VDBC-MAMBA-2: VISUAL-DELTA BAND CONDITIONING FOR CAUSAL AUDIO-VISUAL SPEECH ENHANCEMENT

Boyu Hou

University of Bristol, United Kingdom
vx21242@bristol.ac.uk

ABSTRACT

In real-time audio-visual speech enhancement (AVSE), the speaker's visual information is used as a complementary signal under causal constraints. Feature-wise linear modulation (FiLM), used for visual conditioning, does not guarantee content-driven responses over mere stream presence. Visual-Delta Band Conditioning (VDBC) is proposed: a shared MLP is run on both the real visual embedding and a zero vector for each frequency band, and the two outputs are subtracted, forcing the visual delta to zero when the visual input is zero. VDBC is combined with a causal BlazeNet64 front-end and six causal, frequency-bidirectional Mamba-2 modules to form VDBC-Mamba-2. On LRS2 with DEMAND noise, VDBC-Mamba-2 achieves PESQ 2.022, STOI 0.890, and SI-SDR 11.07 dB with 4.293 M parameters, vs. 8.873 M for CausalRAVEN. Under all-black input, degradation forms a strict gradient: VDBC (-0.09 dB), FiLM (-0.39 dB), BandFiLM (-17.00 dB), and CausalRAVEN (-65.7 dB); the BandFiLM ablation isolates the delta subtraction as the robustness factor.

***Index Terms*—** audio-visual speech enhancement, Mamba-2, causal modelling, visual condition modulation, robustness

1. INTRODUCTION

1.1. Background and problem statement

Target speech in a deployed system is degraded by environmental noise, reverberation, and competing speakers. Audio-visual speech enhancement (AVSE) responds to this condition by incorporating face, lip, or scene visual information as a complementary cue for the separation and recovery of the target signal. This cross-modal information is of particular value at low signal-to-noise ratios and in multi-speaker conditions.

Real-time deployment imposes a further requirement: strict temporal causality. A zero look-ahead definition is adopted in this paper, under which the output at time t depends only on audio and visual inputs at or before t. State-space propagation in the time direction and the visual front-end are therefore constrained from accessing future frames, although bidirectional modelling along the frequency axis remains permitted within the same time frame. Systems described elsewhere as online or causal may still admit a small look-ahead window and are not always equivalent to the definition adopted here.

A second question addressed in this paper is whether the visual condition is genuinely driven by visual content. Standard FiLM generates the scaling coefficient γ and bias β directly from the visual vector. Even when training produces an effective cross-modal mapping, there is no structural mechanism that prevents the model from exploiting frequency-band identity, visual-branch bias, or even the weak cue that a video stream is present as a shortcut. When the visual input becomes anomalous, the fusion module may then produce large-scale and systematic erroneous modulation.

1.2. Related work

Mamba-based state-space models offer linear-cost long-sequence modelling: SEMamba applied Mamba to speech enhancement [1], AVSEMamba added full-face visual information [2], and SAV-SE used scene video for noise-category cues [3]. Online AV-CrossNet reduced overhead via causal layers and compression [4], a related autoregressive method used a lightweight visual front-end [5], and RAVEN used pre-trained AVSR/ASD visual representations with a CPU streaming implementation [6]. None of these systems addresses whether the fusion module falls back interpretably when visual input degrades or goes missing.

1.3. Contributions

The explicit encoding of content dependency into the structure of visual condition modulation is taken as the starting point of this work. VDBC is proposed and validated on a causal Mamba-2 backbone.

Four contributions are made: the VDBC band-level modulation mechanism, which computes the real and zero-visual MLP outputs separately per frequency unit and takes their difference; a compact causal VDBC-Mamba-2 system combining a causal BlazeNet64 front-end with six causal, frequency-bidirectional Mamba-2 modules; a forced identity-modulation experiment showing a 10.55 dB average SI-SDR drop when visual modulation is disabled; and anomalous-visual-input stress tests across four conditioning mechanisms revealing a strict robustness gradient, with the BandFiLM ablation separating the delta operation from band-level sharing. Negative and unresolved findings, including insufficient training epochs, the absence of a shared benchmark, insensitivity to visual frame order, and abnormally large first-layer modulation, are also reported.

2. VDBC-MAMBA-2

2.1. Problem definition

Let the noisy speech observation at time t be y, the target clean speech be s, and the visual sequence be v. Causality, as defined in Sec. 1, constrains the prediction at time t to audio and visual information at or before t. Both the state-space path and the visual encoder are held to this constraint.

Short-time Fourier analysis converts the noisy waveform to a complex spectral representation. After amplitude compression, the result is fed into a densely connected two-dimensional convolutional

encoder. The encoded time-frequency features are then passed sequentially through six causal TF-Mamba-2 modules. VDBC conditioning is applied after each of the six modules, so that visual information is distributed across the full depth of the network rather than fused at a single point. Enhanced amplitude and phase are estimated separately by an amplitude decoding head D_M and a phase decoding head D_Φ, such that $\hat{M} = M_y \cdot D_M(X)$ and $\hat{\Phi} = D_\Phi(X)$. The estimated complex spectrum is given by $\hat{S} = \hat{M} \odot (\cos\hat{\Phi} + j\sin\hat{\Phi})$, where M_y is the noisy amplitude spectrum and X is the backbone time-frequency feature. The time-domain waveform is recovered via the inverse STFT (iSTFT).

2.2. Causal BlazeNet64 visual front-end

The visual branch receives a lip-region sequence sampled at 25 fps. Each frame is first converted to greyscale according to ITU-R BT.601 and then passed through the causal BlazeNet64 visual encoder [5]. The temporal receptive field of this encoder covers only the current and past frames. A 64-dimensional visual embedding $V_{\text{raw}}(t) \in \mathbb{R}^{64}$ is produced at each time step. Feature projection is completed by a 1×1 causal convolution, and the visual frame rate is then upsampled to match the STFT frame rate by nearest-neighbour interpolation, avoiding the introduction of future information.

A distinction between two input conditions must be maintained. At the pixel level, a fully black video frame is not equivalent to a zero vector at the VDBC module input, since both the visual encoder and the projection layer contain learnable bias parameters. A fully black image therefore typically produces non-zero, channel-varying embeddings. The zero-point property of VDBC applies in the input space of the conditioning module, not in the original pixel space (Sec. 3.4).

2.3. Visual-delta band conditioning

Let $f \in \{1,\dots,F\}$ index frequency units, with $F=100$. The STFT with $N_{\text{fft}}=400$ produces 201 raw frequency bins; the dense encoder's stride-2 frequency-axis convolution reduces this to the $F=100$ working resolution used throughout the network. A 64-dimensional learnable band embedding E_f is associated with each unit. A two-layer MLP F_v, shared across all frequency units, receives the concatenation $[E_f, V(t)]$ and produces intermediate conditioning features. Rather than generating FiLM parameters directly from this concatenation, VDBC computes two forward passes of F_v—one with the real visual input and one with a zero input—and takes the difference:

$$h_f(t) = F_v([E_f, V(t)]) - F_v([E_f, \mathbf{0}]). \tag{1}$$

Because both terms in Eq. (1) share all parameters exactly, $h_f(t) = \mathbf{0}$ when $V(t) = \mathbf{0}$, by construction and independently of training. The subtraction removes the baseline response produced by E_f and the shared network's inherent biases.

Two linear projections, each carrying its own trainable bias term b_γ and b_β (initialised to 1 and 0 respectively), then produce the per-band scaling and bias:

$$\gamma_f(t) = W_\gamma h_f(t) + b_\gamma, \quad \beta_f(t) = W_\beta h_f(t) + b_\beta \tag{2}$$

$$\tilde{X}(t,f) = \gamma_f(t) \cdot X(t,f) + \beta_f(t) \tag{3}$$

Because $h_f(t) = \mathbf{0}$ whenever $V(t) = \mathbf{0}$ (Eq. (1)), $\gamma_f(t)$ and $\beta_f(t)$ reduce to b_γ and b_β at that point; Eq. (3) reduces to the identity $\tilde{X}(t,f) = X(t,f)$ only while those bias terms remain at their initialised values, as detailed in Sec. 2.4. VDBC is applied once after each of the six Mamba-2 modules.

Two ablations separate the roles of the delta operation and band-level parameter sharing. BandFiLM retains the same band embedding, shared MLP, architecture, and initialisation as VDBC, but removes the zero-input subtraction of Eq. (1) and sets $h_f(t) = F_v([E_f, V(t)])$. The FiLM ablation removes the band-embedding structure entirely and generates standard FiLM parameters directly from $V(t)$. VDBC's construction differs from simply learning a per-band bias: the reference point is anchored at the conditioning module's own input-space zero, which is fixed by construction, rather than at a value learned freely during training.

2.4. Zero-input property and trained-bias drift

The linear heads W_γ and W_β each retain a trainable bias term, initialised to 1 and 0 respectively. When $V(t) = \mathbf{0}$, the differential latent variable $h_f(t)$ is exactly zero by the algebraic construction of Eq. (1), independently of training. $\gamma_f(t)$ and $\beta_f(t)$ then take the current trained value of those bias terms, not their initialisation value; in the reported checkpoint, the bias terms of one VDBC layer have drifted to $\gamma_f=1.10$ and $\beta_f=-0.03$ after training. The exact-zero property of the delta term is therefore unconditional, but the full conditioning layer reduces to an identity mapping only while the bias terms remain at their initialised values.

All-black pixels passed through the complete visual front-end instead produce a non-zero embedding, since the visual encoder and the projection layer carry learnable bias parameters. Because $V(t)$ is not zero under all-black input, the delta term $h_f(t)$ is also not zero, and the full modulation path remains active; one layer shows a mean γ of approximately 2.96 under this condition. Visual dropout augmentation was not used during training, so all-black frames lie outside the training distribution, and the following sections treat this condition as a structural stress test rather than as a model of natural camera-failure degradation.

2.5. Training objective

The model is trained with a three-component joint loss covering amplitude, phase, and complex spectrum:

$$\mathcal{L} = \lambda_{\text{mag}}\|\hat{M} - M\|_1 + \lambda_{\text{pha}}(\mathcal{L}_{\text{IP}} + \mathcal{L}_{\text{GD}} + \mathcal{L}_{\text{IAF}}) + \lambda_{\text{com}}\|\hat{M}e^{j\hat{\Phi}} - Me^{j\Phi}\|_1 \tag{4}$$

The weights are set to $\lambda_{\text{mag}}=0.9$, $\lambda_{\text{pha}}=0.3$, and $\lambda_{\text{com}}=0.1$; $\mathcal{L}_{\text{IP}}$, $\mathcal{L}_{\text{GD}}$, and $\mathcal{L}_{\text{IAF}}$ are the phase-loss terms as defined in [1]. No auxiliary losses are introduced.

3. EXPERIMENTS AND RESULTS

3.1. Setup

The acoustic front-end uses STFT with $N_{\text{fft}}=400$, hop=100, window length=400, sampling rate $f_s=16$ kHz, and amplitude compression exponent 0.3. The backbone hidden dimension is $C=96$, with expand=2, head dimension=32, chunk size=128, and state dimension $d_{\text{state}}=64$. The speech data follow the official LRS2 partitions [7]: pretrain+main-train pooled for training (142,157 segments), 400 segments from the official validation split, and the full official test set (1,243 segments). Noise is drawn from DEMAND [8]; 14 of 17 environments train, and DLIVING, PCAFETER, TMETRO are held out entirely for validation/test. Training SNR is drawn from $\mathcal{U}[-5,20]$ dB. Test metrics are averaged over $\{-5,0,5,10,15\}$ dB.

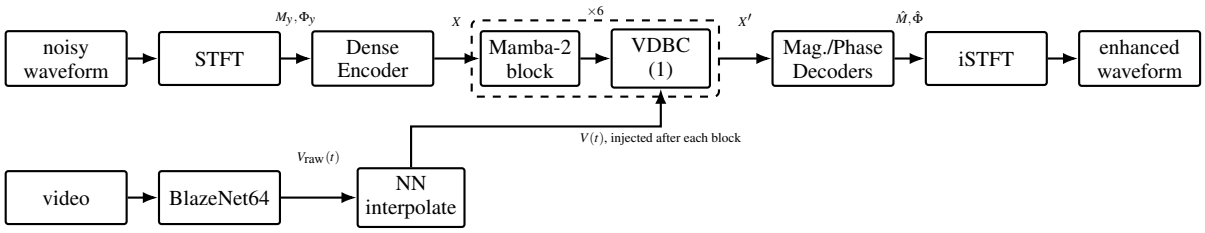

Fig. 1. VDBC-Mamba-2 causal pipeline. The audio branch runs through six Mamba-2 blocks; the visual embedding is injected via VDBC conditioning (Eq. (1)) after *every* block, not once at input or output. All arrows are causal (time $\leq t$ only). VDBC's internal delta computation is detailed separately in Fig. 2.

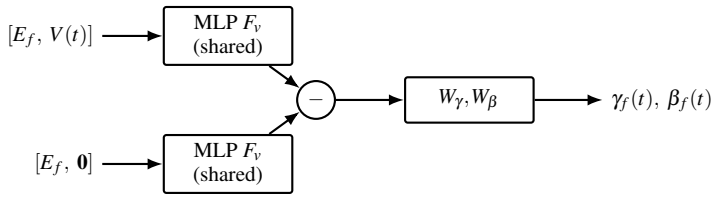

Fig. 2. VDBC's delta computation (Eq. (1)), one bin f. The shared MLP F_v runs twice, on the real visual embedding and on the zero vector; only the *difference* drives the FiLM heads. At $V(t)$=**0** both branches match, so the delta itself is exactly zero, independent of training; γ_f, β_f then equal the heads' trained bias values, not necessarily $1, 0$ (Sec. 2.4).

The four systems compared are VDBC-Mamba-2, standard FiLM, BandFiLM, and CausalRAVEN. CausalRAVEN follows the RAVEN architecture [6] with the bidirectional LSTM replaced by a unidirectional LSTM, sharing the visual front-end, audio encoder, and decoder with VDBC-Mamba-2; only its fusion mechanism (unconstrained feature concatenation) differs. CausalRAVEN therefore uses the same BlazeNet64 visual embeddings as VDBC-Mamba-2 rather than RAVEN's original pre-trained audio-visual-speech-recognition and active-speaker-detection representations, so these results characterise the fusion mechanism in isolation rather than the original RAVEN system. All four systems train for three epochs (53,307 steps, batch 8, AdamW, peak LR 5×10^{-4}, cosine warm-up); the best-validation-SI-SDR checkpoint is used for evaluation. VDBC, FiLM, and BandFiLM peak at epoch 0 and then decline, whereas CausalRAVEN improves monotonically, so absolute numbers are read as relative comparisons within this budget. All four systems receive bit-for-bit identical mini-batches (verified via per-step data-content checksums), and validation/test use a fixed, per-utterance crop, noise, and SNR configuration (verified via input hashing).

Four visual conditions are evaluated: *target* (video aligned with the speech), *zero* (all-black pixels), *scrambled* (randomly permuted frames, same appearance distribution, disrupted temporal order), and *reversed* (chronologically flipped frames). The zero condition refers to black frames at the pixel level, not to a zero vector presented directly to the VDBC module.

3.2. Parameter count and target-condition performance

Table 1 lists the parameter counts of the four systems. CausalRAVEN is approximately 2.07× larger than VDBC-Mamba-2, equivalently 51.6% fewer parameters for VDBC-Mamba-2. The FiLM ablation increases slightly to 4.359 M owing to its modulation-head structure; BandFiLM matches VDBC-Mamba-2 exactly.

Table 2 reports performance under target conditions. CausalRAVEN, VDBC-Mamba-2, and standard FiLM differ by at most 0.04 dB SI-SDR, and BandFiLM trails by approximately 0.20 dB. VDBC-Mamba-2 therefore matches the larger LSTM-based baseline at roughly half the parameter count under this budget.

Table 1. Parameter counts.

Model	Parameters
VDBC-Mamba-2	4.293 M
FiLM (ablation)	4.359 M
BandFiLM (ablation)	4.293 M
CausalRAVEN	8.873 M

Table 2. Test-set results, target condition (mean over five SNRs; best-validation-SI-SDR checkpoint).

Model	PESQ	STOI	SI-SDR	SI-SDRi
CausalRAVEN	2.050	0.889	11.08	6.07
VDBC-Mamba-2	2.022	0.890	11.07	6.06
FiLM	2.015	0.890	11.04	6.02
BandFiLM	2.008	**0.891**	10.87	5.86

3.3. Does the visual condition genuinely contribute to enhancement?

Target-condition performance alone does not establish that the model depends on visual content. To test this, all VDBC layers in the trained VDBC-Mamba-2 are forced to γ_f=1 and β_f=0 at inference, completely bypassing visual conditioning at the module level. With forced identity modulation, the model achieves PESQ 1.170, SI-SDR 0.52 dB, and SI-SDRi −4.49 dB, compared with PESQ 2.022, SI-SDR 11.07 dB, and SI-SDRi 6.06 dB with conditioning active. The resulting 10.55 dB mean gap widens monotonically with input SNR, from 5.23 dB at −5 dB to 16.81 dB at 15 dB—the op-

Table 3. SI-SDR (dB) by visual condition.

Model	target	zero	scrambled	reversed
CausalRAVEN	11.08	−54.64	10.99	11.06
VDBC-Mamba-2	11.07	10.98	11.05	11.05
FiLM	11.04	10.64	11.03	11.03
BandFiLM	10.87	−6.12	10.90	10.87

posite of the usual expectation that visual cues help most when the audio is noisiest. Combined with the 0.09 dB all-black-frame result (Sec. 3.4), this pattern is also consistent with a different explanation: the modulation pathway itself, rather than the visual content passing through it, may be what the rest of the network has adapted to during training. The pathway is indispensable to the trained network either way, but the present experiments do not distinguish between the two explanations.

3.4. Visual content perturbation and robustness gradient

Table 3 presents SI-SDR under all four conditions. Under scrambled and reversed conditions, all four systems change by no more than 0.09 dB relative to target; this near-zero change is a null result, and no claim is made that VDBC is superior in exploiting fine-grained lip-movement dynamics. The all-black-frame condition, by contrast, clearly differentiates the four mechanisms: VDBC-Mamba-2 degrades by 0.09 dB (95% CI $[-0.12, -0.06]$), standard FiLM by 0.39 dB ($[-0.45, -0.34]$), and BandFiLM by 17.00 dB ($[-17.21, -16.78]$). CausalRAVEN drops from 11.08 dB to −54.64 dB, a mean degradation of 65.7 dB ($[-66.4, -65.0]$), with PESQ falling to approximately 1.03. Its concatenation-based fusion feeds the visual embedding into a causal LSTM at every time step with no gating or bound on its contribution, so an out-of-distribution embedding can accumulate across an utterance's full causal recurrence; this offers a plausible, architecture-grounded explanation for the magnitude of its collapse, though the hidden-state trajectory is not directly measured here. The smaller-degradation system is preferred in more than 9,999 of every 10,000 paired bootstrap resamples (10^4 resamples, 1,243 utterances × 5 SNRs), for every adjacent pair.

BandFiLM shares the greatest structural overlap with VDBC, differing only in the removal of Eq. (1)'s subtraction, yet it is more fragile under anomalous input than standard, non-shared FiLM: band-level parameter sharing alone does not reproduce VDBC's stability, plausibly because the shared MLP then produces a consistently directed error across the spectrum that the delta operation would otherwise cancel, a hypothesis not established by further experiments here. The robustness gradient across all four systems may instead, or in addition, reflect how strongly each fusion mechanism depends on the visual input's magnitude rather than on its content: all-black frames remove essentially all visual content yet cost VDBC-Mamba-2 only 0.09 dB, so the present experiments do not distinguish a content-driven mechanism from a magnitude-dependent one.

3.5. Computational efficiency

VDBC-Mamba-2 contains 4.293 M parameters, substantially fewer than the 8.873 M of CausalRAVEN. On a single NVIDIA A100-SXM4-80GB GPU, the complete pipeline (visual encoding, STFT, six paired Mamba-2 and VDBC blocks, decoding, iSTFT) was timed over 30 runs following 5 warm-up runs, with batch size 1 and 2-second speech segments. The mean elapsed time is 82.9 ms, giving a real-time factor (RTF) of 0.041. This measures offline throughput on a fixed 2-second segment rather than per-frame streaming latency, which a chunked or frame-by-frame deployment would need to report separately.

4. LIMITATIONS

Seven limitations constrain the conclusions of this study. First, all four systems train for only three epochs; absolute PESQ values of approximately 2.0 fall well short of the literature optimum, and VDBC, FiLM, and BandFiLM decline after epoch 0 for reasons that remain unexplained. Second, no unified benchmark comparison with AVSEMamba, SAV-SE, and Online AV-CrossNet is reproduced here, as those systems differ in dataset, task, or protocol, so the tables here do not support a state-of-the-art claim. Third, SI-SDR changes under permutation and time reversal do not exceed 0.09 dB, so no system is shown to have learned fine-grained lip-movement timing. Fourth, the mean deviation of γ_f from 1 in the first VDBC layer is approximately 7.79, versus 0.28–0.95 in subsequent layers; this disparity is unexplained and may warrant inter-layer regularisation. Fifth, a single random seed is used per system, so training variance is not characterised. Sixth, no audio-only model trained from scratch under the same budget is compared against the inference-time result in Sec. 3.3, which bounds the pathway's necessity rather than substituting for a dedicated baseline. Seventh, no FiLM variant trained with visual dropout is evaluated, so the all-black test remains out-of-distribution for every system here.

5. CONCLUSION

A structural question in real-time audio-visual speech enhancement is addressed here: whether the visual fusion module can bind modulation responses to visual content changes without relying on training contingency. VDBC is proposed as a response, combined with a causal BlazeNet64 visual front-end and a six-layer causal TF-Mamba-2 backbone to form VDBC-Mamba-2. By running both the real visual embedding and a zero reference through the same shared MLP and subtracting the outputs, a differential latent variable h is constructed that is exactly zero under zero visual input, providing an interpretable reference point for modulation degradation.

VDBC-Mamba-2 matches CausalRAVEN on target-condition SI-SDR (11.07 vs. 11.08 dB) at roughly half the parameter count, and forcing identity modulation costs 10.55 dB, more at high SNR than low and far more than the 0.09 dB cost of all-black visual input, indicating that the modulation pathway is indispensable to the trained network, though whether this reflects genuine use of visual content or dependence on an adapted pathway is not resolved here. Anomalous-frame degradations of 0.09, 0.39, 17.00, and 65.7 dB for VDBC, FiLM, BandFiLM, and CausalRAVEN establish a clear robustness gradient, and the BandFiLM ablation indicates that the delta operation, not band-level parameter sharing alone, is the critical factor. The more defensible conclusion is therefore not that VDBC has resolved causal AVSE, but that this differential reference point is an interpretable, parameter-efficient mechanism that performs stably under specific out-of-distribution stress tests, with its generalisability and mechanistic boundaries left to more extensive training, shared benchmarks, and realistic visual-failure scenarios.

Acknowledgements

The author thanks the University of Bristol for research resources, the BBC for releasing LRS2, and the anonymous reviewers for their constructive feedback.

Compliance with Ethical Standards

This study did not involve human participants, human data collection, or animal experiments beyond secondary use of the publicly available LRS2 dataset [7] and the DEMAND noise corpus [8], each accessed and used under its own published licence and terms.

6. REFERENCES

[1] Rong Chao, Wen-Huang Cheng, Moreno La Quatra, Sabato Marco Siniscalchi, 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)*, 2024, pp. 302–308, arXiv:2405.06573.

[2] Rong Chao, Wenze Ren, You-Jin Li, Kuo-Hsuan Hung, Sung-Feng Huang, Szu-Wei Fu, Wen-Huang Cheng, and Yu Tsao, "Leveraging mamba with full-face vision for audio-visual speech enhancement," in *Proc. Interspeech*, 2025, arXiv:2508.13624.

[3] Xinyuan Qian, Jiaran Gao, Yaodan Zhang, Qiquan Zhang, Hexin Liu, Leibny Paola Garcia, and Haizhou Li, "SAV-SE: Scene-aware audio-visual speech enhancement with selective state space model," *IEEE J. Sel. Topics Signal Process.*, 2025, arXiv:2411.07751.

[4] Cheng Yu, Vahid Ahmadi Kalkhorani, Buye Xu, and DeLiang Wang, "Online AV-CrossNet: A causal and efficient audiovisual system for speech enhancement and target speaker extraction," in *Proc. Interspeech*, 2025, pp. 2985–2989.

[5] Zexu Pan, Wupeng Wang, Shengkui Zhao, Chong Zhang, Kun Zhou, Yukun Ma, and Bin Ma, "Online audio-visual autoregressive speaker extraction," in *Proc. Interspeech*, 2025, arXiv:2506.01270.

[6] Teng Ma, Sile Yin, Li-Chia Yang, and Shuo Zhang, "Real-time audio-visual speech enhancement using pre-trained visual representations," in *Proc. Interspeech*, 2025, arXiv:2507.21448.

[7] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *arXiv:1809.02108*, 2018.

[8] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "DEMAND: A collection of multi-channel recordings of acoustic noise in diverse environments," Zenodo, `https://doi.org/10.5281/zenodo.1227121`, 2013.