BAT: LIGHTWEIGHT CAUSAL BONE-CONDUCTION-GUIDED SPEECH SEPARATION WITH CROSS-MODAL CONDITIONING AND ADAPTIVE MASK BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the recovery of wearer and interlocutor speech from a noisy air-conduction (AC) mixture and a synchronized bone-conduction (BC) reference available only from the wearer. BC provides a speaker-specific cue, but its high-frequency attenuation and locally varying reliability complicate its use for separation. We introduce BAT, a compact frame-causal framework that adapts BC conditioning throughout feature fusion and spectral reconstruction. Its Cross-Band Harmonic Reliability Fusion (CHRF) module aggregates cross-band BC context and conditions AC features according to local cross-sensor reliability. The Confidence-Adaptive Mask (CAM) decoder combines separator features with cross-sensor agreement and BC activity to adapt the allowable complex-mask ranges. We also introduce LS-Bone, English AC–BC pairs obtained by controlled re-recording of LibriSpeech, and VibroMix, which combines these pairs with Chinese ABCS speech and environmental noise. On VibroMix-2Mix, BAT achieves 17.3 dB mean SI-SNRi, exceeding the two-output Robust Fusion adaptation by 2.6 dB. Mean SDRi and PESQ improve by 2.5 dB and 0.2, respectively. Under the common causal profiling configuration, BAT uses approximately 46.2% fewer parameters and 34.6% fewer multiply–accumulate operations than the original single-target Robust Fusion model. Code and audio demos are available at `https://anonymous.4open.science/w/BAT-Demo-3603/`.

1 INTRODUCTION

Wearable speech interfaces, including earbud voice assistants and hands-free communication devices, must recover the wearer's speech amid nearby talkers and environmental noise (He et al., 2023; Heitkaemper et al., 2025). Their air-conduction (AC) microphones capture mixtures of wearer speech, interfering speech, and ambient noise. During overlapping conversation, recovering the wearer's speech requires distinguishing it from the other sources in this mixture.

Target speaker extraction (TSE) uses a target-specific reference to guide speech recovery. Enrollment-based methods such as TD-SpeakerBeam and SpEx+ derive a speaker representation from a separately recorded utterance (Delcroix et al., 2020; Ge et al., 2020). Subsequent systems strengthen enrollment–mixture interaction or update the reference during inference (Hao et al., 2025; Liu et al., 2026). Differences in recording conditions or vocal state between the enrollment and current speech can nevertheless cause reference mismatch and increase the risk of extracting an interfering speaker (Sato et al., 2022; Li et al., 2024a; Liu et al., 2026). Audio-visual methods such as CTCNet and IIANet instead use synchronized lip or facial video to guide extraction (Li et al., 2024b;c), but these cues depend on camera visibility and can be disrupted by occlusion, head motion, or a limited field of view.

Bone conduction (BC) offers a synchronized wearer-specific reference by capturing speech-induced vibrations through a contact sensor (He et al., 2023; Wang et al., 2022). Joint AC–BC systems combine this reference with the broadband acoustic mixture through learned cross-modal representations and adaptive fusion (Li et al., 2025; Liu et al., 2025; Kim & Chung, 2025; Wu et al., 2026). MBENet applies AC–BC fusion to TSE and uses BC bandwidth extension as an auxiliary training

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.

objective (Zhang et al., 2025). These studies establish BC as a useful reference for speech enhancement and target speaker extraction.

Exploiting BC requires accounting for its uneven spectral content and locally varying reliability. The body and sensor transmission path attenuate high-frequency components, concentrating BC information in the lower frequency bands (Wang et al., 2022; Li et al., 2025); Figure 5 in Appendix A.1 illustrates this contrast for a synchronized pair. Its usefulness also varies across time–frequency locations with the wearer's speech activity and the local quality of the two observations (Kim & Chung, 2025). The resulting challenge is to propagate informative low-frequency BC cues across bands while adapting their influence on both feature fusion and speech reconstruction to local reliability.

To support low-latency interaction in real-world wearable settings, we introduce BAT, a lightweight causal framework that uses local BC reliability to control both cross-modal conditioning and spectral reconstruction. Given an AC mixture and a synchronized BC reference available only from the wearer, BAT reconstructs both the wearer's and the interlocutor's speech. Its *Cross-Band Harmonic Reliability Fusion (CHRF)* module aggregates sparse cross-frequency BC context, motivated by the harmonic structure of voiced speech, and uses cross-sensor agreement and spectral-power statistics to regulate its contribution to the AC representation. A causal separator then models spectrotemporal dependencies using current and past frames. The *Confidence-Adaptive Mask (CAM) decoder* carries agreement and BC-activity cues into reconstruction, where they combine with latent features to adjust the allowable real and imaginary ranges of the complex mask at each time–frequency bin.

Publicly accessible paired English AC–BC datasets remain scarce (He et al., 2023; Hosain et al., 2024). To expand these resources, we used a Timekettle W4 headset[1] to synchronously re-record controlled acoustic and vibration-platform playback of LibriSpeech utterances (Panayotov et al., 2015), creating LS-Bone, an English paired AC–BC dataset. We then combined LS-Bone with the Chinese ABCS corpus (Wang et al., 2023b) and environmental noise to construct VibroMix, whose VibroMix-2Mix benchmark supports two-speaker mixture evaluation. On VibroMix-2Mix, BAT achieved 17.3 dB mean SI-SNRi, a 2.6 dB improvement over the two-output Robust Fusion adaptation (Liu et al., 2025), the strongest baseline by mean SI-SNRi in our causal comparison. BAT also improved mean SDRi and PESQ by 2.5 dB and 0.2, respectively. Under the common causal profiling configuration, BAT used approximately 46.2% fewer parameters and 34.6% fewer MACs than the original single-target Robust Fusion model.

Our main contributions are as follows:

- We introduce BAT, a lightweight causal framework that recovers both wearer and interlocutor speech from an AC mixture and a BC reference available only from the wearer. Its CHRF module aggregates cross-band BC context and uses local cross-sensor reliability to condition the AC representation.
- We design CAM to extend reliability-aware conditioning from feature fusion to spectral reconstruction. It combines separator features, cross-sensor agreement, and BC-activity cues to adapt real and imaginary complex-mask ranges at each time–frequency bin.
- We construct LS-Bone with English AC–BC pairs and combine it with ABCS and environmental noise to create VibroMix-2Mix for BC-guided two-speaker speech separation.

2 Related Work

2.1 Causal blind speech separation

Causal blind speech separation reconstructs individual speakers from an acoustic mixture using current and past frames, without a target-speaker reference. Retaining useful historical context at low computational cost is a central challenge in this setting. The causal configuration of SkiM propagates recurrent states across segments with unidirectional LSTMs for low-latency continuous separation (Li et al., 2022). ReSepFormer reduces transformer cost through non-overlapping chunks and compact chunk summaries; its causal variant masks attention to future time steps (Della Libera et al., 2024). TF-ACM combines cache memory with causal attention refinement to retain and

[1] https://www.timekettle.co/products/w4-ai-interpreter-earbuds

exploit historical information in the time–frequency domain (Chen et al., 2025). Identifying the wearer's stream remains necessary, motivating a synchronized wearer-specific reference.

2.2 Target-guided multimodal speech separation

Enrollment-based TSE methods such as DSINet use a prerecorded speaker reference to guide extraction (Hao et al., 2025), but differences between enrollment and deployment conditions can cause reference mismatch (Li et al., 2024a). Audio-visual methods such as CTCNet and IIANet use synchronized lip or facial video through hierarchical fusion and intra- and inter-modality attention (Li et al., 2024b;c); in wearable applications, occlusion, head motion, and a limited camera field of view can make these visual cues incomplete or unavailable. VibVoice, bone-conducted multimodal enhancement, and bone-conduction-guided earbud enhancement show that BC can provide a synchronized cue for wearer speech (He et al., 2023; Wang et al., 2022; Heitkaemper et al., 2025). For AC–BC fusion, BiNet (Li et al., 2025) combines feature concatenation with contrastive learning, while BAF-Net (Kim & Chung, 2025) and Robust Fusion (Liu et al., 2025) use mask-driven fusion and adaptive time–frequency attention, respectively. FlowDense (Wu et al., 2025) organizes enhancement within a subband framework, while DBMIF (Wu et al., 2026) uses iterative cross-modal fusion. MBENet (Zhang et al., 2025) applies AC–BC fusion to target speaker extraction with BC bandwidth extension as an auxiliary training objective. BAT uses cross-band BC context and local reliability to guide fusion and spectral reconstruction in causal BC-guided speech separation.

3 Method

3.1 Task definition

In real-world wearable conversation systems, device wearing patterns depend on the application; in some scenarios, only one participant wears a device such as smart glasses or a head-mounted communication device (Žmolíková et al., 2024; Pang et al., 2024). The device can then acquire a synchronized BC observation only from the wearer, while the interlocutor may not wear a compatible device and therefore has no corresponding BC reference. The AC microphone simultaneously captures the wearer, the interlocutor, and environmental noise, and the two speakers may overlap in conversation (He et al., 2023). The system must therefore use the BC observation to identify the wearer's speech while retaining the interlocutor's audio. Accordingly, we study bone-conduction-guided speech separation from an AC mixture and a synchronized BC reference available only from the wearer, with the two recovered speech signals kept in a fixed wearer–interlocutor order.

3.2 Overall pipeline

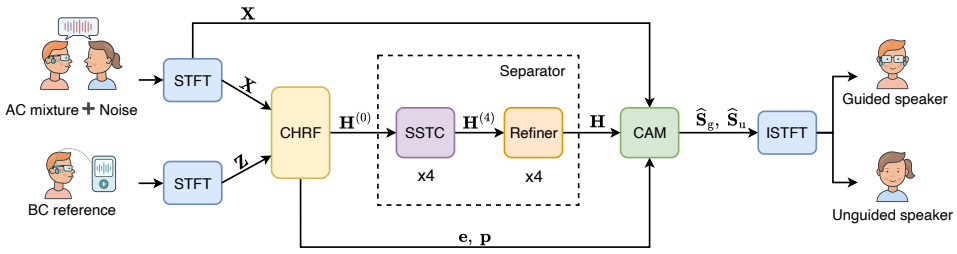


Figure 1: The overall pipeline of BAT.

BAT comprises four components: a shared audio encoder, the CHRF module, a separator, and a CAM decoder, as illustrated in Figure 1. The inputs are an AC mixture $\boldsymbol{x} \in \mathbb{R}^L$ and a synchronized BC recording $\boldsymbol{b} \in \mathbb{R}^L$ from the wearer, where L is the number of waveform samples. The shared audio encoder applies the same fixed STFT $\mathcal{F}$ (Appendix A.2) to both signals, yielding spectra $\boldsymbol{X}, \boldsymbol{Z} \in \mathbb{C}^{T\times F}$ via $\boldsymbol{X} = \mathcal{F}(\boldsymbol{x})$ and $\boldsymbol{Z} = \mathcal{F}(\boldsymbol{b})$, where T and F are the numbers of frames and frequency bins. The spectra $\boldsymbol{X}$ and $\boldsymbol{Z}$ are retained for cross-sensor statistics and final reconstruction.

Given the AC and BC spectra $\boldsymbol{X}$ and $\boldsymbol{Z}$, CHRF produces a BC-conditioned AC representation $\boldsymbol{H}^{(0)} \in \mathbb{R}^{C\times T\times F}$ for the separator, where C is the number of feature channels. It also provides

an agreement map $\boldsymbol{e} \in \mathbb{R}^{T\times F}$ and a BC-activity cue $\boldsymbol{p} \in \mathbb{R}^{T}$ to the CAM decoder. The separator receives $\boldsymbol{H}^{(0)}$ and models its temporal and frequency dependencies through stacked Selective Spectrotemporal Contextualizer (SSTC) modules. A Refiner then performs local feature refinement before producing the separator representation $\boldsymbol{H} \in \mathbb{R}^{C\times T\times F}$. CAM uses $\boldsymbol{H}$ together with $\boldsymbol{e}$ and $\boldsymbol{p}$ from CHRF to predict real and imaginary margin maps $\boldsymbol{m}^{\mathrm{R}}, \boldsymbol{m}^{\mathrm{I}} \in \mathbb{R}^{T\times F}$ that define the allowable complex-mask ranges. Source-specific heads then predict mask values within these ranges. Finally, the predicted masks are applied to the AC mixture spectrum $\boldsymbol{X}$, and the inverse STFT $\mathcal{F}^{-1}$ maps the masked spectra to the guided and unguided speech waveforms $\widehat{\boldsymbol{s}}_{\mathrm{g}}, \widehat{\boldsymbol{s}}_{\mathrm{u}} \in \mathbb{R}^{L}$, respectively.

3.3 Cross-band harmonic reliability fusion module

As illustrated in Figure 2, the CHRF module conditions AC features on cross-band BC context, using local reliability to control the modulation strength. A shared causal convolutional encoder maps the stacked real and imaginary components of $\boldsymbol{X}$ and $\boldsymbol{Z}$ to features $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{C\times T\times F}$, respectively.

Cross-sensor similarity (CosSim). AC and BC follow different propagation paths, so their feature magnitudes can differ. CosSim compares feature directions through cosine similarity, reducing the influence of feature-norm differences. At time index t and frequency index f, let $\boldsymbol{A}_{t,f}, \boldsymbol{B}_{t,f} \in \mathbb{R}^{C}$ denote the channel vectors. CosSim computes

$$e_{t,f} = \frac{\langle \boldsymbol{A}_{t,f}, \boldsymbol{B}_{t,f} \rangle}{\max(\|\boldsymbol{A}_{t,f}\|_2, \epsilon_e)\max(\|\boldsymbol{B}_{t,f}\|_2, \epsilon_e)}. \tag{1}$$

Here, $\epsilon_e > 0$ prevents division by zero. The resulting $\boldsymbol{e}$ describes cross-sensor feature agreement and is passed to both the BC Activity–Reliability Estimator (BC-ARE) and CAM. Numerical settings are provided in Appendix A.3.

Cross-band frequency context. BC high-frequency attenuation makes its information unevenly distributed across frequency (Wang et al., 2022). Restricting fusion to nearby frequency bins limits how its low-frequency cues can inform other bands. Harmonic correlations in speech can span nonlocal frequency regions (Yin et al., 2020). F-DConv aggregates this context with two parallel depthwise convolutions at different frequency dilations, yielding $\boldsymbol{C} \in \mathbb{R}^{C\times T\times F}$:

$$\boldsymbol{C} = \mathcal{N}(\boldsymbol{B} + \mathcal{D}_8(\boldsymbol{B}) + \mathcal{D}_{32}(\boldsymbol{B})). \tag{2}$$

Here, $\mathcal{D}_8$ and $\mathcal{D}_{32}$ use 1×5 kernels with frequency dilation rates of 8 and 32, respectively, and $\mathcal{N}$ denotes channel-wise layer normalization (Ba et al., 2016). These sparse connections reach separated frequency bins with few parameters, while the residual term retains the original BC features. All operations remain within the current frame, introducing no future-frame dependence.

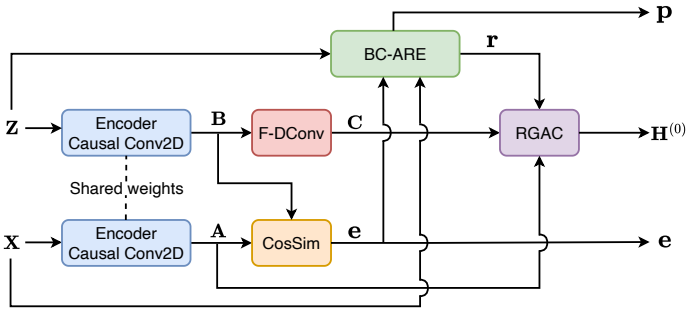


Figure 2: The structure of the Cross-Band Harmonic Reliability Fusion (CHRF) module.

BC activity and reliability estimation (BC-ARE). Cross-band aggregation extends BC context, but modulation strength should also reflect local conditions. BC-ARE receives $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{e}$ and computes a frame-level low-frequency cue p_t and a time–frequency reliability gate $\boldsymbol{r} \in (0,1)^{T\times F}$. Let $P^{b}_{t,f} = |Z_{t,f}|^2$ denote BC spectral power. Because BC speech is concentrated more strongly at low frequencies (Wang et al., 2022), BC-ARE compares low-band and full-band BC power:

$$p_t = \log\left(1 + \frac{\frac{1}{F_b}\sum_{f=0}^{F_b-1} P^{b}_{t,f}}{\frac{1}{F}\sum_{f=0}^{F-1} P^{b}_{t,f} + \epsilon}\right), \tag{3}$$

where F_b is the number of low-frequency bins and $\epsilon > 0$ prevents numerical instability. The ratio measures relative low-frequency concentration and is less sensitive to the overall recording level.

At each time–frequency bin, BC-ARE forms a four-dimensional vector $\boldsymbol{\rho}_{t,f} \in \mathbb{R}^4$. Its entries, in order, are feature agreement $e_{t,f}$, the normalized difference between BC and AC log-magnitudes, the normalized deviation of BC log-magnitude from its causal running mean, and the low-frequency cue $\tanh(p_t)$. The running mean uses only the current and preceding frames. A learned pointwise network $\mathcal{G}_r : \mathbb{R}^4 \to \mathbb{R}$ maps this vector to a scalar, yielding the reliability gate $r_{t,f} = \sigma(\mathcal{G}_r(\boldsymbol{\rho}_{t,f})) \in (0, 1)$, where σ denotes sigmoid. The gate values $r_{t,f}$ form $\boldsymbol{r}$, which controls the strength of BC conditioning in RGAC. Appendix A.3 provides the statistical definitions and network configuration.

Reliability-gated affine conditioner (RGAC). The AC mixture retains broadband information needed for reconstruction, while BC provides a synchronized auxiliary cue. RGAC therefore receives $\boldsymbol{A}$, $\boldsymbol{C}$, and $\boldsymbol{r}$, preserves a direct AC path, and uses the BC context $\boldsymbol{C}$ to adjust it. A 1×1 projection of $\boldsymbol{C}$ produces $2C$ channels, which are split into gain and shift proposals $\boldsymbol{\Gamma}, \boldsymbol{\beta} \in \mathbb{R}^{C\times T\times F}$. The gain reweights existing features, and the shift adds a correction derived from BC context. Following feature-wise affine conditioning (Perez et al., 2018), RGAC uses the local gate to control both contributions and applies channel-wise normalization and PReLU to obtain its output:

$$\boldsymbol{H}^{(0)} = \mathrm{PReLU}\big(\mathcal{N}\big[\boldsymbol{A} \odot \big(1 + \eta \boldsymbol{r} \odot \tanh \boldsymbol{\Gamma}\big) + \eta \boldsymbol{r} \odot \boldsymbol{\beta}\big]\big)\,. \tag{4}$$

Here, η is a learnable scale, $\odot$ denotes elementwise multiplication, and $\boldsymbol{r}$ is broadcast over channels. The $\tanh$ bounds the gain proposal, while η sets the overall modulation scale. The direct $\boldsymbol{A}$ term is retained before normalization and activation. The RGAC output $\boldsymbol{H}^{(0)}$ enters the separator, while the CosSim output $\boldsymbol{e}$ and BC-ARE output $\boldsymbol{p}$ are passed to CAM so that these cross-sensor cues also inform reconstruction. Projection and initialization details are given in Appendix A.3.

3.4 Separator

The separator models temporal and frequency dependencies in the AC mixture representation after BC conditioning. It maps $\boldsymbol{H}^{(0)} \in \mathbb{R}^{C\times T\times F}$ through stacked SSTC modules and a Refiner to produce $\boldsymbol{H} \in \mathbb{R}^{C\times T\times F}$ for the CAM decoder. Each SSTC module first models temporal context at each frequency and then integrates frequencies within each frame (Figure 3).

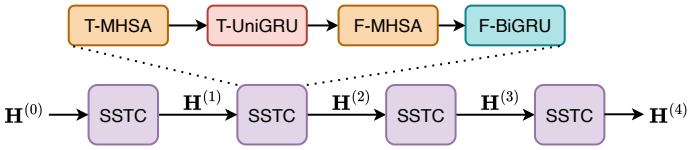


Figure 3: Separator architecture and the structure of an SSTC module.

Specifically, in the l-th SSTC module, the temporal operator $\mathcal{A}_t^{(l)}$ maps $\boldsymbol{H}^{(l-1)} \in \mathbb{R}^{C\times T\times F}$ to $\boldsymbol{U}^{(l)} \in \mathbb{R}^{C\times T\times F}$. The frequency operator $\mathcal{A}_f^{(l)}$ then maps $\boldsymbol{U}^{(l)}$ to $\boldsymbol{W}^{(l)} \in \mathbb{R}^{C\times T\times F}$. A weighted residual sum of the input and both outputs is normalized to obtain $\boldsymbol{H}^{(l)} \in \mathbb{R}^{C\times T\times F}$:

$$\begin{aligned} \boldsymbol{U}^{(l)} &= \mathcal{A}_t^{(l)}(\boldsymbol{H}^{(l-1)}), \qquad \boldsymbol{W}^{(l)} = \mathcal{A}_f^{(l)}(\boldsymbol{U}^{(l)}), \\ \boldsymbol{H}^{(l)} &= \mathcal{N}\Big(\boldsymbol{H}^{(l-1)} + \lambda_t^{(l)}\boldsymbol{U}^{(l)} + \lambda_f^{(l)}\boldsymbol{W}^{(l)}\Big)\,, \end{aligned} \tag{5}$$

where $\lambda_t^{(l)}, \lambda_f^{(l)} \in \mathbb{R}$ are learnable residual scales. Each operator contains a multi-head self-attention sublayer (Vaswani et al., 2017) followed by a recurrent sublayer comprising a gated recurrent unit (GRU) (Cho et al., 2014), PReLU, and a linear projection back to C channels. Both sublayers use pre-normalization and residual connections, and each operator ends with channel-wise normalization. The temporal operator processes a sequence in $\mathbb{R}^{C\times T}$ at each frequency using causally masked attention and a unidirectional GRU. The frequency operator processes a sequence in $\mathbb{R}^{C\times F}$ at each frame using full-frequency attention and a bidirectional GRU. This bidirectional processing traverses frequency bins within the current frame and therefore requires no future frames.

The Refiner maps the final SSTC output to $\boldsymbol{H}$ through four residual blocks, each operating independently at a time–frequency bin. Each block applies channel-wise normalization, a 1×1 convolution

from C to a hidden dimension, PReLU, and a second 1×1 convolution back to C, followed by residual addition. These operations introduce no additional temporal look-ahead. The separator configuration is detailed in Appendix A.4.

3.5 Confidence-Adaptive Mask (CAM) decoder

Fixed mask ranges impose the same correction limits at every time–frequency bin. CAM adapts these limits to local reliability using separator features $\boldsymbol{H}$, the agreement map $\boldsymbol{e}$, and the BC-activity cue $\boldsymbol{p}$ (Figure 4). Its Causal Confidence Estimator (CCE) controls range scaling, while the Mask Parameter Heads (MPH) predict preliminary margins and source-specific logits for complex masks (Williamson et al., 2016).

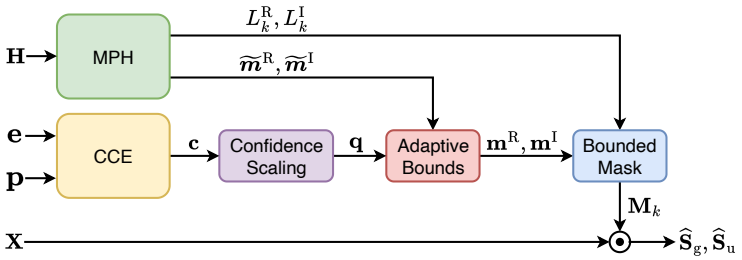


Figure 4: The structure of the Confidence-Adaptive Mask (CAM) decoder.

Causal confidence estimation. CCE centers the cues to make confidence depend on changes relative to their causal history. Let $\widetilde{e}_{t,f} = e_{t,f} - \overline{e}_{t,f}$ and $\widetilde{p}_t = p_t - \overline{p}_t$, where the running means $\overline{e}_{t,f}$ and $\overline{p}_t$ use frames 0 through t. With learnable scalar weights w_e, w_p, CCE and Confidence Scaling define the confidence map $\boldsymbol{c}$ and boundary-scale map $\boldsymbol{q}$ entrywise as

$$c_{t,f} = \sigma(w_e \widetilde{e}_{t,f} + w_p \widetilde{p}_t), \qquad q_{t,f} = 1 + \alpha(c_{t,f} - \tfrac{1}{2}), \tag{6}$$

where σ denotes sigmoid, $\alpha > 0$ controls the scaling strength, and $\widetilde{p}_t$ is broadcast over frequency. The running means and initialization are detailed in Appendix A.5.1.

Mask parameter prediction. In parallel, MPH uses separate branches on $\boldsymbol{H}$ so that the correction range can be adjusted independently of source-specific mask values. The margin branch predicts real and imaginary boundary offsets, then converts them to shared preliminary margins $\widetilde{\boldsymbol{m}}^{\mathrm{R}}, \widetilde{\boldsymbol{m}}^{\mathrm{I}} \in \mathbb{R}^{T \times F}$ before confidence scaling. The logit branches predict $\boldsymbol{L}_k^{\mathrm{R}}, \boldsymbol{L}_k^{\mathrm{I}} \in \mathbb{R}^{T \times F}$ for $k \in \{\mathrm{g}, \mathrm{u}\}$, indexing the guided and unguided outputs. All branches are causal. Their parameterization and layer settings are given in Appendix A.5.2.

Adaptive mask bounds. Adaptive Bounds adjusts the preliminary margins to local confidence and clips them to keep the correction ranges finite. The final shared margins $\boldsymbol{m}^{\mathrm{R}}, \boldsymbol{m}^{\mathrm{I}} \in \mathbb{R}^{T \times F}$ are

$$m_{t,f}^{\mathrm{R}} = \mathcal{K}(q_{t,f} \widetilde{m}_{t,f}^{\mathrm{R}}, 0, m_{\max}), \qquad m_{t,f}^{\mathrm{I}} = \mathcal{K}(q_{t,f} \widetilde{m}_{t,f}^{\mathrm{I}}, \epsilon_m, m_{\max}), \tag{7}$$

where $\mathcal{K}(z, l, u) = \min(\max(z, l), u)$, $\epsilon_m > 0$ is the imaginary-margin floor, and $m_{\max} > 0$ is the maximum margin. Confidence above or below $1/2$ respectively expands or contracts the preliminary margins before clipping. Because both outputs use the same mixture and cross-sensor cues, we share their margins while retaining independent logits for source-specific estimates. Hyperparameter values are given in Appendix A.5.3.

Masking and reconstruction. To allow amplitude correction around the identity response at one and phase corrections of either sign, Bounded Mask uses asymmetric real and symmetric imaginary ranges. It maps the logits to mask components $\boldsymbol{M}_k^{\mathrm{R}}, \boldsymbol{M}_k^{\mathrm{I}} \in \mathbb{R}^{T \times F}$ as

$$\begin{aligned} \boldsymbol{M}_k^{\mathrm{R}} &= (1 + 2\boldsymbol{m}^{\mathrm{R}}) \odot \sigma(\boldsymbol{L}_k^{\mathrm{R}}) - \boldsymbol{m}^{\mathrm{R}}, \\ \boldsymbol{M}_k^{\mathrm{I}} &= \boldsymbol{m}^{\mathrm{I}} \odot \tanh\left(\tfrac{1}{2}\boldsymbol{L}_k^{\mathrm{I}}\right), \end{aligned} \tag{8}$$

where $\odot$ denotes elementwise multiplication and $\tanh$ is the hyperbolic tangent. The resulting real and imaginary ranges are $[-m_{t,f}^{\mathrm{R}}, 1 + m_{t,f}^{\mathrm{R}}]$ and $[-m_{t,f}^{\mathrm{I}}, m_{t,f}^{\mathrm{I}}]$, respectively. The complex mask $\boldsymbol{M}_k = \boldsymbol{M}_k^{\mathrm{R}} + \mathrm{j}\boldsymbol{M}_k^{\mathrm{I}}$, with $\mathrm{j}^2 = -1$, gives $\widehat{\boldsymbol{s}}_k = \mathcal{F}^{-1}(\boldsymbol{M}_k \odot \boldsymbol{X})$, where $\mathcal{F}^{-1}$ is the inverse STFT. Since $\boldsymbol{X}$ contains noise, the clean-speech estimates are not constrained to sum to the mixture.

4 EXPERIMENTAL SETUP

4.1 DATASET SETUP

To expand public English AC–BC data, we built LS-Bone by replaying LibriSpeech utterances (Panayotov et al., 2015) through an artificial mouth and a vibration platform and synchronously recording both channels with a Timekettle W4 headset. LS-Bone contains 9,866 recording pairs (34 hours) from 1,171 source speakers. Appendix B.1 details the setup and procedure.

We constructed VibroMix from LS-Bone, Chinese speech from ABCS (Wang et al., 2023b), and environmental noise from WHAM! (Wichern et al., 2019) and DNS Challenge 5 (Dubey et al., 2024). All systems were evaluated on VibroMix-2Mix, its two-speaker benchmark comprising Chinese–Chinese, English–English, and Chinese–English mixtures. Overlap ratios (relative to the shorter segment) and relative speaker levels were sampled uniformly from $[0, 1]$ and $[-5, 5]$ dB, respectively. Noise was added with probability 0.8 at a signal-to-noise ratio (SNR) sampled uniformly from $[-5, 20]$ dB relative to the speech mixture. The 115,000 three-second mixtures at 16 kHz (approximately 95.83 hours) were split into 100,000 training, 10,000 development, and 5,000 test examples, each paired with two clean AC targets and corresponding BC references. Appendix B.2 details the data sources, mixture construction, and splits.

4.2 MODEL CONFIGURATION

For BAT, the shared audio encoder used a fixed centered STFT with a 256-sample Hann window and a 128-sample hop at 16 kHz. We set the feature dimension to $C = 64$. The CHRF module used two 1×5 depthwise frequency convolutions with dilation rates of 8 and 32. The separator comprised four SSTC modules; each SSTC module used four attention heads with dropout 0.1. The separator then used four pointwise Refiner blocks with hidden dimension 128. The CAM decoder used feature-dependent real and imaginary mask bounds with causal confidence scaling. Detailed layer configurations, hyperparameters, and initialization are provided in Appendix A.

4.3 TRAINING AND EVALUATION

All models were retrained on the same VibroMix-2Mix training split using three-second segments at 16 kHz. BAT was optimized with the negative signal-to-noise ratio (SNR) loss (Kavalerov et al., 2019), averaged over its two outputs, using AdamW with an initial learning rate of 10^{-3} and weight decay of 0.01. The learning rate was halved after 10 consecutive validation epochs without improvement, and training stopped after 20 such epochs, with a maximum of 1,000 epochs. Training was conducted on eight NVIDIA GeForce RTX 5090 GPUs.

We assessed signal fidelity using scale-invariant signal-to-noise ratio improvement (SI-SNRi) (Le Roux et al., 2019) and signal-to-distortion ratio improvement (SDRi) (Vincent et al., 2006), and perceptual quality using perceptual evaluation of speech quality (PESQ) (Rix et al., 2001). ViSQOL (Chinen et al., 2020) and speaker similarity complement these metrics by assessing perceptual quality and speaker identity preservation, respectively. Further training and evaluation details are provided in Appendix C.

5 RESULTS AND ANALYSIS

5.1 MAIN RESULTS

For BC-guided systems, single-target extraction estimates only the speaker associated with the BC reference, whereas two-output separation jointly estimates the guided and unguided speakers from the same AC mixture and a single BC reference. Table 1 compares BAT with five blind separators and two-output adaptations of five BC-guided baselines on VibroMix-2Mix, with each score averaged over the two recovered sources. Appendix D reports per-source results for BAT and the original single-target baselines; for the latter, the unguided estimate is obtained by subtracting the guided estimate from the mixture.

Table 1: Comparison with SOTA methods on VibroMix-2Mix. Each entry averages the two fixed-order outputs. The BC-guided baseline rows are two-output adaptations; the original single-target configurations are evaluated separately in Appendix D. **Bold** and underlining indicate the best and second-best values per column.

Method	SI-SNRi (dB) ↑	SDRi (dB) ↑	PESQ ↑	ViSQOL ↑	Speaker sim. (%) ↑
Blind separation					
DPRNN	11.6	12.4	2.5	2.1	56.0
SkiM	11.2	12.1	2.5	2.0	52.0
ReSepFormer	12.1	12.9	2.6	2.2	58.6
TF-GridNet	12.6	13.2	2.5	1.9	52.5
TF-ACM	11.8	12.5	2.4	1.5	48.7
BC-guided baselines (two-output)					
BiNet	7.3	8.6	1.1	1.3	30.0
BAF-Net	2.2	5.1	1.2	1.6	36.6
FlowDense	12.6	13.1	1.7	1.9	50.8
DBMIF	11.4	12.0	1.5	1.6	41.6
Robust Fusion	<u>14.7</u>	<u>15.0</u>	<u>2.8</u>	<u>2.8</u>	<u>61.3</u>
BAT	**17.3**	**17.5**	**3.0**	**2.9**	**61.6**

BAT achieves the highest mean scores across all five metrics in Table 1. Compared with the two-output Robust Fusion adaptation, BAT improves SI-SNRi and SDRi by 2.6 dB and 2.5 dB, respectively, and PESQ by 0.2. These gains persist when the baseline also directly predicts both speakers, establishing BAT's advantage in the two-output setting. ViSQOL and speaker similarity also increase, although the speaker-similarity gain is modest at 0.3 percentage points. The representative examples on our demo page also illustrate the improvement in perceived speech quality. Qualitative spectrogram comparisons in Appendix E show the same trend, with BAT more closely following the clean time–frequency structure than Robust Fusion.[2]

5.2 Computational Efficiency

We profile all models in causal configurations using one-second inputs at 16 kHz, with batch size 1 and 32-bit floating-point precision (FP32). CPU and GPU inference times are measured on an Intel Xeon Gold 6530 and an NVIDIA GeForce RTX 5090, respectively. Blind separators receive AC mixtures, while BC-guided systems receive paired AC–BC inputs. BC-guided baselines retain their original single-target configurations, whereas BAT and the blind separators produce two outputs. Further profiling details are provided in Appendix C.4.

Table 2 shows that BAT has the lowest measured GPU inference time (9.2 ms), RAM usage (123.2 MB), and VRAM usage (10.9 MB). Compared with the original single-target Robust Fusion (Liu et al., 2025), BAT uses approximately 46.2% fewer parameters and 34.6% fewer multiply–accumulate operations (MACs), while reducing GPU inference time by 33.8% (13.9 to 9.2 ms). On the CPU, BAT takes 228.7 ms, comparable to Robust Fusion's 232.5 ms but longer than the 58.3 ms of ReSepFormer, the fastest CPU baseline. Relative to ReSepFormer, BAT improves mean SI-SNRi, SDRi, and PESQ by 5.2 dB, 4.6 dB, and 0.4, respectively, at the cost of longer CPU inference time.

5.3 Ablation Study

All ablations used the same VibroMix-2Mix split, training budget, and evaluation protocol as BAT. Each comparison changed the specified component while retaining the remaining model. SI-SNRi, SDRi, and PESQ were averaged over both output sources.

CHRF. To assess how the fusion operator affected AC–BC separation, we replaced CHRF with addition, concatenation, cross-attention, or feature-wise linear modulation (FiLM), while retaining the shared encoder, separator, and CAM decoder. Table 3 compares these operators. CHRF achieved the highest SI-SNRi, SDRi, and PESQ among the fusion operators. Relative to cross-attention, it improved the three metrics by 0.3 dB, 0.4 dB, and 0.1, respectively, while reducing fusion-module parameters from 62.9K to 47.6K, or approximately 24.3%. Appendix F evaluates the contributions of cross-band context and local reliability weighting in the CHRF component ablation.

[2]Demo page: `https://anonymous.4open.science/w/BAT-Demo-3603/`.

Table 2: Computational efficiency under the common profiling configuration.

Method	Params (M) ↓	MACs (G) ↓	CPU (ms) ↓	RAM (MB) ↓	GPU (ms) ↓	VRAM (MB) ↓
Blind separation						
DPRNN	0.7	1.0	99.0	144.0	9.7	11.8
SkiM	3.7	1.6	93.2	166.6	12.3	22.3
ReSepFormer	8.0	12.3	**58.3**	323.5	14.5	185.0
TF-GridNet	1.3	1.7	468.5	135.0	18.8	13.3
TF-ACM	**0.5**	22.4	819.1	133.1	39.5	12.2
BC-guided baselines (single-target)						
BiNet	7.7	1.9	87.3	130.7	28.1	37.7
BAF-Net	15.8	14.7	90.8	164.1	13.4	74.6
FlowDense	3.0	1.0	64.1	128.6	15.8	11.7
DBMIF	10.6	**0.6**	118.0	149.0	40.9	40.8
Robust Fusion	1.3	23.1	232.5	131.8	13.9	13.2
BAT	0.7	15.1	228.7	**123.2**	**9.2**	**10.9**

Table 3: Ablation of AC–BC fusion operators. Parameter counts cover only the fusion module.

Variant	SI-SNRi (dB) ↑	SDRi (dB) ↑	PESQ ↑	Params (K) ↓
Fusion operators				
Add	16.7	16.9	2.9	**38.4**
Concat	16.9	17.1	2.9	56.8
Cross-attention	17.0	17.1	2.9	62.9
FiLM	16.8	17.1	2.9	46.7
CHRF (BAT)	**17.3**	**17.5**	**3.0**	47.6

CAM. To evaluate the contribution of adaptive mask ranges, we replaced CAM's Adaptive Bounds with fixed base ranges, removing both the learned boundary offsets and confidence modulation. We also evaluated sigmoid and tanh mask parameterizations while retaining the remaining decoder. Table 4 compares these mask parameterization and range variants. CAM improved over fixed bounds by 0.5 dB SI-SNRi, 0.4 dB SDRi, and 0.1 PESQ, supporting the use of adaptive ranges for spectral reconstruction. It also outperformed both the sigmoid and tanh variants across the three metrics. The boundary-design ablation is reported in Appendix G.

Table 4: Ablation of CAM mask parameterization and mask ranges.

Variant	SI-SNRi (dB) ↑	SDRi (dB) ↑	PESQ ↑
Mask parameterization and bounds			
Sigmoid mask	14.7	15.0	2.8
Tanh mask	16.7	17.0	2.9
Fixed mask bounds	16.8	17.1	2.9
Full CAM (BAT)	**17.3**	**17.5**	**3.0**

Additional separator ablations showed that four SSTC modules provided a trade-off between separation quality and computational cost, and that GRU matched LSTM within 0.1 dB SI-SNRi and SDRi with fewer parameters and MACs. Removing multi-head self-attention or the Refiner reduced mean SI-SNRi by 0.7 dB or 0.3 dB, respectively. The SSTC studies are detailed in Appendix H, and the Refiner study in Appendix I.

6 Conclusion

We presented BAT, a compact frame-causal framework that recovers wearer and interlocutor speech from an AC mixture and a single BC reference from the wearer. BAT uses cross-band BC context and local reliability to guide feature fusion, and adapts complex-mask bounds to cross-sensor evidence during reconstruction. On VibroMix-2Mix, BAT improves mean SI-SNRi by 2.6 dB over the two-output Robust Fusion adaptation, with ablations supporting the contributions of reliability-aware fusion and adaptive mask bounds. LS-Bone adds controlled English AC–BC recordings, while VibroMix supports evaluation on noisy same-language and cross-language two-speaker mixtures.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript drafting, sentence rewriting, grammar checking, translation, language polishing, and improving readability and clarity. We did not use generative AI tools for research conceptualization, methodology development, experimental design, data collection, data analysis, or the generation or modification of experimental data, model outputs, figures, or reported results; other required disclosure categories are not applicable. We reviewed all AI-assisted work against the manuscript, source code, experimental records, equations, citations, tables, figures, and reported results. All scientific concepts, methodological decisions, analyses, and conclusions were developed and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The task and model architecture are specified in Sections 3 and 4, while the complete shared encoder, CHRF, separator, and CAM configurations are given in Appendix A. Appendix B describes the paired-recording setup, mixture construction, and dataset splits. Appendix C provides the optimization objective, input–output correspondence, and metric implementation and aggregation protocol, and Appendices H and I document the additional component studies. The implementation, model configuration files, and data-generation scripts will be provided as supplementary material or through an anonymous repository to support reproduction.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Guo Chen, Kai Li, Runxuan Yang, and Xiaolin Hu. Time-frequency-based attention cache memory model for real-time speech separation. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2025.

Michael Chinen, Felicia S. C. Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. ViSQOL v3: An open source production ready objective speech and audio metric. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734. Association for Computational Linguistics, 2014.

Marc Delcroix, Tsubasa Ochiai, Katerina Zmolikova, Keisuke Kinoshita, Naohiro Tawara, Tomohiro Nakatani, and Shoko Araki. Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 691–695. IEEE, 2020.

Luca Della Libera, Cem Subakan, Mirco Ravanelli, Samuele Cornell, Frédéric Lepoutre, and François Grondin. Resource-efficient separation transformer. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 761–765. IEEE, 2024.

Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, Alex Ju, Mehdi Zohourian, Min Tang, Mehrsa Golestaneh, et al. Icassp 2023 deep noise suppression challenge. *IEEE Open Journal of Signal Processing*, 5:725–737, 2024.

Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang, and Haizhou Li. SpEx+: A complete time domain speaker extraction network. In *Proceedings of Interspeech 2020*, pp. 1406–1410, 2020.

Fengyuan Hao, Andong Li, Xiao-Dong Li, and Chengshi Zheng. DSINet: Towards real-time target speaker extraction with dynamic speaker information fusion. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Lixing He, Haozheng Hou, Shuyao Shi, Xian Shuai, and Zhenyu Yan. Towards bone-conducted vibration speech enhancement on head-mounted wearables. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, pp. 14–27. ACM, 2023.

Jens Heitkaemper, Joe Caroselli, Max McKinnon, Arun Narayanan, and Nathan Howard. Bone conducted signal guided speech enhancement for voice assistant on earbuds. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025.

Md. Sarwar Hosain, Yosuke Sugiura, M. Shahidur Rahman, and Tetsuya Shimamura. EmoBone: A multinational audio dataset of emotional bone conducted speech. *IEEJ Transactions on Electrical and Electronic Engineering*, 19(9):1492–1506, 2024.

ITU-T. Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs. Technical Report P.862.2, International Telecommunication Union, 2007.

Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin W. Wilson, Jonathan Le Roux, and John R. Hershey. Universal sound separation. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

Yunsik Kim and Yoonyoung Chung. Modality-specific speech enhancement and noise-adaptive fusion for acoustic and body-conduction microphone framework. In *Proceedings of Interspeech 2025*, pp. 3833–3837, 2025.

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)*, 2019.

Changtao Li, Feiran Yang, and Jun Yang. Bone conduction-aided speech enhancement with two-tower network and contrastive learning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:163–174, 2025.

Chenda Li, Lei Yang, Weiqin Wang, and Yanmin Qian. SkiM: Skipping memory LSTM for low-latency real-time continuous speech separation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 681–685. IEEE, 2022.

Junjie Li, Ke Zhang, Shuai Wang, Haizhou Li, Man-Wai Mak, and Kong Aik Lee. On the effectiveness of enrollment speech augmentation for target speaker extraction. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 325–332. IEEE, 2024a.

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, 2024b.

Kai Li, Runxuan Yang, Fuchun Sun, and Xiaolin Hu. IIANet: An intra- and inter-modality attention network for audio-visual speech separation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29181–29200. PMLR, 2024c.

Zhenglong Liu, Zhe Chen, and Fuliang Yin. Robust fusion of bone and air-conducted sensors for speech enhancement with adaptive temporal-frequency attention. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Zi-Kai Liu, Zi-Qian Wang, Xing-Chen Li, Yi-Ke Zhu, Shuai Wang, Long-Shuai Xiao, and Lei Xie. EvoTSE: Evolving enrollment for target speaker extraction. *arXiv preprint arXiv:2604.06810*, 2026.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Cong Pang, Feifei Xiong, Ye Ni, Lin Zhou, and Jinwei Feng. The SEUEE system for the CHiME-8 MMCSG challenge. In *8th International Workshop on Speech Processing in Everyday Environments (CHiME 2024)*, pp. 75–80, 2024.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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)*, 2001.

Hiroshi Sato, Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Takafumi Moriya, Naoki Makishima, Mana Ihori, Tomohiro Tanaka, and Ryo Masumura. Strategies to improve robustness of target speech extraction to enrollment variations. In *Proceedings of Interspeech 2022*, pp. 996–1000, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. WeSpeaker: A research and production oriented speaker embedding learning toolkit. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Mou Wang, Junqi Chen, Xiaolei Zhang, Zhiyong Huang, and Susanto Rahardja. Multi-modal speech enhancement with bone-conducted speech in time domain. *Applied Acoustics*, 200:109058, 2022.

Mou Wang, Junqi Chen, Xiao-Lei Zhang, and Susanto Rahardja. End-to-end multi-modal speech recognition on an air and bone conducted speech corpus. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:513–524, 2023b.

Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. Wham!: Extending speech separation to noisy environments. In *Proceedings of Interspeech*, pp. 1368–1372, 2019. doi: 10.21437/Interspeech.2019-2821.

Donald S Williamson, Yuxuan Wang, and DeLiang Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 24(3): 483–492, 2016.

Yilei Wu, Changyan Zheng, Xingyu Zhang, and Yakun Zhang. Flowdense: A subband framework for air–and bone–conduction speech enhancement. In *2025 6th International Conference on Machine Learning and Computer Application (ICMLCA)*, pp. 1094–1097. IEEE, 2025.

Yilei Wu, Changyan Zheng, Xingyu Zhang, Yakun Zhang, Chengshi Zheng, Shuang Yang, Ye Yan, and Erwei Yin. Dbmif: a deep balanced multimodal iterative fusion framework for air-and bone-conduction speech enhancement. *Applied Intelligence*, 56(5):165, 2026.

Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. PHASEN: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9458–9465, 2020.

Chen Zhang, Linfeng Feng, Zhi Liu, Xiao-Lei Zhang, and Xuelong Li. MBENet: Bone-conduction and air-conduction fusion network for target speaker extraction. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–5. IEEE, 2025.

Kateřina Žmolíková, Simone Merello, Kaustubh Kalgaonkar, Ju Lin, Niko Moritz, Pingchuan Ma, Ming Sun, Honglie Chen, Antoine Saliou, Stavros Petridis, Christian Fuegen, and Michael Mandel. The CHiME-8 MMCSG challenge: Multi-modal conversations in smart glasses. In *8th International Workshop on Speech Processing in Everyday Environments (CHiME 2024)*, pp. 7–12, 2024.

A Detailed Model Configurations

A.1 Qualitative AC–BC Spectral Comparison

Figure 5 provides a qualitative comparison of synchronized bone-conducted (BC) and air-conducted (AC) recordings for the same speech content. The BC spectrogram retains the main low-frequency speech structure but shows weaker high-frequency detail, whereas the AC spectrogram preserves speech energy over a broader frequency range. This contrast motivates treating BC as a frequency-selective auxiliary cue and using cross-band context to transfer informative BC evidence to related AC frequency regions. The spectrogram illustrates the spectral characteristics qualitatively; quantitative separation results are reported in the main text.

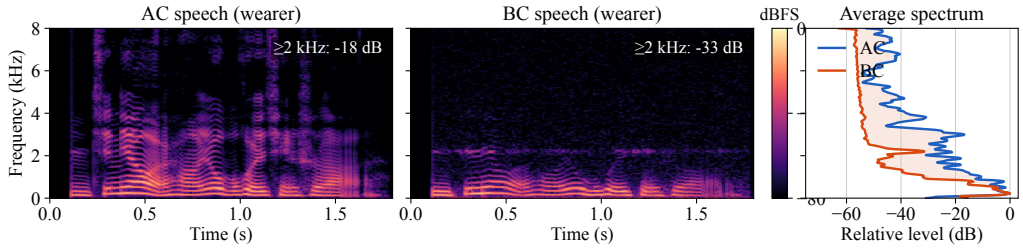


Figure 5: Qualitative comparison of synchronized bone-conducted (BC) and air-conducted (AC) speech spectrograms for the same speech content. BC retains the main low-frequency structure but exhibits weaker high-frequency detail than AC. Color denotes log magnitude in dB.

A.2 Shared audio encoder

The shared audio encoder applies the same fixed centered STFT to the AC and BC waveforms with a 256-sample Hann window and a 128-sample hop at 16 kHz, giving $F = 129$ frequency bins. This corresponds to a 16 ms window and an 8 ms frame advance and requires analysis-window buffering; end-to-end latency also depends on computation. The resulting stacked real and imaginary components are processed by the same two 3×3 convolutional stages with $C = 64$ output channels, each followed by channel-wise layer normalization (Ba et al., 2016) and PReLU (He et al., 2015). The convolutional parameters are shared across modalities, including the PReLU parameters.

A.3 CHRF module

F-DConv configuration. F-DConv uses two parallel depthwise frequency convolutions with 1×5 kernels and dilation rates $d_1 = 8$ and $d_2 = 32$. At 62.5 Hz per bin, adjacent taps are separated by 500 Hz and 2 kHz, respectively, with a maximum offset of 64 bins. The convolutions preserve feature dimensions; their outputs are added to the original BC features and normalized channel-wise.

BC-ARE implementation details. The low-frequency cue p_t is defined in Equation 3. Let $P^x_{t,f} = |X_{t,f}|^2$ and $P^b_{t,f} = |Z_{t,f}|^2$ denote AC and BC spectral power, respectively. To represent changes in BC level relative to its causal history, BC-ARE defines the BC log-magnitude $\ell_{t,f} = \frac{1}{2}\log(P^b_{t,f} + \epsilon)$

and its running mean $\overline{\ell}_{t,f} = (t+1)^{-1} \sum_{\tau=0}^{t} \ell_{\tau,f}$. The mean uses the current and preceding frames only. At each time–frequency bin, BC-ARE concatenates four normalized statistics:

$$\boldsymbol{\rho}_{t,f} = \left[e_{t,f},\ \frac{\mathcal{K}\left(\frac{1}{2} \log \frac{P^{b}_{t,f}+\epsilon}{P^{x}_{t,f}+\epsilon}, -\kappa_p, \kappa_p \right)}{\kappa_p},\ \frac{\mathcal{K}(\ell_{t,f} - \overline{\ell}_{t,f}, -\kappa_\ell, \kappa_\ell)}{\kappa_\ell},\ \tanh(p_t) \right]^{\top}, \tag{9}$$

where $\mathcal{K}(z, l, u) = \min(\max(z, l), u)$ clips values to $[l, u]$, with clipping thresholds κ_p and κ_ℓ.

We used $F_b = 11$, $\epsilon = 10^{-8}$, $\epsilon_e = 10^{-6}$, $\kappa_p = 6$, and $\kappa_\ell = 4$. The pointwise network $\mathcal{G}_r$ consists of two 1×1 convolutions with channel widths $4 \to 8 \to 1$, a PReLU activation between them (He et al., 2015), and biases. A sigmoid maps its scalar output to the reliability gate $r_{t,f}$. The final projection was initialized to zero, giving an initial gate of 0.5.

RGAC projection and initialization. RGAC maps $\boldsymbol{C}$ through a 1×1 convolution with bias from C to $2C$ channels, then splits the result into gain and shift proposals $\boldsymbol{\Gamma}$ and $\boldsymbol{\beta}$. The gain proposal is bounded by $\tanh$, with a fixed multiplier of 1.0. The gate $\boldsymbol{r}$ and learnable scale η control both proposals as in Equation 4, followed by channel-wise layer normalization and a C-channel PReLU. These operations act independently at each time–frequency location. We initialized the learnable fusion scale to $\eta = 0.1$. The F-DConv weights and the RGAC projection's weights and biases were initialized to zero, giving no initial affine modulation.

A.4 Separator

The separator contains four SSTC modules, each with four attention heads and dropout 0.1. Temporal GRUs have 64 hidden units, while frequency GRUs are bidirectional with 32 units per direction. The residual scales $\lambda_t^{(l)}$ and $\lambda_f^{(l)}$ were initialized to one. Four pointwise Refiner blocks follow the SSTC stack; each uses a $64 \to 128 \to 64$ projection with PReLU, channel-wise normalization, and an input skip connection. The second pointwise projection was zero-initialized, so the Refiner started as an identity mapping.

A.5 CAM decoder

A.5.1 CCE and confidence scaling

CCE computes the causal running means $\overline{\boldsymbol{e}} \in \mathbb{R}^{T \times F}$ and $\overline{\boldsymbol{p}} \in \mathbb{R}^{T}$ as

$$\overline{e}_{t,f} = \frac{1}{t+1} \sum_{\tau=0}^{t} e_{\tau,f}, \qquad \overline{p}_t = \frac{1}{t+1} \sum_{\tau=0}^{t} p_\tau. \tag{10}$$

Both means include the current frame and use no future frames. Subtracting them from $\boldsymbol{e}$ and $\boldsymbol{p}$ gives the centered cues $\widetilde{\boldsymbol{e}}$ and $\widetilde{\boldsymbol{p}}$ used in Equation 6, which produces $\boldsymbol{c} \in (0,1)^{T \times F}$ and $\boldsymbol{q} \in \mathbb{R}^{T \times F}$. We initialized the scalar weights w_e, w_p to zero, giving $c_{t,f} = 0.5$ and $q_{t,f} = 1$. We set $\alpha = 0.8$, so the confidence scale lies between 0.6 and 1.4.

A.5.2 Mask parameter heads

Boundary-offset branch. The boundary-offset branch applies channel normalization $\mathcal{N}$ to $\boldsymbol{H}$ and uses a pointwise network $\mathcal{G}_m$ with widths $64 \to 16 \to 2$, channel normalization, and ReLU between projections. Its real and imaginary offsets $\boldsymbol{\Delta}^{\mathrm{R}}, \boldsymbol{\Delta}^{\mathrm{I}} \in \mathbb{R}^{T \times F}$ and preliminary margins are

$$\begin{aligned} (\boldsymbol{\Delta}^{\mathrm{R}}, \boldsymbol{\Delta}^{\mathrm{I}}) &= \zeta(\mathcal{G}_m(\mathcal{N}(\boldsymbol{H}))) - \log 2, \\ \widetilde{\boldsymbol{m}}^{\mathrm{R}} &= \max(b_{\mathrm{R}} + \boldsymbol{\Delta}^{\mathrm{R}}, 0), \qquad \widetilde{\boldsymbol{m}}^{\mathrm{I}} = \max(b_{\mathrm{I}} + \boldsymbol{\Delta}^{\mathrm{I}}, \epsilon_m), \end{aligned} \tag{11}$$

where ζ is the elementwise softplus, $b_{\mathrm{R}}, b_{\mathrm{I}} > 0$ are base margins, and $\epsilon_m > 0$ is the imaginary-margin floor. Subtracting $\log 2 = \zeta(0)$ makes zero projection output correspond to zero offset. The output projection was initialized to zero.

Logit branches. The four logit heads are independent of one another and of the boundary-offset branch. Each uses two causal 3×3 convolutions with channel normalization and pointwise nonlinear processing, followed by a pointwise projection to one channel. Real-logit heads use dilation 1 in both layers; imaginary-logit heads use dilation 2 followed by 1. Their outputs $\boldsymbol{L}_k^{\mathrm{R}}, \boldsymbol{L}_k^{\mathrm{I}}$ enter the bounded mapping in Equation 8.

A.5.3 Adaptive bounds and hyperparameters

Adaptive Bounds applies the clipping operator in Equation 7 elementwise. The two outputs share $\boldsymbol{m}^{\mathrm{R}}$ and $\boldsymbol{m}^{\mathrm{I}}$ while retaining independent logits. We set $b_{\mathrm{R}} = 1.0$, $b_{\mathrm{I}} = 1.5$, $\epsilon_m = 10^{-4}$, and $m_{\max} = 4.0$, giving global real and imaginary mask bounds of $[-4, 5]$ and $[-4, 4]$. The neutral confidence scale and zero boundary offsets at initialization recover the base margins.

B Dataset Construction Details

B.1 LS-Bone Data Collection

LibriSpeech (Panayotov et al., 2015) is an English read-speech corpus derived from LibriVox audiobooks. It contains approximately 1,000 hours of audio at 16 kHz, with utterance-level transcriptions and speaker identities, and is distributed under CC BY 4.0. We randomly selected utterances from this corpus for controlled physical re-recording.

The resulting LS-Bone corpus comprised 9,866 paired AC–BC recordings totaling 34 hours and spanning 1,171 source speakers. The two channels were acquired synchronously and stored at 16 kHz. These recordings paired English speech content with measurements from a wearable device's sensing and processing chain, and served as the English source pairs in VibroMix.

The original LibriSpeech waveform drove the artificial mouth, while its 0–2 kHz low-pass version drove the bone-conduction vibration platform. The two paths were played synchronously and recaptured by a Timekettle W4 headset. The headset's built-in environmental-noise-cancellation (ENC) processing remained part of the capture chain.

The apparatus used a CRY1125 vibration platform, a CRY609 artificial mouth, an Alctron first-stage adjustable amplifier, a CRY584D second-stage amplifier, and a custom marble base and test fixture. The artificial mouth was positioned approximately 10 cm from the headset. Recordings were made in a $4.4 \times 4.2 \times 2.6$ m reverberation room with approximately 35 dB background noise and a 0.28 s reverberation time. Figure 6 shows the recording apparatus.

B.2 VibroMix-2Mix Construction

Speech and noise sources. ABCS was introduced as a synchronized AC–BC corpus for multimodal Mandarin speech recognition (Wang et al., 2023b). Its public release contains 47,182 paired utterances from 100 speakers, totaling approximately 42 hours, with individual utterances lasting 1–5 s.[3] Participants wore a headset and read text prompts in an anechoic chamber while the two channels were recorded synchronously. We used these Mandarin speech pairs together with the English pairs in LS-Bone to construct same-language and cross-language mixtures.

WHAM! extends the wsj0-2mix speech-separation benchmark with real ambient noise (Wichern et al., 2019). Its original noise collection comprised nearly 80 hours recorded at 44 locations in the San Francisco Bay Area, including coffee shops, restaurants, bars, offices, and parks. We used its environmental recordings as additive noise for the paired AC–BC speech.

The fifth DNS Challenge provided 48 kHz resources for joint noise suppression, dereverberation, and interfering-talker suppression (Dubey et al., 2024). It reused the noise corpus from the fourth challenge; the challenge paper reported 181 hours of training noise, including 60,000 AudioSet clips covering 152 sound classes. We used the public AudioSet and Freesound noise recordings.[4]

[3] ABCS corpus release: `https://github.com/wangmou21/abcs`.

[4] DNS Challenge data release: `https://github.com/microsoft/DNS-Challenge`.

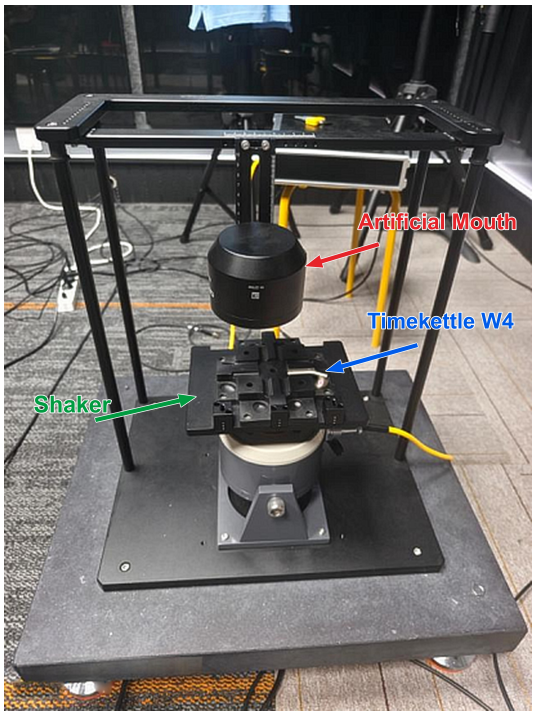


Figure 6: LS-Bone recording apparatus: the vibration platform, artificial mouth, and Timekettle W4 headset used to recapture synchronized air-conduction and bone-conduction signals.

We filtered the DNS pool by removing Freesound recordings with speech-related filename keywords and applying Silero voice activity detection to AudioSet recordings. An AudioSet recording was excluded when detected speech lasted at least 0.30 s and occupied at least 5% of its duration. This filtering retained 7,665 Freesound and 2,097 AudioSet files, giving 9,762 DNS noise recordings.

Mixture generation. Each mixture occupied a three-second window of 48,000 samples. Chinese–Chinese, English–English, and Chinese–English mixtures received equal allocations up to integer rounding; the cross-language group balanced the two source orderings. To distribute speaker usage, each draw selected uniformly from the least-used quarter of the relevant speaker pool. Same-corpus mixtures used distinct speakers.

For each selected utterance, the segment duration was sampled uniformly between 1.5 s and the smaller of 2.8 s and the utterance duration. Shorter utterances were excluded, and random cropping retried up to five times to avoid silent regions. AC and BC shared crop boundaries and placement. The generator sampled a requested overlap ratio uniformly from zero to one relative to the shorter segment and randomized which speaker started first.

The relative speech level was sampled uniformly from -5 to 5 dB. The second speaker's contribution to the AC mixture was scaled using the two AC energies, and its BC signal received the same scale. A common peak-normalization factor was applied to the speech mixture, clean AC signals, and BC signals, followed by a shared global gain sampled uniformly from -3 to 3 dB. Noise was added with probability 0.8. Conditional on noise addition, WHAM! and DNS were chosen with equal probability, and the noise SNR relative to the speech mixture was sampled uniformly from -5 to 20 dB. Noise recordings were randomly cropped to three seconds or repeated cyclically when shorter. Noise affected only the AC mixture; clean targets and BC references were unchanged.

Each example stored 16-bit PCM audio: a mixture, two clean AC targets, and two BC references.

Dataset splits. Table 5 summarizes the mixture counts and duration. Generation used random seed 42. The WHAM! noise pool combined training and validation partitions; testing used the test set.

Table 5: VibroMix-2Mix scale. Duration counts each three-second mixture once.

Split	Mixtures	Duration (h)
Training	100,000	83.33
Development	10,000	8.33
Test	5,000	4.17
Total	115,000	95.83

C Training and Evaluation Details

C.1 Optimization and objective

All models were retrained using 16 kHz mixtures and the corresponding clean AC targets. BAT was trained on three-second segments using AdamW with an initial learning rate of 10^{-3} and weight decay of 0.01 on eight NVIDIA GeForce RTX 5090 GPUs. The learning rate was halved after 10 consecutive validation epochs without improvement, and training stopped after 20 such epochs, with a maximum of 1,000 epochs. The configured batch size was 8.

BAT minimized the negative waveform SNR (Kavalerov et al., 2019), averaged over the two source estimates. Let $\boldsymbol{s}_k \in \mathbb{R}^L$ denote the clean AC target for source k. The mean-centered target and estimate are $\widetilde{\boldsymbol{s}}_k, \widetilde{\widehat{\boldsymbol{s}}}_k \in \mathbb{R}^L$, respectively. Omitting the batch index, the per-example loss $\mathcal{L} \in \mathbb{R}$ is

$$\mathcal{L} = -\frac{1}{2} \sum_{k \in \{\mathrm{g},\mathrm{u}\}} 10 \log_{10} \left(\frac{\|\widetilde{\boldsymbol{s}}_k\|_2^2}{\|\widetilde{\widehat{\boldsymbol{s}}}_k - \widetilde{\boldsymbol{s}}_k\|_2^2 + \epsilon} + \epsilon \right). \tag{12}$$

The loss was averaged over each minibatch. The guided and unguided targets retained their fixed order throughout training and validation. Validation used negative scale-invariant signal-to-distortion ratio (SI-SNR) (Le Roux et al., 2019) in the same order.

C.2 Input and output correspondence

The dataset fields `spk1` and `spk2` correspond to the guided and unguided speakers, respectively. The guided speaker is the source associated with the BC input. This correspondence was retained across training, development, and testing without exchanging the BC reference. Blind separators received only the AC mixture; BC-guided systems additionally received the guided speaker's BC recording. Each test mixture was evaluated once, giving 5,000 test examples.

For BAT, the blind separators, and the two-output BC-guided adaptations, the first and second output channels were compared directly with the clean guided and unguided targets, respectively, without permutation matching. For the original single-target BC baselines, the guided estimate was $\widehat{\boldsymbol{s}}_{\mathrm{g}} \in \mathbb{R}^L$, and the unguided estimate was

$$\widehat{\boldsymbol{s}}_{\mathrm{u}} = \boldsymbol{x} - \widehat{\boldsymbol{s}}_{\mathrm{g}} \in \mathbb{R}^L. \tag{13}$$

Both estimates were evaluated against the corresponding clean AC targets. The mixture-subtracted estimate was computed from the observed mixture, including any environmental noise, without access to either clean target.

C.3 Evaluation Metrics

We computed SI-SNR, SDR, and PESQ using TorchMetrics.[5] We evaluated complete three-second waveforms at 16 kHz, comparing each estimate with its corresponding clean AC target.

SI-SNRi. SI-SNR measures how much of a separated waveform is explained by a scaled version of the clean target (Le Roux et al., 2019). We removed the temporal mean from both waveforms before evaluation. For source $k \in \{\mathrm{g}, \mathrm{u}\}$, the least-squares scaling coefficient is $\alpha_k = \langle \widetilde{\widehat{\boldsymbol{s}}}_k, \widetilde{\boldsymbol{s}}_k \rangle / \|\widetilde{\boldsymbol{s}}_k\|_2^2$,

[5] `https://github.com/Lightning-AI/torchmetrics`

using the mean-centered waveforms defined above. The corresponding score is

$$\text{SI-SNR}(\widehat{\boldsymbol{s}}_k, \boldsymbol{s}_k) = 10 \log_{10} \left(\frac{\|\alpha_k \widetilde{\boldsymbol{s}}_k\|_2^2}{\|\widetilde{\widehat{\boldsymbol{s}}}_k - \alpha_k \widetilde{\boldsymbol{s}}_k\|_2^2} \right). \tag{14}$$

The numerator measures the energy of the projected target component, while the denominator measures the residual error, including interfering speech, noise, and reconstruction artifacts. A uniform gain applied to the estimate therefore leaves the score unchanged. We computed the improvement as $\text{SI-SNRi}_k = \text{SI-SNR}(\widehat{\boldsymbol{s}}_k, \boldsymbol{s}_k) - \text{SI-SNR}(\boldsymbol{x}, \boldsymbol{s}_k)$, expressed in decibels. This difference quantifies the gain in separation fidelity over the unprocessed mixture for the same clean target.

SDRi. SDR evaluates the energy of the target component relative to the residual distortion under a specified class of allowed target deformations (Vincent et al., 2006). We used a time-invariant finite impulse response filter with 512 coefficients to fit the clean reference to the estimate, without subtracting the waveform means. SDR is then the logarithmic ratio between the energy of the fitted reference and that of the remaining error. The error includes residual interference, noise, and artifacts that cannot be explained by the fitted reference. This definition permits linear filtering of the target, whereas SI-SNR permits a single scaling coefficient (Le Roux et al., 2019). We computed $\text{SDRi}_k = \text{SDR}(\widehat{\boldsymbol{s}}_k, \boldsymbol{s}_k) - \text{SDR}(\boldsymbol{x}, \boldsymbol{s}_k)$ in decibels, applying the same SDR definition to the separated waveform and the input mixture.

PESQ. PESQ is a full-reference objective measure that predicts perceived speech quality by comparing a processed signal with its clean reference (Rix et al., 2001). Its perceptual model accounts for the frequency-dependent sensitivity of hearing and the audibility of differences between the two signals. The resulting disturbance measurements are pooled into a single quality score, so the evaluation reflects perceptual degradation beyond waveform error energy. We used the wideband extension at 16 kHz (ITU-T, 2007) and reported the absolute score for each recovered AC waveform against its clean AC target. Higher PESQ indicates better predicted listening quality.

ViSQOL. ViSQOL is a full-reference perceptual quality metric based on spectrotemporal similarity (Chinen et al., 2020). It converts the reference and processed waveforms into gammatone spectrograms, aligns local time–frequency patches, and compares the aligned patches using the neurogram similarity index measure. These similarities are aggregated and mapped to a mean opinion score–listening quality objective (MOS-LQO), with higher values indicating better predicted listening quality. This provides a perceptual assessment based on the preservation of local spectral and temporal structure. We computed ViSQOL using Google's official implementation.[6]

Speaker similarity. Speaker similarity assesses whether the recovered speech retains the identity of the corresponding clean speaker. We used pretrained ResNet34-LM speaker encoders from WeSpeaker (Wang et al., 2023a), with the VoxCeleb checkpoint for English speech[7] and the CN-Celeb checkpoint for Chinese speech.[8] Each complete estimate and its clean reference were mapped to a 256-dimensional utterance-level embedding. We L2-normalized the two embeddings and computed their cosine similarity, and reported 100 times this value as a percentage. Scores were averaged over the two fixed-order sources and all test examples before rounding. Comparing the embeddings of the estimate and its clean reference focuses on speaker characteristics and complements the assessment of waveform fidelity and perceptual quality.

C.4 Computational Efficiency Profiling

We profiled each model using one-second inputs at 16 kHz with batch size 1 and 32-bit floating-point precision (FP32). CPU and GPU inference were measured on an Intel Xeon Gold 6530 processor and an NVIDIA GeForce RTX 5090 GPU, respectively. Runtime measurements used 500–1,000 repeated inference runs per model. Table 2 reports total-model parameter counts, MACs for one input, CPU and GPU inference times, and the corresponding memory usage.

[6] https://github.com/google/visqol

[7] https://huggingface.co/Wespeaker/wespeaker-voxceleb-resnet34-LM

[8] https://huggingface.co/Wespeaker/wespeaker-cnceleb-resnet34-LM

D PER-SOURCE COMPARISON WITH SINGLE-TARGET BC BASELINES

D.1 COMPARISON PROTOCOL

Table 6 separates guided-speaker extraction from unguided-speaker recovery on VibroMix-2Mix. The BC-guided baselines retain their original single-target configurations. S1 denotes their directly estimated guided speaker, and S2 is obtained by subtracting S1 from the observed mixture, as in Equation 13. BAT predicts both sources directly. All entries are evaluated against the corresponding clean AC targets in the fixed guided–unguided order, and each Mean column is the arithmetic mean of S1 and S2. The input correspondence and metric definitions are given in Appendices C.2 and C.3.

Table 6: Per-source results for BC-guided baselines and BAT. BC-guided baselines use their original single-target configurations: S1 is the guided output and S2 is obtained by subtracting S1 from the mixture. BAT directly separates both sources and retains the fixed guided–unguided order.

Method	**SI-SNRi (dB) ↑**			**SDRi (dB) ↑**			**PESQ ↑**			**ViSQOL ↑**			**Speaker sim. (%) ↑**		
	S1	**S2**	**Mean**	**S1**	**S2**	**Mean**	**S1**	**S2**	**Mean**	**S1**	**S2**	**Mean**	**S1**	**S2**	**Mean**
BiNet	8.5	5.0	6.8	9.5	5.3	7.4	2.4	1.2	1.8	1.4	1.4	1.4	32.8	47.8	40.4
BAF-Net	7.4	3.5	5.4	8.4	4.4	6.4	2.1	1.2	1.7	2.0	1.5	1.8	48.0	46.4	47.2
FlowDense	13.9	7.9	10.9	14.2	8.2	11.2	3.0	1.7	2.4	2.3	1.7	2.0	54.3	52.5	53.4
DBMIF	13.1	7.5	10.3	13.5	7.8	10.7	2.9	1.6	2.3	2.0	1.6	1.8	50.2	51.6	50.9
Robust Fusion	17.1	8.9	13.0	17.2	9.0	13.1	3.0	1.8	2.4	**3.0**	2.0	2.5	**63.6**	55.8	59.7
BAT	**18.1**	**16.5**	**17.3**	**18.2**	**16.7**	**17.5**	**3.1**	**2.9**	**3.0**	**3.0**	**2.8**	**2.9**	62.8	**60.4**	**61.6**

D.2 GUIDED AND UNGUIDED RECOVERY

For the guided speaker, BAT achieved the highest SI-SNRi, SDRi, and PESQ in Table 6, reaching 18.1 dB, 18.2 dB, and 3.1, respectively. Relative to Robust Fusion, the strongest single-target baseline by guided SI-SNRi, these gains were 1.0 dB, 1.0 dB, and 0.1. Both systems obtained a guided ViSQOL of 3.0. Guided speaker similarity was higher for Robust Fusion (63.6%) than BAT (62.8%).

The larger gains occurred for the unguided speaker. BAT achieved 16.5 dB SI-SNRi, 16.7 dB SDRi, and 2.9 PESQ, exceeding the mixture-subtracted Robust Fusion estimates by 7.6 dB, 7.7 dB, and 1.1, respectively. Its unguided ViSQOL increased from 2.0 to 2.8, and speaker similarity increased from 55.8% to 60.4%, a gain of 4.6 percentage points. BAT thus led all five unguided-source metrics, with a larger waveform-separation improvement than for the guided source.

The output construction helps interpret this difference. Writing the mixture as $\boldsymbol{x} = \boldsymbol{s}_{\mathrm{g}} + \boldsymbol{s}_{\mathrm{u}} + \boldsymbol{n}$, where $\boldsymbol{s}_{\mathrm{g}}$ and $\boldsymbol{s}_{\mathrm{u}}$ are the clean speech signals and $\boldsymbol{n}$ is environmental noise, Equation 13 gives

$$\widehat{\boldsymbol{s}}_{\mathrm{u}} - \boldsymbol{s}_{\mathrm{u}} = \boldsymbol{n} - (\widehat{\boldsymbol{s}}_{\mathrm{g}} - \boldsymbol{s}_{\mathrm{g}}) . \tag{15}$$

The residual error thus depends on both mixture noise and guided-speaker estimation error. Even a perfect guided estimate leaves any mixture noise in S2. BAT learns a separate clean-speech estimate for the unguided speaker, so its second output is not fixed by mixture subtraction. The reported gains compare complete systems with different architectures and output constructions.

D.3 SOURCE BALANCE AND AGGREGATE PERFORMANCE

BAT reduced the SI-SNRi gap between S1 and S2 from 8.2 dB for Robust Fusion to 1.6 dB. The corresponding SDRi gap decreased from 8.2 dB to 1.5 dB, and the PESQ gap decreased from 1.2 to 0.2. These smaller gaps accompanied improvements in both sources, indicating more balanced recovery of the guided and unguided speakers.

Averaged over the two sources, BAT improved SI-SNRi, SDRi, and PESQ over the original single-target Robust Fusion configuration by 4.3 dB, 4.4 dB, and 0.6, respectively. Mean ViSQOL increased from 2.5 to 2.9, and mean speaker similarity increased from 59.7% to 61.6%. The larger unguided-source gains account for most of the improvement in mean separation quality. The 4.3 dB mean SI-SNRi margin here is specific to the single-target configuration with mixture subtraction. In Table 1, the two-output Robust Fusion adaptation achieved 14.7 dB, giving the 2.6 dB improvement reported in the main comparison.

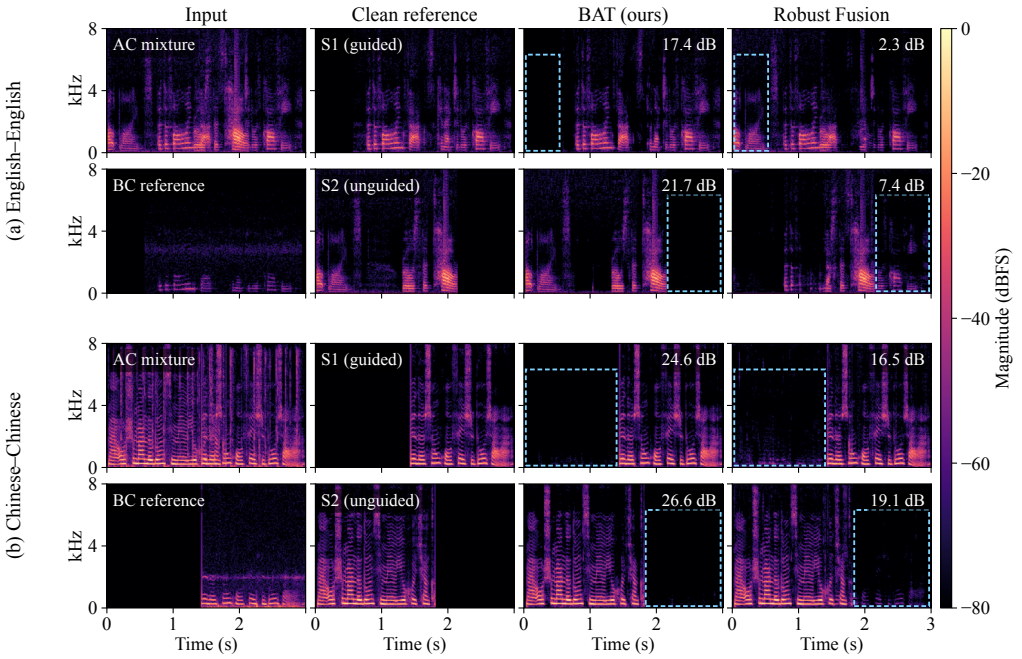


Figure 7: Qualitative spectrogram comparison on representative English (a) and Chinese (b) VibroMix-2Mix examples. From left to right, the columns show the input signals, clean references, BAT estimates, and Robust Fusion estimates. The input column contains the AC mixture and BC reference, while the clean-reference column contains the guided source $S1$ and unguided source $S2$. Cyan dashed boxes mark representative regions for visual comparison. The annotations in the estimate panels report per-source SI-SNRi values. All panels use the same log-magnitude scale.

E Qualitative Spectrogram Analysis

Figure 7 compares representative English and Chinese VibroMix-2Mix examples. The AC mixture preserves speech structure over a broad frequency range, whereas the synchronized BC reference contains predominantly low-frequency speech energy with weaker high-frequency components. This contrast motivates using BC as a complementary cue and aggregating its information across frequency bands.

Across both language examples, BAT follows the clean S1 and S2 references more closely than Robust Fusion. In the highlighted regions, BAT suppresses residual energy that is absent from the corresponding clean reference, while Robust Fusion retains more interference or spurious components. The per-example SI-SNRi values are higher for BAT in all four cases: 17.4 and 21.7 dB versus 2.3 and 7.4 dB for the English guided and unguided sources, and 24.6 and 26.6 dB versus 16.5 and 19.1 dB for the Chinese sources. These examples provide qualitative evidence for the benefits of reliability-aware fusion and adaptive mask reconstruction.

F CHRF Component Ablation

To isolate the contributions of cross-band context and local reliability weighting, we removed the dilated depthwise frequency branches or the per-time–frequency reliability field from CHRF. Removing harmonic context reduced SI-SNRi and SDRi by 0.2 dB, while removing reliability gating reduced both metrics by 0.3 dB and PESQ by 0.1. The corresponding results are reported in Table 7.

Table 7: CHRF component ablation. Params: fusion-module parameter counts; –: unreported.

Variant	SI-SNRi (dB) ↑	SDRi (dB) ↑	PESQ ↑	Params (K) ↓
Without harmonic context	17.1	17.3	**3.0**	47.0
Without reliability gating	17.0	17.2	2.9	47.6
CHRF (BAT)	**17.3**	**17.5**	**3.0**	47.6

G CAM Boundary-Design Ablation

We evaluated two boundary variants in addition to Full CAM. The first omitted causal historical centering of the agreement and BC-activity cues. The second predicted independent real and imaginary bounds for each output instead of sharing them, while retaining the source-specific mask logits. Removing historical centering reduced SI-SNRi by 0.9 dB, while assigning independent output ranges reduced it by 0.8 dB. Both variants lost 0.8 dB SDRi and 0.1 PESQ. The results support the combination of historical centering and shared bounds used in CAM.

Table 8: CAM boundary-design ablation.

Variant	SI-SNRi (dB) ↑	SDRi (dB) ↑	PESQ ↑	Params (K) ↓
Without historical centering	16.4	16.7	2.9	1.2
Independent output ranges	16.5	16.7	2.9	1.3
Full CAM (BAT)	**17.3**	**17.5**	**3.0**	1.2

H SSTC Ablation Details

The following studies isolated the separator choices while keeping the CHRF module, CAM decoder, loss, optimizer, training schedule, and VibroMix split fixed. The tables follow the averaging convention stated in the main text. The depth study changed the number of stacked SSTC modules; the recurrent-operator study changed the recurrent sublayer in both temporal and frequency operators; and the attention study removed the multi-head self-attention sublayers while retaining the recurrent paths. The four-module GRU separator with multi-head self-attention is BAT.

Table 9: Effect of SSTC depth. Params and MACs refer to the complete model.

SSTC depth	SI-SNRi (dB) ↑	SDRi (dB) ↑	PESQ ↑	Params (M) ↓	MACs (G) ↓
1 module	15.6	15.9	2.8	**0.5**	**9.3**
8 modules	**17.9**	**18.1**	**3.0**	1.1	23.0
4 modules (BAT)	17.3	17.5	**3.0**	0.7	15.1

Increasing depth from one to four modules raised SI-SNRi by 1.7 dB and SDRi by 1.6 dB, while increasing the total cost from 0.5M/9.3G to 0.7M/15.1G. Eight modules obtained a further 0.6 dB gain on both separation metrics, but required 1.1M parameters and 23.0G MACs. The four-module configuration is therefore the lower-cost operating point used by BAT, while the eight-module result quantifies the additional quality available with a larger separator.

The LSTM variant gave the highest SI-SNRi and SDRi in this subset, exceeding BAT by 0.1 dB on each metric, but it increased the total cost to 0.8M parameters and 16.6G MACs. BAT retained the same rounded PESQ as LSTM with lower cost, while the RNN was cheaper but lost 0.5 dB SI-SNRi and 0.5 dB SDRi relative to BAT.

Removing multi-head self-attention lowered SI-SNRi by 0.7 dB, SDRi by 0.7 dB, and PESQ by 0.1. The comparison isolated the contribution of the attention interactions because the temporal and frequency recurrent paths remained unchanged.

Table 10: SSTC recurrent operators. Params and MACs refer to the complete model.

Recurrent operator	SI-SNRi (dB) ↑	SDRi (dB) ↑	PESQ ↑	Params (M) ↓	MACs (G) ↓
RNN	16.8	17.0	2.9	**0.6**	**13.2**
LSTM	**17.4**	**17.6**	**3.0**	0.8	16.6
GRU (BAT)	17.3	17.5	**3.0**	0.7	15.1

Table 11: Contribution of multi-head self-attention within SSTC.

Attention setting	SI-SNRi (dB) ↑	SDRi (dB) ↑	PESQ ↑
Without self-attention	16.6	16.8	2.9
With self-attention (BAT)	**17.3**	**17.5**	**3.0**

I Refiner Ablation

To assess the contribution of local feature refinement, we removed the Refiner after the SSTC stack while retaining CHRF, SSTC, CAM, and the training configuration. Table 12 compares this variant with the full separator using the same two-source evaluation protocol.

Table 12: Contribution of the Refiner after the SSTC stack.

Separator variant	SI-SNRi (dB) ↑	SDRi (dB) ↑	PESQ ↑
Without Refiner	17.0	17.2	2.9
Full separator (BAT)	**17.3**	**17.5**	**3.0**

Removing the Refiner reduced mean SI-SNRi and SDRi by 0.3 dB each and PESQ by 0.1. This result supports local feature refinement after the SSTC stack.