Band-Gated Identity-Disentangled Training for Cross-Subject Auditory Attention Decoding

Siying Tao[1], Jingjing Hu[1], Jinyang Huang[1], Fengqi Cui[1], Xueliang Liu[1], Dan Guo[1,2,3]

[1]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui, 230601, China
[2]The Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, China [3]The Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, 230026, China

Abstract

EEG-based auditory attention decoding (AAD) seeks to identify which speaker a listener attends to in multi-talker "cocktail party" settings. A central challenge is cross-subject generalization: neural responses vary substantially across individuals, inducing distribution shifts across subjects; consequently, models trained in subject-dependent or mixed-subject regimes may latch onto subject-specific cues that hinder transfer and interpretation. To address this, we propose a band-gated multi-band framework that decomposes EEG into low- and high-frequency pathways and adaptively fuses them at the sample level to learn attention-discriminative representations while accommodating inter-individual spectral variability. We further introduce an identity-disentangled objective that leverages confidence-filtered pseudo-labels to perform alignment in an auxiliary bottleneck space, encouraging a more subject-invariant bottleneck representation while mitigating subject-specific variability. Evaluated on KUL, DTU, and AVED under leave-one-subject-out protocols and two decision-window settings, our approach achieves the best or highly competitive performance against strong baselines. Analyses of the learned gate and representation geometry provide qualitative support for the roles of adaptive band reweighting and identity suppression in improving robustness in cross-subject AAD. Code is available at https://github.com/siyingtao/BDGI_for_AAD.

Keywords: Auditory Attention Decoding (AAD); EEG; Cross-subject Generalization; Identity-Disentangled Learning

Introduction

In multi-speaker environments, listeners can focus on a target talker while suppressing competing sources — a phenomenon known as the cocktail party effect (Cherry, 1953). This ability relies on auditory attention to prioritize goal-relevant input and on working memory to maintain and continuously update the target stream under interference, offering a naturalistic window into how cognitive control shapes auditory perception. Importantly, selective listening is often difficult in real-world communication and is central to assistive hearing and clinical assessment; these challenges are amplified by substantial inter-individual variability in auditory attention (Tune et al., 2021) and may be especially pronounced for older adults or individuals with hearing loss. EEG-based auditory attention decoding (AAD) aims to infer the attended speech stream from neural recordings, providing a noninvasive and temporally precise readout of attention (Dai et al., 2018). Such neural readouts could enable future assistive hearing and objective clinical assessment of auditory attention.

From a modeling perspective, decoding auditory attention from EEG is challenging because attention-related signatures are subtle and distributed: they depend on dynamic listening context, involve cross-channel/region interactions, and unfold over time as attention is maintained and updated. Recent AAD systems (Cai et al., 2024; Fan et al., 2024, 2025; Tune et al., 2021) leverage deep networks to learn stronger spatio-temporal representations, often complemented by subject-dependent training or calibration. Building on this trend, end-to-end spatio-temporal models such as IFNet, DBPNet, and DARNet (Ni et al., 2024; Wang et al., 2023; Yan et al., 2024) further improve the ability to capture nonlinear and complex neural dynamics beyond earlier linear pipelines. These advances have led to strong performance in within-subject or subject-dependent scenarios, especially when some form of per-subject calibration is available, but they do not directly address whether attention-relevant structure transfers across listeners without relying on listener-specific cues.

However, a complementary and scientifically crucial goal is cross-subject generalization, where the held-out listener remains label-free during training while unlabeled target EEG may be used for adaptation. Cross-subject EEG decoding faces substantial challenges due to inter-individual variability (*e.g.*, physiology, sensor placement, and recording conditions), which induces distribution shifts across subjects (Guo et al., 2024; Huang et al., 2026; Quiñonero-Candela et al., 2008). More critically, AAD models trained in subject-dependent or mixed-subject regimes may inadvertently latch onto subject-specific cues that correlate with supervision, limiting cross-subject transfer and making interpretation less consistent across individuals. These observations motivate a closer examination of how to learn subject-invariant, attention-specific representations, leading to two key challenges: (1) building a backbone that captures transferable attention patterns under inter-individual spectral variability, and (2) designing training objectives that suppress subject-specific nuisance variability and reduce cross-subject distribution gaps without eroding task-relevant structure.

To address these two challenges, we propose a band-gated multi-band framework with an identity-disentangled training objective for cross-subject AAD. First, we decompose EEG into complementary frequency pathways and use a lightweight BandGate to adaptively fuse them at the sample level, en-

Corresponding author: Jingjing Hu

abling the model to capture attention-related dynamics spanning time scales and frequency bands while accommodating inter-individual spectral variability. This design is also supported by the view that functionally relevant EEG activity often manifests as band-specific rhythmic dynamics, making multi-band modeling a natural way to capture complementary cues (Pfurtscheller & Da Silva, 1999; Wang et al., 2023). Second, we adopt distribution alignment to mitigate cross-subject shifts and extend it with within-source subject-invariant regularization, reducing identity-related variability without collapsing task-relevant distinctions in the classifier space (Ben-David et al., 2010; Li et al., 2018; Liang et al., 2025). Our contributions are threefold:

- We introduce a band-gated multi-band backbone with sample-adaptive fusion to model attention dynamics across complementary frequency pathways, improving the extraction of transferable patterns.
- We propose an identity–disentangled objective that strengthens cross-subject generalization by suppressing subject-specific nuisance variability from task-relevant structure.
- We provide systematic LOSO evaluations and analyses (*e.g.*, BandGate visualizations, ablations, and representation-geometry inspection) across three public datasets (KUL, DTU, AVED) and two decision windows. The results consistently support the roles of identity suppression and band-adaptive fusion in improving cross-subject decoding.

Method

We consider cross-subject EEG-based auditory attention decoding (AAD), where a model must decode the attended speaker for held-out listeners whose ublabeled EEG. In this setting, robust decoding requires both (i) extracting generic neural signatures of auditory attention that are consistent across subjects, and (ii) adapting to subject-specific variability in how these patterns are expressed across cortical regions. To address the band-specific and temporally extended nature of EEG dynamics as well as identity-induced distribution shifts across subjects, we propose a framework with two components: (i) a band-specific spatio-temporal backbone that focuses on **learning generic, frequency- and scale-resolved attention-related patterns**, and (ii) an auxiliary identity-disentangled alignment branch that helps the **model adaptively handle subject-dependent deviations** (Figure 1).

Problem Definition

Let $\mathbf{X} \in \mathbb{R}^{C\times T}$ denote an EEG segment with C channels and T time points, and $y \in \{0,1\}$ the attended-speaker label. Each trial is associated with a subject identity $s \in \mathcal{S}$, and subject s induces a data-generating distribution $\mathcal{P}_s$ over $(\mathbf{X}, y)$. We observe labeled samples from source subjects $\mathcal{S}_S$ and unlabeled samples from held-out target subjects $\mathcal{S}_T$: $\mathcal{D}_S = \{(\mathbf{X}_i, y_i, s_i)\}_{i=1}^{N_S}$, $\mathcal{D}_T = \{(\mathbf{X}_j, s_j)\}_{j=1}^{N_T}$, where $s_i \in \mathcal{S}_S$, $s_j \in \mathcal{S}_T$, $\mathcal{S}_S \cap \mathcal{S}_T = \emptyset$. We aim to learn $f_\theta : \mathbb{R}^{C\times T} \to [0,1]$ that predicts $p_\theta(y = 1 \mid \mathbf{X})$ and generalizes from source to unseen target subjects by extracting subject-invariant yet attention-discriminative representations.

Band-Specific Spatio-Temporal Backbone

Band-specific temporal encoder. EEG attention responses are known to be organized in frequency-specific, temporally extended oscillatory patterns. To better capture such generic neural dynamics, we first decompose raw EEG $\mathbf{X} \in \mathbb{R}^{C\times T}$ into B band-limited components via $\Phi_{\text{band}} : \mathbb{R}^{C\times T} \to \mathbb{R}^{B\times C\times T}$, thus $\mathbf{X}^{(1:B)} = \Phi_{\text{band}}(\mathbf{X})$, where each $\mathbf{X}^{(b)} \in \mathbb{R}^{C\times T}$ corresponds to a pre-defined frequency band (zero-phase band-pass filter bank, 4th-order Butterworth). This band-wise "split-screen" design allows the backbone to focus on band-specific yet subject-agnostic oscillatory signatures of auditory attention. For each band b, the temporal encoder $\Phi_{\text{temp}}^{(b)} : \mathbb{R}^{C\times T} \to \mathbb{R}^{C\times T}$ stacks multi-scale depth-wise 1D convolutions with residual refinement to jointly model short-, medium-, and long-range dependencies. Let $\{k_1, k_2, k_3\} = \{15, 31, 63\}$ and $\mathbf{F}_0^{(b)} = \mathbf{X}^{(b)}$; under our sampling rates, these kernels cover progressively longer temporal contexts, allowing the encoder to capture both transient responses and slower attentional modulations in a unified manner. For $m = 1, 2, 3$,

$$\begin{aligned}
\widetilde{\mathbf{F}}_m^{(b)} &= \mathcal{T}_{k_m}\big(\mathbf{F}_{m-1}^{(b)}\big),\\
\mathcal{T}_k(\mathbf{U}) &= \phi\Big(\mathrm{BN}\Big(\mathrm{Conv}_{1\times1}\Big(\big\|_{r\in\{\frac{k}{4},\frac{k}{2},k\}} \mathrm{DWConv}_r(\mathbf{U})\Big)\Big)\Big),\\
\mathbf{F}_m^{(b)} &= \widetilde{\mathbf{F}}_m^{(b)} + \sum_{\ell=1}^{L} \mathcal{R}_\ell\big(\widetilde{\mathbf{F}}_m^{(b)}\big),
\end{aligned} \tag{1}$$

where DWConv_r is a depth-wise 1D convolution of kernel size r along time, BN is batch normalization, ϕ is a pointwise nonlinearity, $\|$ denotes channel-wise concatenation, and $\mathcal{R}_\ell$ are L stacked 1D residual blocks of kernel size 3. After $M = 3$ blocks we obtain $\mathbf{F}^{(b)} = \mathbf{F}_M^{(b)}$ and $\mathbf{F}^{(1:B)} = \Phi_{\text{temp}}\big(\mathbf{X}^{(1:B)}\big)$, denoted as $\mathbf{F}_{\text{low}}$ and $\mathbf{F}_{\text{high}}$ in the two-band case.

BandGate and spatial attention embedding. While the band-specific multi-scale encoder focuses on capturing generic oscillatory patterns of auditory attention, different subjects may emphasize these patterns in different frequency ranges and spatial locations. To adaptively fuse bands and aggregate spatial information in a sample-adaptive manner, we couple a BandGate module with a spatial attention embedding. Given band-specific features $\mathbf{F}_{\text{low}}, \mathbf{F}_{\text{high}} \in \mathbb{R}^{C\times T}$ (two-band case), BandGate first computes a global summary vector and a gate, then interpolates between bands:

$$\begin{aligned}
\mathbf{u} &= \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix}\mathbf{F}_{\text{low}}(:,t)\\ \mathbf{F}_{\text{high}}(:,t)\end{bmatrix} \in \mathbb{R}^{2C}, \quad \mathbf{g} = \sigma(\mathrm{MLP}(\mathbf{u})) \in (0,1)^d,\\
\mathbf{F} &= \mathbf{g}\odot\mathbf{F}_{\text{low}} + (\mathbf{1}-\mathbf{g})\odot\mathbf{F}_{\text{high}},
\end{aligned} \tag{2}$$

where $d = 1$ (global gating) or $d = C$ (channel-wise gating), $\odot$ denotes element-wise multiplication, and $\mathbf{1}$ is an all-ones tensor broadcast to the shape of $\mathbf{g}$. Intuitively, BandGate lets the network softly shift emphasis between low- and high-frequency components, providing a simple mechanism to ac-

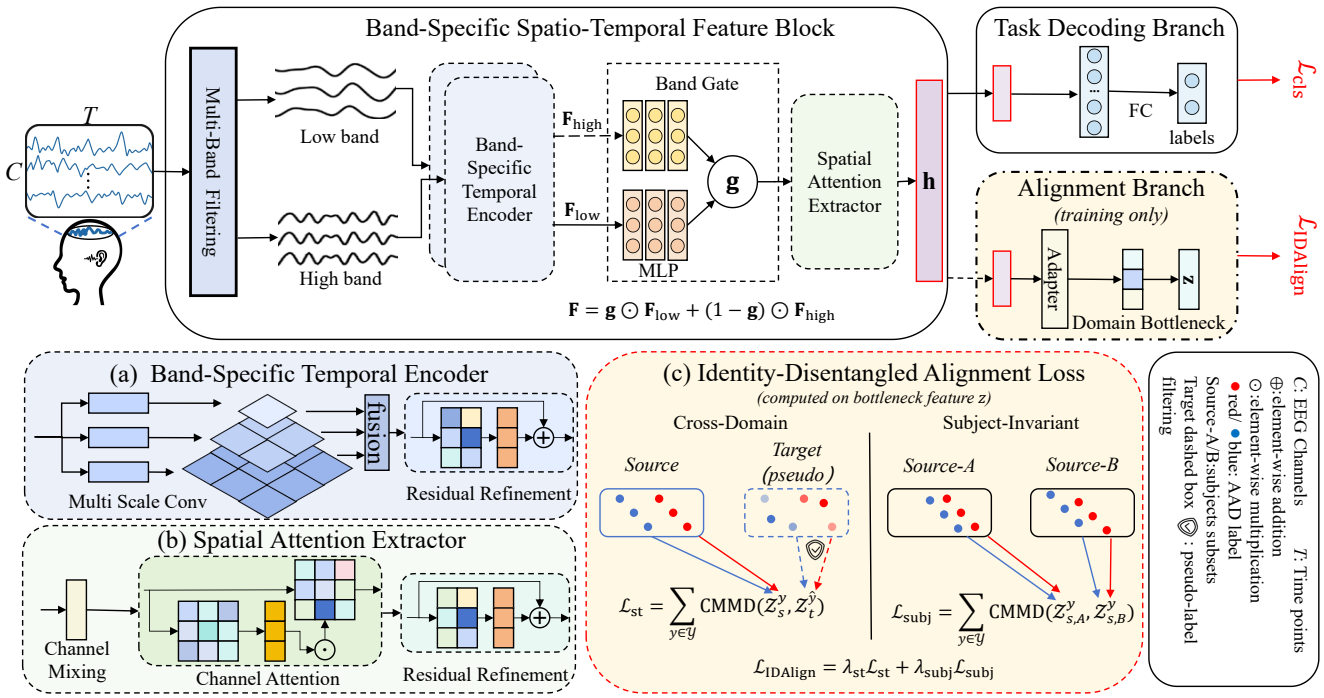

Figure 1: Overview of the proposed band-gated, identity-disentangled training framework for cross-subject auditory attention decoding. EEG is decomposed into low/high-frequency bands, encoded by band-specific temporal encoders, and adaptive fused via a sample-wise BandGate to form the classifier embedding $\mathbf{h}$. During training only, an auxiliary alignment branch maps $\mathbf{h}$ to a bottleneck representation $\mathbf{z}$ and optimizes (i) cross-domain class-conditional alignment between labeled source and pseudo-labeled target samples, and (ii) within-domain subject-invariant alignment across source subject splits.

commodate inter-individual spectral variability in how attentional oscillations distribute across frequencies. On top of the fused representation $\mathbf{F} \in \mathbb{R}^{C\times T}$, we apply a residual spatial encoder $\Phi_{\text{spat}} : \mathbb{R}^{C\times T} \to \mathbb{R}^{d_h}$ with grouped 1×1 convolutions and channel attention. Let $\mathbf{Z}_0 = \mathbf{F}$. A spatial block performs

$$\begin{aligned}\widetilde{\mathbf{Z}} &= \text{Dropout}\Big(\text{CA}\big(\phi_2(\text{BN}_2(\text{Conv}^{(2)}_{1\times1}(\phi_1(\text{BN}_1(\text{Conv}^{(1)}_{1\times1}(\mathbf{Z}_0)))))\big)\Big),\\ \mathbf{Z} &= \phi_{\text{out}}(\mathbf{Z}_0 + \widetilde{\mathbf{Z}}),\end{aligned} \tag{3}$$

where $\text{Conv}^{(1,2)}_{1\times1}$ are grouped 1×1 convolutions across channels, BN_1, BN_2 are normalization layers, $\phi_1, \phi_2, \phi_{\text{out}}$ are pointwise nonlinearities, and CA denotes channel attention. Temporal log-power pooling and flattening then yield the pre-classifier embedding $\mathbf{h} = \Phi_{\text{spat}}(\mathbf{F}) \in \mathbb{R}^{d_h}$, and the backbone as a whole computes $\mathbf{h} = f_{\text{backbone}}(\mathbf{X}) = \Phi_{\text{spat}} \circ \Phi_{\text{gate}} \circ \Phi_{\text{temp}} \circ \Phi_{\text{band}}(\mathbf{X})$, on top of which the task branch applies a classifier $f_{\text{cls}} : \mathbb{R}^{d_h} \to [0, 1]$.

Identity-Disentangled Alignment Branch

To regularize subject-specific factors without corrupting task-discriminative features, we attach an alignment branch on top of the pre-classifier embedding $\mathbf{h} \in \mathbb{R}^{d_h}$ and obtain a bottleneck feature $\psi : \mathbb{R}^{d_h} \to \mathbb{R}^{d_z}, \mathbf{z} = \psi(\mathbf{h})$, which is used only for alignment. The overall identity-disentangled loss is $\mathcal{L}_{\text{IDAlign}} = \lambda_{\text{st}}\mathcal{L}_{\text{st}} + \lambda_{\text{subj}}\mathcal{L}_{\text{subj}}$ with both terms computed in the bottleneck space to constrain identity-sensitive variation instead of directly distorting $\mathbf{h}$. We use class-conditional maximum mean discrepancy (CMMD) (Ren et al., 2019) with Gaussian kernel $k(\mathbf{a}, \mathbf{b}) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{a} - \mathbf{b}\|_2^2\right)$, and its standard empirical estimator on class-specific bottleneck sets. Following neighboring EEG domain-alignment work (Ge et al., 2023; Ma et al., 2025), $\mathcal{L}_{\text{st}}$ is a conventional class-conditional source–target CMMD computed on $\mathbf{z}$ with high-confidence pseudo-labels for target samples.

Within-source subject-invariant alignment. We further propose a *within-source* subject-invariant alignment term that directly targets identity-related variability inside the source domain. Let $\mathcal{S}_S$ be the set of source subjects, and randomly split it into two disjoint subsets A, B at each iteration. Given a labeled source mini-batch $\mathcal{B}_S = \{(\mathbf{z}_i^s, y_i^s)\}$, for each class $y \in \mathcal{Y}$ we form $\mathcal{Z}^y_{s,A} = \{\mathbf{z}_i^s \mid y_i^s = y,\ \text{subj}(i) \in A\}$, $\mathcal{Z}^y_{s,B} = \{\mathbf{z}_i^s \mid y_i^s = y,\ \text{subj}(i) \in B\}$ and define

$$\mathcal{L}_{\text{subj}} = \sum_{y\in\mathcal{Y}} \text{CMMD}\big(\mathcal{Z}^y_{s,A}, \mathcal{Z}^y_{s,B}\big). \tag{4}$$

By enforcing class-conditional similarity between independently sampled subject groups A and B, $\mathcal{L}_{\text{subj}}$ pushes different source subjects to share a common, subject-invariant bottleneck representation, while the classifier still operates on $\mathbf{h}$ and retains fine-grained attention-discriminative information.

Training and Inference

The task branch applies f_{cls} on $\mathbf{h}$: $\hat{y} = f_{\text{cls}}(\mathbf{h})$, and $\mathcal{L}_{\text{cls}} = \mathbb{E}_{(\mathbf{X},y)\sim\mathcal{D}_S}\left[\ell_{\text{CE}}(f_{\text{cls}}(f_{\text{backbone}}(\mathbf{X})), y)\right]$, with binary

cross-entropy loss ℓ_{CE}. The alignment branch maps $\mathbf{h}$ to $\mathbf{z} = \psi(\mathbf{h})$ and is optimized by $\mathcal{L}_{\text{IDAlign}}$. The full objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{IDAlign}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{st}}\mathcal{L}_{\text{st}} + \lambda_{\text{subj}}\mathcal{L}_{\text{subj}}. \quad (5)$$

During training, both branches are jointly updated; at inference, the alignment branch is removed and predictions are made solely by $\hat{y} = f_{\text{cls}}(f_{\text{backbone}}(\mathbf{X}))$, so deployment adds no extra overhead beyond the backbone classifier.

Experiments

Datasets and Preprocessing

We primarily focus on KUL (Das et al., 2019), a well-established AAD benchmark in this work, and include DTU (Fuglsang et al., 2018) and AVED (ZHANG, ZHANG, et al., 2024) as complementary testbeds to assess robustness under more challenging acoustic and multimodal conditions. In all datasets, subjects are instructed to attend to one of two competing speech streams. **KUL** and **DTU** are audio-only attention decoding datasets with 16 and 18 subjects, respectively, both recorded with 64-channel EEG and around 50 min of data per subject. **AVED** is a newer and relatively small-scale dataset that includes audio-only and audio-visual settings (10 subjects each; 32 channels), where the latter adds visual cues by showing the attended narrator. Across datasets, we apply a consistent preprocessing pipeline (Yan et al., 2024) including 0.1–50 Hz band-pass filtering, standard artifact removal, and re-referencing, and use dataset-specific resampling rates (128 Hz for KUL and AVED, and 64 Hz for DTU).

Evaluation Protocol

We adopt a cross-subject leave-one-subject-out (LOSO) protocol. In each fold, one subject is held out as the target domain and all remaining subjects form the source domain. Only the unlabeled EEG of the target subject is used during training (*e.g.*, for computing $\mathcal{L}_{\text{st}}$), and its ground-truth labels are used solely for final evaluation. Following (Ni et al., 2024; Yan et al., 2024), we use classification accuracy as the evaluation metric. For each dataset and each decision window, we report the mean and standard deviation across all LOSO folds. Consistent with common AAD protocols, we consider both 1-second (1-s) and 2-second (2-s) decision windows.

Baselines

To assess the effectiveness of the proposed method in the challenging cross-subject setting, we compare against four representative state-of-the-art AAD models with publicly available implementations: **SSF-CNN** (Cai et al., 2021): Fuses EEG alpha-band features with a CNN backbone for low-latency auditory spatial attention detection. **MBSSFCC** (Jiang et al., 2022): Combines multi-band differential-entropy features with a ConvLSTM network for high-precision auditory spatial attention detection without requiring auditory stimuli. **DBPNet** (Ni et al., 2024): Uses a dual-branch parallel architecture (temporal-attention branch and frequency-residual branch) to jointly model temporal, spectral, and spatial EEG characteristics. **DARNet** (Yan et al., 2024): Captures spatio-temporal patterns and long-range dependencies in EEG via spatio-temporal construction, dual-attention refinement, and multi-level feature fusion modules.

Implementation Details

For the **backbone**, we adopt a two-band (B=2), zero-phase filter-bank front-end to obtain low- and high-band EEG signals. Specifically, we use a 4th-order Butterworth band-pass filter with forward–backward filtering (filtfilt). The low band is set to [4, 16] Hz and the high band to [16, 40] Hz; for DTU, the high band is adjusted to [16, 31] Hz to avoid a cutoff too close to the Nyquist limit under its sampling rate. Unless otherwise specified, we instantiate BandGate as a global scalar gate (d=1) in all experiments. The temporal encoder uses multi-scale kernels $\{15, 31, 63\}$, and the spatial extractor outputs a d_h=48-dimensional embedding $\mathbf{h}$. Dropout with rate 0.2 is applied in the spatial block. For the **loss and optimization**, we set the bottleneck dimension to d_z (default d_z=48 unless otherwise noted) and use class-conditional MMD losses $\mathcal{L}_{\text{st}}$ and $\mathcal{L}_{\text{subj}}$ with a Gaussian kernel. Training uses a fixed configuration across all datasets for fairness and efficiency: batch size of 128, up to 100 epochs, and the Adam optimizer with a learning rate of 1×10^{-3} and weight decay of 3×10^{-4}. The IDAlign loss is activated after epoch 2 using a warm-up schedule that gradually ramps its weight to λ_{st} and λ_{subj}. All experiments are conducted on a workstation equipped with NVIDIA A40 GPUs (48 GB memory).

Results

Comparison with Prior Art

Table 1 summarizes cross-subject AAD accuracy. Overall, our band-gated, identity-disentangled framework achieves the best or highly competitive performance across datasets and window settings with a compact parameter budget, transferring well from earlier audio-only benchmarks (KUL, DTU) to the more recent AVED dataset. (1) On the two **audio-only** datasets **KUL** and **DTU**, our method consistently outperforms prior deep-learning baselines. On KUL, it improves over **DARNet** for both 1-s and 2-s windows (72.5% *vs*. 69.9%, 73.8% *vs*. 71.9%) with only 0.335M parameters. On DTU, it attains the highest mean accuracy for both 1-s and 2-s windows (58.3% and 58.4%), significantly outperforming the best prior baseline while remaining much smaller than MBSSFCC. These results indicate strong cross-subject transfer with a small backbone and highlight the benefit of combining band-specific modeling with identity-disentangled alignment. (2) On the more recent **AVED (audio-only)** split, our model is on par with the best baseline DBPNet, being slightly better at the 1-s window (52.5% *vs*. 52.1%) and slightly worse at the 2-s window (52.6% *vs*. 53.3%), suggesting that the gain of the proposed alignment is smaller under the current setup. (3) On **AVED (audio-visual)**, a more challenging setting with additional visual stimulation, our method achieves the best performance at both 1-s and 2-s windows (54.8% and 54.3%),

Table 1: Cross-subject AAD accuracy (%, mean±std over LOSO subjects) on KUL, DTU, and AVED (audio-only / audio-visual) under 1-second and 2-second decision windows. "Params" denotes trainable parameters. Best results are highlighted in **bold**. *: $p < 0.05$ vs. the best prior baseline across LOSO folds.

Method	Venue	Params (M)	KUL (audio-only)		DTU (audio-only)		AVED (audio-only)		AVED (audio-visual)	
			1-second	2-second	1-second	2-second	1-second	2-second	1-second	2-second
SSF-CNN	*EMBC'21*	4.21	59.3±6.69	60.8±8.40	52.3±3.50	53.4±4.16	51.7±0.85	52.5±1.55	52.4±2.29	53.8±2.27
MBSSFCC	*Neural Eng'22*	83.91	62.7±8.08	64.7±8.62	52.5±4.35	53.9±5.80	52.2±1.52	52.7±1.87	52.8±1.57	54.1±1.86
DBPNet	*IJCAI'24*	0.91	61.1±8.26	62.3±7.37	55.5±6.33	55.8±6.11	52.1±1.19	**53.3**±1.88	53.3±2.39	54.0±1.61
DARNet	*NeurIPS'24*	0.08	69.9±11.82	71.9±13.01	55.6±4.13	55.6±4.04	51.3±0.21	52.1±1.54	51.4±0.32	52.6±0.29
Ours (full)	–	0.335	**72.5***±10.64	**73.8***±11.06	**58.3***±3.95	**58.4***±3.71	**52.5**±1.10	52.6±1.43	**54.8***±1.92	**54.3***±2.97
Backbone	① Ours w/o low band		71.6±10.41	73.2±11.10	57.6±3.71	57.8±4.17	51.9±1.16	51.6±1.95	53.0±2.16	53.0±2.16
	② Ours w/o high band		61.8±4.96	62.8±4.99	57.3±4.58	57.2±3.85	51.6±0.07	52.1±1.04	53.3±1.98	53.4±2.13
	③ Ours w/o gate		71.2±10.74	73.2±10.62	57.3±4.65	57.8±4.65	51.4±0.85	51.8±1.55	53.7±2.45	53.1±2.17
Domain loss	④ Ours w/o IDAlign		71.5±10.26	72.4±10.29	55.9±3.70	56.4±5.01	51.7±1.45	52.1±1.29	53.5±1.18	53.8±1.79
	⑤ Ours w/o subj		72.4±10.38	71.4±11.42	57.5±3.68	57.1±4.02	52.4±1.38	52.4±1.51	53.8±2.24	52.9±2.07
	⑥ Ours w/o st		71.2±10.33	71.7±10.48	57.5±3.81	57.8±3.50	52.2±1.50	51.3±2.01	53.9±2.16	53.6±2.89

outperforming all baselines and suggesting that the method remains competitive when additional visual cues are present. Across **all** datasets, the relative ranking of methods remains largely stable between 1-s and 2-s windows, suggesting that our gains are not window-specific but instead reflect more reliable cross-subject generalization.

Ablation Analysis

Network structure ablations. Unless stated otherwise, all ablations follow the same LOSO protocol, preprocessing, and optimization settings as the full model; we only remove the specified component/loss. Table 1 reports results for backbone ablations (①–③) and domain-loss ablations (④–⑥).

- **Backbone ablations (①–③).** We assess the low-/high-band pathways and BandGate fusion. On KUL, removing the high-band pathway output $\mathbf{F}_{\text{high}}$ (②) yields a pronounced drop (72.5/73.8 → 61.8/62.8 for 1-s/2-s), while removing the low-band pathway output $\mathbf{F}_{\text{low}}$ (①) slightly affects performance (72.5/73.8 → 71.6/73.2), indicating that high-band high-band cues appear particularly informative under the current preprocessing and evaluation setting. Across DTU and AVED, ablating either band (① or ②) causes moderate but consistent degradations (typically ∼0.5–1.8 points; *e.g.*, DTU 58.3/58.4 → 57.6/57.8 or 57.3/57.2), suggesting complementary contributions from low and high frequencies for cross-subject robustness. Moreover, replacing adaptive gating fusion (Eq. 2) with a non-gated variant uniform fusion (equal weight for both bands) (③) further reduces accuracy across datasets (*e.g.*, DTU: 58.3/58.4 → 57.3/57.8; AVED-AV: 54.8/54.3 → 53.7/53.1), supporting sample-adaptive reweighting between bands.
- **Domain-loss ablations (④–⑥).** We analyze IDAlign loss design. Removing $\mathcal{L}_{\text{IDAlign}}$ entirely (④) consistently harms cross-subject generalization across all settings, including KUL (72.5/73.8 → 71.5/72.4) and with the most evident drop on DTU (58.3/58.4 → 55.9/56.4), confirming that reducing identity-related bias provides a large empirical gain.

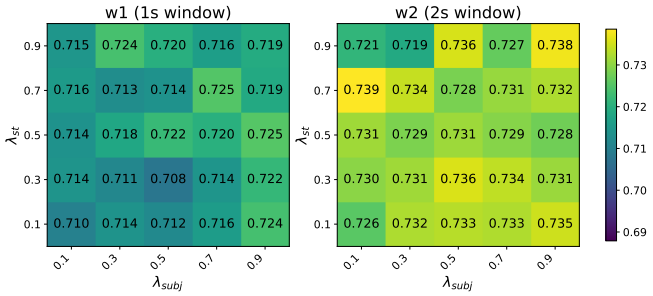

Figure 2: Grid search of IDAlign weights (λ_{st}, λ_{subj}) on KUL. Heatmaps show mean accuracy under 1s and 2s windows.

Ablating either term alone—subject-invariant alignment term $\mathcal{L}_{\text{subj}}$ (⑤) or source–target alignment term $\mathcal{L}_{\text{st}}$ (⑥), both recovers part of the gain but remains below the full objective (*e.g.*, DTU: 57.5/57.1 and 57.5/57.8 *vs.* 58.3/58.4), indicating complementary effects across datasets and decision windows. Thus, IDAlign loss branch provides the primary cross-subject benefit, while the band-specific backbone and gated fusion contribute additional, consistent improvements.

IDAlign loss sensitivity ablation. We analyze the sensitivity to the IDAlign weights ($\lambda_{\text{st}}, \lambda_{\text{subj}}$) in Eq. 5, which scale the two alignment terms. On KUL (LOSO), Figure 2 reports a 5×5 grid search for w1 (1s window, left) and w2 (2s window, right). Overall, the heatmaps exhibit a broad region of competitive performance, indicating that IDAlign is not overly sensitive to a specific weight choice. For the w1, the best accuracy is achieved with moderate-to-large weights (*e.g.*, around λ_{st}=0.5–0.7 and λ_{subj}=0.7–0.9, while the worst setting is only moderately lower (0.708 *vs.* 0.725), suggesting a relatively smooth response surface. For the w2, the optimum occurs at $(\lambda_{\text{st}}, \lambda_{\text{subj}}) = (0.7, 0.1)$ with 0.739 accuracy, and several nearby configurations yield similarly strong results. Comparing w1 and w2, w2 shows a slightly flatter landscape with multiple near-optimal regions, implying reduced sensitivity

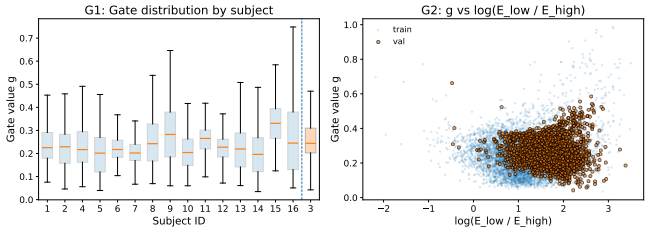

Figure 3: BandGate on KUL (LOSO): (G1) subject-wise gate distribution; (G2) gate *vs*. log band-energy ratio.

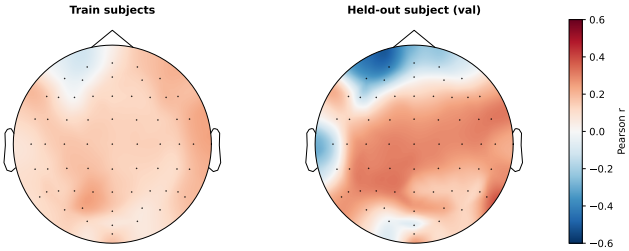

Figure 4: Electrode-level Pearson correlation between the learned BandGate coefficient **g** and the log band-energy ratio r_c on KUL (left: training, right: held-out).

when more temporal evidence is available.

Visualization Analysis

BandGate Visualization. To interpret the learned band-fusion behavior, we visualize the *learned* BandGate coefficient **g** (Eq. 2). Overall, rather than acting as a fixed preference for a specific band, BandGate adaptively reweights low/high-band features conditioned on sample-level characteristics. (1) In Figure 3, **G1: Subject-wise distribution.** Gate values vary systematically across training subjects yet remain in a bounded mid-range without collapsing to 0 or 1, indicating stable soft fusion rather than a hard switch. **G2: Beyond a power-ratio proxy.** The relationship between **g** and the log band-energy ratio $\log(E_{\text{low}}/E_{\text{high}})$ is correlated but clearly non-monotonic, with substantial variability of g even at similar ratios, suggesting that gating is not a trivial power-ratio heuristic. The held-out subject largely overlaps with the training range in both G1 and G2, implying that the learned gating policy generalizes beyond training identities. (2) In Figure 4, electrode-wise correlations between **g** and $r_c = \log(E^{(c)}_{\text{low}}/E^{(c)}_{\text{high}})$ exhibit non-uniform but spatially structured scalp patterns and remain smooth on the held-out subject, indicating that gating leverages localized spatio-spectral variations rather than global power alone. Differences between training and held-out topographies suggest subject-adaptive yet bounded fusion rather than noise-driven decisions.

Domain Bottleneck and Identity Suppression. To verify that our alignment branch suppresses identity nuisance factors without disrupting the task-discriminative geometry, we visualize the pre-classifier embedding **h** and the domain bottleneck representation **z** on KUL using UMAP (Figure 5). The four

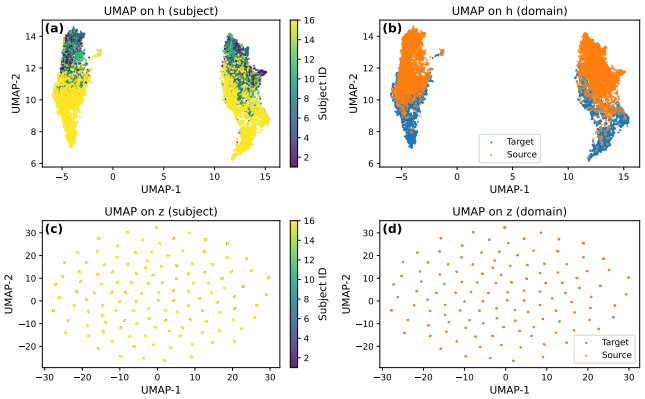

Figure 5: UMAP on KUL contrasting **h** (a,b) and **z** (c,d). Subject coloring is shown in (a,c) and domain coloring in (b,d), illustrating stronger subject-wise structure in **h** and more mixed subject structure in **z**.

panels contrast subject-wise structure and source/target mixing in **h** *vs*. **z**. In Figure 5(a), UMAP on **h** colored by subject shows clear subject-dependent structure, indicating that h still carries identity-related cues. Consistently, Figure 5(b) colored by domain reveals an evident separation between source and target samples, suggesting distribution shift is present in the discriminative space. In contrast, when projecting the bottleneck **z**, subject-wise clustering largely disappears (Figure 5(c)), and source/target samples become well mixed (Figure 5(d)). These results support our design choice: **z** serves as an identity-suppressed alignment space to reduce domain bias, while **h** remains responsible for task discrimination, thereby decoupling alignment from classifier embedding.

Conclusion

In this paper, we study cross-subject EEG-based auditory attention decoding (AAD), where generalization remains challenging due to substantial inter-subject variability and identity-related confounds. We propose a band-gated, identity-disentangled framework that couples sample-adaptive band fusion with an alignment objective to improve cross-subject transfer while reducing identity-related discrepancies. Experiments on three public AAD datasets (KUL, DTU, and AVED) under the LOSO protocol demonstrate strong and consistent cross-subject performance across decision windows. Ablation and visualization analyses further provide qualitative support for the proposed design, suggesting that the alignment objective contributes a large empirical gain and that identity-related structure is reduced in the bottleneck representation **z** while **h** remains the classifier-facing representation. Overall, our results suggest that explicitly reducing identity-related variability in an auxiliary bottleneck space can improve cross-subject transfer in AAD and provide preliminary qualitative evidence for more interpretable band-selection behavior. A natural next step is to explore test-time adaptation for cross-subject AAD under realistic distribution shifts.

Acknowledgements

This work is supported by Natural Science Foundation of China (72188101), National Key R&D Program of China (NO.2024YFB3311602), Natural Science Foundation of China (62272144), the Anhui Provincial Natural Science Foundation (2408085J040), and the Major Project of Anhui Provincial Science and Technology Breakthrough Program (202423k09020001), the Anhui Provincial Graduate Quality Engineering Program (2024cxcysj002), the Fundamental Research Funds for the Central Universities (JZ2024AHST0337), and the New Cornerstone Science Foundation through the XPLORER PRIZE.

References

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, *79*(1), 151–175.

Cai, S., Sun, P., Schultz, T., & Li, H. (2021). Low-latency auditory spatial attention detection based on spectro-spatial features from eeg. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 5812–5815.

Cai, S., Zhang, R., & Li, H. (2024). Robust decoding of the auditory attention from eeg recordings through graph convolutional networks. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2320–2324.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the acoustical society of America*, *25*, 975–979.

Dai, B., Chen, C., Long, Y., Zheng, L., Zhao, H., Bai, X., Liu, W., Zhang, Y., Liu, L., Guo, T., et al. (2018). Neural mechanisms for selectively tuning in to the target speaker in a naturalistic noisy situation. *Nature communications*, *9*(1), 2405.

Das, N., Francart, T., & Bertrand, A. (2019). Auditory attention detection dataset kuleuven. *Zenodo*.

Fan, C., Yang, X., Zhang, H., Chen, Y., Li, L., Zhou, J., & Lv, Z. (2025). Listennet: A lightweight spatio-temporal enhancement nested network for auditory attention detection. *arXiv preprint arXiv:2505.10348*.

Fan, C., Zhang, H., Huang, W., Xue, J., Tao, J., Yi, J., Lv, Z., & Wu, X. (2024). Dgsd: Dynamical graph self-distillation for eeg-based auditory spatial attention detection. *Neural Networks*, *179*, 106580.

Fuglsang, S. A., Wong, D., & Hjortkjær, J. (2018). Eeg and audio dataset for auditory attention decoding. *Zenodo*.

Ge, P., Ren, C.-X., Xu, X.-L., & Yan, H. (2023). Unsupervised domain adaptation via deep conditional adaptation network. *Pattern Recognition*, *134*, 109088.

Guo, D., Li, K., Hu, B., Zhang, Y., & Wang, M. (2024). Benchmarking micro-action recognition: Dataset, methods, and applications. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(7), 6238–6252.

Huang, J., Feng, Y., Cui, F.-Q., Zhang, X., Liu, Z., Liu, X., Liu, J., Zhang, F., & Li, M. (2026). Identifying who you are no matter what you write through abstracting handwriting style. *IEEE Transactions on Dependable and Secure Computing*, 1–15. https://doi.org/10.1109/TDSC.2026.3668275

Jiang, Y., Chen, N., & Jin, J. (2022). Detecting the locus of auditory attention based on the spectro-spatial-temporal analysis of eeg. *Journal of neural engineering*, *19*(5), 056035.

Li, D., Yang, Y., Song, Y.-Z., & Hospedales, T. (2018). Learning to generalize: Meta-learning for domain generalization. *Proceedings of the AAAI conference on artificial intelligence*, *32*(1).

Liang, J., He, R., & Tan, T. (2025). A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, *133*(1), 31–64.

Ma, J., Zhang, J., Yang, Y., Yang, B., & Shan, C. (2025). Motor imagery classification based on temporal-spatial domain adaptation for stroke patients. *Cognitive Computation*, *17*(3), 119.

Ni, Q., Zhang, H., Fan, C., Pei, S., Zhou, C., & Lv, Z. (2024). Dbpnet: Dual-branch parallel network with temporal-frequency fusion for auditory attention detection. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2024)*.

Pfurtscheller, G., & Da Silva, F. L. (1999). Event-related eeg/meg synchronization and desynchronization: Basic principles. *Clinical neurophysiology*, *110*(11), 1842–1857.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2008). *Dataset shift in machine learning*. Mit Press.

Ren, C.-X., Ge, P., Dai, D.-Q., & Yan, H. (2019). Learning kernel for conditional moment-matching discrepancy-based image classification. *IEEE transactions on cybernetics*, *51*(4), 2006–2018.

Tune, S., Alavash, M., Fiedler, L., & Obleser, J. (2021). Neural attentional-filter mechanisms of listening success in middle-aged and older individuals. *Nature Communications*, *12*(1), 4533.

Wang, J., Yao, L., & Wang, Y. (2023). Ifnet: An interactive frequency convolutional neural network for enhancing motor imagery decoding from eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 1900–1911.

Yan, S., Fan, C., Zhang, H., Yang, X., Tao, J., & Lv, Z. (2024). Darnet: Dual attention refinement network with spatiotemporal construction for auditory attention detection. *Advances in Neural Information Processing Systems*, *37*, 31688–31707.

ZHANG, H., ZHANG, J., et al. (2024). Based on audio-video evoked auditory attention detection electroencephalogram dataset. *Journal of Tsinghua University (Science and Technology)*, *64*(11), 1919–1926.