample operation; FM indicates frequency modeling module in SSGM.

H. Results on WER and DNSMOS

We further evaluated the performance of our enhanced speech using the pre-trained Whisper [21] model for speech recognition and DNSMOS as a subjective quality metric. As shown in Table 14, our proposed method significantly improves both word error rate (WER) and DNSMOS compared to the noisy input across various SNR conditions. Specifically, the enhanced speech achieves a lower WER and a higher DNSMOS score, demonstrating substantial improvements in both intelligibility and perceptual quality.

SNR	-15 dB		-10 dB		-5 dB		0 dB	
	WER(%)↓	DNSMOS↑	WER(%)↓	DNSMOS↑	WER(%)↓	DNSMOS↑	WER(%)↓	DNSMOS↑
Noisy	> 100	2.459	> 100	2.470	> 100	2.497	> 100	2.527
CogCM(Ours)	**43.2**	**2.828**	**40.9**	**2.863**	**35.3**	**2.877**	**57.7**	**2.871**

Table 14. Comparison of WER and DNSMOS on the LRS3 + DNS dataset. Lower WER and higher DNSMOS values indicate better performance.

I. Visualization of spectrograms

Figure 2 displays the spectrograms of three samples, arranged from top to bottom as follows: noisy speech, clean speech, DualAVSE enhancement results, and our CogCM enhancement results. The figure clearly illustrates that our CogCM enhancement results yield a spectrum with clearer and richer details. Particularly in extreme noise conditions, as seen in sample 3, the spectrogram enhanced by DualAVSE appears very blurry, while the CogCM enhancement results are clearer.

J. Notation

In this section, we summarize the key symbols and notations used throughout this paper. Table 15 provides an overview of each symbol, its corresponding dimensions, and a brief description of its role in our model.

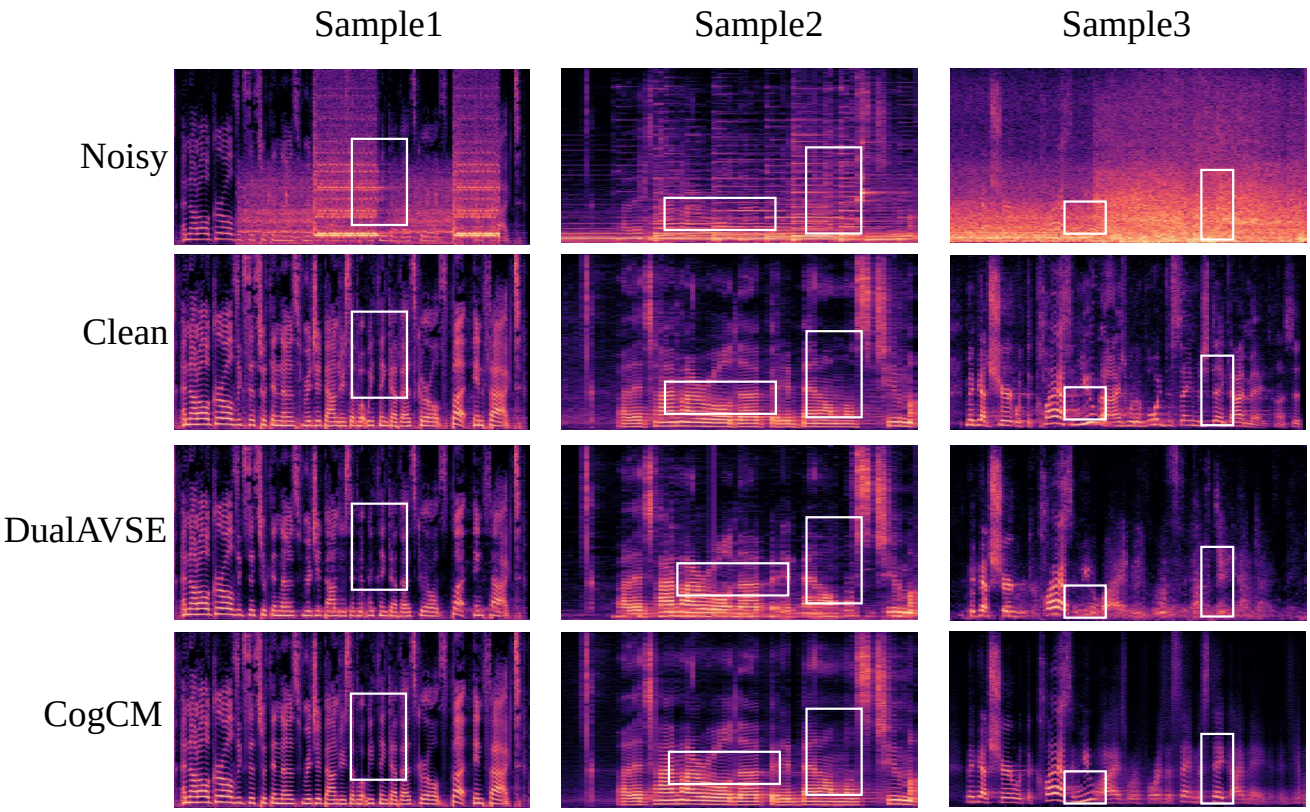

Figure 2. Visualization of the spectrum for samples.

Symbol	Dimension	Description
x	$\mathbb{R}^{T_a}$	Noisy speech waveform of length T_a
s	$\mathbb{R}^{T_a}$	Clean (target) speech waveform
V	$\mathbb{R}^{H\times W\times T_v}$	Video frame sequence with height H, width W, and T_v frames
X	$\mathbb{R}^{2\times T_x\times F_x}$	Complex spectrogram of the noisy speech (e.g., real+imaginary)
X'	$\mathbb{R}^{3\times T_x\times F_x}$	Input spectral features, consisting of magnitude, real, and imaginary parts
S	$\mathbb{R}^{2\times T_x\times F_x}$	Clean speech spectrogram (real+imaginary)
X_m, X_p, X_r, X_i	$\mathbb{R}^{1\times T_x\times F_x}$	Magnitude, phase, real, and imaginary components of the noisy spectrogram
S_m, S_r, S_i	$\mathbb{R}^{1\times T_x\times F_x}$	Magnitude, real, and imaginary components of the clean spectrogram
E_a	$\mathbb{R}^{C\times T_x\times F'_x}$	High-dimensional audio embeddings from the Audio Encoder
E_{se}	$\mathbb{R}^{C_{se}\times T_{se}}$	Semantic context embeddings extracted by SeCM
E'_{se}	$\mathbb{R}^{C\times T_x\times F'_x}$	Semantic context embeddings after TF-Upsampler
H_1, H_2	$\mathbb{R}^{C\times T_x}/\mathbb{R}^{C\times F'_x}$	Input features of CSBlock
H'_1, H'_2	$\mathbb{R}^{C\times T_x}/\mathbb{R}^{C\times F'_x}$	Output features of CSBlock
H''_1, H''_2	$\mathbb{R}^{C\times T_x}/\mathbb{R}^{C\times F'_x}$	Signal context features from H'_1, H'_2
W	$\mathbb{R}^{C\times T_x}/\mathbb{R}^{C\times F'_x}$	Weights vector in CGBlock
$\hat{S}_m$	$\mathbb{R}^{1\times T_x\times F_x}$	Predicted magnitude spectrogram from the Magnitude Decoder
$\hat{S}'_r, \hat{S}'_i$	$\mathbb{R}^{1\times T_x\times F_x}$	Predicted real and imaginary components from the Complex Decoder
$\hat{S}_r, \hat{S}_i$	$\mathbb{R}^{1\times T_x\times F_x}$	Predicted real and imaginary components of the enhanced spectrogram
$\hat{s}$	$\mathbb{R}^{T_a}$	Final enhanced speech waveform after ISTFT

Table 15. Summary of Notations

References

[1] S Balasubramanian, R Rajavel, and Asutosh Kar. Estimation of ideal binary mask for audio-visual monaural speech enhancement. *Circuits, Systems, and Signal Processing*, pages 1–25, 2023. 3

[2] S Balasubramanian, R Rajavel, and Asuthos Kar. Ideal ratio mask estimation based on cochleagram for audio-visual monaural speech enhancement. *Applied Acoustics*, 211:109524, 2023. 7

[3] Honglie Chen, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Rt-la-voce: Real-time low-snr audio-visual speech enhancement. In *Interspeech 2024*, pages 2215–2219, 2024. 6

[4] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. In *INTERSPEECH*, pages 3291–3295. ISCA, 2020. 3, 6

[5] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018. 4

[6] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. *Conference of the International Speech Communication Association*, 2018. 4

[7] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE, 2021. 6

[8] Xiaofeng Ge, Jiangyu Han, Yanhua Long, and Haixin Guan. Percepnet+: A phase and SNR aware percepnet for real-time speech enhancement. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 916–920. ISCA, 2022. 7

[9] Mandar Gogate, Kia Dashtipour, Ahsan Adeel, and Amir Hussain. Cochleanet: A robust language-independent audio-visual model for real-time speech enhancement. *Information Fusion*, 63:273–285, 2020. 7

[10] Ali Golmakani, Mostafa Sadeghi, and Romain Serizel. Audio-visual speech enhancement with a deep kalman filter generative model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3, 4, 5

[11] SN Graetzer, Jon Barker, Trevor J Cox, Michael Akeroyd, John F Culling, Graham Naylor, Eszter Porter, R Viveros Munoz, et al. Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing. In *INTERSPEECH*, pages 686–690. International Speech Communication Association (ISCA), 2021. 4

[12] Naomi Harte and Eoin Gillen. TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech. *IEEE Transactions on Multimedia*, 17 (5):603–615, 2015. 3

[13] Zhiqi Kang, Mostafa Sadeghi, Radu Horaud, Xavier Alameda-Pineda, Jacob Donley, and Anurag Kumar. The impact of removing head movements on audio-visual speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7302–7306. IEEE, 2022. 4, 5

[14] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019. 4

[15] Simon Leglaive, Laurent Girin, and Radu Horaud. A variance modeling framework based on variational autoencoders for speech enhancement. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2018. 5

[16] Nan Li, Longbiao Wang, Qiquan Zhang, and Jianwu Dang. Dual-stream noise and speech information perception based speech enhancement. *Expert Systems with Applications*, 261:125432, 2025. 7

[17] Shubo Lv, Yanxin Hu, Shimin Zhang, and Lei Xie. DCCRN+: channel-wise subband DCCRN with SNR estimation for speech enhancement. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 2816–2820. ISCA, 2021. 7

[18] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-avsr: Audio-visual speech recognition with automatic labels. In *ICASSP*, pages 1–5. IEEE, 2023. 3

[19] Rodrigo Mira, Buye Xu, Jacob Donley, Anurag Kumar, Stavros Petridis, Vamsi Krishna Ithapu, and Maja Pantic. La-voce: Low-snr audio-visual speech enhancement using neural vocoders. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3, 5, 6

[20] Zexu Pan, Ruijie Tao, Chenglin Xu, and Haizhou Li. Muse: Multi-modal target speaker extraction with visual cues. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6678–6682. IEEE, 2021. 6

[21] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 28492–28518. PMLR, 2023. 10

[22] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. Mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, pages 367–372, 2014. 1

[23] Karthik Ramesh, Chao Xing, Wupeng Wang, Dong Wang, and Xiao Chen. Vset: A multimodal transformer for visual speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6658–6662. IEEE, 2021. 4

[24] Chandan K.A. Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan. INTERSPEECH 2021 Deep Noise Suppression Challenge. In *Interspeech 2021*, pages 2796–2800. ISCA, 2021. 3

[25] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, pages 749–752. IEEE, 2001. 3

[26] Mostafa Sadeghi and Xavier Alameda-Pineda. Mixture of inference networks for vae-based audio-visual speech enhancement. *IEEE Transactions on Signal Processing*, 69:1899–1909, 2021. 5

[27] Mostafa Sadeghi, Simon Leglaive, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Audio-visual speech enhancement using conditional variational auto-encoders. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 28:1788–1800, 2020. 5

[28] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *ICLR*. OpenReview.net, 2022. 3

[29] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011. 2

[30] Ke Tan and DeLiang Wang. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:380–390, 2019. 6

[31] Andrew Varga and Herman J. M. Steeneken. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.*, 12(3):247–251, 1993. 4

[32] Fexiang Wang, Shuang Yang, Shiguang Shan, and Xilin Chen. Dual attention for audio-visual speech enhancement with facial cues. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. 3, 4, 5

[33] Wupeng Wang, Chao Xing, Dong Wang, Xiao Chen, and Fengyu Sun. A robust audio-visual speech enhancement model. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7529–7533. IEEE, 2020. 4

[34] Haitao Xu, Liangfa Wei, Jie Zhang, Jianming Yang, Yannan Wang, Tian Gao, Xin Fang, and Lirong Dai. A multi-scale feature aggregation based lightweight network for audio-visual speech enhancement. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 4

[35] Xinmeng Xu, Yang Wang, Jie Jia, Binbin Chen, and Dejun Li. Improving Visual Speech Enhancement Network by Learning Audio-visual Affinity with Multi-head Attention. In *Proc. Interspeech 2022*, pages 971–975, 2022. 4

[36] Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Li-Rong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Trans. Multim.*, 26: 1055–1064, 2024. 3

[37] Zhehui Zhu, Lijun Zhang, Kaikun Pei, and Siqi Chen. Endpoint-aware audio-visual speech enhancement utilizing dynamic weight modulation based on snr estimation. *Neural Networks*, 185:107152, 2025. 4, 7