BIGMNET: BIDIRECTIONAL REPRESENTATIONS WITH PREDICTIVE SUPERVISION FOR EEG-GUIDED TARGET SPEAKER EXTRACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain-assisted target speaker extraction aims to recover the attended target speech from competing speech, using electroencephalography (EEG) as the sole auxiliary cue. However, architectures relying on local post-fusion mappings and waveform supervision lack explicit coupling between temporal organization and direct supervision of target-speech dynamics, while uniform elementwise updates overlook the distinct roles of internal matrices and other parameters. To address these issues, this paper proposes Bidirectional Griffin Adapter with Multi-Frame Prediction Network (BiGMNet), an EEG-conditioned encoder–mask–decoder framework for target speaker extraction. The method introduces three coordinated components: the Bidirectional Griffin Adapter (BiGriffinAdapter) for bidirectional temporal refinement, direction-aware multi-frame prediction (DA-MFP) for supervising directional intermediate states, and Hybrid Muon–AdamW optimization for parameter-role-aware updates. BiGriffinAdapter combines bidirectional gated recurrence and local attention to refine fused representations for mask estimation, and training-only DA-MFP supervises its pre-concatenation forward and backward states through sequential prediction of future and past continuous clean-speech latents, respectively. Hybrid Muon–AdamW approximately orthogonalizes updates to selected internal matrices and applies elementwise adaptive updates to the remaining parameters according to their computational roles. This paper evaluates subject-specific models on MM-AAD and AVED with ten participants per dataset through EEG-only baseline comparisons and component ablations. BiGMNet achieves 9.76 and 10.90 dB SI-SDR, exceeding IFENet's reported scores by 1.57 and 2.14 dB, respectively, with 12.31 G MACs, 15.5% fewer than IFENet. These results support directly supervising EEG-conditioned temporal states to improve target speaker extraction, with the optimization rule chosen according to parameter roles.

1 INTRODUCTION

In complex acoustic environments, humans can selectively attend to a target speaker despite competing talkers. Target speaker extraction (TSE) aims to emulate this ability by recovering target speech from a multi-talker mixture (Mesgarani & Chang, 2012; Ephrat et al., 2018; Žmolíková et al., 2019; Xu et al., 2020; Pan et al., 2022; Liu et al., 2023; Lin et al., 2023; Hao et al., 2024; Tao et al., 2025; Kim et al., 2025). Electroencephalography (EEG) provides neural signals associated with auditory attention, from which attention to competing speech can be decoded (O'Sullivan et al., 2015). This ability makes EEG attention cues a natural conditioning signal for EEG-guided TSE.

In a typical EEG-guided TSE system, the acoustic mixture and EEG are encoded into an acoustic representation and an EEG target-attention representation (Han et al., 2019; Ceolini et al., 2020), respectively. Frame-level temporal alignment and cross-modal fusion then produce a fused audio–EEG representation for target-mask estimation and waveform reconstruction. This fused representation must retain attentional information and acoustic detail, while its target-speech temporal structure is critical to separating the attended speaker from competing speech. Recent EEG-TSE systems (Hosseini et al., 2021; 2022; Zhang et al., 2023; Fan et al., 2024; Pan et al., 2024a; Fan et al., 2025a) have improved modality encoding, cross-modal interaction, and representation alignment through

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.

advanced acoustic modeling, attention mechanisms, and EEG–speech alignment strategies. These advances improve representation learning and fusion quality, but do not directly address target-speech temporal structure after fusion.

Two related gaps therefore remain before target-mask estimation. First, standard audio–EEG fusion provides cross-modal conditioning but does not itself explicitly refine the fused audio–EEG representation through bidirectional temporal modeling. Second, waveform reconstruction only indirectly constrains target-speech temporal structure, while attention supervision focuses on attentional selection. Even when bidirectional refinement constructs forward and backward directional states, architectural directionality alone does not ensure direction-specific target-speech temporal structure. Explicit supervision is therefore needed to impose future- and past-oriented target-speech temporal structure on the forward and backward directional states before combination.

To address these gaps, we propose the **B**idirectional **G**riffin Adapter with **M**ulti-Frame Prediction **N**etwork (BiGMNet). Bidirectional Griffin Adapter (BiGriffinAdapter) performs post-fusion bidirectional temporal refinement, producing forward and backward directional states and a refined fused representation for mask estimation. Direction-Aware Multi-Frame Prediction (DA-MFP) supervises these states before combination: the forward directional state predicts future clean-target latent frames, whereas the backward directional state predicts past clean-target latent frames. During training, DA-MFP sequentially predicts multiple clean-target latent frames under clean-target guidance, while the auxiliary prediction branches are removed at inference. Hybrid Muon–AdamW is further adopted to optimize heterogeneous parameter groups with role-aware update rules.

The contributions of this paper are summarized as follows:

- This paper proposes BiGriffinAdapter for post-fusion bidirectional temporal refinement, which models forward and backward contexts to enhance the fused audio–EEG representation before target-mask estimation.
- This paper proposes Direction-Aware Multi-Frame Prediction (DA-MFP), which supervises the forward and backward directional states by predicting future and past clean-target latent frames, respectively.
- This paper introduces Hybrid Muon–AdamW, a parameter-role-aware optimization strategy that applies different update rules to heterogeneous parameter groups according to their structural roles.

2 Related Works

EEG-guided target speaker extraction (EEG-TSE) exploits neural responses associated with auditory attention to recover the attended speaker from multi-talker mixtures. Early approaches such as BISS (Ceolini et al., 2020) first reconstruct the attended-speech envelope from EEG and then use it to guide speech extraction. Later end-to-end systems directly incorporate EEG into the extraction network. BESD and U-BESD (Hosseini et al., 2021; 2022) introduce EEG-conditioned feature modulation, while BASEN (Zhang et al., 2023) strengthens audio–EEG interaction through cross-attention. NeuroHeed (Pan et al., 2024a) further learns an EEG-derived neural attractor for target extraction, and NeuroHeed+ (Pan et al., 2024b) introduces auxiliary auditory-attention detection to alleviate speaker confusion.

Recent studies increasingly focus on enhanced temporal modeling and cross-modal representation learning. MSFNet (Fan et al., 2024) combines multi-scale acoustic features with graph-based EEG representations, while IFENet (Fan et al., 2025b) employs bidirectional state-space modeling for speech representation together with dedicated EEG encoding. M3ANet (Fan et al., 2025a) further addresses EEG–speech temporal misalignment through multi-scale modeling and contrastive alignment. GALENet (Fan et al., 2026) improves audio–EEG interaction using gated cross-attention while combining local speech modeling with efficient global sequence modeling. TIDENet (Lv et al., 2026) further investigates the limitations of EEG–speech alignment and representation discrimination by introducing trainable EEG interpolation and structure-sharing dual-path encoders, enabling adaptive temporal alignment and more effective extraction of target-related and complementary representations from both modalities. Collectively, these works advance EEG-TSE through improved modality representation learning, acoustic modeling, multimodal alignment, and feature fusion.

However, existing methods mainly focus on modality encoding, cross-modal fusion, and alignment strategies, while comparatively less attention has been paid to the temporal refinement of the fused representation before mask estimation. Moreover, bidirectional temporal modules are generally optimized through the final extraction objective without explicitly exploiting direction-specific supervision of the corresponding latent states. These observations motivate our post-fusion bidirectional refinement and direction-aware predictive learning framework.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND MODEL OVERVIEW

BiGMNet couples EEG-conditioned temporal refinement, directional prediction, and parameter-role-based optimization for target speaker extraction. Figure 1 shows the extraction backbone and training-only prediction branches. We model a multi-speaker mixture as

$$x = \sum_{i=1}^{J} s_i + n_{\text{env}}, \tag{1}$$

where J is the number of speakers in the mixture (typically $J = 2$ in target speaker extraction scenarios), s_i is a speaker waveform, and n_{env} is environmental interference. Given x and synchronized EEG e, BiGMNet estimates the attended speaker waveform $s = s_a$ as $\hat{s}$, where a denotes the target speaker selected from the J speakers by the corresponding EEG cue. Clean speech s supplies training supervision.

The model first extracts acoustic and neural representations from the speech mixture and synchronized EEG signals. These modality-specific representations are subsequently integrated through audio–EEG fusion to establish a conditioned representation for target speaker extraction. BiGriffinAdapter then performs temporal refinement by combining bidirectional gated recurrence with local attention, producing enhanced sequential features for mask estimation. A full-sequence mask estimator predicts the separation mask, which is applied to the mixture embedding to preserve the target speaker components. The decoder finally reconstructs the estimated waveform from the masked audio representation.

During training, DA-MFP supervises directional states before concatenation through sequential prediction of future and past clean-speech latents from the shared encoder with stop-gradient targets. Hybrid Muon–AdamW assigns update rules by parameter role: Muon approximately orthogonalizes selected matrix updates, while AdamW adapts the remaining parameters elementwise. Waveform SI-SDR and DA-MFP jointly optimize the model. During inference, only the extraction backbone is retained, requiring the mixture speech and synchronized EEG as inputs.

3.2 EEG ENCODER

The EEG encoder converts the input EEG segment e into a neural conditioning sequence E.

$$E = \text{EEGEncoder}(e) \in \mathbb{R}^{C \times T}, \tag{2}$$

where C and T denote the channel dimension and temporal length of the encoded EEG representation, respectively. It first applies a temporal convolution over overlapping windows to extract local electrode patterns while reducing temporal resolution. The extracted features are projected into frame-level embeddings and processed by a residual Transformer (Vaswani et al., 2023; He et al., 2016) to model temporal dependencies across the EEG sequence. The Transformer incorporates self-attention and feed-forward transformations with residual normalization, while an encoder-level residual connection preserves local convolutional information. The resulting EEG representation is retained on the encoder temporal grid and fused with acoustic features to provide neural guidance for target speaker extraction.

3.3 SPEECH ENCODER AND AUDIO–EEG FUSION

The speech encoder learns a time-domain analysis basis for mixture x, producing a nonnegative acoustic embedding:

$$X = \text{ReLU}(\text{Conv1D}(x)). \tag{3}$$

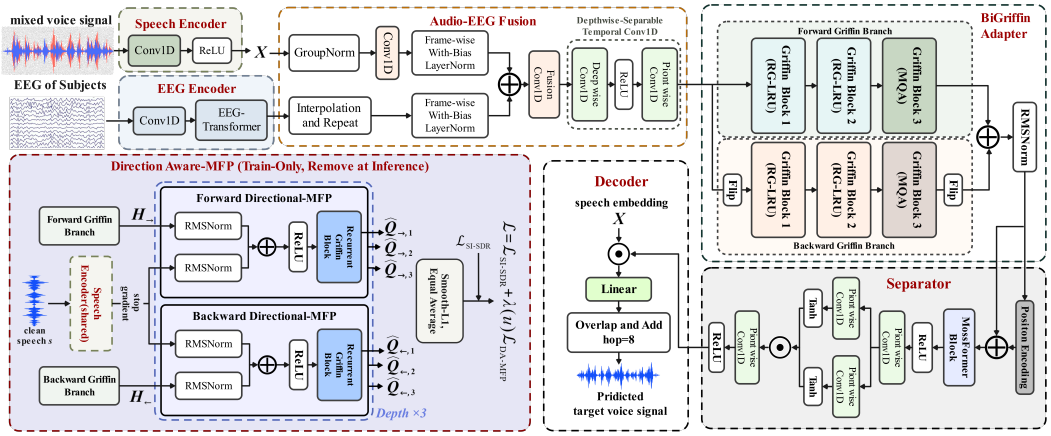


Figure 1: Overall pipeline of BiGMNet. Mixture and EEG representations are fused and temporally refined by BiGriffinAdapter for mask estimation and waveform reconstruction. DA-MFP introduces auxiliary training supervision by predicting clean-speech latents from BiGriffinAdapter's forward and backward directional states. $\odot$ and $\oplus$ indicate element-wise multiplication and feature concatenation.

The same encoder parameters are shared for the clean target representation. The resulting acoustic embedding X is retained as the basis representation for subsequent mask-based reconstruction while also providing the acoustic stream for multimodal fusion.

Fusion aligns the EEG and acoustic representations into a unified feature space. We linearly interpolate E to the acoustic temporal grid and match its feature dimension through repetition and, when needed, truncation. In parallel, X passes through single-group normalization and a pointwise bottleneck to project the acoustic representation into the same feature space as the aligned EEG representation. Separate layer normalizations then normalize the aligned EEG and acoustic features with modality-specific parameters. The normalized streams are concatenated and projected through a pointwise convolution. A depthwise temporal convolution followed by ReLU incorporates local temporal context, and a subsequent pointwise projection further mixes channel responses.

The fusion module combines the aligned EEG and acoustic features to produce an EEG-conditioned acoustic representation:

$$H = \mathrm{AudioEEGFusion}(X, E) \in \mathbb{R}^{D \times T_X}, \tag{4}$$

where $D = 64$ is the acoustic and fused feature width and T_X is the number of frames on the acoustic temporal grid. This EEG-conditioned representation enters BiGriffinAdapter for temporal refinement before mask estimation.

3.4 BiGriffinAdapter: Post-Fusion Temporal Refinement

The audio–EEG fusion stage integrates multimodal cues through local temporal mixing, but target-mask estimation further requires modeling how EEG-conditioned acoustic representations evolve over time. BiGriffinAdapter (De et al., 2024) therefore performs temporal refinement on the fused representation H using bidirectional contextual modeling.

The adapter first projects H with a pointwise layer and transposes it into frame-major order, obtaining a sequence $Z \in \mathbb{R}^{T_X \times D}$. Two independently parameterized directional stacks process Z from opposite temporal directions. The forward stack scans frames chronologically, while the backward stack operates on a reversed sequence:

$$\begin{aligned} H_{\rightarrow} &= \mathcal{G}_{\mathrm{forward}}(Z), \\ H_{\leftarrow} &= \mathrm{TemporalReverse}\left(\mathcal{G}_{\mathrm{backward}}(\mathrm{TemporalReverse}(Z))\right). \end{aligned} \tag{5}$$

After reversing back, $H_{\rightarrow}$ and $H_{\leftarrow}$ are aligned in chronological order, with each containing T_X frames and D channels.

Each directional stack consists of two recurrent blocks followed by a local multi-query attention block (Shazeer, 2019; Beltagy et al., 2020; Su et al., 2024). The recurrent blocks use causal depth-wise convolution and RG-LRU-based state (Orvieto et al., 2023) updates to accumulate directional temporal context. The following attention block performs local feature selection with eight query heads, one shared key/value head, and a causal window of 64 frames. Thus, each direction combines recurrent historical information with local adaptive refinement (Appendix A).

The forward and backward representations provide complementary temporal contexts. Their aligned states are concatenated and normalized to obtain the bidirectional representation:

$$H_{\mathrm{bi}} = \mathrm{RMSNorm}(\mathrm{Concat}(H_{\rightarrow}, H_{\leftarrow})). \tag{6}$$

Finally, an output projection maps the $2D$-dimensional representation back to the original channel dimension and injects the temporal refinement through a residual connection:

$$\widetilde{H} = H + P_{\mathrm{out}}(H_{\mathrm{bi}}^{\top}). \tag{7}$$

Zero-initializing P_{out} preserves the original fused representation at initialization, enabling stable residual refinement. During inference, BiGriffinAdapter remains active and produces the refined representation $\widetilde{H}$ for subsequent mask estimation.

3.5 Mask Estimation and Speech Decoder

The refined representation $\widetilde{H}$ provides EEG-conditioned temporal context for mask estimation. A full-sequence MossFormer-style estimator (Zhao & Ma, 2023; Zhao et al., 2024) further processes $\widetilde{H}$ with positional encoding, FLASH attention, and gated FSMN transformations to generate a non-negative mask M aligned with the acoustic embedding X.

The estimated mask is applied to the mixture representation to extract the target latent representation:

$$\begin{aligned} \hat{S} &= X \odot M, \\ \hat{s} &= \mathrm{Decoder}(\hat{S}). \end{aligned} \tag{8}$$

The decoder reconstructs the target waveform through a learned synthesis basis and overlap-and-add operation, producing the final estimated speech waveform $\hat{s}$ from the masked latent representation $\hat{S}$.

3.6 Direction-Aware Multi-Frame Prediction

The waveform-level SI-SDR objective directly optimizes the final speech reconstruction, but provides limited supervision for intermediate temporal representations. To provide additional temporal guidance, we introduce DA-MFP (DeepSeek-AI et al., 2026; 2025; Liu et al., 2025), an auxiliary latent prediction objective that learns clean speech representations from the directional states of BiGriffinAdapter.

Given clean speech s, the shared speech encoder produces a latent sequence $Q \in \mathbb{R}^{T_X \times D}$. For two scanning directions, we define $Q_{\rightarrow} = Q$ and $Q_{\leftarrow} = \mathrm{TemporalReverse}(Q)$. BiGriffinAdapter provides two directional states, including a forward state $H_{\rightarrow}$ and a backward state $H_{\leftarrow}$, which are used by DA-MFP predictors to recover directional clean speech latent trajectories.The latent target Q_d is detached when used for auxiliary supervision.

Inspired by sequential multi-token prediction, DA-MFP progressively predicts long-range latent segments at different depths. For direction $d \in \{\rightarrow, \leftarrow\}$, the initial prediction states are defined as

$$\begin{aligned} R_{\rightarrow,0} &= H_{\rightarrow}, \\ R_{\leftarrow,0} &= \mathrm{TemporalReverse}(H_{\leftarrow}), \end{aligned} \tag{9}$$

and the predictor input at depth k is constructed as

$$V_{d,k} = (R_{d,k-1}[1 : T_X - k], Q_d[k : T_X - 1]), \quad 1 \le k \le K < T_X. \tag{10}$$

The predictor estimates the corresponding clean latent sequence:

$$\begin{aligned} R_{d,k} &= \mathrm{DirectionalPredictor}_{d,k}(V_{d,k}), \\ \hat{Q}_{d,k} &= \mathrm{SharedHead}(R_{d,k}), \end{aligned} \tag{11}$$

where $\mathrm{DirectionalPredictor}_{d,k}$ denotes the depth-specific directional predictor and $\mathrm{SharedHead}$ is the shared latent projection head. The predictor input and target sequence contain consistent temporal lengths for each prediction depth.

For each direction, the auxiliary objective is defined as the Smooth-L1 distance between predicted and target directional latent sequences:

$$\mathcal{L}_d = \frac{1}{K}\sum_{k=1}^{K} \mathrm{mean}\left[\mathrm{SL1}(\hat{Q}_{d,k} - Q_d[k+1:T_X])\right], \qquad \mathcal{L}_{\mathrm{DA-MFP}} = \frac{\mathcal{L}_{\rightarrow} + \mathcal{L}_{\leftarrow}}{2}. \tag{12}$$

The complete model is optimized by jointly considering waveform-level extraction and latent prediction:

$$\mathcal{L} = \mathcal{L}_{\mathrm{SI-SDR}} + \lambda(u)\mathcal{L}_{\mathrm{DA-MFP}}, \tag{13}$$

where

$$\mathcal{L}_{\mathrm{SI-SDR}} = -\,\mathrm{SI\text{-}SDR}(\hat{s}, s). \tag{14}$$

To stabilize multi-objective optimization, the DA-MFP weight is gradually increased during early training:

$$\lambda(u) = 0.05\min(u/U_w, 1), \tag{15}$$

where u denotes the current training iteration and U_w denotes the number of warm-up iterations. DA-MFP is only introduced during training and is removed during inference, therefore introducing no additional inference overhead.

3.7 Hybrid Muon–AdamW Optimization and Inference

Hybrid Muon–AdamW optimization addresses the heterogeneous optimization behaviors of different parameter types in our framework. We assign Muon (Liu et al., 2025; Jordan et al., 2024) to selected internal linear transformations involved in feature mixing, while AdamW (Kingma & Ba, 2017; Loshchilov & Hutter, 2019) updates convolutional layers, normalization parameters, decoder components, and DA-MFP modules. Both parameter groups receive gradients from the joint objective through a single backward pass, followed by global gradient clipping with a norm of 1.0. Muon applies Nesterov momentum with FP32 Newton–Schulz orthogonalization (Bernstein & Newhouse, 2024) for structured update directions, whereas AdamW provides adaptive element-wise updates.

Both optimizers use decoupled weight decay with a shared warmup and cosine-decay learning-rate schedule. The checkpoint achieving the lowest validation SI-SDR loss is selected. During inference, training-only DA-MFP components are removed, while the backbone performs unchanged estimation from mixture audio and synchronized EEG inputs.

4 Experimental Setup

We evaluate our method on AVED and MM-AAD (Zhang et al., 2024; Fan et al., 2025c), two Mandarin audio–EEG datasets.

AVED. We use the 10 audio-only participants from AVED's cohort of 20 normal-hearing listeners. Each completed 16 trials of 152 s, with stories narrated by a male and a female speaker delivered to opposite ears. Participants were instructed to continuously attend to a designated speaker. Audio was sampled at 44.1 kHz, and synchronized 32-channel EEG was recorded at 1 kHz.

MM-AAD. MM-AAD was designed for auditory attention decoding. We randomly select 10 of its 50 participants and use their speech signals and synchronized EEG recordings. Each participant completed 20 trials, with 32-channel EEG recorded at 2,048 Hz.

Preprocessing and Protocol. For both datasets, EEG signals are notch filtered at 50 Hz, bandpass filtered between 1–50 Hz, downsampled to 128 Hz, and transformed using common-average referencing. Audio signals are resampled to 14.7 kHz. For AVED, two speaker streams are RMS-normalized and linearly mixed, with the attended speaker serving as the target signal. Each trial

is divided into non-overlapping training, validation, and test splits (80%/10%/10%), followed by aligned 2-s EEG–audio segmentation. Models are trained separately for each participant, and the reported results are averaged across participants.

Training. Models are implemented in PyTorch and trained on a single NVIDIA GeForce RTX 3090 GPU with batch size 8 for 400 epochs on AVED and 300 on MM-AAD; other settings are unchanged across datasets. The full model uses hybrid Muon–AdamW (Section 3.7); the AdamW ablation configurations use AdamW. External baseline scores are taken from the literature. The learning rate peaks at 1×10^{-4} after linear warmup over the first 5% of steps, followed by cosine decay.

Metrics. We report signal-to-distortion ratio (SDR), scale-invariant signal-to-distortion ratio (SI-SDR)(Roux et al., 2019; Vincent et al., 2006), perceptual evaluation of speech quality (PESQ)(Rix et al., 2001), short-time objective intelligibility (STOI), and extended short-time objective intelligibility (ESTOI)(Taal et al., 2011), covering signal fidelity, perceptual quality, and intelligibility. SDR compares target and distortion energies, while SI-SDR is invariant to global amplitude scaling. SDR and SI-SDR are measured in decibels (dB); higher values are better for all five metrics.

5 RESULTS

5.1 COMPARISON WITH BASELINE MODELS

Comparison with baseline methods. We first evaluate whether BiGMNet improves EEG-guided target speaker extraction compared with existing EEG-only approaches. The comparison includes BASEN, MSFNet, GALENet, M3ANet, IFENet, and TIDENet on MM-AAD and AVED (Table 1).

Table 1: EEG-TSE results on MM-AAD and AVED. Baseline results are from the literature or our reproductions. Our results are averaged over ten subjects; SDR, SI-SDR (dB), and PESQ are reported as mean±SD, while STOI and ESTOI are reported as mean values. Mixture metrics are single values; dashes denote unavailable or not applicable entries.

Dataset	Model	SDR↑	SI-SDR↑	PESQ↑	STOI↑	ESTOI↑	MACs (G)↓
MM-AAD	Mixture	1.91	1.71	1.63	0.81	0.67	–
	BASEN	8.75±1.16	7.95±1.27	1.98±0.13	0.87	0.75	16.12
	MSFNet	8.05±0.83	7.08±0.86	1.88±0.07	0.85	0.73	76.19
	GALENet	8.63±0.89	7.60±0.95	1.89±0.11	0.86	0.75	39.80
	M3ANet	8.92±0.95	7.90±0.99	1.94±0.07	0.86	0.74	53.54
	IFENet	9.03±0.27	8.19±0.36	2.04±0.05	0.88	0.78	14.57
	TIDENet	10.37±0.42	9.71±0.48	2.16±0.07	0.90	0.81	57.67
	BiGMNet (ours)	**10.62±0.70**	**9.76±0.67**	**2.18±0.18**	**0.91**	**0.81**	**12.31**
AVED	Mixture	1.83	1.43	1.55	0.80	0.68	–
	BASEN	7.34±2.12	6.58±2.34	1.76±0.16	0.85	0.72	16.12
	MSFNet	7.72±1.28	6.78±1.40	1.71±0.11	0.85	0.71	76.19
	GALENet	8.40±1.39	7.45±1.54	1.72±0.12	0.86	0.73	39.80
	M3ANet	8.88±1.47	7.84±1.62	1.84±0.09	0.87	0.74	53.54
	IFENet	9.62±1.13	8.76±1.11	1.91±0.12	0.88	0.77	14.57
	TIDENet	11.21±1.86	10.74±1.92	1.97±0.13	0.88	0.79	57.67
	BiGMNet (ours)	**11.68±2.00**	**10.90±2.01**	**1.98±0.19**	**0.88**	**0.80**	**12.31**

On MM-AAD, BiGMNet achieves 10.62 dB SDR and 9.76 dB SI-SDR, providing consistent improvements over all compared baselines. Compared with TIDENet, the strongest baseline in this setting, BiGMNet further improves SDR and SI-SDR by 0.25 dB and 0.05 dB, respectively. The improvements are also reflected in perceptual and intelligibility metrics, where BiGMNet obtains the highest PESQ (2.18), STOI (0.91), and ESTOI (0.81).

To examine whether the improvement generalizes beyond MM-AAD, we further evaluate BiGMNet on AVED, which contains a different EEG-guided extraction scenario. BiGMNet achieves 11.68 dB SDR and 10.90 dB SI-SDR, exceeding TIDENet by 0.47 dB and 0.16 dB, respectively. Meanwhile, BiGMNet maintains improvements in PESQ and ESTOI, indicating that the proposed architecture

provides more effective target-speech representation across datasets rather than being limited to a specific evaluation setting.

Computational efficiency. Besides separation quality, we compare the computational cost of the baseline models. BiGMNet requires 12.31 G MACs, the lowest among the compared methods. Compared with IFENet, it reduces MACs by 15.5% while improving SDR and SI-SDR by 1.59 and 1.57 dB, respectively. This comparison indicates that the performance gains are achieved without increasing computational cost.

5.2 Ablation Study

Table 2 reports ten-subject averages on MM-AAD and AVED. We first compare AVConv (the original post-fusion 1×1 convolution) with BiGriffinAdapter under AdamW without DA-MFP. Fixing BiGriffinAdapter, we then cross DA-MFP on/off with AdamW/Hybrid Muon–AdamW in a 2×2 design. Appendices B.1, B.2, and B.3 compare fusion placement, AVConv ordering, and prediction depth on MM-AAD. BiGMNet achieves the highest mean SDR, SI-SDR, PESQ, and ESTOI on both datasets; AVED STOI peaks at 0.89 with Hybrid Muon–AdamW without DA-MFP, versus 0.88 for BiGMNet.

Table 2: Temporal-refiner and training ablations, averaged over ten subjects per dataset. SDR and SI-SDR (dB) and PESQ report mean ± SD; STOI and ESTOI report means. ✓/× indicate enabled/disabled components; the all-enabled row is BiGMNet. Bold cells identify the highest mean per dataset.

Dataset	BiGriffin Adapter	DA-MFP	Hybrid Muon–AdamW	SDR↑	SI-SDR↑	PESQ↑	STOI↑	ESTOI↑
MM-AAD	×	×	×	9.78±0.62	8.94±0.62	1.94±0.17	0.88	0.78
	✓	×	×	10.03±0.65	9.19±0.65	2.04±0.14	0.89	0.79
	✓	✓	×	10.45±0.59	9.60±0.60	2.09±0.11	0.90	0.80
	✓	×	✓	10.30±0.55	9.45±0.53	2.05±0.17	0.89	0.80
	✓	✓	✓	**10.62±0.70**	**9.76±0.67**	**2.18±0.18**	**0.91**	**0.81**
AVED	×	×	×	9.62±2.10	8.84±2.09	1.77±0.18	0.88	0.76
	✓	×	×	9.74±2.14	8.96±2.14	1.75±0.13	0.87	0.75
	✓	✓	×	9.79±1.96	9.01±1.93	1.85±0.18	0.88	0.77
	✓	×	✓	10.30±1.80	9.52±1.81	1.87±0.14	**0.89**	0.79
	✓	✓	✓	**11.68±2.00**	**10.90±2.01**	**1.98±0.19**	0.88	**0.80**

Post-fusion temporal refinement. We first isolate the effect of BiGriffinAdapter under AdamW with DA-MFP disabled. Replacing AVConv with BiGriffinAdapter improves SDR and SI-SDR by 0.25 dB on MM-AAD and 0.12 dB on AVED. MM-AAD also shows higher PESQ, STOI, and ESTOI, whereas AVED exhibits small decreases in PESQ, STOI, and ESTOI. These results show that bidirectional post-fusion refinement consistently benefits separation, while its perceptual and intelligibility effects are dataset dependent.

Factorial analysis of DA-MFP and optimization. With BiGriffinAdapter fixed, the remaining four configurations form a 2×2 factorial ablation over DA-MFP and optimizer. Enabling DA-MFP improves SI-SDR, PESQ, and ESTOI in every matched optimizer comparison, with SI-SDR gains of 0.41/0.05 dB on MM-AAD/AVED under AdamW and 0.31/1.38 dB under Hybrid Muon–AdamW. DA-MFP supervises the pre-concatenation directional states, using the forward branch to predict future clean-speech latents and the backward branch to predict past latents, thereby complementing the waveform-level objective. Independently, switching from AdamW to Hybrid Muon–AdamW improves SI-SDR by 0.26/0.56 dB without DA-MFP and 0.16/1.89 dB with DA-MFP. The hybrid optimizer applies Muon only to selected internal linear matrices while retaining AdamW for the remaining parameter groups. The strongest gains occur on AVED when both training components are enabled; STOI is the only local exception, decreasing from 0.89 to 0.88 after enabling DA-MFP under Hybrid Muon–AdamW.

Full configuration. Combining BiGriffinAdapter, DA-MFP, and Hybrid Muon–AdamW yields BiGMNet, which improves SI-SDR by 0.57 dB on MM-AAD and 1.94 dB on AVED over BiGrif-

finAdapter trained with AdamW and without DA-MFP. The full model achieves the highest mean SDR, SI-SDR, PESQ, and ESTOI on both datasets. Together, these results support the intended roles of the two main methodological designs: BiGriffinAdapter addresses insufficient bidirectional temporal refinement after audio–EEG fusion, while DA-MFP explicitly imposes future- and past-oriented target-speech supervision on the directional states before their combination. AVED STOI is the only exception, where Hybrid Muon–AdamW without DA-MFP reaches 0.89 compared with 0.88 for BiGMNet.

Sample-level distributions. Figure 2 illustrates the SI-SDR distributions of Subject 1 under different DA-MFP configurations. DA-MFP consistently improves the median SI-SDR under both optimization strategies, increasing it from 9.64 to 10.29 dB with AdamW and from 9.84 to 10.46 dB with Hybrid Muon–AdamW. Meanwhile, the distributions remain partially overlapped, indicating that DA-MFP mainly shifts the overall performance tendency rather than uniformly improving every sample. These sample-level variations complement the ten-subject aggregated results reported in Table 2, which characterize performance from a different statistical perspective.

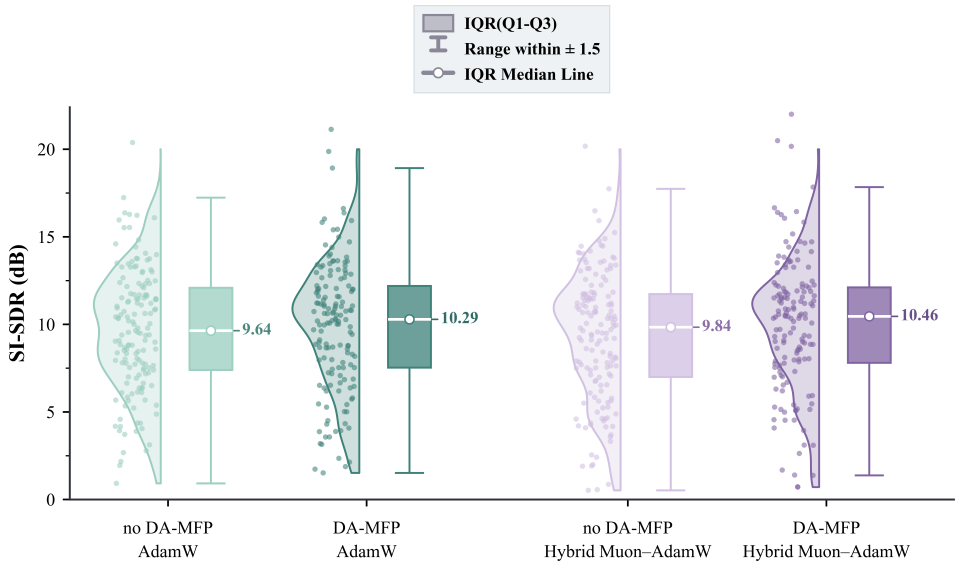


Figure 2: SI-SDR↑ distributions for nonnegative samples of MM-AAD Subject 1 across DA-MFP and optimizer settings; the rightmost configuration is BiGMNet. Boxes span Q1–Q3, white markers denote medians, and whiskers follow the $1.5\times$ IQR rule. Values above 20 dB appear in a compressed upper band.

6 Conclusion

This paper introduces a novel time-domain brain-guided target speaker extraction model, BiGMNet, to address two gaps after audio–EEG fusion: standard fusion lacks explicit bidirectional temporal refinement, and waveform reconstruction does not directly constrain future- and past-oriented target-speech structure in directional states. BiGriffinAdapter refines the fused representation, while training-only DA-MFP supervises forward and backward states before combination by predicting future and past clean-speech latents, respectively. Hybrid Muon–AdamW supports these components with parameter-role-aware updates. In subject-specific experiments on MM-AAD and AVED, BiGMNet improves separation over the compared EEG-only baselines, reaching 10.90 dB SI-SDR on AVED; ablations show SI-SDR gains from each design choice. These findings suggest that explicit temporal supervision can improve EEG-conditioned extraction. Future work will test adaptive prediction horizons, subject-independent conditioning, and learned parameter-role assignment while preserving compact inference.

AI USE STATEMENT

During the preparation of this manuscript, the authors utilized artificial intelligence (AI) tools as auxiliary resources to support the research and writing process. Specifically, AI assistance was employed for writing refinement and language polishing to improve the clarity, readability, and overall presentation of the manuscript. AI tools were also used for retrieval and discovery of related literature, helping the authors efficiently identify relevant research directions and background materials. In addition, AI assistance was used during the research ideation and execution process, including brainstorming potential approaches, organizing research plans, and supporting the exploration of technical solutions. Furthermore, AI tools were used to assist in checking mathematical derivations and verifying the consistency of mathematical expressions.

The use of AI tools was strictly limited to providing supportive suggestions and improving the efficiency of the research workflow. AI-generated content was carefully reviewed, revised, and validated by the authors before being incorporated into the manuscript. AI tools were not used to independently generate research claims, determine experimental conclusions, or replace the authors' scientific reasoning and decision-making process.

All analyses, methodological designs, experimental evaluations, interpretations of results, editorial judgments, and conclusions presented in this paper were independently conducted and determined by the authors. The authors assume full responsibility for the accuracy, originality, validity, and integrity of all content contained in this manuscript.

ETHICS STATEMENT

This study uses publicly available human EEG datasets, including AVED and MM-AAD, for evaluating EEG-based speech-related models. The MM-AAD dataset was collected from human participants who were informed of the experimental procedures and provided written informed consent before participation. The original dataset providers were responsible for participant recruitment and ethical procedures. We do not collect any new human data in this study. All experiments are conducted using released datasets under their corresponding usage conditions. We respect participant privacy and only use anonymized EEG recordings for research purposes. No personally identifiable information is accessed or processed in our experiments.

REFERENCES

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Enea Ceolini, Jens Hjortkjær, Daniel Wong, James O'Sullivan, Vinay Raghavan, Jose Herrero, Ashesh Mehta, Shih-Chii Liu, and Nima Mesgarani. Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception. *NeuroImage*, 223:117282, 2020. doi: 10.1016/j.neuroimage.2020.117282.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 933–941, 2017.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. Griffin: Mixing gated linear recurrences with local attention for efficient language models, 2024.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, and Bingxuan Wang. DeepSeek-V3 technical report, 2025.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, and Bingxuan Wang. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4):1–11, 2018. doi: 10.1145/3197517.3201357.

Cunhang Fan, Jingjing Zhang, Hongyu Zhang, Wang Xiang, Jianhua Tao, Xinhui Li, Jiangyan Yi, Dianbo Sui, and Zhao Lv. MSFNet: Multi-scale fusion network for brain-controlled speaker extraction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1652–1661, 2024. doi: 10.1145/3664647.3681550.

Cunhang Fan, Ying Chen, Jian Zhou, Zexu Pan, Jingjing Zhang, Youdian Gao, Xiaoke Yang, Zhengqi Wen, and Zhao Lv. M3ANet: Multi-scale and multi-modal alignment network for brain-assisted target speaker extraction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 8040–8048, 2025a. doi: 10.24963/ijcai.2025/894.

Cunhang Fan, Youdian Gao, Zexu Pan, Jingjing Zhang, Hongyu Zhang, Jie Zhang, and Zhao Lv. Improved feature extraction network for neuro-oriented target speaker extraction. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025b. doi: 10.1109/ICASSP49660.2025.10888763.

Cunhang Fan, Hongyu Zhang, Qinke Ni, Jingjing Zhang, Jianhua Tao, Jian Zhou, Jiangyan Yi, Zhao Lv, and Xiaopei Wu. Seeing helps hearing: A multi-modal dataset and a mamba-based dual branch parallel network for auditory attention decoding. *Information Fusion*, 118:102946, 2025c. doi: 10.1016/j.inffus.2025.102946.

Cunhang Fan, Junling Li, Jingjing Zhang, Youdian Gao, Ying Chen, Jiangyan Yi, and Zhao Lü. GALENet: Neuro-oriented speaker extraction network fusing gated attention and local feature enhancement. *Journal of Computer Research and Development*, 63(6):1534–1547, 2026. doi: 10.7544/issn1000-1239.202550467.

Cong Han, James O'Sullivan, Yi Luo, Jose Herrero, Ashesh D. Mehta, and Nima Mesgarani. Speaker-independent auditory attention decoding without access to clean speech sources. *Science Advances*, 5(5):eaav6134, 2019. doi: 10.1126/sciadv.aav6134.

Fengyuan Hao, Xiaodong Li, and Chengshi Zheng. X-TF-GridNet: A time–frequency domain target speaker extraction network with adaptive speaker embedding fusion. *Information Fusion*, 112: 102550, 2024. doi: 10.1016/j.inffus.2024.102550.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Maryam Hosseini, Luca Celotti, and Éric Plourde. Speaker-independent brain enhanced speech denoising. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1310–1314, 2021. doi: 10.1109/ICASSP39728.2021.9414969.

Maryam Hosseini, Luca Celotti, and Éric Plourde. End-to-end brain-driven speech enhancement in multi-talker conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1718–1733, 2022. doi: 10.1109/TASLP.2022.3169629.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

Minsu Kim, Rodrigo Mira, Honglie Chen, Stavros Petridis, and Maja Pantic. Contextual speech extraction: Leveraging textual history as an implicit cue for target speech extraction. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887655.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Jiuxin Lin, Xinyu Cai, Heinrich Dinkel, Jun Chen, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Zhiyong Wu, Yujun Wang, and Helen Meng. Av-Sepformer: Cross-attention sepformer for audio-visual target speaker extraction. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10094306.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training, 2025.

Kai Liu, Ziqing Du, Xucheng Wan, and Huan Zhou. X-SEPFORMER: End-to-end speaker extraction network with explicit optimization on speaker confusion. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095609.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Zhao Lv, Haoran Zhou, Ying Chen, Youdian Gao, Xinhui Li, Ruibo Fu, and Cunhang Fan. Trainable EEG interpolation and structure-sharing dual-path encoders for brain-assisted target speaker extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 32392–32400, 2026. doi: 10.1609/aaai.v40i38.40514.

Nima Mesgarani and Edward Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485:233–6, 2012. doi: 10.1038/nature11020.

Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 26670–26698, 2023.

J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral cortex*, 25(7):1697–1706, 2015.

Zexu Pan, Xinyuan Qian, and Haizhou Li. Speaker extraction with co-speech gestures cue. *IEEE Signal Processing Letters*, 29:1467–1471, 2022. doi: 10.1109/LSP.2022.3175130.

Zexu Pan, Marvin Borsdorf, Siqi Cai, Tanja Schultz, and Haizhou Li. NeuroHeed: Neuro-steered speaker extraction using EEG signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4456–4470, 2024a. doi: 10.1109/TASLP.2024.3463498.

Zexu Pan, Gordon Wichern, François G. Germain, Sameer Khurana, and Jonathan Le Roux. NeuroHeed+: Improving neuro-steered speaker extraction with joint auditory attention detection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11456–11460, 2024b. doi: 10.1109/ICASSP48485.2024.10446333.

A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Proceedings of the Acoustics, Speech, and Signal Processing, 200. on IEEE International Conference - Volume 02*, pp. 749–752, 2001. doi: 10.1109/ICASSP.2001.941023.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR – half-baked or well done? In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019. doi: 10.1109/ICASSP.2019.8683855.

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Jianlin Su, Ahmed Murtadha, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011. doi: 10.1109/TASL.2011.2114881.

Ruijie Tao, Xinyuan Qian, Yidi Jiang, Junjie Li, Jiadong Wang, and Haizhou Li. Audio-visual target speaker extraction with selective auditory attention. *IEEE Transactions on Audio, Speech and Language Processing*, 33:797–811, 2025. doi: 10.1109/TASLPRO.2025.3527766.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006. doi: 10.1109/TSA.2005.858005.

Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li. SpEx: Multi-scale time domain speaker extraction network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 1370–1384, 2020. doi: 10.1109/TASLP.2020.2987429.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Hongyu Zhang, Jingjing Zhang, Xingguang Dong, Zhao Lü, Jianhua Tao, Jian Zhou, Xiaopei Wu, and Cunhang Fan. Based on audio-video evoked auditory attention detection electroencephalogram dataset. *Journal of Tsinghua University (Science and Technology)*, 64(11):1919–1926, 2024. doi: 10.16511/j.cnki.qhdxxb.2024.26.024.

Jie Zhang, Qing-Tian Xu, Qiu-Shi Zhu, and Zhen-Hua Ling. BASEN: Time-domain brain-assisted speech enhancement network with convolutional cross attention in multi-talker conditions, 2023.

Shengkui Zhao and Bin Ma. MossFormer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096646.

Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, and Bin Ma. MossFormer2: Combining transformer and RNN-free recurrent network for enhanced time-domain monaural speech separation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10356–10360, 2024. doi: 10.1109/ICASSP48485.2024.10445985.

Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Tsubasa Ochiai, Tomohiro Nakatani, Lukáš Burget, and Jan Černocký. SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):800–814, 2019. doi: 10.1109/JSTSP.2019.2922820.

APPENDIX

A INTERNAL STRUCTURE OF A GRIFFIN RESIDUAL BLOCK

A.1 INPUT AND NOTATION

A Griffin residual block processes a frame sequence $H \in \mathbb{R}^{B\times T\times D}$, where B is the batch size, T is the number of acoustic frames, and $D = 64$ is the feature width in our model. Here, H denotes the input to one residual block in sequence-major order. Each block preserves its input shape, allowing multiple blocks to be stacked within a directional branch of BiGriffinAdapter.

The block contains a temporal mixer followed by a gated MLP. The temporal mixer is either a recurrent module (Orvieto et al., 2023) or a local multi-query attention module. Both variants use the same residual structure and gated MLP.

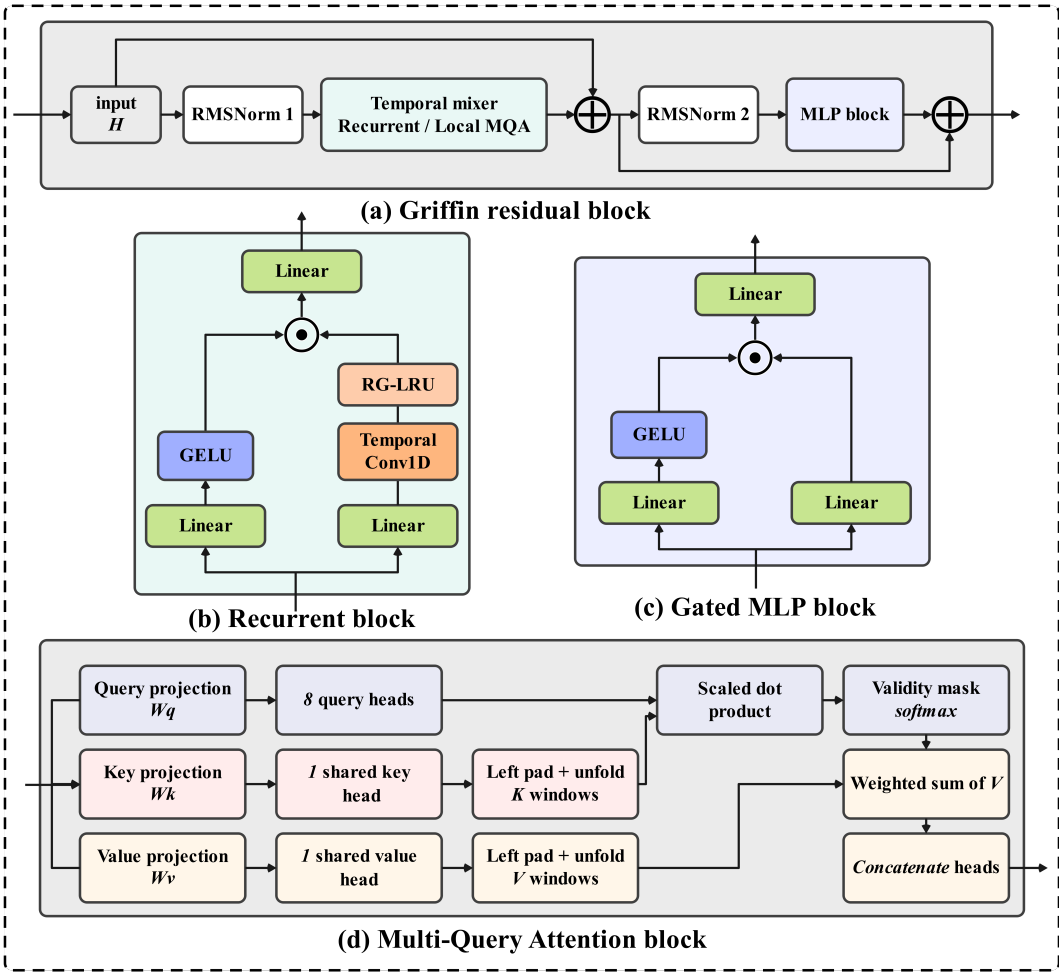


Figure 3: Components of a Griffin residual block: (a) the pre-normalized residual structure, (b) the recurrent temporal mixer, (c) the gated MLP, and (d) the local multi-query attention mixer.

A.2 GRIFFIN RESIDUAL BLOCK: TWO PRE-NORMALIZED RESIDUAL UPDATES

Figure 3(a) shows the two residual paths of a Griffin block. The block applies a separate RMS normalization (Zhang & Sennrich, 2019) before each submodule. For block input H, its two updates are

$$\begin{aligned} U &= H + \mathcal{M}\big(\mathrm{RMSNorm}_1(H)\big), \\ Y &= U + \mathrm{MLP}\big(\mathrm{RMSNorm}_2(U)\big), \end{aligned} \tag{16}$$

where $\mathcal{M}$ denotes the selected temporal mixer and $Y \in \mathbb{R}^{B\times T\times D}$ is the block output. The first residual path preserves the block input while the mixer adds temporal information. The second residual path preserves U while the MLP (Dauphin et al., 2017) transforms its channel features. The two normalization layers and the parameters of both submodules belong to this individual block.

RMS normalization operates over the last, channel-wise dimension of each frame. For a frame vector $z \in \mathbb{R}^D$, it takes the form

$$\mathrm{RMSNorm}(z) = g \odot \frac{z}{\sqrt{D^{-1}\sum_{j=1}^{D} z_j^2 + \epsilon}}, \tag{17}$$

where g is a learned channel-wise scale and ϵ is a numerical constant. The recurrent and attention versions of the block differ only in $\mathcal{M}$; both retain the two residual updates in Eq. (16).

A.3 Recurrent Temporal Mixer

The recurrent mixer in Figure 3(b) expands each normalized input frame from $D = 64$ to two branches of width $d_r = \lfloor 4D/3 \rfloor = 85$. One $64 \to 170$ linear projection produces the two branches, denoted p_t and g_t. The temporal branch p_t passes through a depthwise causal convolution with kernel size four. Its output u_t depends directly on the current frame and up to three preceding frames in the branch's scan order; padding and cropping preserve the sequence length.

The convolution output enters a real-gated linear recurrent unit (RG-LRU) (Orvieto et al., 2023). For each frame t, the unit computes

$$\begin{aligned} r_t &= \sigma(W_a u_t + b_a), \\ i_t &= \sigma(W_x u_t + b_x), \\ a_t &= \exp[-8\ \mathrm{softplus}(-\Lambda) \odot r_t]\,, \\ h_0 &= 0, \\ h_t &= a_t \odot h_{t-1} + \sqrt{1 - a_t^2 + 10^{-6}} \odot (i_t \odot u_t), \end{aligned} \tag{18}$$

where r_t and i_t control state retention and input incorporation, respectively, and the learned $\Lambda \in \mathbb{R}^{85}$ sets the channel-wise decay. The gate projections W_a and W_x are full linear transformations. An associative prefix scan evaluates the recurrence over the sequence.

The second projected branch is transformed by GeLU (Hendrycks & Gimpel, 2016; Shazeer, 2020) and multiplied elementwise with the RG-LRU output. A final $85 \to 64$ linear projection returns the mixer result to the block width:

$$\mathcal{M}_{\mathrm{rec}}(Z)_t = W_{\mathrm{rec,out}}\big(h_t \odot \mathrm{GeLU}(g_t)\big) + b_{\mathrm{rec,out}}, \tag{19}$$

where $Z = \mathrm{RMSNorm}_1(H)$ is the recurrent mixer's input. This result enters the first residual addition in Eq. (16). The recurrent state is initialized to zero for each processed sequence.

A.4 Gated MLP

Figure 3(c) shows the gated MLP. After the temporal mixer's residual addition, the intermediate sequence U passes through the second RMS normalization and the gated MLP. For each normalized frame $z_t \in \mathbb{R}^{64}$, an input projection produces 384 channels, split evenly into $x_{1,t}, x_{2,t} \in \mathbb{R}^{192}$. The MLP computes

$$\mathrm{MLP}(z_t) = W_{\mathrm{mlp,out}}\big[\mathrm{GeLU}(x_{1,t}) \odot x_{2,t}\big] + b_{\mathrm{mlp,out}}. \tag{20}$$

The output projection maps the gated 192-dimensional vector back to 64 dimensions. GeLU is applied to the first branch; the second branch enters the elementwise product directly. The MLP output is then added to U through the second residual path.

A.5 Local Multi-Query Attention Temporal Mixer

When the block uses attention instead of recurrence, the local multi-query attention module (Shazeer, 2019; Beltagy et al., 2020; Su et al., 2024) in Figure 3(d) supplies $\mathcal{M}$ in the first residual

update. With $D = 64$, the query projection forms eight heads of dimension $d_h = 8$. The key and value projections each form one head of dimension eight, shared by all query heads. Thus, the query tensor has shape $B \times 8 \times T \times 8$, whereas each key and value tensor has shape $B \times 1 \times T \times 8$.

At scan position t, the mixer constructs a causal local window

$$\mathcal{W}(t) = \{\max(1, t - 63), \ldots, t\}. \tag{21}$$

Its width of 64 frames includes the current frame. Keys and values are left-padded to form the windows, and a validity mask removes padded positions near the start of the sequence. For query head m, attention within the valid window is

$$\begin{aligned} \alpha_{t,j}^{(m)} &= \mathrm{softmax}_{j \in \mathcal{W}(t)} \left(\frac{q_t^{(m)} \cdot k_j}{\sqrt{8}} \right), \\ o_t^{(m)} &= \sum_{j \in \mathcal{W}(t)} \alpha_{t,j}^{(m)} v_j. \end{aligned} \tag{22}$$

The eight head outputs are concatenated and passed through a $64 \to 64$ output projection to form $\mathcal{M}_{\mathrm{att}}(Z)_t$, where $Z = \mathrm{RMSNorm}_1(H)$. This projected result enters the same first residual addition as the recurrent mixer. The attention window limits the frames read directly by this mixer; when an attention block follows recurrent blocks, its input can already contain information accumulated by their recurrent states.

A.6 Use Within BiGriffinAdapter

Each directional stack of BiGriffinAdapter applies two blocks with the recurrent mixer followed by one block with the local attention mixer. Every block has its own normalization, mixer, and MLP parameters. A single block operates in the temporal order of the sequence it receives: the forward stack receives chronological frames, while the backward stack receives a reversed sequence. This block-level description specifies the computations shared by both directions; the adapter-level fusion is given in Section 3.4.

B Additional Results

B.1 BiGriffin Integration Designs

The main model applies BiGriffinAdapter after Audio–EEG fusion to refine the EEG-conditioned representation before mask estimation. We compare this design with an alternative that incorporates BiGriffin into the fusion stage. All three models use the MM-AAD protocol in Section 4, with DA-MFP disabled and AdamW used for training. Relative to Model #1, integrating BiGriffin into

Table 3: Comparison of BiGriffin integration designs on MM-AAD. Model #1 uses conventional Audio–EEG fusion followed by AVConv; Model #2 uses BiGriffin-based fusion followed by AVConv; Model #3 uses conventional Audio–EEG fusion followed by BiGriffinAdapter. Results are averaged over ten subjects. SDR and SI-SDR are in dB. SDR, SI-SDR, and PESQ report mean ± SD; STOI and ESTOI report means. Higher values are better for all metrics.

Model	SDR↑	SI-SDR↑	PESQ↑	STOI↑	ESTOI↑
#1	9.78 ± 0.62	8.94 ± 0.62	1.94 ± 0.17	0.88	0.78
#2	9.53 ± 0.78	8.55 ± 0.78	1.92 ± 0.17	0.87	0.76
#3	10.03 ± 0.65	9.19 ± 0.65	2.04 ± 0.14	0.89	0.79

the fusion stage in Model #2 reduces mean SDR by 0.25 dB and SI-SDR by 0.39 dB. Its PESQ, STOI, and ESTOI means are also lower by 0.02, 0.01, and 0.02, respectively. In contrast, Model #3 improves SDR and SI-SDR by 0.25 dB each over Model #1, alongside gains of 0.10 in PESQ and 0.01 in both STOI and ESTOI.

Model #3 also exceeds Model #2 by 0.50 dB in SDR, 0.64 dB in SI-SDR, 0.12 in PESQ, 0.02 in STOI, and 0.03 in ESTOI. Under the evaluated MM-AAD configuration, post-fusion BiGriffin

refinement is the better-performing integration design. Its improvement over Model #1 provides performance evidence for refining the fused Audio–EEG representation through bidirectional temporal modeling, addressing the first gap identified in the Introduction. The comparison concerns complete designs: Models #2 and #3 differ in their fusion structures, use of AVConv, and parameter composition.

B.2 Combining and Ordering AVConv and BiGriffinAdapter after Fusion

Table 3 shows that conventional Audio–EEG fusion followed by BiGriffinAdapter performs best among the three integration designs compared there. We retain this post-fusion placement and examine whether adding AVConv improves extraction quality and whether the order of the two modules affects extraction quality. The fusion module already mixes channels through pointwise projections and incorporates local temporal context. AVConv adds a frame-wise 1×1 mapping after fusion, whereas BiGriffinAdapter models bidirectional temporal context.

Following the numbering in Table 3, Model #3 uses conventional Audio–EEG fusion followed by BiGriffinAdapter, without AVConv. Model #4 applies AVConv to the fusion output before BiGriffinAdapter. Model #5 applies BiGriffinAdapter to the fusion output before AVConv. All three use the same conventional fusion, mask estimator, and decoder. On MM-AAD, each configuration is trained separately for the five from the 10 subjects, with DA-MFP disabled and AdamW used for training. Table 4 reports results averaged over these five subject-specific models. The Model #3 results in this table are computed on these five subjects, whereas Table 3 reports ten subjects.

Table 4: Comparison of post-fusion Models #3–#5 on MM-AAD. SDR and SI-SDR are in dB. SDR, SI-SDR, and PESQ report mean $\pm$ SD; STOI and ESTOI report means.

Model	SDR↑	SI-SDR↑	PESQ↑	STOI↑	ESTOI↑
#3	$\mathbf{9.96 \pm 0.91}$	$\mathbf{9.06 \pm 0.96}$	$\mathbf{1.98 \pm 0.08}$	0.882	0.773
#4	9.82 ± 0.88	8.95 ± 0.93	1.96 ± 0.10	**0.889**	**0.784**
#5	9.63 ± 0.77	8.79 ± 0.82	1.93 ± 0.09	0.857	0.754

To complement the averaged results in Table 4, Fig. 4 visualizes the per-subject SI-SDR values of Models #3–#5 on the five MM-AAD subjects. This subject-level view helps reveal whether the relative ordering among the three post-fusion designs is consistent across subjects. Model #4 exceeds Model #5 on the mean of all five metrics, including 0.19 dB higher SDR and 0.16 dB higher SI-SDR. Its STOI and ESTOI means are also higher by 0.031 and 0.030, respectively. As further visualized in Fig. 4, the SI-SDR advantage of Model #4 over Model #5 is consistent across all five subjects.

Relative to Model #3, Model #4 has 0.14 dB lower mean SDR and 0.11 dB lower mean SI-SDR, while PESQ changes from 1.98 to 1.96. Its STOI and ESTOI means increase from 0.882 to 0.889 and from 0.773 to 0.784, respectively. Model #5 is below Model #3 on all five mean metrics. These results indicate that if both AVConv and BiGriffinAdapter are retained, applying AVConv before BiGriffinAdapter gives the better ordering; however, Model #3 still provides the highest separation scores overall in this MM-AAD comparison, whereas Model #4 yields slightly better mean intelligibility scores.

B.3 AVConv Retention across DA-MFP Prediction Depths

Section B.2 compares post-fusion BiGriffinAdapter with AVConv followed by BiGriffinAdapter under AdamW without DA-MFP. We next examine AVConv retention with DA-MFP and Hybrid Muon–AdamW enabled. DA-MFP supervises BiGriffinAdapter's pre-concatenation directional states by predicting future and past clean-speech latents (Section 3.6). The prediction depth K determines how many successive depths contribute to this auxiliary objective: $K = 1$ uses the first depth, whereas $K = 3$ includes the first three, with their losses averaged. The prediction modules are removed at inference. Comparing these depths tests whether additional directional prediction supervision changes extraction quality and the effect of retaining AVConv before BiGriffinAdapter.

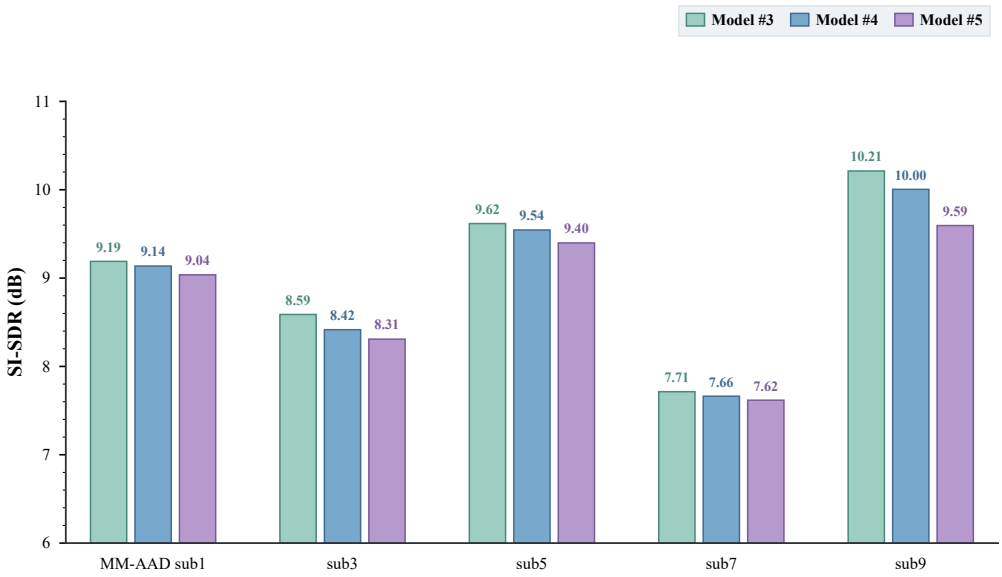


Figure 4: Per-subject SI-SDR comparison of post-fusion Models #3–#5 on the five MM-AAD subjects. Model #3 achieves the highest SI-SDR for each subject, while Model #4 consistently outperforms Model #5, indicating that when both modules are retained, placing AVConv before BiGriffinAdapter is more effective than placing it after BiGriffinAdapter.

Model #6 uses BiGriffinAdapter alone after fusion with $K = 1$. Model #7 also uses $K = 1$, but inserts AVConv before BiGriffinAdapter. Model #8 applies BiGriffinAdapter alone with $K = 3$, whereas Model #9 combines AVConv with BiGriffinAdapter under the same $K = 3$ setting. All four models share the same fusion module, mask estimator, decoder, and five MM-AAD participants as in Section B.2. Table 5 reports their subject-averaged results.

Table 5: Comparison of Models #6–#9 on MM-AAD. SDR, SI-SDR, and PESQ report mean $\pm$ SD; STOI and ESTOI report means.

Model	SDR↑	SI-SDR↑	PESQ↑	STOI↑	ESTOI↑
#6	10.22 ± 0.98	9.37 ± 0.95	2.01 ± 0.08	0.89	0.79
#7	10.21 ± 0.93	9.39 ± 0.94	2.02 ± 0.11	0.88	0.79
#8	$\mathbf{10.76 \pm 1.15}$	$\mathbf{9.86 \pm 1.09}$	$\mathbf{2.04 \pm 0.09}$	**0.90**	**0.80**
#9	10.41 ± 1.10	9.54 ± 1.04	2.03 ± 0.11	0.89	**0.80**

Relative to Model #6, Model #7 has 0.01 dB lower mean SDR and 0.02 dB higher mean SI-SDR. PESQ increases by 0.01, STOI decreases by 0.01, and ESTOI is unchanged. The small mean SI-SDR difference combines increases for two subjects with decreases for three, ranging from -0.15 to $+0.23$ dB. Relative to Model #8, Model #9 has 0.35 dB lower mean SDR and 0.32 dB lower mean SI-SDR. PESQ and STOI each decrease by 0.01, while ESTOI is unchanged. SI-SDR is lower for Model #9 in all five subjects, with differences from -0.49 to -0.13 dB. Figure 5 visualizes these subject-level SI-SDR differences.

Model #8 exceeds Model #6 by 0.54 dB in mean SDR and 0.49 dB in mean SI-SDR, while Model #9 exceeds Model #7 by 0.20 and 0.15 dB, respectively. PESQ, STOI, and ESTOI also increase in both depth comparisons. As visualized in Figure 5, increasing the prediction depth from $K = 1$ to $K = 3$ improves SI-SDR for all five subjects in both architectures, with a larger gain for Model #8 over Model #6 than for Model #9 over Model #7 in every subject. In contrast, retaining AVConv has little consistent effect at $K = 1$, but Model #9 is consistently below Model #8 across all five subjects at $K = 3$. Thus, the deeper prediction setting improves separation scores in both architectures, while the benefit of retaining AVConv is not supported under the deeper DA-MFP setting in this five-subject MM-AAD comparison.

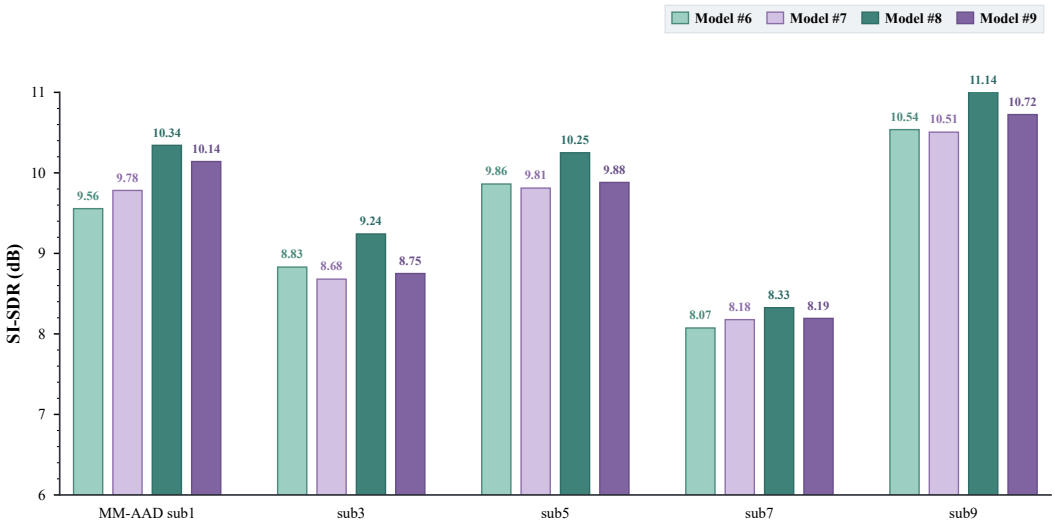


Figure 5: Per-subject SI-SDR comparison of Models #6–#9 on the five MM-AAD subjects. Increasing the DA-MFP prediction depth from $K = 1$ to $K = 3$ improves SI-SDR for every subject in both architectural variants. Model #8 consistently outperforms Model #9, while Models #6 and #7 show mixed subject-level differences, indicating that retaining AVConv provides no consistent benefit and becomes less effective under the deeper $K = 3$ prediction setting.