FROM OBSERVATION-DRIVEN TO CAUSE-GOVERNED: FREQUENCY-COUPLED STATE SPACE RECOVERY FOR AUDIO-VISUAL SPEECH ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-visual speech enhancement (AVSE) has achieved substantial progress by exploiting visual articulatory information to improve speech recovery under acoustic degradation. However, two limitations persist in many existing AVSE methods. Recovery is primarily observation-driven, with vision serving mainly as an auxiliary feature or condition; frequency information is typically modeled within a unified recovery process. These limitations make reliable speech recovery more difficult, particularly at low SNRs. We therefore formulate AVSE as a cause-governed, frequency-coupled state-recovery process. Specifically, a visual articulatory prior not directly corrupted by acoustic noise governs state evolution along an explicit cause $\rightarrow$ content $\rightarrow$ detail dependency. Accordingly, we propose the Frequency-aware State Space Cause-governed Modeling (FSSCM) framework, comprising three key components. First, Visual Spectral Prior Learning (VSPL) models temporal articulatory dynamics from continuous lip movements to construct a temporally structured visual articulatory prior that governs state transitions. Second, Adaptive Progressive Frequency Mamba (APFM) realizes frequency-coupled recovery by using the visual prior to modulate low-frequency Mamba state transitions for content recovery, after which the evolving low-frequency state and the same prior jointly modulate high-frequency state transitions for detail reconstruction. Finally, Prior-anchored Frequency Consistency Learning (PFCL) is designed to enforce cross-frequency consistency under the same visual articulatory prior. Experiments on five AVSE benchmarks demonstrate the superiority of FSSCM over state-of-the-art methods, with its SDR advantage over competing methods increasing as SNR decreases.

1 INTRODUCTION

Speech enhancement (SE) aims to recover clean and intelligible speech from noise-corrupted acoustic observations, with applications in hearing aids, mobile communication, and automatic speech recognition Loizou (2013). Audio-only speech enhancement (AOSE) has progressed from convolutional and recurrent architectures Défossez et al. (2020); Hu et al. (2020) to Transformer-based models Lu et al. (2023). More recently, Mamba-based state-space models have been explored for speech enhancement Chao et al. (2024); Li et al. (2025). Despite these advances, AOSE remains challenging under severe acoustic degradation, particularly at low SNRs and under non-stationary noise Zhang et al. (2026). To address this challenge, audio-visual speech enhancement (AVSE) incorporates visual articulatory information to complement degraded acoustic observations.

Existing AVSE methods realize this complementary role Michelsanti et al. (2021) through several representative strategies, including direct feature fusion, attention-based fusion, and visual conditioning. Direct feature fusion combines visual and acoustic representations to provide complementary information Tiwari et al. (2024); Chao et al. (2025). Attention-based fusion selectively integrates cross-modal representations, enabling adaptive interaction between visual and acoustic information Wang et al. (2023); Saleem et al. (2026a). Visual conditioning uses visual information as a condition or higher-level context for speech reconstruction Ahmed et al. (2025); Wang et al. (2025). Compared with AOSE, these strategies broaden the information available for speech reconstruction beyond the acoustic modality alone. Beyond visual integration, recent studies have

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

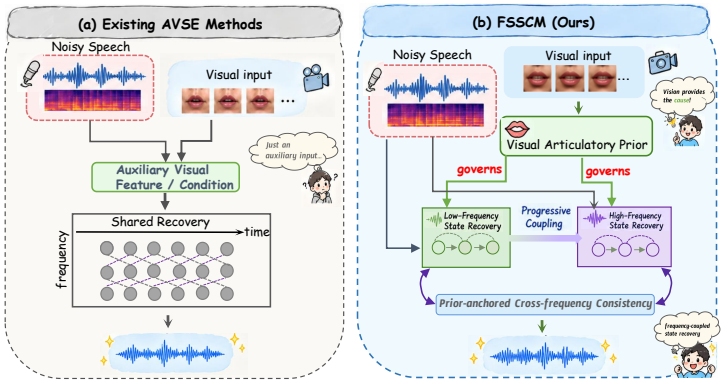


Figure 1: Comparison between existing AVSE methods and our FSSCM.

explored finer-grained spectro-temporal and frequency-aware modeling to better characterize acoustic structure during recovery Xu et al. (2023); Saleem et al. (2026b). Overall, these developments have expanded visual integration strategies and refined acoustic modeling in AVSE.

However, two limitations remain in many existing AVSE methods. First, visual information is typically incorporated at the representation level through feature fusion, modulation, or conditioning. Although effective, these strategies exploit visual information through its interaction with acoustic representations, which may constrain how fully vision contributes to speech recovery. This limitation can become more pronounced as acoustic degradation increases. Second, although recent AVSE methods employ finer-grained spectro-temporal and frequency-aware modeling, different frequency components are still often handled within a unified recovery process, without explicitly accounting for heterogeneity in their recovery characteristics. Such treatment may limit effective recovery across the spectrum.

These limitations lead us to revisit AVSE from a speech-production perspective, guided by two closely related observations. First, continuous lip movements provide an observable trace of the coordinated articulatory process that physically generates the acoustic speech signal Masapollo et al. (2025). Because visual articulatory information captures the dynamics of this upstream production process without being directly corrupted by acoustic noise, it provides a suitable basis for constructing a visual articulatory prior that governs recovery-state evolution. Second, speech information is distributed unevenly across the frequency spectrum: low-frequency components generally provide relatively robust cues to speech content Fogerty & Chen (2016), whereas higher-frequency components contribute finer-grained details and are more vulnerable to acoustic degradation de Jong & Tashiro (2025). Their shared articulatory origin and differing robustness motivate a differentiated yet coupled recovery strategy, in which relatively reliable low-frequency content provides a basis for subsequent high-frequency detail reconstruction. Taken together, these observations motivate using an articulatory prior to organize frequency-dependent speech recovery. This perspective further suggests that such a recovery paradigm becomes increasingly beneficial as acoustic observations become less reliable. Accordingly, we propose the Frequency-aware State Space Cause-governed Modeling (FSSCM) framework. Fig. 1 illustrates the differences between FSSCM and existing AVSE methods.

Our contributions are three-fold:

- We formulate AVSE as a *cause-governed, frequency-coupled state-recovery process*, in which a visual articulatory prior governs state evolution while acoustic observations provide the evidence for speech reconstruction. This formulation establishes an explicit *cause* $\rightarrow$ *content* $\rightarrow$ *detail* recovery principle.

- To instantiate this formulation, we develop FSSCM with three key components. VSPL derives a temporally structured visual articulatory prior from continuous lip dynamics; APFM uses this prior to govern frequency-specific Mamba state transitions and progressively couples low-frequency content recovery with high-frequency detail reconstruction; PFCL anchors cross-frequency consistency to the same visual articulatory prior.

- Extensive experiments across five AVSE benchmarks, together with comprehensive ablation studies, validate the effectiveness of FSSCM and the contribution of its key designs across diverse acoustic conditions, particularly under severe acoustic degradation.

2 Related Work

2.1 Audio-Only Speech Enhancement

Early AOSE methods mainly rely on convolutional and recurrent architectures. DEMUCS Défossez et al. (2020) performs waveform-domain enhancement with an encoder–decoder architecture, while FullSubNet Hao et al. (2021) exploits full- and sub-band information for speech reconstruction. Subsequent methods such as CMGAN Cao et al. (2022) and MP-SENet Lu et al. (2023) further strengthen spectro-temporal modeling through attention- and Transformer-based architectures. More recently, Mamba-based state-space models have been introduced into speech enhancement. SEMamba Chao et al. (2024) explores selective state-space modeling for long-range acoustic dependencies, while CSMamba Li et al. (2025) further incorporates cross- and sub-band processing for frequency-aware acoustic modeling. These studies demonstrate the potential of state-space models for efficient and fine-grained acoustic modeling. Nevertheless, these advances remain within the audio-only setting, where clean speech must be inferred from degraded acoustic observations alone.

2.2 Audio-Visual Speech Enhancement

Existing AVSE studies span direct feature fusion, attention-based fusion, and visual conditioning. A-V DEMUCS Tiwari et al. (2024) extends the audio-only DEMUCS architecture by incorporating visual features into waveform-based enhancement, while AVSEMamba Chao et al. (2025) extends Mamba-based speech enhancement to the audio-visual setting using full-face visual cues. Attention-based methods introduce more adaptive cross-modal interaction. DualAVSE Wang et al. (2023) employs cooperative attention to dynamically integrate facial and acoustic information, VG-MCA-AVSENet Saleem et al. (2026a) introduces viseme-gated cross-attention for adaptive audio-visual interaction, while FA-Mamba-AVSE Saleem et al. (2026b) further integrates frequency-aware state-space modeling with cross-attentive gated fusion. Visual conditioning methods instead use visual information as a condition or higher-level context for speech reconstruction. AV-LocoFiLM Ahmed et al. (2025) conditions acoustic representations on visual lip movements through feature-wise modulation, while AV-DKF Golmakani et al. (2023) incorporates visual information into sequential generative modeling. More recently, CogCM Wang et al. (2025) uses multimodal semantic context to guide fine-grained temporal-frequency signal modeling. Across these strategies, acoustic recovery has also incorporated increasingly fine-grained spectro-temporal and frequency-aware modeling.

Although AVSS addresses a distinct task, representative methods such as VisualVoice Gao & Grauman (2021), CTCNet Li et al. (2024), RTFS-Net Pegg et al. (2024), RAVSS Pan et al. (2024), and Dolphin Li et al. (2026) also exploit visual information for target-speech recovery and provide relevant references for AVSE. Overall, existing AVSE methods leave room to better exploit visual articulatory information in recovery dynamics and to explicitly account for heterogeneous recovery characteristics across frequency components.

3 Method

3.1 Overall Framework

The overall architecture of FSSCM is illustrated in Fig. 2. Given a noisy waveform y, STFT produces the magnitude and phase spectra, which are concatenated and encoded by a dense encoder E into a latent acoustic representation $Z \in \mathbb{R}^{d\times T\times \hat{F}}$, where $\hat{F} = F/2$. In parallel, Visual Spectral Prior Learning (VSPL, Sec. 3.2) encodes the lip video $V \in \mathbb{R}^{T_v\times H\times W}$ into a visual articulatory prior $C_v \in \mathbb{R}^{T\times d}$ and a gating signal $G_v \in \mathbb{R}^{T\times d}$, where C_v governs state-transition modulation and G_v provides complementary input-amplitude modulation. Adaptive Progressive Frequency Mamba (APFM, Sec. 3.3) routes Z into low- and high-frequency branches and performs progressive coupled recovery, producing L_{out} and H_{out}. Prior-anchored Frequency Consistency Learning (PFCL, Sec. 3.4) coordinates the recovered representations under the same visual articulatory prior, yielding Z_{fused}. Magnitude and phase decoders then reconstruct the enhanced spectrum, followed by

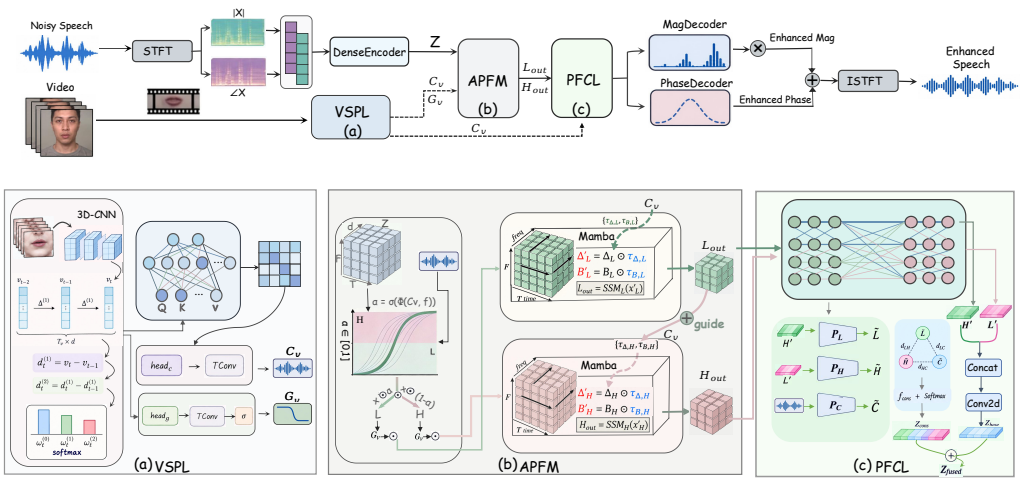


Figure 2: Overall architecture of FSSCM. VSPL derives a visual articulatory prior not directly corrupted by acoustic noise and shares it throughout recovery; APFM realizes visually governed progressive low-to-high frequency state evolution; and PFCL anchors cross-frequency consistency to the same prior.

iSTFT to obtain the enhanced waveform. Overall, FSSCM realizes the *cause* → *content* → *detail* recovery dependency through visual-governed state transition and progressive low-to-high frequency coupling.

3.2 Visual Spectral Prior Learning (VSPL)

VSPL constructs a temporally structured visual articulatory prior from continuous lip movements for subsequent cause-governed recovery. A single visual state mainly describes the current articulatory configuration, whereas speech production involves continuous articulatory evolution over time. We therefore explicitly model both the visual state and its temporal variations before constructing the visual prior. Given a lip video $V \in \mathbb{R}^{T_v \times H \times W}$, a 3D-CNN first extracts spatiotemporal visual features:

$$v = f_{\text{3dcnn}}(V) \in \mathbb{R}^{T_v \times d}. \tag{1}$$

Articulatory Dynamics Modeling. To explicitly characterize the temporal evolution of articulatory states, we compute first- and second-order temporal variations:

$$d_t^{(1)} = v_t - v_{t-1}, \qquad d_t^{(2)} = d_t^{(1)} - d_{t-1}^{(1)}, \tag{2}$$

where the boundary terms are initialized to zero. Here, $d_t^{(1)}$ captures the instantaneous change of the visual state, while $d_t^{(2)}$ further characterizes the variation of this motion tendency. Together with the original visual state, they describe both the current articulatory configuration and its local temporal evolution.

Since the relative contributions of these components may vary over time, VSPL adaptively integrates them as

$$\boldsymbol{\omega}_t = \text{softmax}\Big(f_\omega([v_t; d_t^{(1)}; d_t^{(2)}])\Big), \qquad u_t = \sum_{i=0}^{2} \omega_t^{(i)} W_i d_t^{(i)}, \tag{3}$$

where $d_t^{(0)} \equiv v_t$. The learned weights $\boldsymbol{\omega}_t$ dynamically balance the contributions of the visual state, first-order temporal variation, and second-order temporal variation at each time step, yielding a dynamics-aware visual sequence $U = [u_1, \ldots, u_{T_v}]$.

Role-Decoupled Prior Construction. While the above temporal variations characterize local articulatory changes, speech articulation also exhibits dependencies over longer temporal contexts. We therefore further refine the dynamics-aware sequence using temporal self-attention:

$$\widetilde{U} = U + \text{MHA}(UW_Q, UW_K, UW_V). \tag{4}$$

Based on these representations, VSPL derives two visual signals with different roles:

$$C_v = \mathrm{TConv}\Big(\mathrm{head}_c(\widetilde{U})\Big), \qquad G_v = \sigma(\mathrm{TConv}(\mathrm{head}_g(v))), \tag{5}$$

where the temporal convolutions align the visual sequences with the acoustic time resolution T.

The resulting $C_v \in \mathbb{R}^{T\times d}$ denotes the dynamics-enriched visual articulatory prior, while $G_v \in (0,1)^{T\times d}$ serves as a complementary visual gate. The temporally resolved C_v is subsequently passed to APFM for state-transition modulation, while G_v provides auxiliary amplitude modulation.

3.3 Adaptive Progressive Frequency Mamba (APFM)

APFM instantiates the proposed recovery hierarchy with two coupled frequency-specific state-space branches, where the low-frequency branch first establishes a content-oriented recovery state that subsequently guides high-frequency detail reconstruction together with the visual articulatory prior.

Dynamic Frequency Routing. To obtain an adaptive low-to-high frequency organization, the visual articulatory prior C_v is summarized into a global context $c_v^g = \mathrm{Pool}_t(C_v) \in \mathbb{R}^d$. APFM employs K learnable monotonic spectral bases,

$$\phi_k(f_j) = \sigma(s_k(c_k - f_j)), \qquad k = 1, \dots, K, \tag{6}$$

where $f_j \in [0,1]$ denotes the normalized frequency coordinate, $c_k \in (0,1)$, and $s_k > 0$. The visual context adaptively predicts their mixture weights:

$$\pi = \mathrm{softmax}_K(\mathrm{proj}(c_v^g)), \qquad \alpha_j = \sum_{k=1}^{K} \pi_k \phi_k(f_j). \tag{7}$$

Since each basis decreases monotonically with frequency, α forms a low-frequency-biased mask, while $1-\alpha$ provides its complementary high-frequency-biased mask. The routed representations are therefore obtained as

$$L = Z \odot \alpha, \qquad H = Z \odot (1-\alpha). \tag{8}$$

The visual gate G_v further modulates their amplitudes as $L \leftarrow L \odot G_v$ and $H \leftarrow H \odot G_v$.

Cause-Governed State Transition. For each branch $b \in \{L, H\}$, the acoustic representation determines the base input-dependent Mamba parameters $\Delta_{b,t}$ and $B_{b,t}$. Given a governing signal q_t^b, APFM modulates these parameters as

$$[\tau_{\Delta,b,t}; \tau_{B,b,t}] = 2\sigma\big(f_{\mathrm{trans}}^b(q_t^b)\big), \qquad \Delta'_{b,t} = \Delta_{b,t} \odot \tau_{\Delta,b,t}, \quad B'_{b,t} = B_{b,t} \odot \tau_{B,b,t}. \tag{9}$$

The modulation factors are bounded in $(0,2)$ and centered at 1, yielding multiplicative adaptation of the acoustic-dependent dynamics. The corresponding recovery state evolves as

$$h_{b,t} = \exp(\Delta'_{b,t} \odot A_b) \odot h_{b,t-1} + \Delta'_{b,t} \odot B'_{b,t} \odot x'_{b,t}, \tag{10}$$

where, for each routed branch $X_b \in \{L, H\}$, a branch-level summary is obtained by global time-frequency pooling:

$$r_b = \mathrm{Pool}_{T,F}(X_b), \qquad g_b = \sigma\big(W_g^b r_b + b_g^b\big). \tag{11}$$

The branch input is then modulated as $x'_{b,t} = x_{b,t} \odot g_b$, with g_b shared across temporal and frequency positions. The state matrix A_b remains unchanged; $\Delta'_{b,t}$ and $B'_{b,t}$ control the hidden recovery-state evolution, while G_v and g_b provide visual amplitude modulation and branch input regulation, respectively.

$L \to H$ Progressive Frequency Coupling. For low-frequency content recovery, the temporally resolved visual articulatory prior directly provides the governing signal:

$$q_t^L = C_{v,t}, \qquad L_{\mathrm{out}} = \mathrm{SSM}_L\left(L; \{\Delta'_{L,t}, B'_{L,t}\}_{t=1}^T, g_L\right). \tag{12}$$

The evolving low-frequency recovery is then compressed only along the frequency dimension while retaining its temporal structure:

$$L_{\mathrm{guide},t} = \mathrm{MLP}\left(\mathrm{Pool}_F(L_{\mathrm{out},t})\right). \tag{13}$$

For high-frequency detail reconstruction, the same visual articulatory prior and the temporally aligned low-frequency guidance jointly form the governing signal:

$$q_t^H = C_{v,t} + L_{\mathrm{guide},t}. \tag{14}$$

The resulting q_t^H generates the high-frequency transition modulation according to Eq. 9. The low-frequency guidance is also aggregated into the high-frequency branch summary for input regulation. The high-frequency branch is then computed as

$$H_{\mathrm{out}} = \mathrm{SSM}_H\left(H; \{\Delta'_{H,t}, B'_{H,t}\}_{t=1}^{T}, g_H\right). \tag{15}$$

Thus, q_t^L is driven by the visual articulatory prior, whereas q_t^H additionally incorporates the evolving low-frequency recovery, realizing the *cause* $\rightarrow$ *content* $\rightarrow$ *detail* dependency across the N stacked APFM layers. Within each branch, bidirectional Mamba scanning is performed along the temporal and frequency axes.

3.4 Prior-anchored Frequency Consistency Learning (PFCL)

PFCL refines the evolved low- and high-frequency representations by evaluating their consistency against the shared visual articulatory prior.

Cross-Frequency Interaction. Let $L = L_{\mathrm{out}}$ and $H = H_{\mathrm{out}}$. Bidirectional cross-attention exchanges complementary information between the two branches:

$$L' = L + \mathrm{Out}_L\left(\mathrm{Attn}(Q_L, K_H, V_H)\right), \qquad H' = H + \mathrm{Out}_H\left(\mathrm{Attn}(Q_H, K_L, V_L)\right). \tag{16}$$

The resulting representations are further refined by temporal self-attention.

Prior-Anchored Consistency Estimation. The visual articulatory prior is broadcast along the frequency dimension and projected together with the two acoustic branches into a shared consistency space:

$$\widetilde{L} = P_L(L'), \qquad \widetilde{H} = P_H(H'), \qquad \widetilde{C} = P_C(C_v^{\mathrm{ref}}), \tag{17}$$

where $C_v^{\mathrm{ref}} = \mathrm{Broadcast}_F(C_v)$. Their pairwise deviations are

$$d_{LH} = |\widetilde{L} - \widetilde{H}|, \qquad d_{LC} = |\widetilde{L} - \widetilde{C}|, \qquad d_{HC} = |\widetilde{H} - \widetilde{C}|. \tag{18}$$

The relative consistency of the two branches is estimated as

$$[s_L, s_H] = f_{\mathrm{cons}}\left([\widetilde{L}, \widetilde{H}, \widetilde{C}, d_{LH}, d_{LC}, d_{HC}]\right), \qquad [w_L, w_H] = \mathrm{softmax}([s_L, s_H]). \tag{19}$$

The visual prior further provides branch-specific consistency gates:

$$A_L = \widetilde{L} \odot \sigma\Big(g_L(\widetilde{C})\Big), \qquad A_H = \widetilde{H} \odot \sigma\Big(g_H(\widetilde{C})\Big), \tag{20}$$

and the consistency-aware representation is obtained by

$$Z_{\mathrm{cons}} = w_L \odot A_L + w_H \odot A_H. \tag{21}$$

Residual Reconstruction. The interacted frequency representations are retained through a residual path:

$$Z_{\mathrm{base}} = \mathrm{Conv2d}([L'; H']), \qquad Z_{\mathrm{fused}} = Z_{\mathrm{base}} + Z_{\mathrm{cons}}. \tag{22}$$

4 Experiments

Datasets. We evaluate FSSCM on five AVSE benchmark settings: LRS3 Afouras et al. (2018)+DNS Dubey et al. (2023), TCD-TIMIT Harte & Gillen (2015)+NTCD-TIMIT Abdelaziz (2017), MEAD Wang et al. (2020a)+DEMAND Thiemann et al. (2013), GRID Cooke et al. (2006)+CHiME3 Barker et al. (2015), and AVSpeech Ephrat et al. (2018)+DNS Dubey et al. (2023). For each benchmark, we follow established data construction and evaluation protocols used in prior AVSE studies. We use LRS3+DNS as the primary benchmark for our main experiments Wang et al. (2023; 2025), and evaluate FSSCM at four SNR levels (-15, -10, -5, and 0 dB) using SDR, PESQ,

Table 1: Comparison with state-of-the-art methods on the LRS3+DNS benchmark. Best results are highlighted in **bold**, while second-best results are highlighted in underlined.

Model	-15 dB			-10 dB			-5 dB			0 dB		
	SDR↑	PESQ↑	STOI↑	SDR↑	PESQ↑	STOI↑	SDR↑	PESQ↑	STOI↑	SDR↑	PESQ↑	STOI↑
DEMUCS	2.33	1.210	0.561	5.84	1.297	0.682	9.10	1.443	0.777	11.85	1.631	0.839
SEMamba	3.03	1.213	0.611	6.15	1.314	0.694	9.47	1.483	0.787	11.86	1.666	0.843
DVPD	7.12	1.780	0.720	9.89	2.360	0.816	11.78	2.550	0.882	15.26	2.780	0.937
RTFS-Net	5.58	2.060	0.662	7.49	2.320	0.759	8.27	2.480	0.818	12.45	2.020	0.849
RAVSS	5.76	2.190	0.590	7.62	2.130	0.765	8.76	2.310	0.807	12.57	1.960	0.854
CTCNet	5.26	2.298	0.770	5.94	2.357	0.790	6.17	2.347	0.815	14.15	2.964	0.840
VisualVoice	2.52	1.317	0.643	5.73	1.475	0.735	8.16	1.682	0.808	10.32	1.963	0.865
Dolphin	7.49	2.470	0.680	10.42	2.618	0.862	12.27	2.850	0.914	16.85	3.030	0.944
DualAVSE	4.45	1.435	0.700	7.54	1.643	0.780	9.96	1.909	0.843	12.32	2.241	0.889
AV-DEMUCS	3.12	1.223	0.615	6.08	1.214	0.692	9.27	1.473	0.772	11.76	1.676	0.823
AV-DKF	2.74	1.139	0.530	5.36	1.411	0.590	6.58	1.794	0.716	8.20	1.445	0.795
CogCM	7.28	2.269	0.815	10.39	2.608	0.886	13.00	2.922	0.917	15.34	3.235	0.939
VG-MCA-AVSENet	5.26	1.950	0.660	8.59	1.690	0.731	9.78	2.360	0.798	13.55	2.690	0.869
AVSEMamba	3.97	1.467	0.720	6.39	1.695	0.795	8.83	1.857	0.802	10.75	1.412	0.839
FSSCM (Ours)	**10.27**	**2.949**	**0.830**	**12.58**	**3.098**	**0.921**	**14.98**	**3.326**	**0.930**	**17.30**	**3.496**	**0.954**

and STOI as the primary metrics. Benchmark-specific data construction, preprocessing, evaluation metrics, and implementation details are provided in Sec. A.1–A.3 of the supplementary material.

Baselines. We compare FSSCM primarily with representative AVSE methods as the most direct baselines, while including representative AOSE systems to assess its advantage over audio-only enhancement and representative AVSS methods as relevant references for audio-visual target-speech recovery. Specifically, the AOSE baselines include DEMUCS Défossez et al. (2020), SEMamba Chao et al. (2024), and DVPD Xue et al. (2026). For AVSE, we compare with DualAVSE Wang et al. (2023), AV-DEMUCS Tiwari et al. (2024), AV-DKF Golmakani et al. (2023), CogCM Wang et al. (2025), VG-MCA-AVSENet Saleem et al. (2026a), and AVSEMamba Chao et al. (2025). We further include RTFS-Net Pegg et al. (2024), RAVSS Pan et al. (2024), CTCNet Li et al. (2024), VisualVoice Gao & Grauman (2021), and Dolphin Li et al. (2026) as representative AVSS references. Baseline results are obtained under our evaluation protocol using publicly available implementations where necessary, while directly comparable reported results are adopted when available.

5 Results

5.1 Comparison Results

As shown in Table 1, FSSCM consistently achieves superior performance across all four evaluated SNR levels and three evaluation metrics. The performance advantage becomes more pronounced as the acoustic corruption increases from 0 dB to the most challenging condition at -15 dB. At -15 dB, FSSCM achieves 10.27 dB SDR, 2.949 PESQ, and 0.830 STOI, outperforming the corresponding second-best results by 2.78 dB, 0.479, and 0.015, respectively. Averaged across all evaluated SNR levels, FSSCM further improves over CogCM by 2.28 dB in SDR and 0.459 in PESQ, while consistently achieving higher STOI across all SNR levels. Interestingly, FSSCM achieves a PESQ of 2.949 at -15 dB, which is comparable to or higher than the 0 dB results of many competing methods. This observation further demonstrates its robustness under severe acoustic corruption.

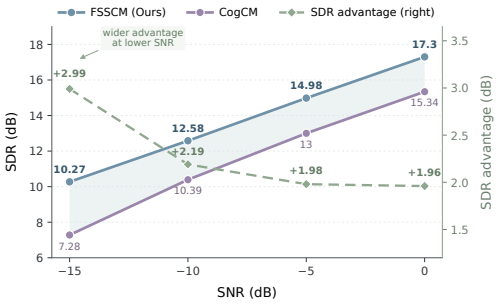


Figure 3: SDR performance and relative advantage of FSSCM over CogCM across SNR levels on LRS3+DNS.

To examine the hypothesis introduced in Sec. 1, we further analyze how the relative advantage changes with acoustic reliability. Compared with CogCM, the best-performing AVSE baseline in Table 1, the SDR advantage of FSSCM widens from 1.96 dB at 0 dB SNR to 2.99 dB at -15 dB SNR, as illustrated in Fig. 3, while its PESQ advantage increases from 0.261 to 0.680. Notably, the

Table 2: Ablation of the three core components of FSSCM on LRS3+DNS. Values in parentheses denote SDR degradation relative to the full model.

Model	−15 dB			−10 dB			−5 dB			0 dB		
	SDR↑	PESQ↑	STOI↑	SDR↑	PESQ↑	STOI↑	SDR↑	PESQ↑	STOI↑	SDR↑	PESQ↑	STOI↑
Full FSSCM	**10.27**	**2.949**	**0.830**	**12.58**	**3.098**	**0.921**	**14.98**	**3.326**	**0.930**	**17.30**	**3.496**	**0.954**
w/o VSPL	8.12 (↓**2.15**)	2.731	0.807	10.41 (↓**2.17**)	2.866	0.849	13.02 (↓**1.96**)	3.104	0.883	15.68 (↓**1.62**)	3.252	0.932
w/o APFM	8.43 (↓**1.84**)	2.775	0.811	10.71 (↓**1.87**)	2.912	0.853	13.35 (↓**1.63**)	3.152	0.887	16.05 (↓**1.25**)	3.298	0.934
w/o PFCL	9.58 (↓**0.69**)	2.881	0.822	11.92 (↓**0.66**)	3.021	0.865	14.41 (↓**0.57**)	3.258	0.895	16.85 (↓**0.45**)	3.424	0.945

stronger advantage under severe acoustic degradation is not limited to CogCM: compared with the best-performing AVSS and AOSE baselines in terms of SDR, Dolphin and DVPD, the SDR margins increase from 0.45 and 2.04 dB at 0 dB SNR to 2.78 and 3.15 dB at −15 dB SNR, respectively. These results are consistent with our hypothesis that cause-governed recovery becomes increasingly beneficial as acoustic reliability decreases. Qualitative spectrogram comparisons at −15 dB are provided in Sec. A.5.3 of the supplementary material.

To further examine whether the effectiveness of FSSCM extends beyond LRS3+DNS, we evaluate it on four additional benchmarks under their respective benchmark-specific protocols. For a concise cross-benchmark comparison, Fig. 4 reports the lowest-SNR condition of each benchmark: STOIi on GRID+CHiME3 (−5 dB), SI-SDRi on TCD-TIMIT+NTCD-TIMIT (−5 dB) and MEAD+DEMAND (−10 dB), and PESQi on AVSpeech+DNS (Cond. 3, −10 dB). FSSCM achieves higher scores than CogCM on all four benchmarks under the reported conditions, demonstrating its effectiveness across diverse datasets and evaluation settings. Together with the multi-SNR results on LRS3+DNS, these cross-benchmark evaluations indicate that FSSCM maintains its advantage under challenging acoustic conditions. Complete comparisons across all metrics and competing methods are provided in Sec. A.4 of the supplementary material.

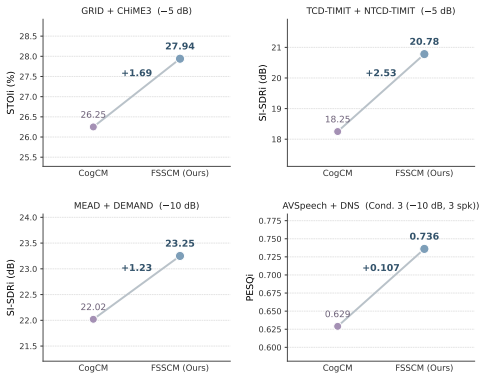


Figure 4: Comparison between FSSCM and CogCM on four additional AVSE benchmarks under their lowest evaluated SNR conditions.

5.2 Ablation Studies

We conduct ablation experiments on LRS3+DNS to evaluate the contributions of the core FSSCM components. Under the same training settings as the full model, each component is replaced by a minimal functional baseline while the remaining architecture is kept unchanged. Specifically, VSPL is replaced by a basic visual encoder preserving the C_v and G_v interfaces, APFM by a plain dual-branch recovery module without adaptive frequency routing, transition modulation, or progressive frequency coupling, and PFCL by direct low-/high-frequency feature interaction. Table 2 reports results across four SNR levels, with SDR drops relative to the full model shown in parentheses. Additional configurations and internal mechanism ablations are provided in the supplementary material.

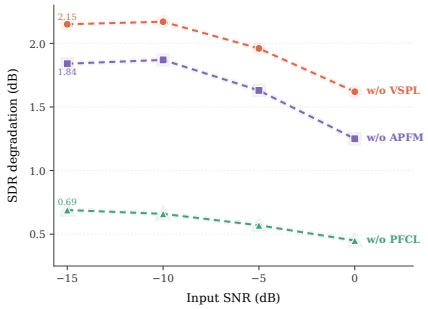


Figure 5: SNR-dependent SDR degradation caused by replacing each core module of FSSCM with its corresponding baseline.

Contribution of Core Modules. Replacing each of the three proposed modules with its corresponding minimal functional baseline consistently degrades performance across all evaluated SNR levels, indicating that each proposed design contributes to the final recovery performance. Replacing VSPL with the basic visual encoder causes the largest degradation, reducing SDR by 2.15 dB at −15 dB and 1.62 dB at 0 dB. Replacing APFM with the plain dual-branch

Table 3: Analysis of visual intervention and frequency recovery organization under varying SNR conditions. SDR (dB) is reported here; complete PESQ and STOI results are provided in the supplementary material.

(a) Visual Intervention

Design	SDR (dB)			
	-15	-10	-5	0
Feature Fusion	8.23	10.57	13.24	15.76
Feature Modulation	8.87	11.34	13.77	16.28
Transition Governance	**10.27**	**12.58**	**14.98**	**17.30**

(b) Frequency Recovery Organization

Organization	SDR (dB)			
	-15	-10	-5	0
Unified	8.46	10.83	13.18	15.72
$L \parallel H$	9.73	12.07	14.43	16.88
$H \rightarrow L$	8.87	11.24	13.67	16.23
$L \rightarrow H$	**10.27**	**12.58**	**14.98**	**17.30**

recovery baseline leads to corresponding drops of 1.84 dB and 1.25 dB, while replacing PFCL with direct frequency-feature fusion reduces SDR by 0.69 dB and 0.45 dB, respectively. These results demonstrate the complementary roles of dynamics-aware visual articulatory prior learning, frequency-coupled state recovery, and prior-anchored cross-frequency consistency.

Trend Across SNR Levels. As shown in Fig. 5, the SDR penalties caused by replacing each core component with its corresponding baseline tend to become larger as SNR decreases. Comparing 0 and -15 dB, the SDR degradation increases from 1.62 to 2.15 dB for VSPL, from 1.25 to 1.84 dB for APFM, and from 0.45 to 0.69 dB for PFCL. This overall trend indicates that the contributions of the three core components become more pronounced under increasingly severe acoustic corruption. These results further support the importance of the key mechanisms underlying cause-governed recovery under severe acoustic degradation. Additional mechanism-level ablations and robustness to missing visual input are provided in Sec. A.5 of the supplementary material.

5.3 Analysis of Recovery Principles

Where Should Visual Information Intervene in Recovery? We compare Feature Fusion, Feature Modulation, and the proposed Transition Governance under the same visual encoder, acoustic backbone, frequency routing, decoder, and training protocol, differing only in how C_v intervenes in recovery. As shown in Table 3(a), Transition Governance achieves the highest SDR at all SNR levels. Its gain over Feature Fusion increases from 1.54 dB at 0 dB to 2.04 dB at -15 dB, while the gain over Feature Modulation increases from 1.02 dB to 1.40 dB. These results demonstrate the effectiveness of transition-level visual governance over feature-level interaction, particularly under severe acoustic degradation.

Frequency Recovery Organization and Directionality. We compare unified full-band recovery, parallel low-/high-frequency recovery ($L \parallel H$), the reverse dependency ($H \rightarrow L$), and the proposed $L \rightarrow H$ organization. Except for unified recovery, all split-frequency variants share the same partitioning mechanism and branch architectures. As shown in Table 3(b), $L \parallel H$ already improves over unified recovery, while $L \rightarrow H$ further outperforms $L \parallel H$ by 0.42–0.55 dB across SNR levels and consistently surpasses $H \rightarrow L$. These results validate the proposed content-to-detail dependency, where evolving low-frequency recovery guides subsequent high-frequency detail reconstruction. Complete PESQ/STOI results and configurations are provided in Sec. A.5.5 and Sec. A.5.6 of the supplementary material.

6 Conclusion

This work reframes AVSE from observation-driven recovery to cause-governed, frequency-coupled state recovery. FSSCM realizes this perspective by using a noise-independent visual articulatory prior to govern recovery dynamics and by explicitly modeling the dependency between low-frequency content and high-frequency details. These principles are instantiated within a unified state-space framework. Experiments on five benchmarks consistently demonstrate the effectiveness and robustness of FSSCM under diverse acoustic conditions. In particular, its advantage on LRS3+DNS widens as SNR decreases, reaching 2.99 dB SDR and 0.68 PESQ improvements over CogCM at -15 dB, supporting our hypothesis that cause-governed recovery becomes increasingly beneficial when acoustic observations become unreliable. These results highlight visual articulatory priors and coupled frequency-state evolution as a promising direction for robust AVSE.

REPRODUCIBILITY STATEMENT

The implementation of FSSCM is developed in Python using PyTorch. To facilitate reproducibility, we provide detailed implementation settings, optimization hyperparameters, data preprocessing procedures, dataset construction protocols, evaluation metrics, and ablation configurations in the supplementary material. The experiments are conducted on five AVSE benchmark settings, and all datasets are obtained from their official sources or following the access procedures described in the cited references. Upon acceptance, we plan to release the source code, configuration files, data preprocessing scripts, and evaluation code required to reproduce the reported experiments. Detailed training settings and hyperparameters are provided in Sec. A.1, evaluation protocols and metric definitions in Sec. A.2, and benchmark-specific dataset details in Sec. A.3 of the supplementary material.

AI USE STATEMENT

In this work, generative AI tools were used for language-related assistance, including manuscript editing, translation, and refinement of technical descriptions. AI-generated suggestions were reviewed and revised by the authors. Generative AI was not used to generate experimental data, results, or scientific conclusions. The authors take full responsibility for the methodology, experiments, reported results, and final content of this work.

REFERENCES

Ahmed Hussen Abdelaziz. NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition. In *Proceedings of Interspeech*, pp. 3752–3756, 2017. doi: 10.21437/Interspeech.2017-860.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: A large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

Shafique Ahmed, Jen-Cheng Hou, and Yu Tsao. Av-locofilm: Audio-visual speech enhancement using film-based fusion and hybrid local–global transformers. In *4th Cogmhear Audio-Visual Speech Enhancement Challenge (AVSEC)*, pp. 42–44, 2025. doi: 10.21437/AVSEC.2025-9.

S. Balasubramanian, R. Rajavel, and Asuthos Kar. Ideal ratio mask estimation based on cochleagram for audio-visual monaural speech enhancement. *Appl. Acoust.*, 211:109524, 2023.

Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third CHiME speech separation and recognition challenge: Dataset, task and baselines. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 504–511, 2015. doi: 10.1109/ASRU.2015.7404837.

Ruizhe Cao, Sherif Abdulatif, and Bin Yang. Cmgan: Conformer-based metric gan for speech enhancement. In *Proc. Interspeech*, pp. 936–940, 2022.

Rong Chao, Wen-Huang Cheng, Moreno La Quatra, Sabato Marco Siniscalco, Chao-Han Huck Yang, Szu-Wei Fu, and Yu Tsao. An investigation of incorporating mamba for speech enhancement. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pp. 302–308, 2024.

Rong Chao, Wenze Ren, You-Jin Li, Kuo-Hsuan Hung, Sung-Feng Huang, Sze-Wei Fu, Wen-Huang Cheng, and Yu Tsao. Leveraging mamba with full-face vision for audio-visual speech enhancement. In *4th Cogmhear Audio-Visual Speech Enhancement Challenge (AVSEC)*, pp. 77–78, 2025.

Honglie Chen, Rodrigo Mira, Stavros Petridis, and Maja Pantic. RT-LA-VocE: Real-time low-snr audio-visual speech enhancement. In *Proceedings of Interspeech 2024*, pp. 2215–2219, 2024. doi: 10.21437/Interspeech.2024-516.

Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006. doi: 10.1121/1.2229005.

Kenneth J. de Jong and Yuka Tashiro. Locating fricative categories in a synthetic acoustic space: Spectral moments measures of a perceptual english fricative system. *J. Acoust. Soc. Am.*, 157 (4_Supplement):A166, 2025. doi: 10.1121/10.0037771.

Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. In *Proc. Interspeech*, pp. 3291–3295, 2020.

Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, Hannes Gamper, Mehrsa Golestaneh, and Robert Aichner. ICASSP 2023 deep noise suppression challenge. In *ICASSP*, 2023.

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*, 37(4):112:1–112:11, 2018. doi: 10.1145/3197517.3201357.

Daniel Fogerty and Fei Chen. Vowel fundamental and formant frequency contributions to english and mandarin sentence intelligibility. In *Proc. Interspeech*, pp. 1382–1386, 2016. doi: 10.21437/Interspeech.2016-28.

Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. In *Proc. Interspeech*, pp. 1170–1174, 2018.

Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *Proc. CVPR*, pp. 15490–15500, 2021.

Ali Golmakani, Mostafa Sadeghi, and Romain Serizel. Audio-visual speech enhancement with a deep kalman filter generative model. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 1–5, 2023.

Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6633–6637, 2021. doi: 10.1109/ICASSP39728.2021.9414177.

Naomi Harte and Eoin Gillen. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015. doi: 10.1109/TMM.2015.2407694.

Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. In *Proc. Interspeech*, pp. 2472–2476, 2020. doi: 10.21437/Interspeech.2020-2537.

Jesper Jensen and Cees H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016.

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 *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7302–7306. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9746401.

Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, volume 1, pp. 125–128. IEEE, 1993.

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

Simon Leglaive, Laurent Girin, and Radu Horaud. A variance modeling framework based on variational autoencoders for speech enhancement. In *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, pp. 1–6, 2018.

Jizhen Li, Weiping Tu, Yuhong Yang, Xinmeng Xu, Yiqun Zhang, and Yanzhen Ren. Improving speech enhancement by cross- and sub-band processing with state space model. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10888886.

Kai Li, Fenghua Xie, Hang Chen, Kexin Yuan, and Xiaolin Hu. An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6637–6651, 2024. doi: 10.1109/TPAMI.2024.3384034.

Kai Li, Kejun Gao, and Xiaolin Hu. Efficient audio-visual speech separation with discrete lip semantics and multi-scale global-local attention. In *International Conference on Learning Representations (ICLR)*, 2026.

Philipos C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, 2 edition, 2013.

Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling. MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra. In *Proc. Interspeech*, pp. 3834–3838, 2023.

Matthew Masapollo, Rosalie Gendron, Erin Wyndham, Ally Marcellus, Allen Shamsi, and Nathan Maxfield. Interarticulator timing relations underlie the production of precise and consistent vocal tract constrictions during speech. *Journal of Speech, Language, and Hearing Research*, 68(6): 2700–2720, 2025. doi: 10.1044/2025_JSLHR-24-00535.

Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 29:1368–1396, 2021.

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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096390.

Tianrui Pan, Jie Liu, Bohan Wang, Jie Tang, and Gangshan Wu. RAVSS: Robust audio-visual speech separation in multi-speaker scenarios with missing visual cues. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4748–4756, 2024. doi: 10.1145/3664647.3681261.

Zexu Pan, Ruijie Tao, Chenglin Xu, and Haizhou Li. Muse: Multi-modal target speaker extraction with visual cues. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 6678–6682, 2021.

Samuel Pegg, Kai Li, and Xiaolin Hu. RTFS-Net: Recurrent time-frequency modelling for efficient audio-visual speech separation. In *International Conference on Learning Representations (ICLR)*, 2024.

Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis. mir_eval: A transparent implementation of common mir metrics. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 367–372, 2014.

Karthik Ramesh, Chao Xing, Wupeng Wang, Dong Wang, and Xiao Chen. VSET: A multi-modal transformer for visual speech enhancement. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6658–6662. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9414053.

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 (ICASSP)*, volume 2, pp. 749–752. IEEE, 2001.

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. doi: 10.1109/TSP.2021.3066038.

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. doi: 10.1109/TASLP.2020.3000593.

Nasir Saleem, Adeel Hussain, Kia Dashtipour, Eamon Sheikh, Aziz Sheikh, Tughrul Arslan, and Amir Hussain. Viseme-gated multilayer cross-attentional feature fusion for cognitively-inspired multimodal speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, 34:469–481, 2026a. doi: 10.1109/TASLPRO.2025.3648801.

Nasir Saleem, Fazal E. Wahab, Sami Bourouis, and Amir Hussain. Frequency-aware selective state-space modeling for audio-visual speech enhancement. *Digital Signal Processing*, 174:105986, 2026b. doi: 10.1016/j.dsp.2026.105986.

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.

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, 2020. doi: 10.1109/TASLP.2019.2955276.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings. *Proceedings of Meetings on Acoustics*, 19(1):035081, 2013. doi: 10.1121/1.4799597.

Utkarsh Tiwari, Mandar Gogate, Kia K. Dashtipour, Eamon Sheikh, Rimjhim Dr. Singh, Tughrul Arslan, and Amir Hussain. Real-time audio visual speech enhancement: Integrating visual cues for improved performance. In *3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC)*, pp. 38–42, 2024. doi: 10.21437/AVSEC.2024-9.

Feixiang 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)*. BMVA, 2023.

Feixiang Wang, Shuang Yang, Shiguang Shan, and Xilin Chen. Cogcm: Cognition-inspired contextual modeling for audio-visual speech enhancement. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 21408–21418, 2025.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pp. 700–717. Springer, 2020a. doi: 10.1007/978-3-030-58589-1_42.

Wupeng Wang, Chao Xing, Dong Wang, Xiao Chen, and Fengyu Sun. A robust audio-visual speech enhancement model. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7529–7533. IEEE, 2020b.

Haitao Xu, Liangfa Wei, Jie Zhang, Jianming Yang, Yannan Wang, Tian Gao, Xin Fang, and Li-Rong Dai. A multi-scale feature aggregation based lightweight network for audio-visual speech enhancement. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096565.

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 *Proceedings of Interspeech 2022*, pp. 971–975, 2022. doi: 10.21437/Interspeech.2022-10041.

Ke Xue, Rongfei Fan, Kai Li, Shanping Yu, Puning Zhao, and Jianping An. Dual-view predictive diffusion: Lightweight speech enhancement via spectrogram-image synergy. In *Proc. International Conference on Machine Learning (ICML)*, 2026.

Tao Zhang, Qianwei Sun, Weiwei Zhang, and Yanzhang Geng. Nrsrnet: Speech enhancement network based on noise reduction and restoration module under extremely low snr conditions. *Applied Acoustics*, 241:111012, 2026. doi: 10.1016/j.apacoust.2025.111012.

A Supplementary Material

A.1 Implementation Details

A.1.1 Data Preprocessing

In all experiments, audio samples are resampled to 16 kHz, and the video frame rate is set to 25 fps. The length of speech segments is set to 2 seconds for GRID+CHiME3 and AVSpeech+DNS, and 1.5 seconds for LRS3+DNS, MEAD+DEMAND, and NTCD-TIMIT. The STFT / iSTFT is performed using a Hann window of 400 samples with a hop size of 100 samples for all datasets except MEAD+DEMAND, where $n_{\text{fft}} = 512$ with the same window length and hop size. For visual inputs, we extract the lip ROI at a resolution of 88×88 grayscale frames. During training, random cropping and horizontal flipping are applied to GRID+CHiME3 and AVSpeech+DNS. For LRS3+DNS, where the ROI is already cropped to 88×88 during preprocessing, only temporal random cropping of the audio is applied. For MEAD+DEMAND and NTCD-TIMIT, only temporal random cropping is applied to both audio and video. During testing, GRID+CHiME3 uses deterministic center cropping, while the other datasets use the full utterance.

A.1.2 Hyperparameter Settings

We use the AdamW optimizer with an initial learning rate of 5×10^{-4}, weight decay of 0.01, and beta coefficients of $(0.8, 0.99)$. The learning rate is decayed by $\gamma = 0.99$ per epoch using an ExponentialLR scheduler. The loss weights are set as $\lambda_{\text{mag}} = 0.9$, $\lambda_{\text{ri}} = 0.1$, $\lambda_{\text{pha}} = 0.3$, $\lambda_{\text{time}} = 0.2$, and $\lambda_{\text{con}} = 0.1$. All experiments are conducted on NVIDIA A800-80G GPUs: LRS3+DNS uses 4 GPUs with a total batch size of 8; GRID+CHiME3, MEAD+DEMAND, and NTCD-TIMIT use 4 GPUs with a total batch size of 4; AVSpeech+DNS uses 2 GPUs with a total batch size of 4. Training proceeds until validation loss converges, with convergence rates varying across datasets.

A.2 Evaluation Metrics

We employ a comprehensive set of evaluation metrics to assess system performance. These metrics are grouped into Absolute Metrics, Improvement Metrics, and Downstream Task Metrics.

Absolute Metrics. Absolute metrics directly evaluate the quality and intelligibility of the enhanced speech:

- **Signal-to-Distortion Ratio (SDR)** Raffel et al. (2014): Measures the distortion between the enhanced speech and the clean reference. Higher values indicate better reconstruction quality.
- **Scale-Invariant Signal-to-Noise Ratio (SI-SNR)** Le Roux et al. (2019): Measures speech reconstruction quality while removing the effect of global signal scaling. Higher values indicate better performance.
- **Short-Time Objective Intelligibility (STOI)** Taal et al. (2011): Measures speech intelligibility based on the short-time correspondence between the enhanced and clean speech. Higher values indicate better intelligibility.
- **Extended Short-Time Objective Intelligibility (ESTOI)** Jensen & Taal (2016): Extends STOI to provide a more robust estimate of speech intelligibility under complex acoustic conditions. Higher values indicate better intelligibility.
- **Perceptual Evaluation of Speech Quality (PESQ)** Rix et al. (2001): Evaluates the perceptual quality of the enhanced speech relative to the clean reference. Higher values indicate better perceptual quality.
- **Mel Cepstral Distortion (MCD)** Kubichek (1993): Measures spectral distortion between the enhanced and clean speech in the mel-cepstral domain. Lower values indicate better spectral fidelity.

For consistency with CogCM, SDR is computed using the same `mir_eval`-based evaluation protocol.

Improvement Metrics quantify the change in performance from the noisy speech to the enhanced speech. This approach is adopted because some methods report these metric gains Mira et al. (2023), 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.

A.3 Dataset Details

We conduct AVSE experiments on five benchmark settings: LRS3+DNS, GRID+CHiME3, TCD-TIMIT+NTCD-TIMIT, MEAD+DEMAND, and AVSpeech+DNS. Across these settings, clean audio-visual speech is paired with either environmental noise recordings or benchmark-specific noisy speech to construct or evaluate corrupted acoustic inputs.

LRS3 + DNS: LRS3 Afouras et al. (2018) contains approximately 438 hours of talking videos collected from TED and TEDx clips. We use its pretrain subset, containing approximately 407 hours of video, and partition it into training, validation, and testing sets with a ratio of 8:1:1. Following Défossez et al. Défossez et al. (2020), we obtain noise from the DNS noise subset, which contains approximately 181 hours of audio from diverse acoustic events. During training and evaluation, these noise samples are added to clean speech to construct synthetic noisy inputs.

GRID + CHiME3: We use 33 speakers from the GRID audio-visual corpus Cooke et al. (2006). Following Balasubramanian et al. (2023), speakers s2 and s22 are used for validation, s1 and s12 for testing, and the remaining 29 speakers for training. Noise is sampled from CHiME3 Barker et al. (2015), including Cafe, Street, Bus, and Pedestrian environments. The CHiME3 noise data are partitioned into training, validation, and testing sets with an 8:1:1 ratio.

TCD-TIMIT + NTCD-TIMIT: We use a 56-speaker subset of TCD-TIMIT Harte & Gillen (2015), containing audio-visual English speech recorded at 16 kHz. Following Harte and Gillen Harte & Gillen (2015), we use 39 speakers for training, 8 for validation, and 9 for testing. Noisy speech is obtained from NTCD-TIMIT Abdelaziz (2017), which is constructed by adding six noise types—Living Room, White, Cafe, Car, Babble, and Street—to TCD-TIMIT speech at different SNRs. Following Golmakani et al. Golmakani et al. (2023), we select five utterances per noise level and noise type for each test speaker, yielding 1,350 test utterances.

MEAD + DEMAND: MEAD Wang et al. (2020a) is a multi-view emotional audio-visual dataset containing eight emotions, three intensity levels, and seven camera viewpoints. Following Kang et al. (2022), we use the 46 publicly available participants and select frontal-view videos at the highest emotion intensity (level 3). Approximately 5 hours are used for training, 0.7 hours for validation, and 0.7 hours for testing. Noise is sampled from DEMAND Thiemann et al. (2013), which contains recordings from diverse real-world acoustic environments.

AVSpeech + DNS: AVSpeech Ephrat et al. (2018) is a large-scale audio-visual dataset containing approximately 4,700 hours of 3–10 second video segments collected from about 290,000 YouTube videos and covering nearly 150,000 speakers. Each clip contains a single visible speaker, with substantial diversity in speakers, languages, and facial poses.

A.4 Additional Benchmark Results

We further evaluate FSSCM on the GRID+CHiME3 benchmark. Following Wang et al. (2020b), we use CHiME3 noises to synthesize noisy inputs and evaluate the models at -5 and 0 dB SNR, reporting STOI and PESQ improvements. We compare with L2L (Ephrat et al., 2018), VSE (Gabbay et al., 2018), OVA (Wang et al., 2020b), VSET (Ramesh et al., 2021), MHCA-AVCRN (Xu et al., 2022), M3Net (Xu et al., 2023), DualAVSE (Wang et al., 2023), VG-MCA-AVSENet (Saleem et al., 2026a), and CogCM (Wang et al., 2025). We include previously reported

Table 4: Performance comparison on the GRID+CHiME3 benchmark. The best performance is shown in **boldface**, and the second best is shown in underlined.

Method	−5 dB		0 dB	
	STOIi(%)↑	PESQi↑	STOIi(%)↑	PESQi↑
L2L	11.14	0.54	8.86	0.62
VSE	–	0.45	–	0.60
OVA	–	0.40	–	0.66
VSET	–	0.50	–	0.75
MHCA-AVCRN	13.51	0.76	11.25	0.88
M3Net	13.42	0.75	11.31	0.89
DualAVSE	15.79	0.76	13.56	0.92
VG-MCA-AVSENet	20.27	0.72	15.12	0.92
CogCM	26.25	1.01	**20.50**	1.21
FSSCM (Ours)	**27.94**	**1.11**	20.16	**1.23**

results under this benchmark setting and additionally reproduce several recent AVSE methods under the same evaluation protocol. As shown in Table 4, FSSCM achieves the best PESQi at both evaluated SNR levels while maintaining competitive STOIi performance.

A.4.1 Comparison Results on TCD-TIMIT + NTCD-TIMIT

We further evaluate FSSCM on the TCD-TIMIT+NTCD-TIMIT benchmark. We compare with A-VAE and AV-VAE (Sadeghi & Alameda-Pineda, 2021), A-DKF and AV-DKF (Golmakani et al., 2023), DualAVSE (Wang et al., 2023), and CogCM (Wang et al., 2025). Following the evaluation protocol of Golmakani et al. (2023), we report improvements in SI-SDR, PESQ, and STOI for comparison. As shown in Table 5, FSSCM achieves competitive performance across the evaluated metrics and SNR levels.

Table 5: Performance comparison on the TCD-TIMIT+NTCD-TIMIT benchmark. The best performance is shown in **boldface**, and the second best is shown in underlined.

Method	SI-SDRi (dB) ↑					PESQi ↑					STOIi ↑				
	−5 dB	0 dB	5 dB	10 dB	15 dB	−5 dB	0 dB	5 dB	10 dB	15 dB	−5 dB	0 dB	5 dB	10 dB	15 dB
A-VAE	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	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	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	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	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	18.25	18.95	**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**
FSSCM (Ours)	**20.78**	**18.96**	18.29	15.68	**14.18**	**1.23**	1.29	**1.79**	**1.98**	**2.12**	**0.28**	0.21	0.21	**0.19**	**0.11**

A.4.2 Comparison Results on MEAD + DEMAND

We further evaluate FSSCM on the MEAD+DEMAND benchmark. We compare with A-VAE (Leglaive et al., 2018), AV-CVAE (Sadeghi et al., 2020), AV-CVAE-WithHM and AV-CVAE-RFF (Kang et al., 2022), DualAVSE (Wang et al., 2023), VG-MCA-AVSENet (Saleem et al., 2026a), and CogCM (Wang et al., 2025). We include prior AVSE methods with previously reported results under this benchmark setting and additionally reproduce several recent AVSE methods under the same evaluation protocol. Following the benchmark configuration described in Sec. A.3, we report improvements in SI-SDR, PESQ, and STOI across multiple SNR conditions. As shown in Table 6, FSSCM achieves competitive performance across the evaluated metrics and SNR levels, demonstrating its effectiveness under diverse acoustic conditions.

Table 6: Performance comparison on the MEAD+DEMAND benchmark. The best performance is shown in **boldface**, and the second best is shown in underlined.

Method	SI-SDRi (dB) ↑					PESQi ↑					STOIi ↑				
	−10 dB	−5 dB	0 dB	5 dB	10 dB	−10 dB	−5 dB	0 dB	5 dB	10 dB	−10 dB	−5 dB	0 dB	5 dB	10 dB
A-VAE	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	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	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	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	16.06	15.21	14.09	12.98	11.27	0.35	0.54	0.74	0.92	1.01	0.12	0.11	0.08	0.05	**0.03**
VG-MCA-AVSENet	16.42	15.38	14.27	14.45	10.98	0.80	1.37	1.49	1.34	1.34	0.12	0.14	0.09	**0.07**	0.02
CogCM	22.02	**21.27**	19.22	**16.90**	14.13	1.26	**1.57**	1.68	1.62	1.33	**0.17**	0.14	0.09	0.06	**0.03**
FSSCM (Ours)	**23.25**	20.21	**19.28**	16.84	**15.17**	**1.32**	1.48	**1.76**	**1.77**	**1.58**	0.16	**0.15**	**0.12**	**0.07**	**0.03**

Table 7: Comparison results on the AVSpeech+DNS benchmark under different noise conditions. The best performance is shown in **boldface**, and the second best is shown in underlined. Ties at the reported precision are highlighted equally.

Method	Input	Metrics			
		MCDi↓	PESQi↑	STOIi↑	ESTOIi↑
Noise condition 1 (1 background noise at 0 dB SNR + 1 interfering speaker at 0 dB SIR)					
GCRN	A	0.410	0.044	-0.052	-0.038
AV-GCRN	AV	-1.193	0.394	0.220	0.235
AV-Demucs	AV	-5.581	0.738	0.270	0.298
MuSE	AV	-5.528	0.787	0.276	0.299
VisualVoice	AV	-3.781	0.606	0.249	0.270
LA-VocE	AV	-6.653	0.931	0.294	0.333
RT-LA-VocE	AV	-6.157	0.653	0.255	0.282
DualAVSE	AV	-6.164	0.723	0.286	0.213
CogCM	AV	**-13.155**	1.466	0.307	0.445
FSSCM (Ours)	AV	-12.854	**1.482**	**0.314**	**0.452**
Noise condition 2 (3 background noises at −5 dB SNR + 2 interfering speakers at −5 dB SIR)					
GCRN	A	0.416	-0.010	-0.015	-0.015
AV-GCRN	AV	-1.354	0.096	0.234	0.214
AV-Demucs	AV	-5.548	0.274	0.308	0.300
MuSE	AV	-5.314	0.297	0.308	0.289
VisualVoice	AV	-3.388	0.164	0.253	0.237
LA-VocE	AV	-6.863	0.511	**0.379**	0.397
RT-LA-VocE	AV	-6.313	0.297	0.315	0.321
DualAVSE	AV	-7.258	1.056	0.342	0.346
CogCM	AV	**-15.391**	1.056	0.343	**0.446**
FSSCM (Ours)	AV	-13.680	**1.192**	0.352	0.416
Noise condition 3 (5 background noises at −10 dB SNR + 3 interfering speakers at −10 dB SIR)					
GCRN	A	-0.015	0.045	-0.020	-0.005
AV-GCRN	AV	-1.263	0.043	0.171	0.139
AV-Demucs	AV	-5.170	0.013	0.262	0.230
MuSE	AV	-4.418	0.011	0.231	0.218
VisualVoice	AV	-3.125	0.045	0.181	0.166
LA-VocE	AV	-6.170	0.159	**0.371**	0.206
RT-LA-VocE	AV	-5.671	0.039	0.287	0.272
DualAVSE	AV	**-17.705**	0.629	0.323	0.373
CogCM	AV	**-17.705**	0.629	0.342	0.373
FSSCM (Ours)	AV	-17.368	**0.736**	0.368	**0.381**

A.4.3 Comparison Results on AVSpeech + DNS

We further evaluate FSSCM on the AVSpeech+DNS benchmark. Following LA-VocE (Mira et al., 2023), we adopt the same evaluation protocol and randomly select 1% of the test set (approximately 1,500 clips) for evaluation. We compare with GCRN and AV-GCRN (Tan & Wang, 2020), AV-Demucs (Défossez et al., 2020), MuSE (Pan et al., 2021), VisualVoice (Gao & Grauman, 2021), LA-VocE (Mira et al., 2023), RT-LA-VocE (Chen et al., 2024), DualAVSE (Wang et al., 2023), and CogCM (Wang et al., 2025). We include previously reported results under this benchmark setting and additionally reproduce recent AVSE methods under the same evaluation protocol.

We conduct comparisons under three noise conditions. Condition 1 comprises 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. Condition 3 comprises five background noises at -10 dB SNR and three interfering speakers at -10 dB SIR.

As shown in Table 7, we report MCDi, PESQi, STOIi, and ESTOIi. FSSCM achieves competitive performance across all three noise conditions, demonstrating its effectiveness under complex acoustic scenarios.

A.5 Additional Ablation Details and Mechanism Analysis

This section complements the ablation studies in the main paper with detailed replacement configurations and additional mechanism analyses. We first specify the controlled module-level replacements used for the *w/o VSPL*, *w/o APFM*, and *w/o PFCL* variants, where only the target module is replaced while the remaining architecture is kept unchanged. We then further dissect the internal mechanisms of VSPL, APFM, and PFCL, analyze robustness to missing visual input, and provide

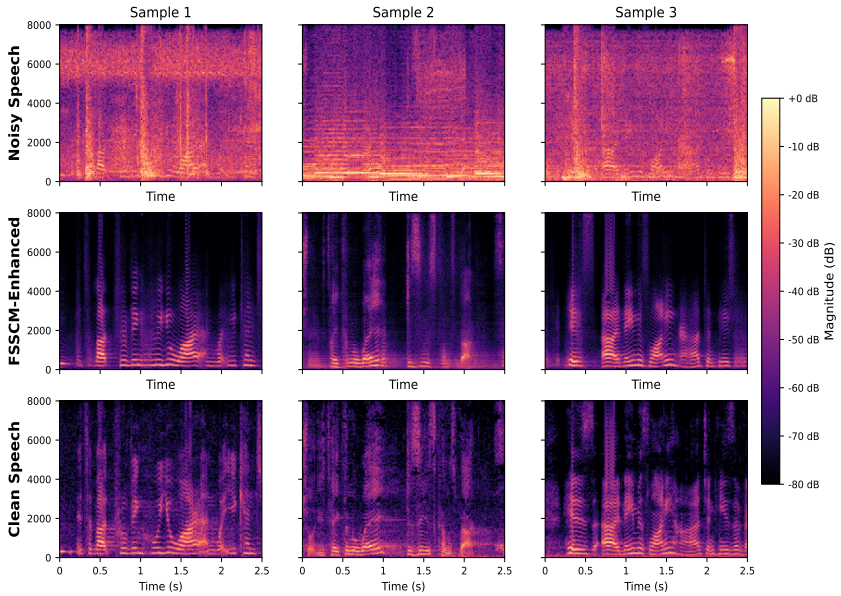


Figure 6: Spectrogram comparison on three randomly selected samples from the LRS3 test set at -15 dB. From top to bottom: noisy speech, FSSCM-enhanced speech, and clean speech.

complete results and implementation details for the visual-intervention and frequency-recovery analyses reported in the main paper.

A.5.1 Module-level Ablation Configurations

To ensure controlled module-level comparisons, each proposed component is replaced by a minimal functional baseline while all remaining components, training settings, and output interfaces are kept unchanged. This design avoids confounding the contribution of a target module with structural changes in its downstream components.

w/o VSPL. We replace the proposed dynamics-aware VSPL with a basic visual encoder. Specifically, the lip sequence is encoded by the same 3D-CNN backbone, while the first- and second-order articulatory dynamics, adaptive multi-order fusion, and temporal self-attention in VSPL are removed. The resulting visual features are directly projected and temporally aligned to produce C_v and G_v with the same dimensionality as in the full model. APFM and PFCL are kept unchanged. Therefore, this variant evaluates the contribution of the proposed dynamics-aware visual articulatory prior construction rather than the presence or absence of visual information itself.

w/o APFM. We replace APFM with a plain dual-branch recovery baseline while preserving the same branch depth and hidden dimensionality. The acoustic representation is divided into low- and high-frequency components using a fixed frequency partition and processed by two independent vanilla Mamba branches. Adaptive frequency routing, visual-prior-driven state-transition modulation, and progressive $L \rightarrow H$ frequency coupling are removed. VSPL and PFCL remain unchanged. This variant therefore isolates the contribution of the proposed adaptive and cause-governed recovery mechanism from that of the underlying dual-branch state-space backbone.

w/o PFCL. We remove the prior-anchored cross-frequency consistency mechanism and replace it with direct feature fusion:

$$Z_{\text{fused}} = \text{Conv2d}\left([L_{\text{out}}; H_{\text{out}}]\right). \tag{23}$$

All VSPL and APFM components are retained unchanged. This variant evaluates whether the proposed prior-anchored cross-frequency interaction and consistency modeling provides additional benefit over simple low-/high-frequency feature fusion.

Table 8: Detailed internal mechanism ablations of FSSCM on LRS3+DNS. Values in parentheses denote SDR degradation relative to the full model.

Variant	−15 dB			−10 dB			−5 dB			0 dB		
	SDR↑	PESQ↑	STOI↑	SDR↑	PESQ↑	STOI↑	SDR↑	PESQ↑	STOI↑	SDR↑	PESQ↑	STOI↑
Full FSSCM	**10.27**	**2.949**	**0.830**	**12.58**	**3.098**	**0.921**	**14.98**	**3.326**	**0.930**	**17.30**	**3.496**	**0.954**
VSPL												
w/o articulatory dynamics	8.45 (↓**1.82**)	2.748	0.806	10.70 (↓**1.88**)	2.942	0.901	13.10 (↓**1.88**)	3.198	0.916	15.80 (↓**1.50**)	3.382	0.944
w/o 2nd-order dynamics	9.30 (↓**0.97**)	2.857	0.819	11.65 (↓**0.93**)	3.027	0.913	14.10 (↓**0.88**)	3.263	0.922	16.50 (↓**0.80**)	3.425	0.949
w/o adaptive dynamics weighting	9.55 (↓**0.72**)	2.883	0.825	11.88 (↓**0.70**)	3.054	0.914	14.32 (↓**0.66**)	3.276	0.927	16.70 (↓**0.60**)	3.447	0.951
w/o temporal self-attention	9.10 (↓**1.17**)	2.833	0.814	11.50 (↓**1.08**)	2.998	0.909	14.00 (↓**0.98**)	3.242	0.921	16.45 (↓**0.85**)	3.408	0.947
w/o G_v gate	9.75 (↓**0.52**)	2.907	0.827	12.10 (↓**0.48**)	3.072	0.916	14.60 (↓**0.38**)	3.303	0.925	17.05 (↓**0.25**)	3.468	0.950
APFM												
fixed frequency partition	9.12 (↓**1.15**)	2.838	0.817	11.52 (↓**1.06**)	3.013	0.908	14.05 (↓**0.93**)	3.247	0.923	16.55 (↓**0.75**)	3.422	0.946
global C_v governance	9.40 (↓**0.87**)	2.873	0.820	11.82 (↓**0.76**)	3.036	0.914	14.38 (↓**0.60**)	3.273	0.924	16.85 (↓**0.45**)	3.443	0.950
w/o Δ/B modulation	8.89 (↓**1.38**)	2.842	0.819	11.18 (↓**1.40**)	2.997	0.904	14.02 (↓**0.96**)	3.223	0.919	16.63 (↓**0.67**)	3.377	0.945
PFCL												
w/o cross-frequency interaction	9.72 (↓**0.55**)	2.886	0.822	12.06 (↓**0.52**)	3.052	0.915	14.53 (↓**0.45**)	3.282	0.926	16.95 (↓**0.35**)	3.469	0.950
w/o prior anchor	9.82 (↓**0.45**)	2.902	0.825	12.16 (↓**0.42**)	3.055	0.920	14.63 (↓**0.35**)	3.299	0.927	17.02 (↓**0.28**)	3.473	0.953
w/o consistency weighting	9.97 (↓**0.30**)	2.913	0.829	12.30 (↓**0.28**)	3.071	0.918	14.76 (↓**0.22**)	3.308	0.930	17.12 (↓**0.18**)	3.488	0.951

A.5.2 Internal Mechanism Ablations

Table 8 further examines the internal mechanisms of VSPL, APFM, and PFCL. All variants are trained under the same settings as the full model, and SDR degradations relative to FSSCM are shown in parentheses.

Internal Ablation Configurations. Each internal ablation modifies only the targeted mechanism while keeping the remaining architecture and training settings unchanged.

For VSPL, *w/o articulatory dynamics* removes $d_t^{(1)}$ and $d_t^{(2)}$ and their multi-order fusion; *w/o 2nd-order dynamics* removes only $d_t^{(2)}$; *w/o adaptive dynamics weighting* replaces the learned $\boldsymbol{\omega}_t$ with equal weighting; *w/o temporal self-attention* sets $\widetilde{U} = U$; and *w/o G_v gate* sets $G_v = \mathbf{1}$. All remaining VSPL outputs and downstream modules are kept unchanged.

For APFM, *fixed frequency partition* replaces the adaptive mask α with a fixed low-/high-frequency split; *global C_v governance* replaces $C_{v,t}$ with $c_v^g = \mathrm{Pool}_t(C_v)$ for transition governance; and *w/o Δ/B modulation* uses the original acoustic-dependent $\Delta_{b,t}$ and $B_{b,t}$ without visual transition modulation. All other APFM mechanisms are retained.

For PFCL, *w/o cross-frequency interaction* removes only the bidirectional $L \leftrightarrow H$ cross-attention; *w/o prior anchor* removes C_v from consistency estimation and prior-conditioned gating while retaining the acoustic consistency framework; and *w/o consistency weighting* replaces the learned weights with $w_L = w_H = 0.5$. All other PFCL operations remain unchanged.

VSPL Internal Design. Removing articulatory dynamics causes the largest degradation within VSPL, reducing SDR by 1.82 dB at -15 dB and 1.50 dB at 0 dB, demonstrating that explicitly modeling temporal articulatory evolution is important for constructing the visual articulatory prior. Removing only the second-order dynamics still reduces SDR by 0.97 and 0.80 dB, respectively, indicating that motion variation provides complementary information beyond first-order articulatory changes. Replacing the adaptive dynamics weighting with non-adaptive fusion leads to corresponding drops of 0.72 and 0.60 dB, confirming the benefit of dynamically balancing visual state and multi-order motion information. Removing temporal self-attention further decreases SDR by 1.17 and 0.85 dB, showing that explicit local articulatory dynamics and longer-range temporal modeling provide complementary cues. Finally, removing the G_v gate produces smaller but consistent degradations of 0.52 and 0.25 dB, supporting its complementary role in visual input-amplitude modulation.

APFM Internal Mechanisms. Replacing the adaptive frequency partition with a fixed partition reduces SDR by 1.15 dB at -15 dB and 0.75 dB at 0 dB, demonstrating the benefit of sample-adaptive frequency organization. Replacing temporally resolved $C_{v,t}$ with a globally pooled visual representation for transition governance causes corresponding drops of 0.87 and 0.45 dB, indicating that time-varying articulatory information provides more effective recovery-state guidance than globally pooled visual governance. Removing Δ/B modulation causes particularly substantial degradation, with SDR drops of 1.38 dB at -15 dB and 0.67 dB at 0 dB, providing direct evidence that modulating

Table 9: Robustness of FSSCM to complete visual failure on LRS3+DNS. Values in parentheses denote the SDR degradation relative to the full model.

Setting	SDR (dB) ↑			
	-15 dB	-10 dB	-5 dB	0 dB
FSSCM (Full)	**10.27**	**12.58**	**14.98**	**17.30**
Black-frame visual input	8.21 (↓2.06)	10.55 (↓2.03)	13.15 (↓1.83)	15.82 (↓1.48)
w/o VSPL (audio-only)	8.12 (↓2.15)	10.41 (↓2.17)	13.02 (↓1.96)	15.68 (↓1.62)

the state-transition parameters is a key mechanism underlying cause-governed recovery. Together, these results validate the complementary roles of adaptive frequency routing, temporally resolved visual governance, and transition-level modulation within APFM.

PFCL Internal Design. Removing cross-frequency interaction reduces SDR by 0.55 dB at -15 dB and 0.35 dB at 0 dB, confirming the benefit of exchanging complementary information between the evolved low- and high-frequency representations. Removing the visual prior anchor while retaining cross-frequency interaction results in corresponding drops of 0.45 and 0.28 dB, demonstrating that evaluating cross-frequency consistency relative to the shared visual articulatory prior provides additional benefit beyond acoustic cross-frequency interaction alone. Finally, replacing the learned consistency weighting with uniform integration reduces SDR by 0.30 and 0.18 dB, respectively, supporting the use of consistency-aware adaptive integration. These results show that PFCL benefits from the combined effects of cross-frequency interaction, prior-anchored consistency estimation, and adaptive consistency weighting.

A.5.3 Qualitative Visualization.

Figure 6 presents spectrograms of three randomly selected LRS3 test utterances at -15 dB SNR. From top to bottom, we show the noisy input, FSSCM-enhanced speech, and clean reference. Under severe acoustic corruption, the enhanced outputs exhibit clearer harmonic structures and reduced background interference compared with the noisy inputs, showing spectral patterns closer to the clean references. These examples qualitatively complement the quantitative improvements reported in the main experiments.

A.5.4 Robustness to Missing Visual Input.

To examine model behavior when visual information is unavailable, we replace all visual frames with black frames during inference. As shown in Table 9, the resulting performance remains close to the audio-only variant (w/o VSPL) across all evaluated SNR levels, with the SDR difference remaining below 0.15 dB. This result indicates that removing usable visual information causes FSSCM to fall back toward audio-only recovery rather than introducing substantial additional degradation.

A.5.5 Frequency Recovery Organization and Directionality.

To isolate the effect of frequency recovery organization, we compare Unified Recovery, parallel low- and high-frequency recovery ($L \parallel H$), the reverse dependency ($H \rightarrow L$), and the proposed $L \rightarrow H$ organization while keeping the remaining architecture and training settings unchanged. Except for Unified Recovery, all variants use the same frequency partitioning mechanism and frequency-specific recovery branches as FSSCM, and differ only in how the low- and high-frequency branches interact. Unified Recovery processes the full-band representation through a single shared recovery path without explicit low-/high-frequency separation. In $L \parallel H$, the two branches are recovered independently; in $H \rightarrow L$, the high-frequency recovery informs subsequent low-frequency recovery; and in $L \rightarrow H$, the evolving low-frequency recovery guides high-frequency detail reconstruction. As shown in Table 11, separating low- and high-frequency recovery already improves over Unified Recovery, while $L \rightarrow H$ consistently performs best across all evaluated SNR levels and metrics. Compared with $L \parallel H$, it improves SDR by 0.54, 0.51, 0.55, and 0.42 dB at -15, -10, -5, and 0 dB, respectively, while also consistently outperforming $H \rightarrow L$. These results support the proposed directional content-to-detail recovery dependency.

Table 10: Comparison of visual intervention roles under varying SNR conditions. All variants share the same backbone and differ only in the primary intervention of C_v.

SNR	Feature Fusion	Feature Modulation	**Transition Governance (FSSCM)**
	SDR / PESQ / STOI	SDR / PESQ / STOI	SDR / PESQ / STOI
-15 dB	8.23 / 2.684 / 0.788	8.87 / 2.793 / 0.803	**10.27 / 2.949 / 0.830**
-10 dB	10.57 / 2.856 / 0.842	11.34 / 2.958 / 0.849	**12.58 / 3.098 / 0.921**
-5 dB	13.24 / 3.124 / 0.876	13.77 / 3.198 / 0.882	**14.98 / 3.326 / 0.930**
0 dB	15.76 / 3.304 / 0.933	16.28 / 3.382 / 0.938	**17.30 / 3.496 / 0.954**

Table 11: Comparison of frequency recovery organizations under varying SNR conditions.

Recovery Organization	-15 dB	-10 dB	-5 dB	0 dB
	SDR / PESQ / STOI	SDR / PESQ / STOI	SDR / PESQ / STOI	SDR / PESQ / STOI
Unified Recovery	8.46 / 2.776 / 0.808	10.83 / 2.927 / 0.856	13.18 / 3.184 / 0.887	15.72 / 3.356 / 0.940
$L \parallel H$	9.73 / 2.908 / 0.821	12.07 / 3.046 / 0.868	14.43 / 3.276 / 0.900	16.88 / 3.452 / 0.946
$H \rightarrow L$	8.87 / 2.845 / 0.813	11.24 / 2.986 / 0.856	13.67 / 3.223 / 0.893	16.23 / 3.394 / 0.939
$L \rightarrow H$ **(FSSCM)**	**10.27 / 2.949 / 0.830**	**12.58 / 3.098 / 0.921**	**14.98 / 3.326 / 0.930**	**17.30 / 3.496 / 0.954**

A.5.6 Where Should Visual Information Intervene in Recovery?

To isolate the effect of visual intervention, we compare Feature Fusion, Feature Modulation, and the proposed Transition Governance while keeping the visual encoder, acoustic recovery backbone, frequency routing, decoder, and training settings unchanged. All variants use the same visual representation and differ only in how it intervenes in acoustic recovery. Feature Fusion introduces visual information directly into the acoustic feature representation, Feature Modulation uses visual information to modulate acoustic features, whereas Transition Governance employs the visual articulatory prior to modulate state-transition dynamics. As shown in Table 10, Transition Governance consistently outperforms both feature-level alternatives across all evaluated SNR conditions and metrics. Its SDR gain over Feature Fusion increases from 1.54 dB at 0 dB to 2.04 dB at -15 dB, while the gain over Feature Modulation increases from 1.02 to 1.40 dB. Consistent improvements are also observed in PESQ and STOI. These results support the proposed cause-governed recovery formulation.

A.6 Training Objectives

FSSCM is trained end-to-end with five reconstruction loss components following MP-SENet. Let s denote the clean speech waveform, $|S|$ the magnitude spectrum, $\angle S$ the phase spectrum, and S_r, S_i the real and imaginary parts of the complex spectrum. Hatted variables denote the corresponding model outputs.

The magnitude loss and complex loss constrain the accuracy of spectral estimation:

$$\mathcal{L}_{\text{mag}} = \||S| - |\hat{S}|\|_2^2, \quad \mathcal{L}_{\text{ri}} = \|S_r - \hat{S}_r\|_2^2 + \|S_i - \hat{S}_i\|_2^2. \tag{24}$$

The phase loss consists of instantaneous phase, group delay, and instantaneous angular frequency terms, following MP-SENet:

$$\mathcal{L}_{\text{pha}} = \mathcal{L}_{\text{ip}} + \mathcal{L}_{\text{gd}} + \mathcal{L}_{\text{iaf}}. \tag{25}$$

The time-domain and consistency losses are defined as

$$\mathcal{L}_{\text{time}} = \|s - \hat{s}\|_1, \quad \mathcal{L}_{\text{con}} = \|\hat{S}_c - \text{STFT}(\hat{s})\|_2^2. \tag{26}$$

The overall objective is

$$\mathcal{L} = \lambda_{\text{mag}}\mathcal{L}_{\text{mag}} + \lambda_{\text{ri}}\mathcal{L}_{\text{ri}} + \lambda_{\text{pha}}\mathcal{L}_{\text{pha}} + \lambda_{\text{time}}\mathcal{L}_{\text{time}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{27}$$

where $\lambda_{\text{mag}} = 0.9$, $\lambda_{\text{ri}} = 0.1$, $\lambda_{\text{pha}} = 0.3$, $\lambda_{\text{time}} = 0.2$, and $\lambda_{\text{con}} = 0.1$.