UGPA-AAD: UNCERTAINTY-AWARE GAUSSIAN GRAPH AND PROTOTYPE ADAPTATION FOR CROSS-SUBJECT AUDITORY ATTENTION DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Auditory Attention Decoding (AAD) plays a vital role in enabling intelligent auditory brain-computer interfaces by decoding attention-related neural responses from EEG signals. Currently, cross-subject AAD has gained increasing attention, aiming to learn subject-invariant attention-related representations that generalize to target subjects. Despite recent advancements, cross-subject AAD still faces two core challenges under subject-specific variations and signal disturbances: unreliable spatial dependency modeling and cross-subject semantic inconsistency. To address these challenges, we propose an **U**ncertainty-aware **G**aussian Graph and **P**rototype **A**daptation framework (UGPA-AAD), which improves representation reliability and cross-subject semantic consistency through uncertainty-aware structural learning and prototype-guided semantic alignment. Specifically, we first introduce a Multi-Scale Feature Encoder (MSFE) to capture informative neural representations by integrating complementary representations across multiple temporal receptive fields. Then, to mitigate unreliable channel interactions under subject variability and signal disturbances, we design an Uncertainty-Aware Gaussian Graph Encoder (UGGE) that leverages channel-wise uncertainty to modulate spatial dependencies and construct robust structural representations. Finally, Prototype-Guided Cross-Subject Learning (PCSL) is designed to alleviate class-wise semantic inconsistency across subjects by constructing global class prototypes as semantic anchors to align representations across subjects toward consistent class structures, thereby enhancing semantic discriminability. Experimental results on four benchmarks demonstrate that our method consistently outperforms existing approaches in both audio-only and audio-visual scenarios. Our code will be released upon acceptance.

1 INTRODUCTION

The "cocktail party effect" refers to the ability of normal-hearing listeners to selectively attend to a target speaker while suppressing interference from competing sound sources in complex acoustic environments (O'Sullivan et al., 2015; Geirnaert et al., 2021). This selective auditory attention is fundamental to human auditory cognition and has important implications for neuro-engineering applications aimed at assisting individuals with hearing impairments (Choudhari et al., 2026). Auditory Attention Decoding (AAD) aims to decode selective auditory attention from recorded neural signals by identifying the sound stream attended by a listener. It addresses a fundamental question in neural rehabilitation: ***How to restore natural auditory scene perception for individuals with hearing impairments.*** Recent advances in non-invasive electroencephalography (EEG)-based AAD have demonstrated promising performance in decoding listeners' attentional focus from cortical responses (Tanveer et al., 2024). Existing methods (Vandecappelle et al., 2021; Jiang et al., 2022) typically learn discriminative neural representations from EEG recordings to characterize attention-related cortical patterns, leading to promising advances in auditory attention decoding.

Despite these advancements, existing AAD methods predominantly follow the subject-dependent paradigm, requiring subject-specific labeled data and model calibration, which limits their practical deployment in real-world hearing assistance systems (Geirnaert et al., 2022; Nguyen et al., 2025).

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.

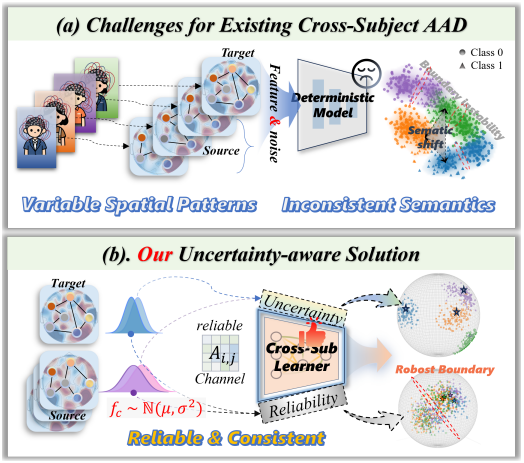


Figure 1: Challenges and Solution for Cross-Subject AAD: (a) Existing methods suffer from variable spatial patterns, resulting in inconsistent semantics across subjects. (b) Our uncertainty-aware framework learns reliable and consistent representations for robust generalization.

In contrast, cross-subject AAD aims to learn subject-invariant representations from labeled source subjects for generalization to target subjects. However, cross-subject AAD still faces two key limitations, as shown in Figure 1(a). **Limitation 1: Unreliable Spatial Dependency Modeling.** Inter-subject variability and signal disturbances can induce channel-wise variations in the reliability of EEG responses across subjects and segments. Existing deterministic relation modeling can be sensitive to such response perturbations, potentially introducing unreliable spatial dependencies and propagating perturbations during spatial aggregation. This ultimately compromises the robustness of learned EEG representations across subjects. **Limitation 2: Cross-Subject Semantic Inconsistency.** Subject-specific variations can cause the same attention state to exhibit inconsistent class-conditional structures across subjects. Existing cross-subject learning methods mainly reduce inter-subject discrepancies at the representation or distribution level, which may not preserve class-wise semantic consistency. Such inconsistency weakens cross-subject intra-class compactness and inter-class separability, making the learned class structures less transferable to target subject.

To address these limitations, we propose **U**ncertainty-aware **G**aussian Graph and **P**rototype **A**daptation for cross-subject auditory attention decoding (UGPA-AAD), a unified framework for reliable structural modeling and cross-subject semantic consistency, as illustrated in Figure 1(b). Technically, the Multi-Scale EEG Feature Encoder (MSFE) first captures complementary EEG dynamics across multiple temporal scales, characterizing transient EEG variations and sustained attention-related patterns for robust feature learning. Then, we design the Uncertainty-Aware Gaussian Graph Encoder (UGGE) for reliable spatial dependency modeling. UGGE models EEG features as Gaussian distributions to disentangle stable representations from subject-dependent uncertainty and leverages the estimated uncertainty to guide adaptive inter-channel dependency learning. Finally, Prototype-Guided Cross-Subject Learning (PCSL) establishes shared attention prototypes as semantic anchors for cross-subject alignment. Through prototype-guided contrastive learning, PCSL organizes representations from different subjects into consistent class-wise structures, fostering intra-class compactness and inter-class separability. This structure-to-semantics learning framework promotes robust cross-subject representations through reliable spatial dependency modeling and consistent class-wise semantics. Our main contributions are summarized as follows:

- This paper proposes UGPA-AAD, a cross-subject AAD framework that integrates uncertainty-aware structural modeling with prototype-guided semantic learning to jointly address unreliable spatial dependencies and cross-subject semantic inconsistency.
- We develop an UGGE that explicitly models subject-dependent uncertainty and incorporates it into adaptive inter-channel dependency learning, suppressing unreliable spatial dependencies and yielding robust spatial structures under cross-subject variability.
- We introduce PCSL, which establishes shared class-level semantic anchors and employs prototype-guided contrastive learning to alleviate cross-subject semantic inconsistency by improving intra-class compactness and inter-class separability.

2 Related Work

2.1 Auditory Attention Decoding

Existing EEG-based AAD methods mainly explore spatiotemporal modeling and time-frequency representation learning (Yan et al., 2024; Chen et al., 2026; Fan et al., 2024; Su et al., 2022).

Spatiotemporal methods capture inter-channel dependencies and temporal dynamics, whereas time-frequency methods exploit temporal evolution and spectral characteristics to characterize attention-related neural patterns. Many AAD methods are developed under subject-dependent settings and rely on subject-specific training or calibration (Li et al., 2025; Ni et al., 2024). Recently, cross-subject AAD has attracted increasing attention, with current approaches (Wang et al., 2025; Fan et al., 2025a) primarily emphasizing transferable representation learning across subjects, while the reliability of spatial dependency modeling and cross-subject class-wise semantic consistency remain underexplored.

2.2 Graph Convolutional Network

Graph Convolutional Networks (GCNs) (Kipf & Welling, 2017) provide an effective paradigm for modeling relational dependencies in non-Euclidean data (Levie, 2023). In EEG-based AAD, GCNs have been employed to capture spatial dependencies among EEG channels (Cai et al., 2024b; Zhou et al., 2025; Fan et al., 2024). Unlike deterministic AAD graph modeling, we develop an uncertainty-aware GCN for cross-subject AAD to capture more reliable inter-channel dependencies under subject variability and signal disturbances.

2.3 Uncertainty Estimation

Uncertainty estimation characterizes the variability or confidence associated with model predictions or latent representations, providing an indication of their reliability (Kendall & Gal, 2017; He et al., 2026). Due to its utility in improving predictive reliability and robustness, uncertainty estimation has been widely explored across various tasks (Hu et al., 2025; Choi et al., 2026). Recent studies have explored uncertainty modeling for characterizing noise-induced variability and cross-subject variations. Some methods exploit uncertainty to guide cross-subject adaptation (Shen et al., 2025) or calibration under domain shifts (Zheng et al., 2026), while other methods (Xu et al., 2026) model latent representations probabilistically to enhance robustness to physiological noise. However, existing methods mainly use uncertainty for calibration, adaptation, or robust representation learning, with limited attention to inter-channel structural modeling. In contrast, we leverage representation uncertainty to guide adaptive inter-channel dependency learning.

3 Methodology

3.1 Preliminary and Overview.

Preliminary. Given a multi-subject window-level EEG dataset $\mathcal{D} = \{(X_i^k, y_i^k)_{i=1}^{W_k}\}_{k=1}^{K+1}$, constructed following the standard AAD data pipeline, where $X_i^k \in \mathbb{R}^{C\times T}$ denotes the i-th EEG decision window from subject k with C channels and T time points, $y_i^k \in \{0, 1\}$ is the attended-speaker label, and W_k is the number of windows from subject k. Under the standard Leave-One-Subject-Out (LOSO) protocol for cross-subject AAD, K subjects are used for training, while the remaining subject is held out for evaluation. Accordingly, the source set is defined as $\{X_1^s, X_2^s, \ldots, X_K^s\}$, and the target set is X^t. The goal of cross-subject AAD is to learn a classifier $F(\cdot;\Theta)$ from the source subjects and generalize it to the target subject:

$$\min_{\Theta} -\frac{1}{N_s}\sum_{s=1}^{K-1}\sum_{i=1}^{W_k}\log\frac{\exp\big(F(X_i^s;\Theta)_{y_i^s}\big)}{\sum_{c=0}^{1}\exp\big(F(X_i^s;\Theta)_c\big)}, \tag{1}$$

where N_s denotes the total number of source windows.

Overview. Our core insight is to learn stable and noise-robust EEG representations under cross-subject variations for reliable auditory attention decoding. Based on this insight, we propose UGPA-AAD, a reliable cross-subject learning framework, as illustrated in Figure 2. It comprises three key components: Multi-Scale Feature Encoder (MSFE), Uncertainty-Aware Gaussian Graph Encoder (UGGE), and Prototype-Guided Cross-Subject Learner (PCSL). For an input EEG window $\mathbf{X} \in \mathbb{R}^{C\times T}$, MSFE first captures multi-scale temporal patterns over multiple receptive fields, yielding an information-rich representation $f_m \in \mathbb{R}^{D\times C\times T}$, where D denotes the latent feature dimension. Subsequently, UGGE reparameterizes f_m into channel-wise Gaussian representations $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, where μ captures subject-invariant attention features and σ estimates subject-dependent uncertainty. Based on $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, UGGE further modulates inter-channel dependencies through

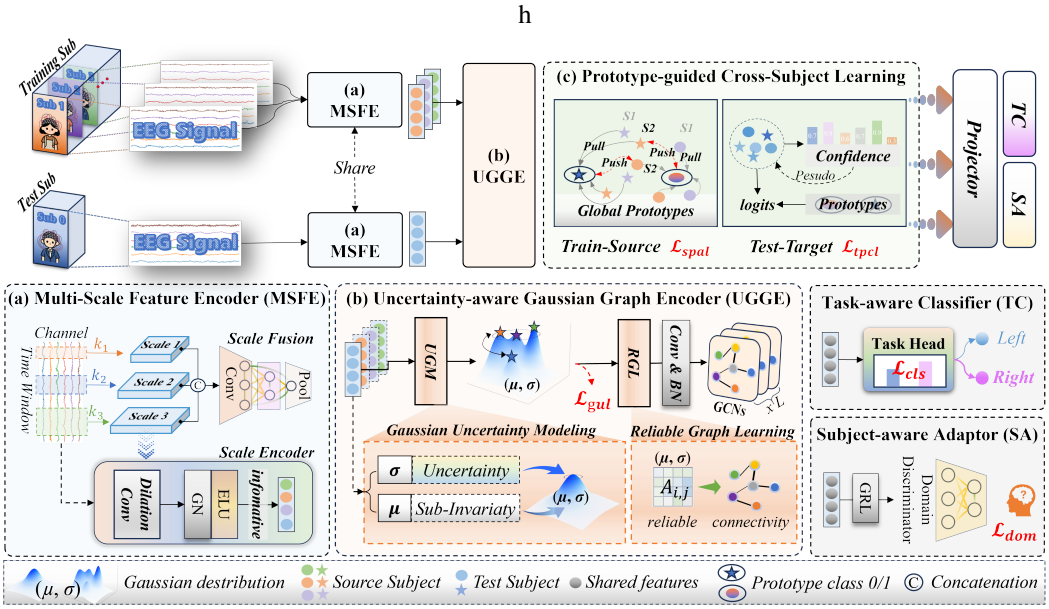


Figure 2: The overview of the proposed **UGPA-AAD**, consisting of three components: (a) Multi-Scale Feature Encoder (MSFE) enriches neural representations through multiple receptive fields; (b) Uncertainty-Aware Gaussian Graph Encoder (UGGE) modulates channel-wise spatial dependencies and produces reliable spatial features; (c) Prototype-Guided Cross-Subject Learning (PCSL) aligns class-wise semantics across subjects for discriminative AAD representations.

functional GCNs to produce a reliable structure-aware representation $f_{\mu,c}$. Next, $f_{\mu,c}$ is aggregated into a compact embedding $z \in \mathbb{R}^{D'}$ by PCSL through shared class-wise prototypes $\mathbf{p}_c$ and cross-subject consistency learning. Finally, z is passed through a linear classifier to obtain the attended-speaker label $\hat{y}$.

3.2 Multi-Scale Feature Encoder

EEG signals exhibit diverse oscillatory Pradeepkumar et al. (2024) and transient patterns, such as spikes Chen et al. (2022). Attention-related EEG responses evolve over multiple temporal scales, where short-term fluctuations provide transient cues, while relatively sustained responses reflect more stable attentional states. In AAD, these dynamic temporal patterns Fan et al. (2025a); Cai et al. (2024a) are coupled with coordinated responses across EEG channels. Based on these observations, we introduce MSFE to capture neural patterns across multiple receptive fields, constructing informative initial EEG representations while preserving fine-grained neural details.

As illustrated in Figure 2(a), MSFE first applies B parallel convolution branches with distinct receptive fields $\{k_1, \cdots, k_B\}$ to the reshaped input $\mathbf{X}'$, producing multi-scale temporal features f_b:

$$f_b = \mathrm{ELU}\left(\mathrm{GN}\left(\mathrm{DConv}_{(1\times k_b, d_b)}(\mathbf{X}')\right)\right), \quad b = 1, \ldots, B, \tag{2}$$

where $\mathrm{DConv}_{(1\times k_b, d_b)}(\cdot)$ denotes a dilated convolution with kernel size $1 \times k_b$ and dilation factor d_b, and $\mathrm{GN}(\cdot)$ and $\mathrm{ELU}(\cdot)$ denote group normalization and nonlinear activation, respectively. The multi-scale features are then concatenated as $f_c = f_1 \| \cdots \| f_B$ and fused by a lightweight block $\mathcal{F}$ to obtain enriched initial EEG representations f_m:

$$f_m = \mathcal{F}(f_c), \qquad \mathcal{F} = \mathcal{D} \circ \mathrm{AvgPool} \circ \mathrm{ELU} \circ \mathrm{GN} \circ \mathrm{Conv}_{1\times 1}, \tag{3}$$

where $\mathrm{Conv}_{1\times 1}(\cdot)$ denotes pointwise convolution, $\mathrm{AvgPool}(\cdot)$ is average pooling, and $\mathcal{D}$ is dropout.

3.3 Uncertainty-Aware Gaussian Graph Encoder

To overcome the limitations of unreliable channel-level spatial relationships under subject-variations, we propose UGGE to learn reliable structure-aware EEG representations via a two-stage

design: Gaussian Uncertainty Modeling for structural representation learning and Reliable Graph Learning for uncertainty-guided relationship modulation. The two stages are detailed below.

Gaussian Uncertainty Modeling. At the representation stage, we model f_m as channel-wise Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ to characterize relatively subject-invariant neural representations $\boldsymbol{\mu}$ from subject-specific uncertainty $\boldsymbol{\sigma}$. Specifically, we first project f_m into a latent channel space as $f_c = \mathcal{M}_t(\psi(f_m))$, where $\psi(\cdot)$ consists of a 1×1 convolution followed by batch normalization and ELU activation, and $\mathcal{M}_t$ denotes temporal averaging. Given the i-th channel feature $f_c^{(i)} \in \mathbb{R}^D$, separate branches $g_\mu(\cdot)$ and $g_\sigma(\cdot)$ estimate its mean and log-variance, yielding the Gaussian representation $q^{(i)}$ as follows,

$$\boldsymbol{\mu}^{(i)} = f_c^{(i)} + g_\mu\left(f_c^{(i)}\right), \quad \log\left((\boldsymbol{\sigma}^{(i)})^2\right) = g_\sigma\left(f_c^{(i)}\right), \quad q^{(i)} = \mathcal{N}\left(\boldsymbol{\mu}^{(i)}, \mathrm{diag}\left((\boldsymbol{\sigma}^{(i)})^2\right)\right). \tag{4}$$

To regularize the learned distributions and prevent degenerate uncertainty estimation, we impose a KL-divergence (Chang et al., 2020) constraint toward the standard Gaussian prior $\mathcal{N}(0, I)$:

$$\mathcal{L}_{gul} = D_{\mathrm{KL}}\left(\mathcal{N}\left(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)\right) \,\|\, \mathcal{N}(0, I)\right). \tag{5}$$

Reliable Graph Learning. Based on the learned Gaussian representations, we construct spatial dependencies by jointly considering distributional similarity and channel reliability. For channels i and j, we measure their distributional discrepancy as $d_{ij}^{(l)}$ and then modulate the distribution-based affinity $\widetilde{A}_{ij}^{(l)}$ by derived channel-wise reliability $r^{(l)}$ from the estimated uncertainty as flollows,

$$d_{ij}^{(l)} = \frac{1}{D}\left\|\boldsymbol{\mu}_i^{(l)} - \boldsymbol{\mu}_j^{(l)}\right\|_2^2 + \frac{1}{D}\sum_{d=1}^{D}\left((\boldsymbol{\sigma}_{i,d}^{(i)})^2 + (\boldsymbol{\sigma}_{j,d}^{(i)})^2 - 2\sqrt{(\boldsymbol{\sigma}_{i,d}^{(i)})^2(\boldsymbol{\sigma}_{j,d}^{(i)})^2}\right), \tag{6}$$

$$r^{(l)} = \exp\left(-\frac{1}{D}\sum_{d=1}^{D}\sigma_d^{2(l)}\right), \qquad \widetilde{A}_{ij}^{(l)} = \mathrm{Softmax}_j\left(-\frac{d_{ij}^{(l)}}{\tau}\right) r_i^{(l)} r_j^{(l)}, \tag{7}$$

where τ is a learnable temperature parameter. The K strongest neighboring dependencies are retained to form a sparse reliable graph with self-connections introduced by the identity matrix $\mathbf{I}$, over which the Gaussian representations are subsequently refined:

$$\mathbf{A}^{(l)} = (1-\beta)\,\mathrm{TopK}\left(\widetilde{\mathbf{A}}^{(l)}, K\right) + \beta\mathbf{I}, \qquad \boldsymbol{\mu}^{(l+1)} = \boldsymbol{\mu}^{(l)} + \mathcal{G}\left(\boldsymbol{\mu}^{(l)}, \mathbf{A}^{(l)}\right), \tag{8}$$

where $\mathcal{G}(\cdot)$ denotes a GCN layer (Razin et al., 2023) and β controls the contribution of self-connections. After stacking multiple reliable graph layers, we apply the reparameterization trick to the refined Gaussian parameters $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$ to preserve differentiability while enabling uncertainty-aware modulation of spatial dependencies, producing reliable spatial representations:

$$f_{\mu,c} = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\sigma}} \odot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \tag{9}$$

3.4 Prototype-Guided Cross-Subject Learning

To promote class-wise semantic consistency across subjects while preserving inter-class separation, we propose Prototype-Guided Cross-Subject Learning (PCSL), comprising Subject-Balanced Class Prototypes for semantic anchoring, Subject-Invariant Learning to cross-subject semantic alignment, and Confidence-Aware Adaptation for reliable transfer of class semantics to the target subject.

Subject-Balanced Class Prototypes. We first aggregate $f_{\mu,c}$ into a compact embedding $z = \mathrm{MLP}(f_{\mu,c}) \in \mathbb{R}^{D'}$ and construct subject-balanced global class prototypes $\mathbf{p}_c$ as semantic anchors to promote consistent semantic structures for the same attention state across subjects. For stability, we update the global prototypes using the batch-wise prototypes $\widehat{\mathbf{p}}_c$ with momentum $\rho \in [0, 1)$. Formally, the process is defined as

$$\widehat{\mathbf{p}}_c = \frac{1}{|\mathcal{S}_c|}\sum_{k \in \mathcal{S}_c}\mathcal{M}(z^k \mid y^k = c), \qquad \mathbf{p}_c \leftarrow \rho\mathbf{p}_c + (1-\rho)\widehat{\mathbf{p}}_c, \tag{10}$$

where $\mathcal{M}$ denotes mean aggregation, $c \in \{0, 1\}$ indexes the attention class, and $\mathcal{S}_c$ denotes the set of source subjects containing class c.

Subject-Invariant Learning. PCSL further learns subject-invariant semantic structures by aligning the K source embeddings $\{z_1^s, z_2^s, \ldots, z_K^s\}$ with the global semantic anchors $\{\mathbf{p}_0, \mathbf{p}_1\}$ through a prototype contrastive objective as follows:

$$\mathcal{L}_{spal} = -\frac{1}{K}\sum_{i=1}^{K}\log\frac{\exp\left(\mathrm{sim}\left(z_i^s, \mathbf{p}_{y_i^s}\right)/\tau_p\right)}{\sum_{c=0}^{1}\exp\left(\mathrm{sim}\left(z_i^s, \mathbf{p}_c\right)/\tau_p\right)}, \tag{11}$$

where τ_p denotes the prototype temperature. To further enhance subject-invariant representation learning, we introduce a subject discriminator $\mathcal{D}(\cdot)$ with a gradient reversal layer $\mathrm{GRL}(\cdot)$ on the source embeddings z_i^s. The discriminator learns to distinguish source subjects, while the reversed gradients encourage the EEG encoder to learn subject-invariant representations. The objective is:

$$\mathcal{L}_{dom} = -\frac{1}{K}\sum_{i=1}^{K}\log\mathcal{D}_{d_i}\left(\mathrm{GRL}\left(z_i^s\right)\right), \tag{12}$$

where d_i denotes the subject identity of the i-th source sample.

Confidence-aware Adaptation. We exploit confident target predictions to progressively align target representations with the shared class structures learned from source subjects. For each target embedding z_i^t, the pseudo label and prediction confidence are obtained as $\hat{y}_i^t = \arg\max_c p(c \mid z_i^t)$ and $s_i = \max_c p(c \mid z_i^t)$. Only samples with $s_i \geq \delta$ are selected, and their contributions are weighted by the corresponding confidence scores. The selected target embeddings are then aligned with the shared class prototypes:

$$\ell_i^t = -\log\frac{\exp\left(\mathrm{sim}\left(z_i^t, \mathbf{p}_{\hat{y}_i^t}\right)/\tau_p\right)}{\sum_{c=0}^{1}\exp\left(\mathrm{sim}\left(z_i^t, \mathbf{p}_c\right)/\tau_p\right)}, \qquad \mathcal{L}_{tpcl} = \frac{1}{N_t'}\sum_{i=1}^{N_t'} s_i\,\ell_i^t. \tag{13}$$

where δ is the confidence threshold and N_t' denotes the number of selected target samples.

3.5 Training Objective

Following the conventional AAD paradigm, we apply a cross-entropy loss to the learned robust spatiotemporal representation f_m to preserve task-relevant semantics:

$$\mathcal{L}_{cls} = -\frac{1}{N_s}\sum_{i=1}^{N_s}\log p\left(y_i^s \mid f_{m,i}^s\right), \tag{14}$$

where N_s denotes the total number of labeled source samples. Together with uncertainty-aware Gaussian modeling and prototype-guided cross-subject learning, the overall training objective of UGPA-AAD is formulated as

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{gul}\mathcal{L}_{gul} + \lambda_{dom}\mathcal{L}_{dom} + \lambda_{spal}\mathcal{L}_{spal} + \lambda_{tpcl}\mathcal{L}_{tpcl}, \tag{15}$$

where the λ terms denote the trade-off coefficients for balancing the corresponding objectives.

4 Experiments

4.1 Experiments Setting

We evaluate UGPA-AAD on four public AAD datasets: DTU Fuglsang et al. (2018; 2017) under the audio-only setting, AV-GC Rotaru et al. (2024b;a) under the audio-visual setting, and MM-AAD Fan et al. (2025b) and AVED ZHANG et al. (2024) under both audio-only and audio-visual settings. These datasets cover diverse acoustic and visual conditions, inter-subject variability, and signal disturbances, allowing UGPA-AAD to be comprehensively evaluated across challenging cross-subject AAD scenarios. More details are provided in the Appendix A.1.

Implementation Details. UGPA-AAD is trained for up to 100 epochs with a batch size of 256. We employ AdamW with an initial learning rate of 3×10^{-4} and a weight decay of 3×10^{-4}, together with a cosine annealing learning-rate scheduler. MSFE adopts temporal convolution kernels of sizes $\{1, 3, 5, 7\}$, with corresponding dilation rates $\{1, 2, 2, 2\}$. In UGGE, the number of GCN layers is

set to $N = 3$, and the adaptive graph $\mathbf{A}$ retains the strongest 80% of inter-channel connections. The prototype temperature is set to 0.1, while the target confidence threshold is gradually reduced from 0.95 to 0.90 during training. The classification loss is weighted by 1.0, the target entropy loss by 0.01, and all remaining loss terms by 0.1. All experiments are implemented in PyTorch and conducted on an NVIDIA RTX 4090 GPU.

Evaluation Protocol. We follow the leave-one-subject-out (LOSO) cross-subject evaluation protocol adopted in Zhong et al. (2022) to evaluate all methods. For each fold, the remaining subjects provide labeled source data for training, while unlabeled data from the target subject are used only for unsupervised adaptation. Target labels are used exclusively for final evaluation and are not involved in training or adaptation. AAD is formulated as a binary classification task to decode left- and right-attended speech. All methods use identical preprocessing, source-target data partitioning, and windowing protocols to ensure a fair comparison. We report the mean decoding accuracy (Acc.) and standard deviation (SD) across subjects under 1-second and 2-second decision windows, with results presented as Acc. $\pm$ SD. We reproduce and compare seven publicly available baseline models: DBPNet (Ni et al., 2024), DARNet (Yan et al., 2024), MHANet (Li et al., 2025), ListenNet (Fan et al., 2025a), DHGCN (Zhou et al., 2025), S^{2}Former (Wang et al., 2025), and PCRNet (Chen et al., 2026). Details in Appendix A.2.

4.2 Comparison with State-of-the-Art Methods

Performance Comparison under the Audio-Only Setting. We evaluate UGPA-AAD on three audio-only AAD datasets under the cross-subject setting, as shown in Table 1. UGPA-AAD achieves either the best or second-best performance across all datasets under both decision windows. Specifically, on MM-AAD, UGPA-AAD outperforms the second-best method by 2.9% and 2.6% under the 1-second and 2-second windows, respectively. In addition, our method consistently surpasses PCRNet (Chen et al., 2026) across all three audio-only datasets. PCRNet mainly improves robustness through phase-guided spectral filtering, which suppresses noise-dominated frequencies and refines complex-valued spectral representations. In contrast, PCSL leverages global attention-class prototypes to guide subject-specific representations toward consistent class structures, thereby promoting cross-subject semantic consistency. UGGE further employs uncertainty-aware Gaussian modeling to quantify uncertainty in channel representations and uses the estimated uncertainty to adapt spatial dependencies.

Table 1: Comparison with state-of-the-art methods under the **audio-only** setting (in %).The best results are highlighted in **bold**, while the second-best results are underlined.

Methods	Publication	DTU		MM-AAD		AVED	
		1-second	2-second	1-second	2-second	1-second	2-second
DBPNet (Ni et al., 2024)	IJCAI 2024	55.5 ± 6.33	55.8 ± 6.11	52.6 ± 2.05	51.9 ± 3.10	52.1 ± 1.19	53.3 ± 1.88
DARNet (Yan et al., 2024)	NeurIPS 2024	55.6 ± 4.13	55.6 ± 4.04	52.8 ± 2.67	53.1 ± 3.25	51.3 ± 0.21	52.1 ± 1.54
MHANet (Li et al., 2025)	IJCAI 2025	53.6 ± 3.77	55.6 ± 4.16	53.6 ± 3.40	54.2 ± 3.57	51.9 ± 1.40	52.5 ± 1.71
ListenNet (Fan et al., 2025a)	IJCAI 2025	56.8 ± 7.32	57.2 ± 5.83	$\underline{55.2 \pm 5.12}$	$\underline{55.9 \pm 4.19}$	$\underline{52.8 \pm 1.30}$	$\underline{53.8 \pm 1.98}$
DHGCN (Zhou et al., 2025)	ACMMM 2025	51.6 ± 6.30	51.2 ± 7.10	51.1 ± 3.28	51.1 ± 3.47	50.0 ± 0.60	50.2 ± 1.40
S^{2}Former (Wang et al., 2025)	NeurIPS 2025	$\underline{57.7 \pm 4.21}$	$\mathbf{59.7 \pm 5.25}$	55.0 ± 3.00	55.3 ± 2.63	52.0 ± 1.25	53.4 ± 1.42
PCRNet (Chen et al., 2026)	ICML 2026	54.4 ± 4.07	56.8 ± 4.88	54.9 ± 3.64	54.4 ± 3.78	52.2 ± 1.58	53.2 ± 1.51
UGPA-AAD	**Ours**	$\mathbf{58.8 \pm 3.73}$	$\underline{59.2 \pm 3.44}$	$\mathbf{58.1 \pm 5.42}$	$\mathbf{58.5 \pm 5.18}$	$\mathbf{54.1 \pm 1.25}$	$\mathbf{54.0 \pm 1.44}$

Performance Comparison under the Audio-Visual Setting. We further evaluate UGPA-AAD under the audio-visual setting, where both auditory and visual cues are available for attention decoding. As shown in Table 2, UGPA-AAD maintains consistently strong performance across all three datasets and both decision windows. Notably, on AV-GC, UGPA-AAD surpasses the second-best method by 2.7% and 3.8% under the 1- and 2-second decision windows, respectively. These consistent improvements across different datasets and decision windows further support the effectiveness of UGPA-AAD in enhancing representation reliability and class-wise semantic consistency.

4.3 Ablation Study

We select the DTU and AV-GC datasets for ablation and further analyses, as both provide 64-channel EEG recordings with fine-grained spatial information while representing audio-only and audio-visual scenarios, respectively. To evaluate the contribution of each component in UGPA-AAD, we conduct a series of ablation studies, with the results reported in Table 3. Removing **MSFE** and directly feeding the input features into UGGE decreases the decoding accuracy from

Table 2: Comparison with state-of-the-art methods under the **audio-visual** setting (in %). The best results are highlighted in **bold**, while the second-best results are underlined.

Methods	Publication	AV-GC		MM-AAD		AVED	
		1-second	2-second	1-second	2-second	1-second	2-second
DBPNet (Ni et al., 2024)	IJCAI 2024	48.8 ± 5.31	48.9 ± 5.12	50.4 ± 1.59	48.5 ± 4.50	53.3 ± 2.39	54.0 ± 1.61
DARNet (Yan et al., 2024)	NeurIPS 2024	49.1 ± 6.72	48.4 ± 5.82	50.4 ± 2.88	49.2 ± 4.14	51.4 ± 0.32	52.6 ± 0.29
MHANet (Li et al., 2025)	IJCAI 2025	55.5 ± 4.88	56.6 ± 6.23	53.1 ± 3.33	53.7 ± 3.59	52.1 ± 1.95	52.8 ± 2.07
ListenNet (Fan et al., 2025a)	IJCAI 2025	53.2 ± 4.90	55.0 ± 7.06	50.4 ± 1.63	50.7 ± 2.71	53.7 ± 1.60	54.1 ± 1.83
DHGCN (Zhou et al., 2025)	ACMMM 2025	50.1 ± 3.28	50.2 ± 3.01	50.8 ± 4.25	50.3 ± 4.64	49.2 ± 0.98	49.9 ± 1.11
S^2Former (Wang et al., 2025)	NeurIPS 2025	54.0 ± 3.78	53.2 ± 3.89	54.5 ± 2.45	54.7 ± 2.51	52.1 ± 1.44	52.5 ± 1.32
PCRNet (Chen et al., 2026)	ICML 2026	57.9 ± 5.41	55.1 ± 5.48	53.9 ± 3.40	54.4 ± 3.62	53.0 ± 1.40	53.6 ± 1.22
UGPA-AAD	**Ours**	$\mathbf{60.6 \pm 7.58}$	$\mathbf{60.4 \pm 8.68}$	$\mathbf{57.0 \pm 4.35}$	$\mathbf{57.2 \pm 4.26}$	$\mathbf{57.0 \pm 2.74}$	$\mathbf{57.2 \pm 3.43}$

58.8%/59.2% to 57.4%/58.2% on DTU and from 60.6%/60.4% to 59.1%/59.4% on AV-GC under the 1-second/2-second settings, respectively. This confirms the benefit of multi-scale temporal modeling for capturing complementary EEG dynamics. We remove **UGGE** while keeping the remaining architecture unchanged. This modification leads to a notable decrease in decoding accuracy on both datasets and decision windows. The consistent degradation across both scenarios supports the contribution of modeling subject-specific uncertainty and incorporating it into spatial dependency learning. Removing **PCSL** also consistently degrades performance, yielding 57.1%/58.0% on DTU and 58.6%/58.3% on AV-GC. Without prototype-guided learning, cross-subject class-wise structures become less constrained, highlighting the role of PCSL in maintaining semantic consistency across subjects. We further remove **SA**, resulting in accuracy decreases to 58.3%/58.5% on DTU and 59.3%/59.6% on AV-GC, demonstrating its contribution to cross-subject generalization. Notably, jointly removing UGGE and PCSL results in the largest performance degradation, with accuracy decreasing to 54.9%/55.7% on DTU and 56.2%/56.8% on AV-GC. This result highlights their complementary roles in spatial dependency modeling and cross-subject semantic regularization. In summary, the ablation results verify the complementary contributions of the proposed components to robust cross-subject AAD. More ablation results are provided in the Appendix B.1.

Table 3: Ablation study of different components on the DTU and AV-GC datasets (in %).

	MSFE	UGGE	PCSL	SA	DTU		AV-GC	
					1-second	2-second	1-second	2-second
(a)	×	✓	✓	✓	57.4 ± 4.12	58.2 ± 3.68	59.1 ± 7.71	59.4 ± 8.31
(b)	✓	×	✓	✓	56.3 ± 4.08	56.8 ± 3.81	57.7 ± 8.16	58.0 ± 8.44
(c)	✓	✓	×	✓	57.1 ± 3.94	58.0 ± 3.73	58.6 ± 7.84	58.3 ± 8.91
(d)	✓	✓	✓	×	58.3 ± 3.86	58.5 ± 3.59	59.3 ± 7.69	59.6 ± 8.77
(e)	✓	×	×	✓	54.9 ± 4.47	55.7 ± 4.02	56.2 ± 9.24	56.8 ± 9.18
(f)	✓	✓	✓	✓	$\mathbf{58.8 \pm 3.73}$	$\mathbf{59.2 \pm 3.44}$	$\mathbf{60.6 \pm 7.58}$	$\mathbf{60.4 \pm 8.68}$

4.4 Visual Results

Spatial Activation Patterns. We visualize spatial activation patterns using EEG topographic maps to examine the cross-subject consistency of the learned representations. Representative subjects are selected from the DTU and AV-GC datasets, and samples belonging to the same attention class under the 1-second setting are averaged for visualization. As shown in Figure 3, the original EEG responses exhibit substantial spatial variability across subjects, whereas the learned representations show more consistent spatial activation patterns under the same attention direction. Meanwhile, distinguishable spatial characteristics are retained between the left- and right-attended conditions. These observations suggest that UGPA-AAD reduces irregular subject-specific spatial fluctuations while preserving attention-related spatial characteristics, leading to more consistent and class-discriminative representations across subjects.

Feature Distribution. We further employ t-SNE (Van der Maaten & Hinton, 2008) to visualize the representations learned under the 1-second setting, as shown in Figure 4. Compared with the raw state, the full model produces more compact class-wise distributions with clearer inter-class separation, while source and target samples of the same class exhibit more consistent distributions. Removing UGGE leads to more dispersed representations and less distinct class boundaries, indicating its importance in learning robust and discriminative representations. In contrast, without PCSL, although class separation is partially preserved, the class-wise correspondence between source and target samples becomes less consistent, highlighting the role of PCSL in cross-subject semantic

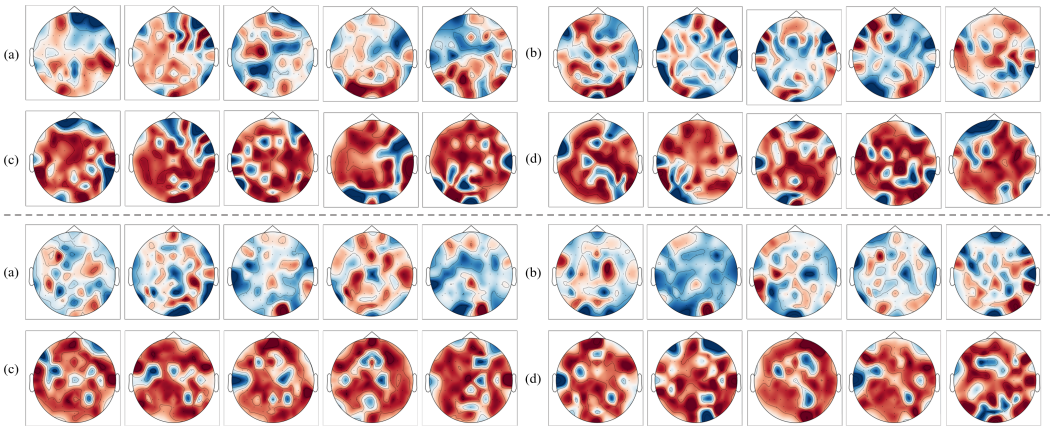


Figure 3: Spatial activation maps comparing original EEG signals and learned representations across subjects under left- and right-attention conditions. The top and bottom panels correspond to the DTU and AV-GC datasets, respectively. (a) Original EEG: left. (b) Original EEG: right. (c) Learned representation: left. (d) Learned representation: right.

alignment. These qualitative observations are consistent with the performance degradation in Table 3, providing further evidence for the effectiveness of UGGE and PCSL in learning robust and semantically consistent cross-subject representations. More visualizations are in the Appendix 4.4.

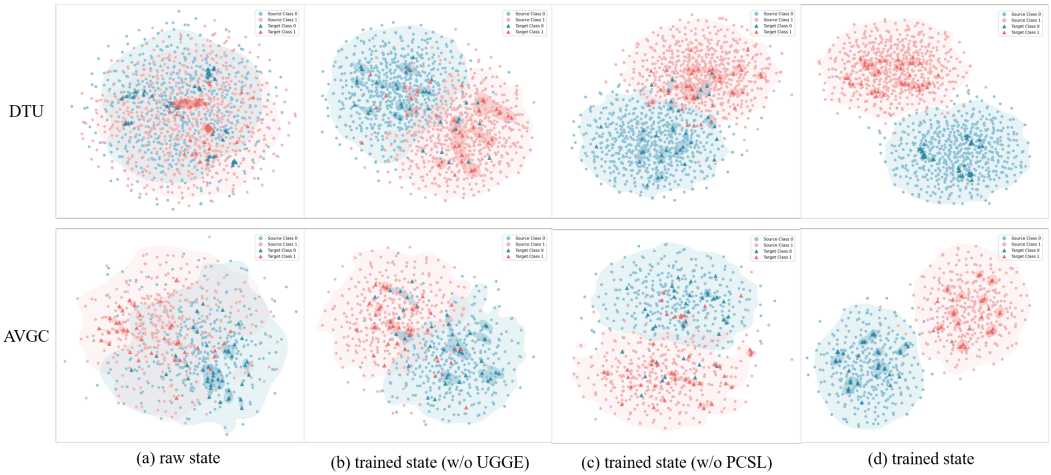


Figure 4: t-SNE visualization of feature distributions under different model configurations on the DTU and AV-GC datasets.

5 Conclusion

In this paper, we introduce UGPA-AAD, an uncertainty-aware and prototype-guided framework designed to improve cross-subject auditory attention decoding under inter-subject variability and signal disturbances. By modeling subject-dependent uncertainty and leveraging the estimated uncertainty to guide adaptive graph learning, UGPA-AAD reduces the influence of unreliable spatial dependencies and promotes robust spatial representation learning across subjects. Furthermore, prototype-guided cross-subject learning alleviates cross-subject semantic inconsistency by organizing representations across subjects into consistent class-wise semantic structures. Experiments on four benchmarks demonstrate the effectiveness of UGPA-AAD across both audio-only and audio-visual scenarios. These results demonstrate the potential of UGPA-AAD for reliable and generalizable cross-subject auditory attention decoding.

AI USE STATEMENT

In this work, generative AI tools were used to assist with language checking and manuscript polishing, including grammar correction, wording refinement, and improving the clarity and readability of the text, as well as literature retrieval and discovery for identifying potentially relevant related work. Generative AI tools were not used for research idea generation, method design, experimental design, data generation, result interpretation, or drawing scientific conclusions. All AI-assisted revisions and retrieved literature were carefully reviewed and verified by the authors to ensure technical accuracy and consistency with the original research content. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. No new human-subject or animal experiments were conducted in this study. All EEG datasets used in our experiments were obtained from publicly available sources and used in accordance with their corresponding usage guidelines. No personally identifiable information was involved in our study, and no additional data collection was performed. We have also taken care to avoid potential privacy, security, bias, or discriminatory concerns during the research process. We are committed to maintaining transparency, integrity, and responsible research practices throughout this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed framework, optimization objectives, data preprocessing procedures, experimental settings, and evaluation protocols in the main paper and supplementary materials. These materials contain the information necessary to reproduce the reported experiments. The source code and corresponding configuration files will be publicly released upon acceptance.

REFERENCES

Siqi Cai, Tanja Schultz, and Haizhou Li. Brain topology modeling with eeg-graphs for auditory spatial attention detection. *IEEE Transactions on Biomedical Engineering*, 71(1):171–182, 2024a.

Siqi Cai, Ran Zhang, Malu Zhang, Jibin Wu, and Haizhou Li. Eeg-based auditory attention detection with spiking graph convolutional network. *IEEE Transactions on Cognitive and Developmental Systems*, 16(5):1698–1706, 2024b.

Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *CVPR*, pp. 5709–5718, 2020.

Xiran Chen, Xiaoke Yang, Jian Zhou, Zhao Lv, and Cunhang Fan. PCRNet: Phase-aware complex refinement network for EEG-based auditory attention decoding. In *Forty-third International Conference on Machine Learning*, 2026.

Zheng Chen, Lingwei Zhu, Ziwei Yang, and Renyuan Zhang. Multi-tier platform for cognizing massive electroencephalogram. *arXiv preprint arXiv:2204.09840*, 2022.

Eunseo Choi, Ho-Yeon Kim, Jaewon Lee, Taeyong jo, Myungjun lee, and Heejin Ahn. Uncertainty estimation via hyperspherical confidence mapping. In *The Fourteenth International Conference on Learning Representations*, 2026.

Vishal Choudhari, Maximilian Nentwich, Sarah Johnson, Jose L. Herrero, Stephan Bickel, Ashesh D. Mehta, Daniel Friedman, Adeen Flinker, Edward F. Chang, and Nima Mesgarani. Real-time brain-controlled selective hearing enhances speech perception in multi-talker environments. *Nature Neuroscience*, may 2026. ISSN 1546-1726.

Cunhang Fan, Hongyu Zhang, Wei Huang, Jun Xue, Jianhua Tao, Jiangyan Yi, Zhao Lv, and Xiaopei Wu. Dgsd: Dynamical graph self-distillation for eeg-based auditory spatial attention detection. *Neural Networks*, 179:106580, 2024.

Cunhang Fan, Xiaoke Yang, Hongyu Zhang, Ying Chen, Lu Li, Jian Zhou, and Zhao Lv. Listennet: A lightweight spatio-temporal enhancement nested network for auditory attention detection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025*, pp. 4137–4145, 2025a.

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, 2025b.

Søren A Fuglsang, DD Wong, and Jens Hjortkjær. Eeg and audio dataset for auditory attention decoding. *Zenodo*, 2018.

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. ISSN 1053-8119.

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.

Simon Geirnaert, Tom Francart, and Alexander Bertrand. Time-adaptive unsupervised auditory attention decoding using eeg-based stimulus reconstruction. *IEEE Journal of Biomedical and Health Informatics*, 26(8):3767–3778, 2022.

Simon Geirnaert, Iustina Rotaru, Tom Francart, and Alexander Bertrand. Linear stimulus reconstruction works on the ku leuven audiovisual, gaze-controlled auditory attention decoding dataset, 2024.

Wenchong He, Zhe Jiang, Tingsong Xiao, Zelin Xu, and Yukun Li. A survey on uncertainty quantification methods for deep learning. *ACM Comput. Surv.*, 58(7), February 2026. ISSN 0360-0300.

Xinyue Hu, Zhibin Duan, Bo Chen, and Mingyuan Zhou. Enhancing uncertainty estimation and interpretability with bayesian non-negative decision layer. In *The Thirteenth International Conference on Learning Representations*, 2025.

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, oct 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR (Poster)*, 2017.

Ron Levie. A graphon-signal analysis of graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64482–64525. Curran Associates, Inc., 2023.

Lu Li, Cunhang Fan, Hongyu Zhang, Jingjing Zhang, Xiaoke Yang, Jian Zhou, and Zhao Lv. Mhanet: Multi-scale hybrid attention network for auditory attention detection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025*, pp. 4173–4181, 2025.

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 Transactions 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 *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2024)*, pp. 3115–3123, 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, 07 2015. ISSN 1047-3211.

Jathurshan Pradeepkumar, Mithunjha Anandakumar, Vinith Kugathasan, Dhinesh Suntharalingham, Simon L Kappel, Anjula C De Silva, and Chamira US Edussooriya. Toward interpretable sleep stage classification using cross-modal transformers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:2893–2904, 2024.

Noam Razin, Tom Verbin, and Nadav Cohen. On the ability of graph neural networks to model interactions between vertices. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 26501–26545. Curran Associates, Inc., 2023.

I Rotaru, S Geirnaert, A Bertrand, and T Francart. Audiovisual, gaze-controlled auditory attention decoding dataset ku leuven (av-gc-aad). *Zenodo, Apr*, 2024a.

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, feb 2024b.

Jian Shen, Lechun You, Yu Ma, Zeguang Zhao, Huajian Liang, Yanan Zhang, and Bin Hu. Uadaan: An uncertainty-aware dynamic adversarial adaptation network for eeg-based depression recognition. *IEEE Transactions on Affective Computing*, 16(3):2130–2141, 2025.

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

M Asjid Tanveer, Martin A Skoglund, Bo Bernhardsson, and Emina Alickovic. Deep learning-based auditory attention decoding in listeners with hearing impairment. *Journal of Neural Engineering*, 21(3):036022, may 2024.

Sarah Tune, Mohsen Alavash, Lorenz Fiedler, and Jonas Obleser. Neural attentional-filter mechanisms of listening success in middle-aged and older individuals. *Nature Communications*, 12(1): 4533, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 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, apr 2021. ISSN 2050-084X.

Jiaqi Wang, Zhengyu Ma, Xiongri Shen, Chenlin Zhou, Leilei Zhao, Han Zhang, Yi Zhong, Siqi Cai, Zhenxi Song, and Zhiguo Zhang. Sˆ2m-former: Spiking symmetric mixing branchformer for brain auditory attention detection. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 126593–126620. Curran Associates, Inc., 2025.

Xu Xu, Junxin Chen, Qiang He, Yongfei Wu, and Yicong Zhou. Eeg emotion recognition with uncertainty-aware contrastive learning and frequency-aware self-attention. *IEEE Transactions on Cybernetics*, 56(4):1845–1858, 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, 2024.

Hongyu ZHANG, Jingjing ZHANG, Xingguang DONG, Zhao LÜ, Jianhua TAO, Jian ZHOU, Xiaopei WU, and Cunhang FAN. Based on audio-video evoked auditory attention detection electroencephalogram dataset. *Journal of Tsinghua University (Science and Technology)*, 64(11): 1919–1926, 2024.

Qiaojie Zheng, Jiucai Zhang, Amy Zhang, and Xiaoli Zhang. Enhancing accuracy of uncertainty estimation in appearance-based gaze tracking with probabilistic evaluation and calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13793–13801, June 2026.

Peixiang Zhong, Di Wang, and Chunyan Miao. Eeg-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3):1290–1301, 2022.

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 *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 612–620, 2025.

A Implementation Details

A.1 Datasets Details

We evaluate our method on four publicly available AAD datasets, including DTU, MM-AAD, AVED, and AV-GC. Their main statistics are summarized in Table 4.

DTU. The DTU dataset (Fuglsang et al., 2018; 2017) contains 64-channel EEG recordings from 18 normal-hearing participants collected in a reverberant environment with dynamic background noise. Two competing speakers were presented with an angular separation of $60°$. Each participant completed 60 trials, with each trial lasting 50 seconds while attending to the designated speech stream.

MM-AAD. The MM-AAD dataset Fan et al. (2025b) contains 32-channel EEG recordings from 50 normal-hearing participants under two-speaker competing scenarios involving both audio-only and audio-visual conditions. The stimuli consisted of 40 short story excerpts narrated by two male and two female speakers, with the competing voices presented at $\pm 90°$. Each participant completed 20 trials per scenario, with each trial lasting 165 seconds.

AVED. The AVED dataset ZHANG et al. (2024) contains 32-channel EEG recordings from 20 normal-hearing participants, including 10 participants under the audio-only condition and 10 under the audio-visual condition. Participants attended to one of two competing speech streams presented at $\pm 90°$. Each participant completed 16 trials, with each trial lasting 152 seconds.

AV-GC. The AV-GC dataset (Rotaru et al., 2024b; Geirnaert et al., 2024; Rotaru et al., 2024a) contains 64-channel EEG recordings from 16 normal-hearing participants under four experimental conditions. Two competing talkers were presented at $\pm 90°$ relative to the participant, while auditory stimuli were delivered through insert earphones using head-related transfer functions to simulate spatial separation. Each participant completed eight trials across the four conditions, with two trials per condition and each trial lasting 600 seconds. Due to unavailable or incomplete recordings, we retain 10 participants with complete data across all four conditions for our experiments.

For preprocessing, the DTU EEG signals were high-pass filtered at 0.1 Hz, with 50 Hz line noise removed, and downsampled to 128 Hz, followed by artifact removal and re-referencing. For MM-AAD and AVED, 50 Hz line noise was removed, followed by band-pass filtering from 0.1 to 50 Hz and downsampling to 128 Hz. For AV-GC, the EEG signals were band-pass filtered from 1 to 40 Hz, re-referenced to the common average, and downsampled to 128 Hz.

A.2 Baseline Models

We compare UGPA-AAD with seven publicly available AAD models covering temporal-frequency modeling, spatiotemporal representation learning, attention mechanisms, graph learning, and lightweight architectures. For a fair comparison, all reproduced baselines follow the same preprocessing, data partitioning, and decision-window settings as our method.

Table 4: Statistics of the four AAD datasets used in our experiments.

Dataset	Subjects	Channels	Scene	Language	Trials	Duration (s)
DTU	18	64	audio-only	Danish	60	50
MM-AAD	50	32	audio-only / audio-visual	Chinese	20 per scenario	165
AVED	20	32	audio-only / audio-visual	Chinese	16	152
AV-GC	10	64	audio-visual	Dutch	8	600

DBPNet (Ni et al., 2024) adopts a dual-branch architecture that separately extracts temporal and frequency-domain representations and subsequently fuses them for auditory attention decoding. This design jointly captures time-varying EEG dynamics and spectral-spatial characteristics.

DARNet (Yan et al., 2024) constructs spatiotemporal EEG representations and employs dual attention refinement to capture spatial distribution patterns and long-range temporal dependencies.

MHANet (Li et al., 2025) combines multi-scale temporal modeling with hybrid attention mechanisms to capture both short- and long-range spatiotemporal dependencies in EEG signals.

ListenNet (Fan et al., 2025a) is a lightweight nested spatiotemporal network that models multi-channel dependencies and multi-scale temporal dynamics while maintaining low computational complexity.

DHGCN (Zhou et al., 2025) introduces dual hypergraph convolutional learning to model high-order spatial and temporal relationships in EEG signals, enabling joint characterization of complex dependencies across channels and time points.

S^{2}M-Former (Wang et al., 2025) introduces a spiking symmetric mixing Branchformer architecture for auditory attention decoding, combining spiking neural computation with symmetric feature mixing to model attention-related EEG representations efficiently.

PCRNet (Chen et al., 2026) performs phase-aware complex-domain refinement by combining temporal context calibration with dual-domain integration, allowing temporal dependencies and phase-sensitive spectral information to be jointly modeled.

B Additional Experimental Results

B.1 Ablation Study

Ablation Study on UGGE. To investigate the effectiveness of the key components in UGGE, we conduct ablation experiments by separately removing Gaussian Uncertainty Modeling (UGM) and Reliable Graph Learning (RGL). Specifically, *w/o UGM* replaces the Gaussian representations with deterministic node features and removes explicit uncertainty estimation. The subsequent graph propagation is retained, while inter-channel relations are constructed based only on deterministic feature discrepancies. In contrast, *w/o RGL* preserves the Gaussian representations learned by UGM but removes uncertainty-derived reliability from graph construction, such that inter-channel relations are determined solely by the discrepancy between mean representations.

As shown in Table 5, removing UGM decreases the accuracy by 1.5% and 1.5% on DTU, and by 1.8% and 1.4% on AV-GC under the 1-second and 2-second settings, respectively. These results indicate that explicitly modeling neural representations together with subject-specific uncertainty benefits cross-subject representation learning. Removing RGL also consistently degrades performance, demonstrating that incorporating distributional discrepancy and uncertainty-derived channel reliability facilitates the reduction of unreliable inter-channel interactions and the learning of stable spatial dependencies.

Effects of Spatial Modeling Strategies. Our ablation study compares three spatial modeling strategies: MLP-based modeling, convolution-based modeling, and conventional graph learning, together with the proposed UGGE, where uncertainty-derived reliability is further introduced to construct more reliable spatial dependencies. The conventional Graph Learning configuration corresponds to the w/o RGL variant in Table 5.

Table 5: Ablation study on the effectiveness of key components in UGGE on DTU and AV-GC in (%).

Configuration	DTU		AV-GC	
	1-second	2-second	1-second	2-second
w/o UGM	57.3 ± 4.02	57.7 ± 3.71	58.8 ± 8.01	59.0 ± 8.92
w/o RGL	58.0 ± 3.86	58.4 ± 3.59	59.5 ± 7.78	59.6 ± 8.86
UGGE	$\mathbf{58.8 \pm 3.73}$	$\mathbf{59.2 \pm 3.44}$	$\mathbf{60.6 \pm 7.58}$	$\mathbf{60.4 \pm 8.68}$

As shown in Table 6, progressively incorporating stronger spatial dependency modeling generally improves decoding performance. Compared with MLP-based modeling, Graph Learning improves the accuracy by 1.2% and 0.9% on DTU, and by 3.3% and 2.9% on AV-GC, indicating the benefit of graph-based spatial dependency modeling. Furthermore, UGGE consistently outperforms conventional Graph Learning by 0.8%, 0.8%, 1.1%, and 0.8% across the four settings, respectively. These results support the effectiveness of reliability-guided graph learning in reducing the influence of uncertain channel interactions and learning more robust spatial dependencies.

Table 6: Comparison of different spatial modeling strategies in UGGE on DTU and AV-GC in (%).

Spatial Modeling Strategy	DTU		AV-GC	
	1-second	2-second	1-second	2-second
MLP-based Modeling	56.8 ± 4.07	57.5 ± 3.74	56.2 ± 8.08	56.7 ± 9.04
Convolution-based Modeling	57.4 ± 3.96	57.9 ± 3.66	57.8 ± 7.95	58.1 ± 8.95
Graph Learning	58.0 ± 3.86	58.4 ± 3.59	59.5 ± 7.78	59.6 ± 8.86
UGGE	$\mathbf{58.8 \pm 3.73}$	$\mathbf{59.2 \pm 3.44}$	$\mathbf{60.6 \pm 7.58}$	$\mathbf{60.4 \pm 8.68}$

Ablation Study on Training Objectives. To verify the role of each loss in the model, we conduct ablation experiments, as shown in Table 7. Specifically, $\mathcal{L}_{\mathrm{spal}}$ organizes source embeddings around shared class prototypes, $\mathcal{L}_{\mathrm{tpcl}}$ adapts target representations via confidence-aware prototype contrastive learning, $\mathcal{L}_{\mathrm{gul}}$ regularizes Gaussian uncertainty modeling, and $\mathcal{L}_{\mathrm{dom}}$ suppresses residual subject-specific information. The results show that removing $\mathcal{L}_{\mathrm{spal}}$ consistently degrades performance, demonstrating the benefit of establishing compact and consistent class structures across source subjects. Removing $\mathcal{L}_{\mathrm{tpcl}}$ results in a more pronounced performance drop, highlighting the importance of transferring the learned class structure to the unlabeled target subject through confidence-aware prototype adaptation. Meanwhile, removing $\mathcal{L}_{\mathrm{gul}}$ also leads to consistent degradation, validating the contribution of Gaussian distribution regularization to uncertainty-aware representation learning. Finally, the performance decrease caused by removing $\mathcal{L}_{\mathrm{dom}}$ confirms the effectiveness of suppressing residual subject-specific information for cross-subject representation learning. The complete objective consistently achieves the best performance across all settings, demonstrating that uncertainty regularization, subject-invariant learning, and prototype-guided source-target adaptation provide complementary supervision for cross-subject auditory attention decoding.

Table 7: Ablation study on different training objectives on DTU and AV-GC in (%).

Configuration	DTU		AV-GC	
	1-second	2-second	1-second	2-second
(a) w/o $\mathcal{L}_{\mathrm{spal}}$	56.8 ± 4.06	57.6 ± 3.72	58.1 ± 8.08	58.7 ± 9.01
(b) w/o $\mathcal{L}_{\mathrm{tpcl}}$	56.2 ± 4.15	56.9 ± 3.81	57.8 ± 8.16	57.6 ± 9.18
(c) w/o $\mathcal{L}_{\mathrm{gul}}$	57.7 ± 3.92	58.0 ± 3.63	59.1 ± 7.89	59.3 ± 8.91
(d) w/o $\mathcal{L}_{\mathrm{dom}}$	58.3 ± 3.86	58.5 ± 3.59	59.3 ± 7.69	59.5 ± 8.77
(e) UGPA-AAD	$\mathbf{58.8 \pm 3.73}$	$\mathbf{59.2 \pm 3.44}$	$\mathbf{60.6 \pm 7.58}$	$\mathbf{60.4 \pm 8.68}$

B.2 HYPERPARAMETER ANALYSIS

Table 8: Performance comparison of different multi-scale convolution configurations on DTU and AV-GC in (%).

Configuration	DTU		AV-GC	
	1-second	2-second	1-second	2-second
(a)	57.6 ± 4.01	58.3 ± 3.64	59.3 ± 7.88	59.5 ± 8.90
(b)	58.2 ± 3.87	58.6 ± 3.55	59.9 ± 7.75	59.8 ± 8.81
(c)	58.5 ± 3.80	$\mathbf{59.3 \pm 3.46}$	60.2 ± 7.67	60.1 ± 8.73
(d) (UGPA-AAD)	$\mathbf{58.8 \pm 3.73}$	59.2 ± 3.44	$\mathbf{60.6 \pm 7.58}$	$\mathbf{60.4 \pm 8.68}$

To analyze the influence of temporal receptive fields in MSFE, we compare four kernel configurations, where (a)–(d) denote $\{3\}$, $\{3, 5\}$, $\{1, 3, 5\}$, and $\{1, 3, 5, 7\}$, respectively. As shown in Table 8, extending the single-scale kernel to multiple receptive fields yields consistent gains across both datasets. Compared with $\{3\}$, the full configuration improves the accuracy by 1.2% and 0.9% on DTU, and by 1.3% and 0.9% on AV-GC. Although $\{1, 3, 5\}$ slightly outperforms the full configuration by 0.1% on the 2-second DTU setting, $\{1, 3, 5, 7\}$ achieves the best overall performance, indicating that complementary temporal scales generally benefit EEG representation learning.

We further analyze the number of GCN layers L in UGGE to investigate its effect on spatial dependency modeling, as shown in Figure 5. The results show that $L = 3$ achieves the best overall performance. Too few layers may lead to insufficient propagation of inter-channel information, while overly deep graph convolution may introduce redundant information and weaken the discriminative representation of EEG channels. Although $L = 4$ slightly outperforms $L = 3$ on the 2-second DTU setting, $L = 3$ performs more consistently across the other settings. Therefore, we set $L = 3$ as the default depth of UGGE.

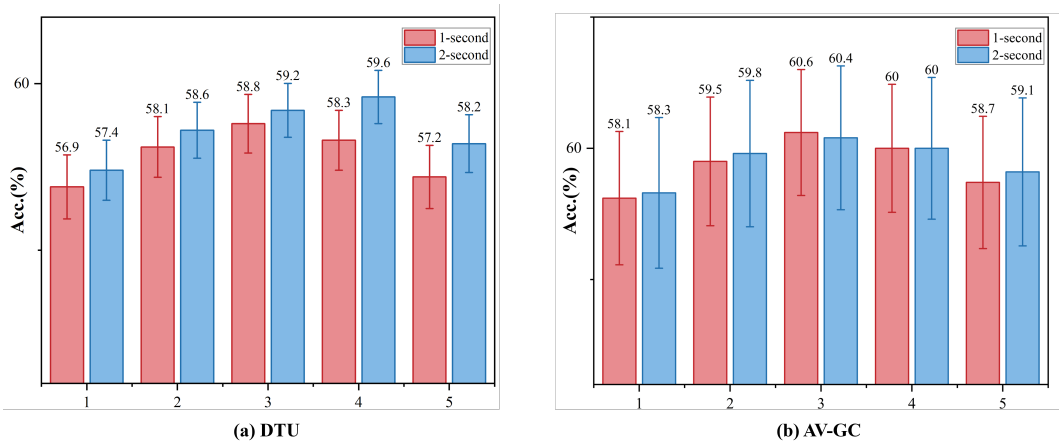


Figure 5: Effect of the number of GCN layers on DTU and AV-GC.

B.3 VISUAL RESULTS

Spatial Activation Patterns. To further examine whether the spatial patterns observed in the main paper generalize across datasets, we provide additional EEG topographic visualizations on MM-AAD (Fan et al., 2025b) and AVED (ZHANG et al., 2024) in Figure 6. Specifically, we use the audio-only setting of MM-AAD and the audio-visual setting of AVED, with all topographic maps constructed from 32 EEG channels. For each dataset, five representative subjects are visualized under both left- and right-attended conditions following the same visualization protocol as in Section 4.4. Consistent with the observations on DTU and AV-GC, the learned representations exhibit more coherent and consistent spatial activation patterns across subjects within the same attention

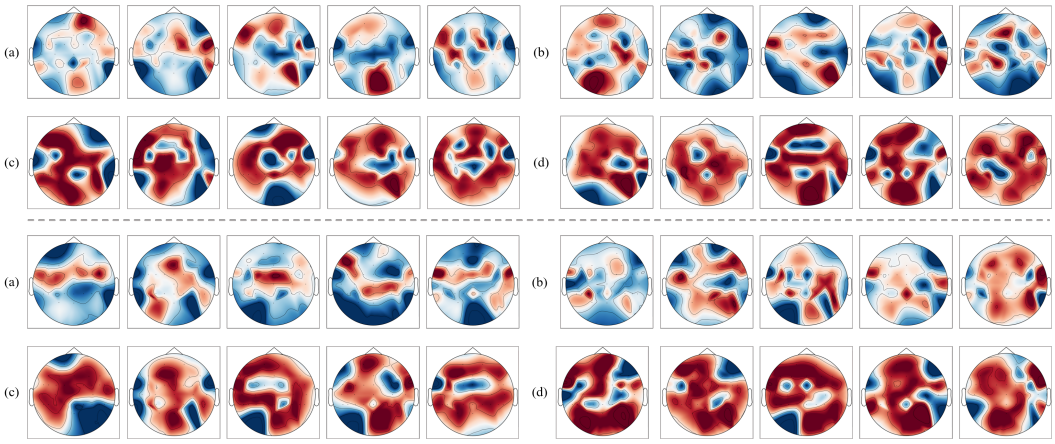


Figure 6: Spatial activation maps comparing original EEG signals and learned representations across representative subjects under left- and right-attended conditions. The upper and lower panels correspond to the audio-only setting of MM-AAD and the audio-visual setting of AVED, respectively. (a) Original EEG under left attention. (b) Original EEG under right attention. (c) Learned representations under left attention. (d) Learned representations under right attention.

condition, while distinct spatial characteristics are retained between the two attention directions. These results indicate that the qualitative spatial patterns observed in the main experiments are also present on MM-AAD and AVED.

Spatial Dependency Visualization. Here, we further visualize the ten strongest inter-channel connections learned for three representative subjects from each of four AAD datasets under the cross-subject. DTU (Fuglsang et al., 2018; 2017) and AV-GC (Rotaru et al., 2024b; Geirnaert et al., 2024; Rotaru et al., 2024a) employ 64-channel EEG configurations, whereas MM-AAD (Fan et al., 2025b) and AVED (ZHANG et al., 2024) use 32 channels. For this visualization, DTU and MM-AAD are evaluated under the audio-only setting, while AV-GC and AVED are evaluated under the audio-visual setting. The nodes arranged along the circumference denote EEG electrodes, and the connecting lines represent the learned spatial dependencies, with colors ranging from blue to red indicating increasing connection strength within each dataset.

As shown in Figure 7, high-weight connections frequently span spatially distant electrode regions, suggesting that UGGE can model long-range inter-channel dependencies beyond local spatial relationships. Despite differences in electrode configurations and stimulus paradigms across datasets, strong connections are frequently observed among electrodes over temporal, fronto-central, and parietal scalp regions. This distribution is broadly consistent with previous findings reporting the involvement of temporal and frontoparietal regions in auditory attention, which have been associated with auditory information processing and attentional control, respectively (Gehmacher et al., 2024; Tune et al., 2021).

C Future Work and Limitation

Our framework employs uncertainty-aware and prototype-guided learning to improve robustness against subject-specific neural variability. However, challenges may remain under extreme conditions where EEG responses are severely corrupted or reliable neural cues are limited. In future work, we will further explore adaptive cross-subject learning, such as zero-shot cross-subject learning, to improve the robustness, generalization, and applicability of AAD in more complex and realistic scenarios.

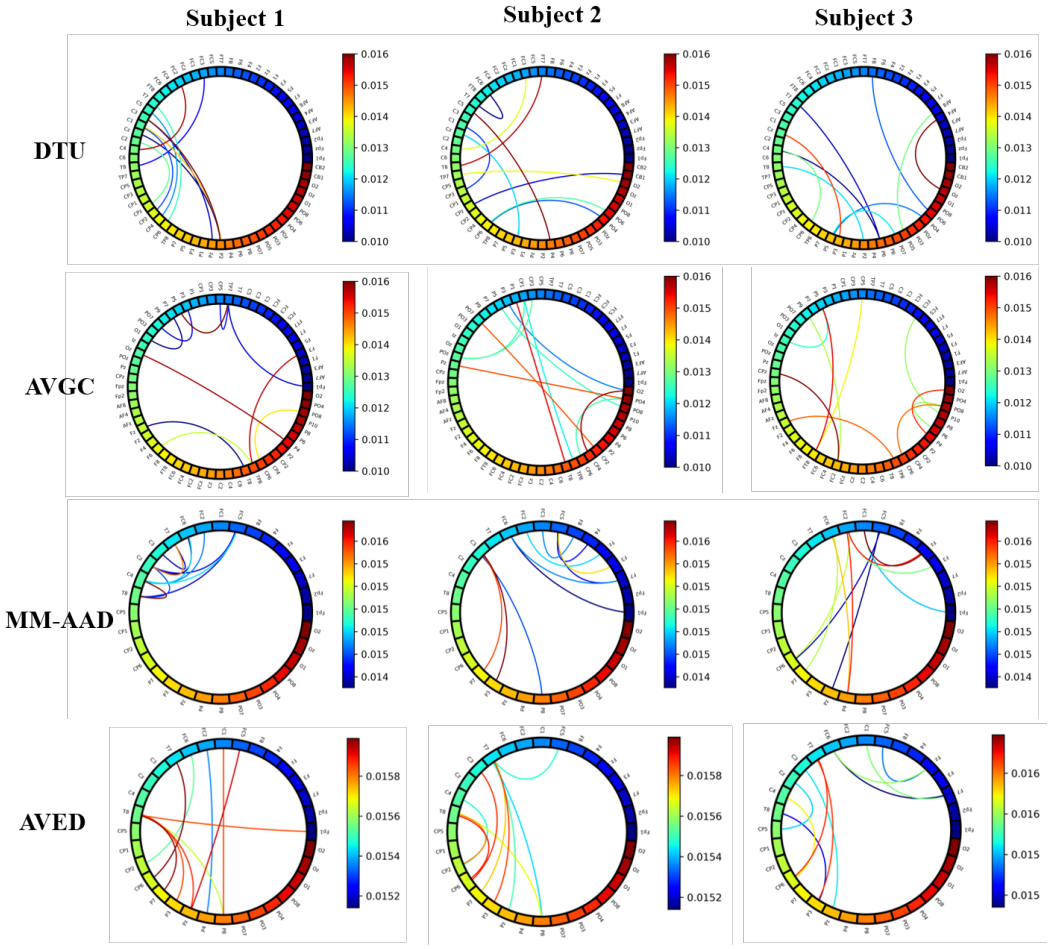


Figure 7: Visualization of the ten strongest learned inter-channel dependencies across the DTU, AVGC, MM-AAD, and AVED datasets.