RAMAMBA-NET: A RELIABILITY-AWARE AND MAMBA-BASED MULTIMODAL FUSION NETWORK FOR AUDITORY ATTENTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Auditory attention decoding (AAD) identifies the attended speaker from physiological signals, supporting neuro-steered hearing devices and natural human-machine interaction. Electroencephalography (EEG) is the dominant modality for AAD but provides incomplete evidence in naturalistic audio-visual scenes, motivating EEG and electrooculography (EOG) fusion. Existing approaches remain limited by weak cross-modal interaction, inefficient temporal modeling, and low robustness to sample variations. To address the limitations, we propose RAMamba-Net, a reliability-aware Mamba-based multimodal fusion network for AAD. RAMamba-Net employs a Mamba-enhanced band-aware convolutional Transformer to capture band-specific EEG patterns and long-range temporal dynamics. A dual-branch temporal-spatial encoder models EOG temporal and inter-channel dependencies. Cross-modal attention enables explicit modality interaction. Then, a reliability-aware module is introduced to estimate sample-wise modality weights for feature and prediction consistency, thereby enhancing multimodal fusion. Experiments on two AAD benchmarks demonstrate that RAMamba-Net effectively exploits complementary EEG-EOG information, yielding accuracy gains of 5.76% over unimodal baselines, together with more robust decoding and discriminative representations. Further analyses show that explicit cross-modal interaction improves multimodal alignment, while the reliability-aware module suppresses unreliable modality evidence and is robust to signal perturbation and parameter variation.

1 INTRODUCTION

Understanding a target speaker in a crowded acoustic scene remains a major challenge for people with hearing impairment (Wilroth et al., 2025a), commonly known as the cocktail-party problem (Cherry, 1953). Auditory attention decoding (AAD) infers the attended speaker from physiological signals (Mesgarani & Chang, 2012; Ding & Simon, 2012; O'sullivan et al., 2015), supporting neuro-steered hearing devices, target-speaker enhancement, and natural human-machine interaction (Alickovic et al., 2019; Geirnaert et al., 2021). These applications require reliable decoding under realistic and non-ideal recording conditions.

Electroencephalography (EEG) is the dominant AAD modality. It reflects cortical activity associated with attentional selection. EEG decoders have progressed from convolutional direction localization (Vandecappelle et al., 2021) to spatio-temporal attention modeling (Su et al., 2022) and end-to-end cross-subject decoding (Nguyen et al., 2025). However, EEG has a low signal-to-noise ratio, is highly susceptible to artifacts (Somers et al., 2018), and often generalizes poorly across subjects and sessions (Puffay et al., 2023). EEG alone may therefore provide incomplete evidence of attentional behavior in naturalistic audio-visual scenes, motivating the integration of complementary physiological signals (Wang et al., 2025).

Eye movements provide behavioral information related to auditory attention, while electrooculography (EOG) directly records ocular activity. Prior work has shown that gaze follows attended natural speech (Gehmacher et al., 2024) and that visual speech cues enhance neural tracking of the attended speaker (Fu et al., 2019). These findings suggest that EOG can complement EEG. Nevertheless, ocular activity may introduce confounding patterns and bias spatial-attention decoding (Rotaru et al.,

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.

2024). EOG should therefore be explicitly modeled and adaptively integrated with EEG rather than discarded as noise or combined through naïve concatenation.

EEG-EOG fusion for AAD still faces key challenges. First, continuous cross-modal dependencies remain insufficiently modeled. Cross-attention has proven effective in related physiological and multimodal tasks (Lu et al., 2015; Yin et al., 2026; Zhuang et al., 2026; Tsai et al., 2019), yet bidirectional token-wise interaction between EEG and EOG remains underexplored in AAD. Second, efficient temporal modeling remains difficult. Convolutions emphasize local patterns, whereas self-attention incurs high computational cost for long sequences. Mamba provides selective state-space modeling with linear-time complexity (Gu & Dao, 2023) and has been introduced into AAD models (Zhang et al., 2024; Fan et al., 2025). However, the band-aware EEG encoder and cross-modal interaction module remain limited. Third, modality quality varies across samples because of artifacts, subject state, and ocular dynamics. Fixed weighting and direct concatenation cannot adapt to these variations, motivating sample-wise reliability estimation.

To address these challenges, we propose RAMamba-Net, a reliability-aware Mamba-based multimodal fusion network for AAD. RAMamba-Net integrates modality-specific encoding, cross-modal interaction, and sample-wise reliability estimation. Our contributions are summarized as follows:

- We propose a unified EEG-EOG multimodal framework for AAD that integrates modality-specific encoding, cross-modal interaction, and reliability-aware fusion into an end-to-end decoding pipeline.
- We design a Mamba-enhanced band-aware EEG encoder that couples band-specific modeling with selective state-space learning to capture spectral patterns and long-range temporal dynamics.
- We introduce cross-modal attention to enable token-wise information exchange between EEG and EOG while preserving their modality-specific characteristics.
- We develop a reliability-aware module (RAM) to estimate sample-wise modality weights from modality preservation and cross-modal consistency in the representation and prediction spaces. Extensive experiments on two AAD benchmarks demonstrate improved decoding accuracy and representation quality, together with robust performance under signal perturbation and parameter variation.

2 Related Work

2.1 EEG-Based Auditory Attention Decoding

EEG-based auditory attention decoding has progressed from handcrafted pipelines to end-to-end models that jointly capture spectral, spatial, and temporal neural patterns. MBSSFCC (Jiang et al., 2022) combines multi-band differential-entropy topographies with ConvLSTM for spatio-temporal attention representation. BSAnet (Cai et al., 2023) adopts a biologically inspired spiking attention network to model attention-related EEG activity. DBPNet (Ni et al., 2024) employs parallel temporal-attentive and frequency-residual branches to capture dynamic and multi-band spectral-spatial information. DARNet (Yan et al., 2024) integrates spatio-temporal construction with dual-attention refinement for spatial distribution and long-range dependency modeling. DHGCN (Zhou et al., 2025) constructs temporal and spatial hypergraphs to learn higher-order dependencies across time points and channels. FAConformer (Wang et al., 2026a) combines band-specific CNN-Transformer encoders, frequency-aware attention, and band-wise supervision to learn within-band representations and adaptive cross-band interactions.

Despite these advances, EEG remains susceptible to low signal-to-noise ratios and environmental artifacts, which hinder stable representation learning under varying recording conditions. Moreover, EEG-only decoding may not fully characterize auditory attention in realistic audio-visual scenes. These limitations motivate multimodal AAD models to exploit complementary information while preserving robust neural representations.

2.2 Multimodal Auditory Attention Decoding

EOG signals provide behavioral evidence complementary to neural activity and have gained increasing attention in multimodal AAD. Gehmacher et al. (2024) combined eye tracking with magnetoencephalography and showed that gaze followed attended natural speech, and was associated with attention-related neural responses. Wilroth et al. (2025b) inferred attended-speech labels from eye-tracking features to supervise portable EEG-based speech-reconstruction models when ground-

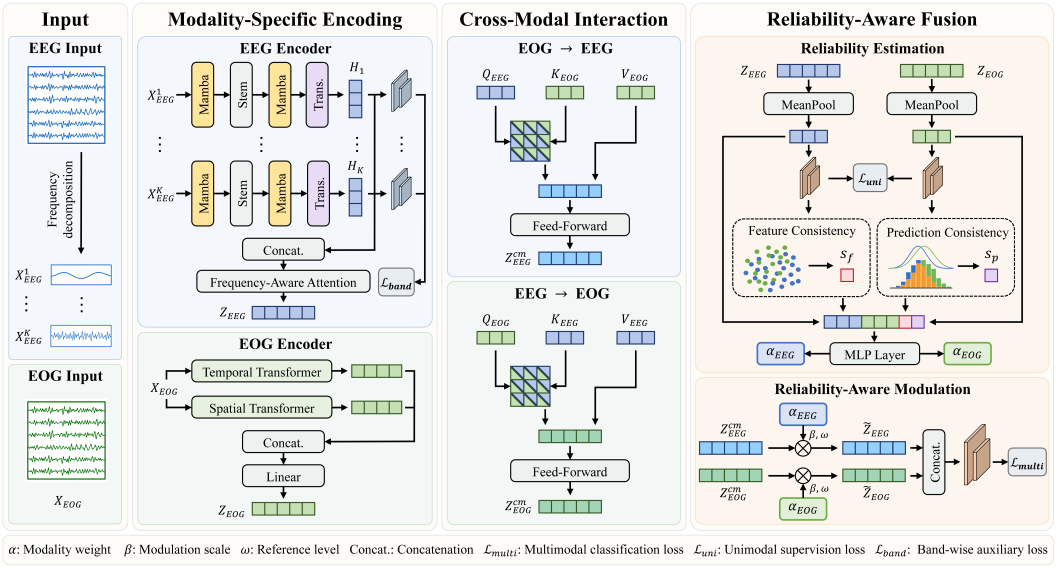


Figure 1: Overview of the proposed RAMamba-Net framework.

truth labels were unavailable. Kosmyna et al. (2022) evaluated low-channel EEG/EOG glasses and reported target-speaker detection performance comparable to a research-grade EEG headset. Xu et al. (2026) further demonstrated that EOG-derived eyeblink sequences contain attention-related information and improve short-window AAD when combined with EEG.

However, existing work remains constrained by indirect eye-movement supervision and simplified EOG modeling. Consequently, the complementary relationship between EEG and EOG signals remains insufficiently modeled. These limitations motivate a unified multimodal AAD framework that explicitly captures modality-specific information and effective cross-modal interactions.

3 Method

This section details the proposed RAMamba-Net, a reliability-aware and Mamba-based multimodal fusion network for multimodal auditory attention decoding. We represent the EEG and EOG signals as $\mathbf{X}_{\text{EEG}} \in \mathbb{R}^{K \times C_{\text{EEG}} \times T}$ and $\mathbf{X}_{\text{EOG}} \in \mathbb{R}^{C_{\text{EOG}} \times T}$, respectively, where K is the number of EEG frequency bands, C_{EEG} and C_{EOG} are the number of channels, and T is the number of time points.

3.1 Model Structure

As illustrated in Figure 1, RAMamba-Net comprises a backbone network and the RAM optimization module. The backbone network contains four components: an EEG encoder, an EOG encoder, a cross-modal fusion module, and classifiers. The two encoders first transform EEG and EOG signals into modality-specific feature representations. The fusion module promotes bidirectional interaction between the two modalities. The final fused feature vectors from each modality are concatenated and fed into the multimodal, unimodal, and band classifiers to identify the attended speaker. This backbone provides the basis for the reliability-aware learning strategy.

EEG Encoder. The EEG encoder takes reference from the band-aware convolution Transformer architecture of FAConformer (Wang et al., 2026a). We further enhance it by introducing the Mamba blocks, thereby constructing a Mamba-enhanced FAConformer. For the k-th EEG frequency band, the band-specific representation is obtained by:

$$\mathbf{H}_k = \mathcal{T}^k_{\text{band}}\left(\mathcal{M}^k_{\text{mid}}\left(\mathcal{S}^k\left(\mathcal{M}^k_{\text{front}}\left(\mathbf{X}^k_{\text{EEG}}\right)\right)\right)\right), \tag{1}$$

where $\mathcal{M}^k_{\text{front}}(\cdot)$ and $\mathcal{M}^k_{\text{mid}}(\cdot)$ denote the front and middle Mamba blocks, respectively. $\mathcal{S}^k(\cdot)$ represents the band-specific stem module, and $\mathcal{T}^k_{\text{band}}(\cdot)$ is the Transformer encoder.

Given an input sequence $\{\mathbf{x}_t\}_{t=1}^{T}$, the Mamba blocks selectively propagate temporal information through:

$$\mathbf{h}_t = \bar{\mathbf{A}}_t \mathbf{h}_{t-1} + \bar{\mathbf{B}}_t \mathbf{x}_t, \;\; \mathbf{y}_t = \mathbf{C}_t \mathbf{h}_t. \tag{2}$$

where $\mathbf{h}_t$ is the hidden state, $\bar{\mathbf{A}}_t$ denotes the discretized state-transition matrix, $\bar{\mathbf{B}}_t$ denotes the discretized input projection matrix, and $\mathbf{C}_t$ denotes the output projection matrix. The front Mamba block $\mathcal{M}^k_{\text{front}}(\cdot)$ reinforces the temporal structure of the raw band-specific EEG sequence, facilitating stable local spectro-temporal feature extraction by the stem, while the middle block $\mathcal{M}^k_{\text{mid}}(\cdot)$ further enhances high-level feature dynamics to provide the Transformer with context-enriched token representations.

We concatenate the band-specific representations $\{\mathbf{H}_k\}_{k=1}^{K}$ along the band dimension and remap the result into a unified EEG token space:

$$\mathbf{Z}_{\text{EEG}} = \mathcal{R}_{\text{EEG}}\left([\mathbf{H}_1 \parallel \cdots \parallel \mathbf{H}_K]\right), \tag{3}$$

where $\mathcal{R}_{\text{EEG}}(\cdot)$ donates the frequency-aware attention module to adaptively determine the contribution of each frequency band to the final decision, promoting the cross-band interaction and dimensionality remapping.

EOG Encoder. The EOG encoder takes reference from DBConformer (Wang et al., 2026b) and comprises parallel temporal and spatial branches. The temporal branch captures fine-grained temporal dependencies through temporal Transformer. The spatial branch models inter-channel dependencies through spatial Transformer. Their outputs are formulated as:

$$\mathbf{Z}^{\text{temporal}}_{\text{EOG}} = \mathcal{E}_t\left(\mathbf{X}_{\text{EOG}}\right), \;\; \mathbf{Z}^{\text{spatial}}_{\text{EOG}} = \mathcal{E}_s\left(\mathbf{X}_{\text{EOG}}\right), \tag{4}$$

where $\mathcal{E}_t(\cdot)$ and $\mathcal{E}_s(\cdot)$ denote the temporal and spatial branches. The temporal and spatial representations are concatenated along the feature dimension and projected into the shared embedding space:

$$\mathbf{Z}_{\text{EOG}} = \left[\mathbf{Z}^{\text{temporal}}_{\text{EOG}} \parallel \mathbf{Z}^{\text{spatial}}_{\text{EOG}}\right] \mathbf{W}_r + \mathbf{b}_r, \tag{5}$$

where $\mathbf{W}_r$ and $\mathbf{b}_r$ are learnable projection parameters. This dual-branch design preserves the temporal and spatial inductive biases of EOG while producing dimension-aligned tokens for subsequent cross-modal interaction.

Cross-Modal Interaction. Given the modality-specific representations, we introduce cross-modal attention to establish explicit token-wise interactions between EEG and EOG. For each modality, its tokens serve as queries, while the other modality provides keys and values.

For modality $m \in \{\text{EEG}, \text{EOG}\}$, let $\bar{m}$ denote the paired modality. The latent representation of modality m is enhanced by retrieving complementary information from modality $\bar{m}$:

$$\mathbf{H}_{m \leftarrow \bar{m}} = \text{softmax}\left(\frac{\mathbf{Q}_m \mathbf{K}_{\bar{m}}^{\top}}{\sqrt{d_h}}\right) \mathbf{V}_{\bar{m}}, \tag{6}$$

where d_h denotes the dimensionality of each attention head. The query, key, and value projections are defined as

$$\mathbf{Q}_m = \mathbf{Z}_m \mathbf{W}_Q^m, \;\; \mathbf{K}_{\bar{m}} = \mathbf{Z}_{\bar{m}} \mathbf{W}_K^{\bar{m}}, \;\; \mathbf{V}_{\bar{m}} = \mathbf{Z}_{\bar{m}} \mathbf{W}_V^{\bar{m}}, \tag{7}$$

where $\mathbf{W}_Q^m$, $\mathbf{W}_K^{\bar{m}}$, and $\mathbf{W}_V^{\bar{m}}$ are learnable projection matrices. The attention output is further refined through residual learning, layer normalization, and a feed-forward network:

$$\mathbf{Z}^{\text{cm}}_m = \mathbf{Z}_m + \mathbf{H}_{m \leftarrow \bar{m}} + \text{FFN}\left(\text{LN}\left(\mathbf{Z}_m + \mathbf{H}_{m \leftarrow \bar{m}}\right)\right). \tag{8}$$

The same update is applied in both directions. Each stream retains its modality-specific information while incorporating complementary evidence from the other modality. This symmetric interaction promotes balanced information exchange and more effective cross-modal alignment.

3.2 RAM Optimization

The backbone enables bidirectional EEG-EOG interaction but does not explicitly account for sample-wise variations in modality quality. RAM therefore infers the relative reliability of EEG and EOG from modality-specific representations that preserve their distinct signal characteristics,

and adaptively regulates their contributions prior to classification. For each modality, the token sequence is mean-pooled into a sample-level descriptor for reliability estimation:

$$\mathbf{z}_m = \mathrm{MeanPool}\big(\mathrm{stopgrad}(\mathbf{Z}_m)\big). \tag{9}$$

Mean pooling summarizes the modality-specific token sequence into a fixed-dimensional representation, while stop-gradient decouples reliability estimation from encoder optimization, preventing the RAM objective from directly reshaping the modality representations.

Feature Consistency. Feature consistency measures the semantic agreement between EEG and EOG representations. It is computed from the pooled modality vectors using cosine similarity and normalized to $[0, 1]$:

$$s_f = \frac{1}{2}\left(1 + \frac{\mathbf{z}_{EEG}^{\top}\mathbf{z}_{EOG}}{\|\mathbf{z}_{EEG}\|\,\|\mathbf{z}_{EOG}\|}\right). \tag{10}$$

A larger s_f indicates stronger feature-level agreement between the two modalities.

Prediction Consistency. The pooled representations are further mapped to modality-specific predictions through lightweight unimodal classifiers $\mathcal{F}_m^{uni}(\cdot)$:

$$\mathbf{P}_m = \mathrm{softmax}\left(\mathcal{F}_m^{uni}(\mathbf{z}_m)\right). \tag{11}$$

Prediction consistency is derived from Jensen-Shannon divergence:

$$s_p = 1 - \frac{D_{\mathrm{KL}}\left(\mathbf{P}_{EEG}\|\mathbf{M}\right) + D_{\mathrm{KL}}\left(\mathbf{P}_{EOG}\|\mathbf{M}\right)}{2\ln 2}, \tag{12}$$

where $\mathbf{M} = (\mathbf{P}_{EEG} + \mathbf{P}_{EOG})/2$, and $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence. A larger s_p indicates stronger decision-level agreement.

Reliability-Aware Fusion. RAM combines the pooled modality-specific descriptors with feature-level and prediction-level consistency scores to construct a sample-wise reliability representation:

$$\mathbf{r} = \left[\mathbf{z}_{\mathrm{EEG}} \,\|\, \mathbf{z}_{\mathrm{EOG}} \,\|\, s_f \,\|\, s_p\right], \tag{13}$$

where s_f and s_p denote the feature and prediction consistency scores, respectively. The resulting representation jointly captures modality-specific information and cross-modal agreement for subsequent reliability estimation. Sample-wise modality weights are then estimated by:

$$[\alpha_{EEG}, \alpha_{EOG}] = \mathrm{softmax}\left(\mathcal{G}_\alpha(\mathbf{r})\right), \tag{14}$$

where $\mathcal{G}_\alpha(\cdot)$ is a reliability mapping MLP layer. The resulting weights regulate the cross-modally interacted representations:

$$\tilde{\mathbf{Z}}_m = \mathbf{Z}_m^{cm} \cdot \left[1 + \beta(\alpha_m - \omega)\right], \tag{15}$$

where β controls the modulation scale, and ω denotes the reference level. The modulated EEG and EOG representations are directly concatenated and classified:

$$\hat{\mathbf{y}} = \mathcal{F}_c\left[\mathrm{vec}(\tilde{\mathbf{Z}}_{EEG}) \,\|\, \mathrm{vec}(\tilde{\mathbf{Z}}_{EOG})\right], \tag{16}$$

where $\mathrm{vec}(\cdot)$ denotes vectorization, and $\mathcal{F}c(\cdot)$ is the norm-constrained linear classifier.

3.3 Training Objectives

RAMamba-Net is optimized with a composite objective consisting of a multimodal classification loss, a unimodal supervision loss, and a band-wise auxiliary loss. The multimodal classification loss supervises the final prediction obtained after cross-modal interaction and reliability-aware fusion:

$$\mathcal{L}_{multi} = \mathrm{CE}(\hat{\mathbf{y}}, \mathbf{y}), \tag{17}$$

where $\mathbf{y}$ denotes the ground-truth label, and $\mathrm{CE}(\cdot,\cdot)$ denotes the cross-entropy loss. To preserve the discriminative capability of each modality-specific encoder, the unimodal loss directly supervises the auxiliary predictions from the EEG and EOG branches:

$$\mathcal{L}_{uni} = \mathrm{CE}(\mathbf{P}_{EEG}, y) + \mathrm{CE}(\mathbf{P}_{EOG}, y). \tag{18}$$

The band-wise auxiliary loss further supervises each EEG frequency band through the independent lightweight classifier:

$$\mathcal{L}_{band} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{CE}\Big(\mathcal{F}_k\big(\mathrm{vec}(\mathbf{H}_k)\big), y\Big), \tag{19}$$

where $\mathcal{F}_k(\cdot)$ denotes the classifier for the k-th band. The overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{multi} + \lambda_{uni}\mathcal{L}_{uni} + \mathcal{L}_{band}, \tag{20}$$

where λ_{uni} controls the contribution of unimodal supervision. More specific details about the model structure of RAMamba-Net are provided in the supplementary material.

4 Experiments and Results

4.1 Datasets

We evaluate the proposed RAMamba-Net on two commonly used public AAD datasets.

AVGC Dataset. AVGC is a public audio-visual dataset for AAD, comprising multimodal physiological and behavioral recordings (Rotaru et al., 2024). Subjects were required to attend to one of two competing speech streams, with the attended speaker located at $\pm 90^{\circ}$. The signals contain 64 EEG channels and 4 EOG channels at 8,192 Hz. Each trial lasted 10 minutes, with the attended direction switching after the first 5 minutes. We excluded subjects without public release permission and subjects with incomplete channel recordings. The final evaluation was conducted on 12 subjects.

DTU Dataset. DTU is a public AAD dataset collected in simulated acoustic environments with different reverberation levels (Fuglsang et al., 2017). It contains recordings from 18 normal-hearing participants listening to two competing speech streams, with two speakers presented from $\pm 60^{\circ}$ spatial directions. Each participant completed 60 dual-speaker trials, with each trial lasting approximately 50 seconds. The signals were recorded at 512 Hz and include 64-channel EEG and 6-channel periocular EOG. All 18 participants were included in the evaluation.

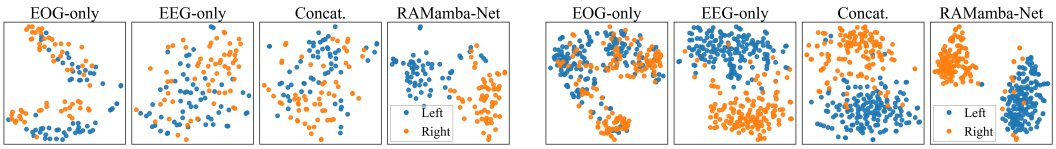
Preprocessing. For AVGC, the recordings were downsampled to 128 Hz and band-pass filtered within 1-40 Hz. For DTU, EEG signals were high-pass filtered at 0.1 Hz, notch filtered at 50 Hz, and then resampled to 64 Hz. EEG additionally underwent drift removal, artifact attenuation, and average referencing, while the 6 EOG channels were directly aligned with the corresponding EEG samples after temporal preprocessing.

4.2 Experimental Settings

To keep each continuous recording segment within an independent subset, each subject's signals were chronologically partitioned before windowing. The first 90% formed the training pool, while the remaining 10% were reserved for testing. A validation set equal in size to the test set was then sampled from the training pool, and the remaining data were used for training. After data partition, each subset was independently segmented into 2-second decision windows. This resulted in an approximately 8:1:1 split across training, validation, and testing sets.

Common spatial pattern (Ang et al., 2008) was applied separately to EEG and EOG data to extract modality-specific spatial patterns. The filters were estimated from the training pool and applied to the validation and test sets. Given the Nyquist frequency after downsampling, AVGC and DTU were divided into 7 and 8 frequency bands, respectively. To comprehensively evaluate decoding performance, we report classification accuracy, balanced accuracy (BCA), Macro-F1, and Cohen's Kappa, covering overall discrimination, class-wise balance, and corrected agreement.

All models were implemented in PyTorch and trained on a single NVIDIA GeForce RTX 3090 GPU. Each experiment was repeated with the seed list $\{41, 42, 43, 44, 45\}$. We used Adam with an initial learning rate of 5×10^{-4} and a weight decay of 3×10^{-4}. The batch size was set to 8 for AVGC and 16 for DTU. Each model was trained for up to 200 epochs, using early stopping with a patience of 10. The trade-off parameter $\mathcal{L}_{uni}$ was set to 0.8 for both datasets. The reliability module coefficients β and ω were set to 0.1 and 0.5, respectively, and the number of Transformer layers and attention heads H_c was set to 2.



(a) AVGC (b) DTU

Figure 2: t-SNE visualization of feature distributions on AVGC and DTU.

4.3 Unimodal Decoding

Before studying EEG-EOG fusion, we first evaluated each modality separately to explore their unimodal discriminative ability. We compared three categories of representative deep AAD models. CNN-based models include EEGNet (Lawhern et al., 2018), SCNN (Schirrmeister et al., 2017), and IFNet (Wang et al., 2023). CNN-Transformer models include CTNet (Zhao et al., 2024), TMSA-Net (Zhao & Zhu, 2025), EEGConformer (Song et al., 2022), MSCFormer (Zhao et al., 2025), MSVTNet (Liu et al., 2024), and DBConformer (Wang et al., 2026b). AAD-specific models include DBPNet (Ni et al., 2024), DARNet (Yan et al., 2024), DHGCN (Zhou et al., 2025), and FAConformer (Wang et al., 2026a). In particular, DBPNet is not included in the EOG-only comparison, as it is specifically designed for 64-channel EEG with frequency band branches.

Table 1: Average unimodal classification accuracies (%) of 13 deep models on AVGC and DTU datasets.

Model	AVGC		DTU	
	EOG	EEG	EOG	EEG
EEGNet	$58.67_{\pm 1.57}$	$60.34_{\pm 0.91}$	$60.81_{\pm 0.48}$	$74.90_{\pm 1.43}$
SCNN	$62.58_{\pm 0.84}$	$67.74_{\pm 0.62}$	$63.38_{\pm 0.54}$	$81.45_{\pm 0.39}$
IFNet	$61.94_{\pm 0.53}$	$69.41_{\pm 0.43}$	$62.73_{\pm 0.30}$	$\underline{82.32}_{\pm 0.34}$
CTNet	$58.05_{\pm 1.32}$	$67.41_{\pm 0.30}$	$60.42_{\pm 0.44}$	$74.84_{\pm 1.04}$
TMSA-Net	$64.36_{\pm 0.51}$	$73.58_{\pm 0.97}$	$65.04_{\pm 0.42}$	$80.47_{\pm 0.39}$
EEGConformer	$63.70_{\pm 0.44}$	$69.84_{\pm 0.68}$	$59.10_{\pm 0.76}$	$66.24_{\pm 0.70}$
MSCFormer	$60.59_{\pm 0.92}$	$69.14_{\pm 2.42}$	$57.66_{\pm 0.54}$	$65.35_{\pm 1.51}$
MSVTNet	$61.30_{\pm 1.04}$	$72.11_{\pm 2.46}$	$59.60_{\pm 0.37}$	$72.01_{\pm 1.80}$
DBConformer	$\mathbf{67.30}_{\pm 1.43}$	$69.06_{\pm 0.64}$	$\mathbf{65.98}_{\pm 0.23}$	$79.81_{\pm 0.31}$
DARNet	$\underline{65.36}_{\pm 0.28}$	$72.83_{\pm 0.67}$	$\underline{65.42}_{\pm 0.50}$	$82.24_{\pm 0.24}$
DBPNet	–	$\underline{73.99}_{\pm 0.94}$	–	$79.83_{\pm 0.30}$
DHGCN	$55.20_{\pm 1.90}$	$63.61_{\pm 0.80}$	$56.34_{\pm 0.26}$	$78.23_{\pm 0.99}$
FAConformer	$58.71_{\pm 1.00}$	$\mathbf{74.23}_{\pm 1.07}$	$62.88_{\pm 0.42}$	$\mathbf{87.03}_{\pm 0.28}$

As reported in Table 1, EEG consistently outperformed EOG across datasets and models, confirming its dominant role in AAD, while EOG remained clearly above chance level, indicating complementary attention-related information. FAConformer and DBConformer achieved the best EEG and EOG decoding performance, respectively, suggesting that frequency-aware temporal-spectral modeling better suits EEG, whereas temporal-spatial dual-branch modeling is more effective for EOG.

Overall, EEG provides stronger neural evidence, while EOG captures complementary ocular dynamics. Their distinct characteristics motivate modality-specific rather than shared encoders.

4.4 Main Results

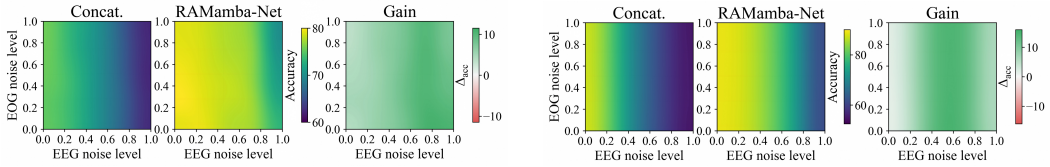
Table 2 compares RAMamba-Net with four unimodal and multimodal baselines, using DBConformer and FAConformer as the EOG-only and EEG-only backbones, respectively. Concat. directly combines the two modality-specific features, while RAMamba-Net w/o RAM retains cross-modal attention but removes the reliability-aware module.

The results showed that EOG provides useful attention-related information, and EEG serves as a strong unimodal reference. Direct concatenation yielded only limited gains, indicating that simple feature fusion cannot fully exploit cross-modal complementary information. In contrast, explicit cross-modal interaction produced more consistent improvements by facilitating information exchange between EEG and EOG.

RAMamba-Net achieved the best results across all metrics and both datasets, with larger gains on AVGC. These results showed that cross-modal interaction and reliability-aware modulation jointly improve the accuracy, balance, and robustness of AAD by adaptively regulating modality contributions. Paired t-tests with adjusted p-values are provided in the supplementary material to verify statistical significance.

Table 2: Average classification performance of different approaches on AVGC and DTU.

Dataset	Approach	Accuracy	BCA	Macro-F1	Cohen's Kappa
AVGC	EOG-only	67.30$_{\pm 1.43}$	67.58$_{\pm 1.23}$	66.60$_{\pm 1.56}$	0.35$_{\pm 0.03}$
	EEG-only	74.23$_{\pm 1.07}$	74.49$_{\pm 1.07}$	77.07$_{\pm 1.14}$	0.58$_{\pm 0.02}$
	Concat.	74.74$_{\pm 1.08}$	75.04$_{\pm 1.12}$	77.59$_{\pm 1.15}$	0.59$_{\pm 0.02}$
	RAMamba-Net w/o RAM	77.16$_{\pm 0.98}$	77.37$_{\pm 0.91}$	79.55$_{\pm 0.96}$	0.61$_{\pm 0.02}$
	RAMamba-Net (Full)	**79.99**$_{\pm 1.03}$	**80.23**$_{\pm 0.66}$	**80.51**$_{\pm 0.56}$	**0.64**$_{\pm 0.01}$
DTU	EOG-only	65.98$_{\pm 0.23}$	65.47$_{\pm 0.32}$	64.06$_{\pm 0.28}$	0.30$_{\pm 0.00}$
	EEG-only	87.03$_{\pm 0.28}$	87.62$_{\pm 0.48}$	87.82$_{\pm 0.31}$	0.75$_{\pm 0.01}$
	Concat.	87.19$_{\pm 0.42}$	87.91$_{\pm 0.35}$	87.98$_{\pm 0.41}$	0.76$_{\pm 0.01}$
	RAMamba-Net w/o RAM	87.90$_{\pm 0.54}$	88.43$_{\pm 0.29}$	88.76$_{\pm 0.57}$	0.78$_{\pm 0.01}$
	RAMamba-Net (Full)	**88.42**$_{\pm 0.15}$	**89.21**$_{\pm 0.16}$	**89.04**$_{\pm 0.18}$	**0.80**$_{\pm 0.00}$



(a) AVGC (b) DTU

Figure 3: Counterfactual robustness under modality-specific noise perturbations on AVGC and DTU.

4.5 Feature Visualization

To further examine how multimodal fusion affects representation learning, we visualized the features before classification using t-SNE (Van der Maaten & Hinton, 2008). We compared EEG-only, Concat., and RAMamba-Net to cover three representative stages: unimodal decoding, direct multimodal fusion, and reliability-aware multimodal fusion. This comparison can assess whether the proposed interaction and reliability mechanisms yield more discriminative feature distributions. The visualization results on AVGC and DTU are shown in Figure 2.

The EEG-only features exhibited limited separability, with attention samples still mixed in the embedding space. Direct concatenation brings EOG information into the representation, but the feature distributions remained insufficiently organized, indicating weak cross-modal alignment.

RAMamba-Net yielded the clearest class structure on both datasets, with learned features forming more compact intra-class clusters and larger inter-class margins. The improvement was more evident on DTU, where the two attention classes are separated into well-structured regions. This suggests that RAMamba-Net produces more discriminative and robust feature representations than unimodal decoding and direct concatenation, effectively enhancing multimodal representation learning.

4.6 Robustness to Noise Injection

To evaluate robustness to modality degradation, we compared Concat. and RAMamba-Net under different EEG and EOG noise levels. The noise level of each modality was varied independently from 0 to 1. A value of 0 denotes a clean signal, while 1 indicates severe corruption with little clean information retained. The first two heatmaps reported classification accuracy under each noise combination. The last heatmap showed the accuracy gain. The results on AVGC and DTU are presented in Figure 3.

With direct feature concatenation, accuracy declined mainly as EEG noise increased. The influence of EOG noise was less pronounced on both datasets. This suggests that the fusion process remains strongly dominated by EEG and is less robust when the primary modality is degraded. In contrast, RAMamba-Net maintained higher accuracy across nearly all noise conditions. The gain maps remained largely positive, even when both modalities were corrupted. This demonstrates that the proposed framework improves not only clean-signal performance but also robustness to changing signal quality. Bidirectional interaction preserves complementary cross-modal information, while RAM suppresses unreliable inputs and adaptively rebalances contributions from each modality. Together, these mechanisms enable more stable decoding under both modality-specific and joint noise perturbations.

4.7 Ablation and Sensitivity Analysis

Ablation Study. To examine the contribution of each reliability component in RAM, we conducted an ablation study on modality preservation, feature consistency, and prediction consistency. Table 3 reports the classification accuracy of different component combinations on AVGC and DTU.

Table 3: Ablation study of the proposed RAM modules on AVGC and DTU.

Modality Preservation	Feature Consistency	Prediction Consistency	AVGC	DTU
×	×	×	$77.16_{\pm 0.98}$	$87.90_{\pm 0.54}$
✓	×	×	$78.40_{\pm 0.82}$	$88.05_{\pm 0.45}$
×	✓	×	$78.54_{\pm 0.64}$	$87.85_{\pm 0.39}$
×	×	✓	$77.89_{\pm 0.96}$	$87.95_{\pm 0.43}$
✓	✓	×	$78.59_{\pm 1.33}$	$\underline{88.30}_{\pm 0.27}$
×	✓	✓	$79.08_{\pm 0.72}$	$88.03_{\pm 0.41}$
✓	×	✓	$\underline{79.63}_{\pm 0.45}$	$88.08_{\pm 0.23}$
✓	✓	✓	$\mathbf{79.99}_{\pm 1.03}$	$\mathbf{88.42}_{\pm 0.15}$

Modality preservation improved the backbone on both datasets, while the gains from feature and prediction consistency varied across datasets. Feature consistency led to the largest term gain on AVGC, whereas modality preservation remained the most effective on DTU. Pairwise combinations provided further improvements, and the complete RAM module achieved the highest accuracy on both datasets. These results show that the three terms capture complementary information of modality reliability. By preserving modality-specific information and jointly measuring consistency in the latent and prediction spaces, RAM enables more reliable sample-wise fusion and stronger multimodal AAD.

Sensitivity Analysis. We analysed the sensitivity of RAMamba-Net to three key hyperparameters: the trade-off parameter λ_{uni}, the reliability module coefficient β, and the number of attention heads H_c in the cross-modal Transformer encoder. Figure 4 presents the parameter sensitivity analysis of RAMamba-Net on AVGC and DTU.

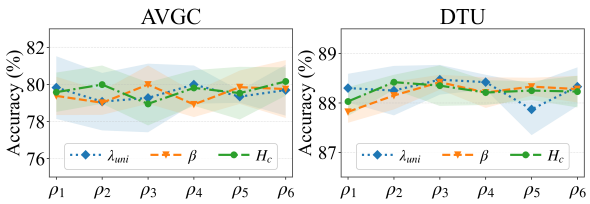


Figure 4: Sensitivity analysis on AVGC and DTU. The trade-off parameter $\lambda_{\text{aux}} \in \{0.1, 0.2, 0.4, 0.8, 1.2, 2.0\}$, the reliability module coefficient $\beta \in \{0.025, 0.05, 0.1, 0.2, 0.4, 0.8\}$, and the number of attention heads $H_c \in \{1, 2, 4, 8, 16, 32\}$.

The accuracy curves remain stable across broad parameter ranges on both datasets. Varying λ_{uni} caused only moderate fluctuations, suggesting that auxiliary supervision does not require precise tuning. The model also maintained competitive performance across different values of β, showing limited sensitivity to the module strength. Similar stability was observed for H_c, as changing the number of attention heads produces no substantial performance degradation. These results demonstrate that RAMamba-Net is robust to hyperparameter variation and requires little dataset-specific tuning. The selected settings were adopted as representative settings within the stable performance ranges.

5 Conclusion

In this work, we proposed RAMamba-Net, a reliability-aware Mamba-based network for multimodal AAD using EEG and EOG. The model employs modality-specific encoders to preserve the distinct characteristics of the two signals. A Mamba-enhanced band-aware encoder captures continuous EEG dynamics, while a temporal-spatial dual-branch encoder models EOG patterns. Cross-modal attention enables explicit information exchange between the modalities. RAM further estimates sample-wise modality reliability, and uses the resulting weights to regulate fusion. Extensive experiments on AVGC and DTU demonstrate that RAMamba-Net consistently improves decoding performance, robustness, and representation quality through explicit cross-modal interaction and sample-wise reliability modeling. Future work will focus on improving the generalizability of EEG decoding under more challenging distribution shifts, including cross-subject and cross-dataset scenarios with diverse users and recording conditions. Moreover, efficient model adaptation and deployment strategies could be explored to reduce computational costs and enable practical multimodal AAD.

AI USE STATEMENT

Generative AI tools were used only for limited language editing to improve the clarity and readability of selected sentences and for assistance with LaTeX formatting and presentation. They were not used for any research task requiring mandatory disclosure under the ICLR 2027 AI Policy, including research ideation, hypothesis formulation, methodology or experiment design, method implementation, data processing or analysis, result interpretation, mathematical claims or proofs, or translation. The authors reviewed and verified all AI-assisted material against the underlying experiments and source literature. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Emina Alickovic, Thomas Lunner, Fredrik Gustafsson, and Lennart Ljung. A tutorial on auditory attention identification methods. *Frontiers in Neuroscience*, 13:153, 2019.

Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *Proc. IEEE Int'l Joint Conf. on Neural Networks*, pp. 2390–2397, Hong Kong, China, Jun. 2008.

Siqi Cai, Peiwen Li, and Haizhou Li. A bio-inspired spiking attentional neural network for attentional selection in the listening brain. *IEEE Trans. on Neural Networks and Learning Systems*, 35 (12):17387–17397, 2023.

Edward Collin Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25:975–979, 1953.

Nai Ding and Jonathan Z. Simon. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. of the National Academy of Sciences*, 109(29):11854–11859, 2012.

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, 2025.

Zhen Fu, Xihong Wu, and Jing Chen. Congruent audiovisual speech enhances auditory attention decoding with EEG. *Journal of Neural Engineering*, 16(6):066033, 2019.

Søren Asp Fuglsang, Torsten Dau, and Jens Hjortkjær. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage*, 156:435–444, 2017.

Quirin Gehmacher, Juliane Schubert, Fabian Schmidt, Thomas Hartmann, Patrick Reisinger, Sebastian Rösch, Konrad Schwarz, Tzvetan Popov, Maria Chait, and Nathan Weisz. Eye movements track prioritized auditory features in selective attention to natural speech. *Nature Communications*, 15(1):3692, 2024.

Simon Geirnaert, Servaas Vandecappelle, Emina Alickovic, Alain De Cheveigne, Edmund Lalor, Bernd T. Meyer, Sina Miran, Tom Francart, and Alexander Bertrand. Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices. *IEEE Signal Processing Magazine*, 38(4):89–102, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Yifan Jiang, Ning Chen, and Jing Jin. Detecting the locus of auditory attention based on the spectro-spatial-temporal analysis of EEG. *Journal of Neural Engineering*, 19(5):056035, 2022.

Nataliya Kosmyna, Arnav Balyan, and Eugene Hauptmann. Target speaker detection with EEG/EOG glasses: A pilot study. In *Proc. of the Future Technologies Conf.*, pp. 433–446, Vancouver, BC, Canada, Oct. 2022.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Nngineering*, 15(5):056013, 2018.

Ke Liu, Tao Yang, Zhuliang Yu, Weibo Yi, Hong Yu, Guoyin Wang, and Wei Wu. MSVTNet: Multi-scale vision Transformer neural network for EEG-based motor imagery decoding. *IEEE Journal of Biomedical and Health Informatics*, 28(12):7126–7137, 2024.

Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. Combining eye movements and EEG to enhance emotion recognition. In *Int'l Joint Conf. on Artificial Intelligence*, volume 15, pp. 1170–1176, Buenos Aires, Argentina, Jul. 2015.

Nima Mesgarani and Edward F. Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236, 2012.

Nhan Duc Thanh Nguyen, Huy Phan, Simon Geirnaert, Kaare Mikkelsen, and Preben Kidmose. AADNet: An end-to-end deep learning model for auditory attention decoding. *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, 33:2695–2706, 2025.

Qinke Ni, Hongyu Zhang, Cunhang Fan, Shengbing Pei, Chang Zhou, and Zhao Lv. DBPNet: Dual-branch parallel network with temporal-frequency fusion for auditory attention detection. In *Int'l Joint Conf. on Artificial Intellignece*, pp. 3115–3123, Jeju Island, South Korea, Aug. 2024.

James A. O'sullivan, Alan J. Power, Nima Mesgarani, Siddharth Rajaram, John J. Foxe, Barbara G. Shinn-Cunningham, Malcolm Slaney, Shihab A. Shamma, and Edmund C. Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, 25(7):1697–1706, 2015.

Corentin Puffay, Bernd Accou, Lies Bollens, Mohammad Jalilpour Monesi, Jonas Vanthornhout, Hugo Van Hamme, and Tom Francart. Relating EEG to continuous speech using deep neural networks: A review. *Journal of Neural Engineering*, 20(4):041003, 2023.

Iustina Rotaru, Simon Geirnaert, Nicolas Heintz, Iris Van de Ryck, Alexander Bertrand, and Tom Francart. What are we really decoding? unveiling biases in EEG-based decoding of the spatial focus of auditory attention. *Journal of Neural Engineering*, 21(1):016017, 2024.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

Ben Somers, Tom Francart, and Alexander Bertrand. A generic EEG artifact removal algorithm based on the multi-channel Wiener filter. *Journal of Neural Engineering*, 15(3):036007, 2018.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional Transformer for EEG decoding and visualization. *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Enze Su, Siqi Cai, Longhan Xie, Haizhou Li, and Tanja Schultz. STAnet: A spatiotemporal attention network for decoding auditory spatial attention from EEG. *IEEE Trans. on Biomedical Engineering*, 69(7):2233–2242, 2022.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for unaligned multimodal language sequences. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, Jul. 2019.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

Servaas Vandecappelle, Lucas Deckers, Neetha Das, Amir Hossein Ansari, Alexander Bertrand, and Tom Francart. EEG-based detection of the locus of auditory attention with convolutional neural networks. *Elife*, 10:e56481, 2021.

Jiaheng Wang, Lin Yao, and Yueming Wang. IFNet: An interactive frequency convolutional neural network for enhancing motor imagery decoding from EEG. *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, 31:1900–1911, 2023.

Ziwei Wang, Siyang Li, and Dongrui Wu. Canine EEG helps human: Cross-species and cross-modality epileptic seizure detection via multi-space alignment. *National Science Review*, 12(6): nwaf086, 2025.

Ziwei Wang, Xingyi He, Tianwang Jia, Hongbin Wang, and Dongrui Wu. FAConformer: Frequency-aware convolutional Transformer for auditory attention decoding. *arXiv preprint arXiv:2606.14120*, 2026a.

Ziwei Wang, Hongbin Wang, Tianwang Jia, Xingyi He, Siyang Li, and Dongrui Wu. DBConformer: Dual-branch convolutional Transformer for EEG decoding. *IEEE Journal of Biomedical and Health Informatics*, 30(5):4134–4147, 2026b.

Johanna Wilroth, Emina Alickovic, Martin A. Skoglund, Carine Signoret, Jerker Rönnberg, and Martin Enqvist. Improving tracking of selective attention in hearing aid users: The role of noise reduction and nonlinearity compensation. *Eneuro*, 12(2), 2025a.

Johanna Wilroth, Oskar Keding, Martin A. Skoglund, Emina Alickovic, and Martin Enqvist. Eye tracking-based speech label estimation for auditory attention decoding with portable EEG. In *Int'l Conf. on Information Fusion*, pp. 1–8, Rio de Janeiro, Brazil, Jul. 2025b.

Xiran Xu, Boda Xiao, Bo Wang, Yujie Yan, Xihong Wu, Heping Cheng, and Jing Chen. Utilizing eyeblink information to improve EEG-based auditory attention decoding. *Biomedical Signal Processing and Control*, 123:110555, 2026.

Sheng Yan, Cunhang Fan, Hongyu Zhang, Xiaoke Yang, Jianhua Tao, and Zhao Lv. DARNet: Dual attention refinement network with spatiotemporal construction for auditory attention detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 31688–31707, Vancouver, BC, Canada, Dec. 2024.

Haolong Yin, Jianming Zhang, Renjie Dai, Weilong Zheng, Qinyu Lv, Zhenghui Yi, and Baoliang Lu. A multimodal EEG-eye movement model for automatic depression detection. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 40, pp. 2218–2226, Singapore, Jan. 2026.

Ziyang Zhang, Andrew Thwaites, Alexandra Woolgar, Brian Moore, and Chao Zhang. SWIM: Short-window CNN integrated with Mamba for EEG-based auditory spatial attention decoding. In *IEEE Spoken Language Technology Workshop*, pp. 1031–1038, Macao, China, Dec. 2024.

Qian Zhao and Weina Zhu. TMSA-Net: A novel attention mechanism for improved motor imagery EEG signal processing. *Biomedical Signal Processing and Control*, 102:107189, 2025.

Wei Zhao, Xiaolu Jiang, Baocan Zhang, Shixiao Xiao, and Sujun Weng. CTNet: A convolutional Transformer network for EEG-based motor imagery classification. *Scientific Reports*, 14(1): 20237, 2024.

Wei Zhao, Baocan Zhang, Haifeng Zhou, Dezhi Wei, Chenxi Huang, and Quan Lan. Multi-scale convolutional Transformer network for motor imagery brain-computer interface. *Scientific Reports*, 15(1):12935, 2025.

Jian Zhou, Yingjie Xie, Cunhang Fan, Huabin Wang, Zhao Lv, and Liang Tao. DHGCN: Dual hypergraph convolutional network for EEG-based auditory attention detection. In *Proc. of the ACM Int'l Conf. on Multimedia*, pp. 612–620, New York, NY, USA, Oct. 2025.

Yifan Zhuang, Calvin Huang, Zepeng Yu, Yongjie Zou, and Jiawei Ju. CAT-Net: A cross-attention tone network for cross-subject EEG-EMG fusion tone decoding. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 40, pp. 18090–18098, Singapore, Jan. 2026.

A Detailed Network Architecture

Table A1: Architecture of RAMamba-Net. B: batch size; C_e, C_o: numbers of EEG and EOG channels; T: number of time points; K_e, K_o: temporal kernel sizes in the EEG and EOG encoders; F_e: number of EEG band-specific filters; F_o: numbers of EOG temporal filters; P_e, P_o: numbers of EEG and EOG temporal patches; N_e: number of EEG frequency bands; D_o: EOG branch embedding dimension; $N_o = P_o + C_o$: number of EOG tokens after temporal-spatial concatenation; D: shared embedding dimension; N_m: number of tokens for modality $m \in \{\mathrm{EEG}, \mathrm{EOG}\}$, where $N_m \in \{N_e, N_o\}$; $N = N_e + N_o$: total number of multimodal tokens; N_c: number of classes.

Stage	Module	Layer	# Kernels	Kernel Size	# Parameters	Output Shape	Options
EEG Encoder	Band Encoder	Mamba Block	–	–	$\sim$	(B, C_e, T)	Temporal refinement
		Stem Layer	F_e	$(K_e, K_e/2, ...)$	$(C_e + 4)F_e$	(B, F_e, P_e)	Band encoding
		Mamba Block	–	–	$\sim$	(B, F_e, P_e)	Band dynamics
		Transformer Encoder	–	–	$\sim$	(B, F_e, P_e)	Within-band context
		Flatten + Band features	–	–	0	$(B, F_e P_e)$	Band representation
	Frequency-Aware Attention	Stacking + Permute	–	–	0	$(N_e, B, F_e P_e)$	Band tokenization
		Transformer Encoder	–	–	$\sim$	$(N_e, B, F_e P_e)$	Cross-band interaction
		Linear	–	–	$F_e P_e D$	(B, N_e, D)	Embedding remapping
EOG Encoder	Temporal Branch	Patch Embedding	F_o	$(K_o,)$	$(K_o + 4)F_o$	(B, P_o, F_o)	Temporal tokenization
		T-Transformer	–	–	$P_o D_o$	(B, P_o, D_o)	Temporal modeling
	Spatial Branch	Patch Embedding	16	25	$16 D_o$	(B, C_o, D_o)	Spatial tokenization
		S-Transformer	–	–	$C_o D_o$	(B, C_o, D_o)	Spatial modeling
	Alignment	Linear	–	–	$N_o N_e$	(B, N_o, D)	Embedding alignment
Cross-Modal Interaction	EEG Enhance	Cross-Modal Attention	–	–	$\sim$	(B, N_e, D)	EOG-guided enrichment
		Feed-Forward	–	–	$(8D + 7)D$	(B, N_e, D)	EEG feature refinement
	EOG Enhance	Cross-Modal Attention	–	–	$\sim$	(B, N_o, D)	EEG-guided enrichment
		Feed-Forward	–	–	$(8D + 7)D$	(B, N_o, D)	EOG feature refinement
Reliability-Aware Fusion	Reliability Estimation	MeanPooling	–	–	0	(B, D)	Detached feature pooling
		MLP	–	–	$\sim$	$(B, 2)$	Reliability-aware scoring
		Softmax	–	–	0	$(B, 2)$	Modality weighting
	Modulation	Feature Rescaling	–	–	0	(B, N_m, D)	Feature modulation
Classifiers	Multimodal Classifier	Flatten	–	–	0	(B, ND)	Main classification
		Max-Norm Linear	–	–	$ND \to N_c$	(B, N_c)	
	Unimodal Classifiers	Linear	–	–	$N_m D \to N_c$	(B, N_c)	Unimodal supervision
		Softmax	–	–			
	Band-Wise Classifiers	Max-Norm Linear	–	–	$F_e P_e \to N_c$	(B, N_c)	Band supervision

Table A1 provides a layer-wise specification of RAMamba-Net. It summarizes the module composition, kernel configuration, parameterization, output shape, and functional role of each operation, complementing the architectural overview and mathematical formulations presented in the main paper.

RAMamba-Net consists of modality-specific EEG and EOG encoders, cross-modal interaction, reliability-aware fusion, and three types of classification heads. The EEG encoder performs band-specific representation learning followed by cross-band integration, while the EOG encoder captures temporal and spatial dependencies through parallel branches. The resulting modality representations are aligned in a shared embedding space and updated through cross-modal attention. RAM subsequently estimates sample-wise modality weights and rescales the interacted representations before multimodal classification.

The multimodal classifier produces the final attended-speaker prediction, whereas the unimodal and band-wise classifiers provide auxiliary supervision for modality-specific and band-specific representations. Together, these components form the complete architecture used in the end-to-end training and inference procedures.

B End-to-End Training and Inference

Based on the network components detailed in Table A1, Algorithm B1 summarizes the end-to-end training and inference procedures of RAMamba-Net. It presents the complete pipeline from modality-specific representation learning and cross-modal interaction to reliability-aware modulation and attended-speaker prediction.

Algorithm B1 RAMamba-Net

Input: Training set $\mathcal{D}_{\text{train}}$; test set $\mathcal{D}_{\text{test}}$
Parameter: Number of EEG frequency bands K; EEG encoder $\{\mathcal{M}^k_{\text{front}}, \mathcal{S}^k, \mathcal{M}^k_{\text{mid}}, \mathcal{T}^k_{\text{band}}\}^K_{k=1}$ and $\mathcal{R}_{\text{EEG}}$; EOG encoder $\mathcal{E}_t$ and $\mathcal{E}_s$; classifiers $\mathcal{F}_c$, $\mathcal{F}_{EEG}$, $\mathcal{F}_{EOG}$, and $\{\mathcal{F}_k\}^K_{k=1}$; trade-off parameter λ_{uni}; modulation scale β; reference level ω
Output: Predicted test labels $\hat{\mathbf{y}}_{\text{test}}$

Initialize all trainable parameters
 # Training
for each training epoch **do**
  for each mini-batch $\{(\mathbf{X}_{i,\text{EEG}}, \mathbf{X}_{i,\text{EOG}}, y_i)\}^N_{i=1} \subset \mathcal{D}_{\text{train}}$ **do**
    Obtain $\mathbf{Z}_{\text{EEG}}$ and $\mathbf{Z}_{\text{EOG}}$ by (3) and (5)
    for $m \in \{\text{EEG}, \text{EOG}\}$ **do**
      Obtain $\mathbf{Z}^{cm}_m$ by (8)
      Obtain $\mathbf{z}_m$ and $\mathbf{P}_m$ by (9) and (11)
    end for
    Compute s_f and s_p by (10) and (12)
    Obtain α_{EEG} and α_{EOG} by (13) and (14)
    Compute $\mathcal{L}_{multi}$, $\mathcal{L}_{uni}$, and $\mathcal{L}_{band}$ by (17)-(19)
    Compute the total loss $\mathcal{L}$ by (20)
    Update all trainable parameters using $\nabla\mathcal{L}$
  end for
end for
 # Inference
for each test sample $(\mathbf{X}^{\text{test}}_{\text{EEG}}, \mathbf{X}^{\text{test}}_{\text{EOG}}) \in \mathcal{D}_{\text{test}}$ **do**
  Obtain $\mathbf{Z}^{cm,\text{test}}_{\text{EEG}}$, $\mathbf{Z}^{cm,\text{test}}_{\text{EOG}}$ by EEG encoder, EOG encoder, and cross-modal attention
  Obtain $\alpha^{\text{test}}_{EEG}$ and $\alpha^{\text{test}}_{EOG}$ by RAM module
  Predict $\hat{\mathbf{y}}_{\text{test}}$
end for

Training. For each mini-batch, the Mamba-enhanced band-aware EEG encoder and the temporal-spatial dual-branch EOG encoder first extract modality-specific representations. Cross-modal attention then enables token-wise information exchange between EEG and EOG while preserving their modality-specific characteristics. RAM module constructs sample-wise reliability representations, and subsequently estimates the modality weights α_{EEG} and α_{EOG}. The resulting weights regulate the cross-modally interacted representations prior to multimodal classification. The entire network is trained end-to-end using the multimodal classification loss, unimodal supervision loss, and band-wise auxiliary loss.

Inference. Given a test sample, the EEG and EOG encoders together with cross-modal attention produce the cross-modally interacted representations $\mathbf{Z}^{\text{cm,test}}_{\text{EEG}}$ and $\mathbf{Z}^{\text{cm,test}}_{\text{EOG}}$. RAM then estimates the sample-wise modality weights from the modality-specific descriptors and the consistency scores. The reliability-modulated EEG and EOG representations are vectorized, concatenated, and fed into the classifier to obtain the predicted test label $\hat{\mathbf{y}}_{\text{test}}$.

C Statistical Significance Analysis

To assess whether the observed improvements remain consistent across different random initializations, we conducted one-sided paired Wilcoxon signed-rank tests between the complete RAMamba-Net and four comparison configurations. For each dataset and evaluation metric, the results obtained with the same five random seeds were paired for statistical testing. The resulting p-values were adjusted using the Benjamini-Hochberg false discovery rate correction to account for multiple comparisons. Table C1 reports the adjusted p-values.

RAMamba-Net significantly outperformed all four comparison approaches on AVGC and consistently surpassed EOG-only, EEG-only, and concatenation across all metrics on DTU, indicating stable and broadly effective performance gains across both datasets. Compared with RAMamba-Net w/o RAM, the complete model achieved significant improvements in BCA and Cohen's Kappa, suggesting that RAM primarily improves balanced and consistent decoding on this dataset.

Table C1: Adjusted p-values between RAMamba-Net and four comparison approaches on AVGC and DTU across four evaluation metrics. Statistically significant results ($p < 0.05$) are marked in bold.

Dataset	Approach	Accuracy	BCA	Macro-F1	Cohen's Kappa
AVGC	EOG-only	**<0.05**	**<0.05**	**<0.05**	**<0.05**
	EEG-only	**<0.05**	**<0.05**	**<0.05**	**<0.05**
	Concat.	**<0.05**	**<0.05**	**<0.05**	**<0.05**
	RAMamba-Net w/o RAM	**<0.05**	**<0.05**	**<0.05**	**<0.05**
DTU	EOG-only	**<0.05**	**<0.05**	**<0.05**	**<0.05**
	EEG-only	**<0.05**	**<0.05**	**<0.05**	**<0.05**
	Concat.	**<0.05**	**<0.05**	**<0.05**	**<0.05**
	RAMamba-Net w/o RAM	0.062	**<0.05**	0.219	**<0.05**

Overall, 30 of the 32 comparisons reached statistical significance. These results demonstrate that the improvements of RAMamba-Net are generally consistent across random initializations rather than being caused by isolated experimental fluctuations. They further support the complementary roles of modality-specific encoding, cross-modal interaction, and reliability-aware fusion in improving multimodal AAD.