HYPERBOLIC REPRESENTATION LEARNING FOR MULTIMODAL-TO-UNIMODAL TRANSFER IN EEG-BASED AUDITORY ATTENTION DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Auditory attention decoding (AAD) from electroencephalography (EEG) remains challenging across subjects, datasets, and recording systems. While auxiliary physiological signals can enrich EEG representations, their availability at inference time cannot be guaranteed. To solve this problem, we propose a codebook-guided pretraining framework that leverages paired EEG and auxiliary signals to learn transferable representations for EEG-only AAD. During pretraining, separate encoders map the two modalities into hyperbolic space, a hyperbolic residual codebook represents shared physiological structure with common codes and modality-specific information with residual codes. The shared center alignment objective further brings the two modalities' common representations into agreement. The framework then masks parts of the EEG representation and trains the EEG encoder to predict codebook targets learned from the paired multimodal data, thereby transferring the shared vocabulary to the EEG modality. For downstream AAD, the pretrained EEG encoder extracts features from EEG, which are passed to a classifier for attention prediction. Cross-subject and cross-dataset experiments on the AVGCAAD and EARAAD datasets demonstrate that the proposed model outperforms a range of state-of-the-art baselines, and ablation studies confirm that each key module contributes to the performance gains.

1 INTORDUCTION

Auditory attention decoding (AAD) aims to infer the sound source or speech stream attended by a listener from neural signals. It is an important technology for auditory brain–computer interfaces, speech enhancement, neuroprosthetic communication, and human–computer interaction. In complex acoustic environments, listeners can selectively attend to a target speaker while suppressing competing sources (Mesgarani & Chang, 2012). AAD seeks to recover this selective auditory state from neural activity, thereby supporting assistive systems to infer a user's auditory intent. Considering that electroencephalography (EEG) provides high temporal resolution, low acquisition cost, and good portability, most existing AAD studies rely on scalp EEG (O'sullivan et al., 2015). They tried to explore spatial filtering, neural response modeling, or deep neural networks for auditory attention decoding (Geirnaert et al., 2021; Vandecappelle et al., 2021; Nguyen et al., 2025).

However, most of the EEG-based AAD methods remain vulnerable to distribution shifts resulted from substantial inter-subject variability or the sensitivity to the changes in recording sessions (Pahuja et al., 2023), environments, hardware, channel configurations, referencing schemes, or signal quality. That is, representations learned under controlled conditions may not transfer reliably to new subjects, datasets, or devices (Wilroth et al., 2023). Specifically, cross-subject decoding faces differences in neural activity across individuals, whereas cross-dataset or cross-device settings are confronted with additional discrepancies in sampling protocols, channel layouts, hardware characteristics, or signal distributions (He & Wu, 2019). As a result, learning generalizable representations across such domains remains challenging .

Beyond scalp EEG, other physiological signals, such as electrooculography (EOG) and ear-EEG, may provide complementary information for auditory attention (Rotaru et al., 2024b; Bleichner et al., 2016). Specifically, EOG captures eye movements, blinks, and gaze-related behavioral states that can reflect attention allocation and task engagement (Xu et al., 2026). While, ear-EEG records

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.

neural activity around the ear and differs from scalp EEG in spatial coverage, signal characteristics, and wearable form factor, providing complementary information for characterizing auditory attention (Kaongoen et al., 2023). Actually, recent comparisons of scalp EEG, around-ear EEG, and in-ear EEG for selective auditory attention decoding further indicate that ear-centered recordings can carry neural information related to auditory attention (Geirnaert et al., 2025). However, most of existing AAD methods rely on fixed EEG channel configurations, making the learned representations be strongly coupled with specific channel numbers and spatial layouts (Sukhbaatar et al., 2026), limiting their generalization across different acquisition systems or recording settings.

The brain's hierarchical cognitive processes are encoded not only in EEG (Sun et al., 2023; Price & Moncrieff, 2021) but also in complementary physiological signals (Búzás et al., 2024). However, the dominant Euclidean embeddings are ill-suited to this: their flat geometry grows only polynomially with radius, mismatching the exponential expansion of tree-like hierarchies (Peng et al., 2022). Hyperbolic spaces, with negative curvature, instead grow exponentially and capture such hierarchies with low distortion and without extra dimensions (Krioukov et al., 2010). As this hierarchical property is shared across modalities, hyperbolic geometry offers a natural unified framework for EEG-based multimodal learning.

Multimodal physiological signal learning methods, which try to take advantage of the complement of heterogeneous biosignals in cognitive tasks have shown great potential (Baltrušaitis et al., 2018), in various downstream tasks including emotion recognition (Shen et al., 2026; Zhang et al., 2026), sleep stage detection (Yin et al., 2026), cognitive assessment (Zhou et al., 2026), and health monitoring (Wang et al., 2025b). Existing studies demonstrate that integrating EEG with auxiliary physiological or behavioral signals can provide richer representations than single-modal learning (Pillalamarri & Shanmugam, 2025). However, these multimodal strategies cannot be applied to AAD. Most existing methods assume that all modalities are available during both training and inference, and learn representations through modality fusion, making them highly dependent on the complete multimodal input. This assumption conflicts with practical EEG-only deployment, where auxiliary signals may be unavailable due to device limitations, acquisition costs, user discomfort, or signal degradation. Moreover, auxiliary modalities in AAD are not consistent across datasets. For example, EOG in AVGCAAD, while ear-EEG in EAR-AAD, since different physiological exhibit different physiological origins and signal characteristics. Therefore, a fusion strategy designed for one auxiliary modality may not generalize well to another. These limitations motivate a new multimodal learning paradigm that exploits auxiliary physiological signals during training while learning transferable representations that can be effectively utilized by an EEG-only model during inference.

To effectively exploit the complementary information from auxiliary physiological signals in AAD, while improving model adaptability under modality missing or heterogeneous modalities, we propose a multimodal pretraining framework based on hyperbolic residual vector quantization. The proposed framework learns shared and complementary representations between EEG and auxiliary modalities, transferring multimodal physiological knowledge into the EEG representation space and thereby improving AAD classification performance and generalization ability under EEG-only settings. The main contributions of this work are summarized as follows:

- **Multimodal-to-unimodal AAD learning formulation.** We formulate a modality-asymmetric AAD setting, where auxiliary physiological signals (any physiological signal except for EEG) are available during pretraining but absent during inference, addressing the practical mismatch between multimodal training and real-world deployment.
- **Hyperbolic residual codebook learning for cross-modal representation modeling.** We introduce a hyperbolic residual vector quantization framework that learns a structured multimodal vocabulary, capturing shared physiological structures through common codes and modality-specific complementary information through residual codes.
- **Multimodal to Unimodal knowledge transfer strategy.** We propose codebook-guided information transferring strategy that transfers multimodal encoder to EEG encoder through masked code prediction, enabling EEG-only inference without auxiliary modalities.
- **Generalization across auxiliary modalities.** We validate the proposed framework on AAD datasets with different auxiliary modalities, including EOG and ear-EEG, demonstrating improved cross-subject and cross-dataset generalization to modality variation.

2 Related Work

Multimodal Physiological Representation Learning. Integrating EEG with complementary physiological modalities has demonstrated promising results in various dounstream tasks, including emotion recognition (Shen et al., 2026; Zhang et al., 2026), sleep staging (Mostafaei et al., 2024), cognitive assessment (Liu et al., 2025), and auditory attention decoding (Stuldreher et al., 2020). By combining heterogeneous physiological signals, multimodal approaches can capture complementary information beyond individual modalities and improve representation quality (Liu et al., 2024). However, due to the inherent differences in physiological sources and signal characteristics, learning unified representations across heterogeneous modalities remains challenging(Baltrušaitis et al., 2018; Yang et al., 2023). Existing EEG-based multimodal methods mainly rely on feature fusion, modality alignment, or joint representation learning to integrate multiple signals (Pillalamarri & Shanmugam, 2025). Although effective when all modalities are available, these approaches typically learn continuous multimodal representations and remain dependent on fixed modality configurations. Recent studies have explored multimodal physiological pretraining to transfer knowledge across modalities (Jiang et al., 2025; Cui et al., 2026), but how to enhance EEG representations with other modalities maintaining its inference capability remains underexplored. To solve this problem, our method learns modality shared and modality-specific discrete representations from multimodal physiological signals and transfers multimodal encoder into an EEG-only encoder.

Discrete Representation Learning. Codebook-based representation learning has recently emerged as an effective approach for physiological signal modeling by discretizing continuous signals into semantic tokens that capture recurring patterns and structured information (Jiang et al., 2024; Ma et al., 2026). Recent studies have explored discrete tokenization for physiological foundation models, including modality-specific codebooks for EEG and ECG signals, such as LaBraM (Jiang et al., 2024) and HeartLang (Jin et al., 2025), as well as frequency-aware tokenization and latent codebook designs for improving representation learning (Pradeepkumar et al., 2026). These approaches demonstrate the potential of discrete vocabularies for capturing meaningful physiological patterns beyond continuous representations. However, existing codebook-based methods mainly focus on single-modality representation learning and construct modality-specific spaces. They do not explicitly model modality shared and modality-specific information across heterogeneous physiological modalities or exploit discrete codes for multimodal-to-unimodal representation transfer. In contrast, our method learns shared and residual codebooks from multimodal physiological signals and leverages shared discrete codes to guide the multimodal to EEG representation transferring.

Hyperbolic Representation Learning. Hyperbolic neural networks (Shimizu et al., 2021) have attracted increasing attention in representation learning due to their ability to model hierarchical and structured information. Actually, it has been successfully applied in natural language processing (Ganea et al., 2018), computer vision (Mettes et al., 2024), and graph learning (Liu et al., 2019). Recent studies have explored hyperbolic representations for EEG analysis (Li et al., 2026; Zhou et al., 2026), demonstrating their potential for capturing complex neural structures. For example, Chang et al. (Chang et al., 2025) perform hyperbolic contrastive pretraining for EEG emotion recognition, while other works investigate hyperbolic embeddings for EEG representation learning (Li et al., 2026). However, existing hyperbolic EEG methods mainly focus on single-modality continuous representations with single geometric structures. They cannot be applied to multimodal physiological signals with diverse geometric structure. In contrast, our method combines Lorentz-manifold-based hyperbolic representation with shared and residual codebooks for multimodal physiological representation learning, enabling structured and discrete cross-modal representation.

3 METHOD

To address the gap between multimodal representation learning and EEG-only deployment, we propose a discrete representation learning framework that discovers shared physiological structures from paired modalities and transfers them to an EEG-only encoder. The framework organizes its discrete vocabulary on a Lorentz manifold, where the underlying geometry naturally captures hierarchical relationships among physiological patterns. It consists of two stages: Multimodal encoder pretraining and Single-modality encoder transferring.

3.1 MULTIMODAL ENCODER PRETRAINING

During multimodal vocabulary learning, we introduce a Hyperbolic Residual Codebook (HRC) to decompose each token into complementary shared and modality-specific components. A shared Lorentz codebook first quantizes the coarse-grained structures common to EEG and eye signals, while modality-specific residual codebooks encode the information not captured by the shared quantization. To encourage consistent cross-modal representations, we construct a shared geometric anchor from selected shared codes using a weighted Fréchet mean and align the EEG and eye-signal distributions toward this anchor at each temporal patch. The shared and residual codes are subsequently composed in a common tangent space for joint reconstruction. This progressive factorization encourages the shared codes to preserve modality-invariant physiological structures while allowing residual codes to retain modality-dependent characteristics.

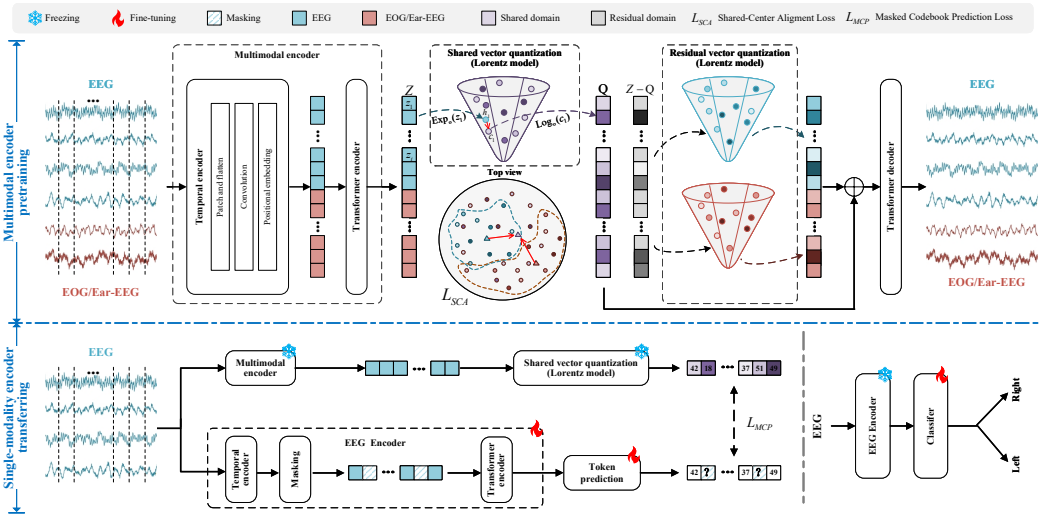

Figure 1: Overview of the proposed framework.

A paired observation contains synchronized EEG and eye-signal windows, denoted by $\mathbf{X}^m$, where $m \in \{E, Y\}$ indexes the EEG and ECG/Ear-EEG modalities, respectively. Each window is divided into non-overlapping temporal patches. We employ only temporal positional embeddings and remove channel-identity embeddings, making the encoder invariant to channel permutations. All patches from different channels and modalities are flattened into a unified token sequence of length L, representing the total number of patches. All tokens are processed jointly by the same modality-agnostic encoder, without distinguishing their modality identities. The resulting contextualized token embeddings are denoted by $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_L]$. These Euclidean token embeddings serve as the inputs to the subsequent hyperbolic residual quantization.

Hyperbolic Residual Codebook. HRC performs discrete quantization to capture cross-modal structure from multimodal information. The shared hyperbolic codebook first captures coarse structural patterns common to different physiological modalities. The residual codebook then quantizes the information that is incompletely represented by the shared code. This progressive decomposition allows HRC to represent each token using complementary shared and residual discrete codes.

Let $\mathcal{C}^S = \{\mathbf{c}_j^S\}_{j=1}^{K_S}$ denote the shared codebook and $\mathcal{C}^{R,m} = \{\mathbf{c}_j^{R,m}\}_{j=1}^{K_R}$ denote the residual codebook for modality $m \in \{E, Y\}$. All code vectors lie on the d-dimensional Lorentz manifold $\mathbb{L}_c^d$ with constant curvature $-c$, where $c > 0$. At each point $\mathbf{p} \in \mathbb{L}_c^d$, the tangent space $T_{\mathbf{p}}\mathbb{L}_c^d$ consists of vectors $\mathbf{v} \in \mathbb{R}^{d+1}$ satisfying $\langle \mathbf{p}, \mathbf{v} \rangle_{\mathcal{L}} = 0$. The exponential map $\exp_{\mathbf{p}}^c : T_{\mathbf{p}}\mathbb{L}_c^d \to \mathbb{L}_c^d$ maps tangent vectors onto the manifold, whereas the logarithmic map $\log_{\mathbf{p}}^c : \mathbb{L}_c^d \to T_{\mathbf{p}}\mathbb{L}_c^d$ provides the inverse mapping. Given a Euclidean token embedding $\mathbf{z}_i^m \in \mathbb{R}^d$, we augment it as $\tilde{\mathbf{z}}_i^m = (0, \mathbf{z}_i^m) \in T_{\mathbf{o}}\mathbb{L}_c^d$, where $\mathbf{o} = (1/\sqrt{c}, \mathbf{0})$ is the Lorentz origin. We then obtain its manifold representation as $\mathbf{h}_i^m = \exp_{\mathbf{o}}^c(\tilde{\mathbf{z}}_i^m)$. Its nearest shared code under the Lorentzian geodesic distance:

$$k_i^S = \underset{1\le j\le K_S}{\arg\min}\, d_{\mathcal{L}}^c\left(\mathbf{h}_i^m, \mathbf{c}_j^S\right), \mathbf{q}_i^S = \log_{\mathbf{o}}^c\left(\mathbf{c}_{k_i^S}^S\right) \tag{1}$$

where $d_{\mathcal{L}}^c$ denotes the geodesic distance on $\mathbb{L}_c^d$. Because subtraction is not intrinsically defined between points on the Lorentz manifold, both the token representation and its assigned shared code are mapped to the common tangent space at $\mathbf{o}$. The resulting tangent-space residual is defined as

$$\widetilde{\mathbf{r}}_i^m = \widetilde{\mathbf{z}}_i^m - \mathbf{q}_i^S \in T_{\mathbf{o}}\mathbb{L}_c^d \tag{2}$$

The residual $\widetilde{\mathbf{r}}_i^m$ explicitly isolates the information not fully represented by the shared code. HRC maps the residual to Lorentz manifold and quantizes it using the codebook dedicated to modality m:

$$k_i^{R,m} = \underset{1\le j\le K_R}{\arg\min}\, d_{\mathcal{L}}^c\left(\exp_{\mathbf{o}}^c(\widetilde{\mathbf{r}}_i^m), \mathbf{c}_j^{R,m}\right), \mathbf{q}_i^{R,m} = \log_{\mathbf{o}}^c\left(\mathbf{c}_{k_i^{R,m}}^{R,m}\right) \tag{3}$$

Definitions of the Lorentz manifold, distance function, exponential map, logarithmic map are provided in Appendix A.

Consequently, the index pair $(k_i^S, k_i^{R,m})$ provides a discrete representation of each token. The shared index selects a code from the shared codebook, while the residual index selects an additional code for information not fully captured by the selected shared code. Thus, HRC decomposes token quantization into sequential shared and residual stages.

Shared Center Alignment. The shared codebook is designed to capture physiological patterns common to both modalities. Since EEG and other signals lack a one-to-one channel correspondence, their representations are aligned at the temporal-patch level rather than the channel level. Each patch-level center is computed intrinsically on the Lorentz manifold via a Fréchet mean, as follows.

We index tokens by a global index i, where each i already encodes its originating modality m, sample b, and temporal patch t. Let

$$\mathcal{I}_{b,t}^m = \{\, i : \text{token } i \text{ belongs to modality } m,\ \text{sample } b,\ \text{patch } t \,\}, \qquad |\mathcal{I}_{b,t}^m| = C_m,$$

denote the set of channel tokens of modality m within patch t of sample b, where C_m is the number of channels in modality m. For each sample b and patch t, we summarize the continuous channel representations of each modality by a Lorentzian center, and construct a shared-code center from the shared codes selected by all channel tokens of both modalities within the same patch:

$$\boldsymbol{\mu}_{b,t}^m = \mathrm{FM}\big(\{\mathbf{h}_i^m\}_{i\in\mathcal{I}_{b,t}^m}\big) \in \mathbb{L}_c^d, \qquad \boldsymbol{\mu}_{b,t}^S = \mathrm{FM}\big(\{\mathbf{c}_{k_i^S}^S : i \in \mathcal{I}_{b,t}^E \cup \mathcal{I}_{b,t}^Y\}\big) \in \mathbb{L}_c^d \tag{4}$$

Here k_i^S is the shared-code index assigned to token i in Eq. equation 1, and all tokens contribute equally to their respective centers. The modality centers and the shared-code center serve as channel-independent, patch-level representatives for alignment. The formal definition of the Fréchet mean is provided in Appendix A.4.

We use the shared-code center as geometric anchor and pull the two modality centers toward it:

$$\mathcal{L}_{\mathrm{SCA}} = \frac{1}{2BN_t}\sum_{b=1}^{B}\sum_{t=1}^{N_t}\sum_{m\in\{E,Y\}}\Big[d_{\mathcal{L}}^c\big(\boldsymbol{\mu}_{b,t}^m,\, \mathrm{sg}(\boldsymbol{\mu}_{b,t}^S)\big)\Big]^2 \tag{5}$$

Here B is the batch size, N_t is the number of temporal patches, and $\mathrm{sg}(\cdot)$ is the stop-gradient operator. Minimizing Eq. equation 5 encourages both modality centers to approach the same patch-level anchor under the Lorentzian geodesic distance. Because the anchor is detached, gradients from this objective act only on the encoder-generated modality representations, leaving the shared codebook unaffected.

Reconstruction and training objective. For each token i, the shared and residual codes are summed in the tangent space at $\mathbf{o}$ to form the quantized representation $\mathbf{z}_{i,\mathrm{q}}^m = \mathbf{q}_i^S + \mathbf{q}_i^{R,m}$. The quantized tokens are then restored to their original order and decoded by a Transformer. Gradients bypass the discrete assignment via a straight-through estimator, while the codebooks are updated by exponential

moving average (EMA). Let $\mathbf{X}_b^m$ and $\widehat{\mathbf{X}}_b^m$ denote the input and reconstructed signals of sample b for modality m. The reconstruction loss is

$$\mathcal{L}_{\text{rec}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{m\in\{E,Y\}} \left\|\widehat{\mathbf{X}}_b^m - \mathbf{X}_b^m\right\|_F^2. \tag{6}$$

Since the codebooks are maintained by EMA, only commitment losses are needed to keep the continuous representations close to their assigned codes:

$$\mathcal{L}_{\text{com}}^S = \frac{1}{L}\sum_i \left[d_{\mathcal{L}}^c\big(\exp_{\mathbf{o}}^c(\widetilde{\mathbf{z}}_i^m),\, \text{sg}(\mathbf{c}_{k_i^S}^S)\big)\right]^2,\ \mathcal{L}_{\text{com}}^R = \frac{1}{L}\sum_i \left[d_{\mathcal{L}}^c\big(\exp_{\mathbf{o}}^c(\widetilde{\mathbf{r}}_i^m),\, \text{sg}(\mathbf{c}_{k_i^{R,m}}^{R,m})\big)\right]^2 \tag{7}$$

where L is the total number of tokens. The complete HRC objective combines the reconstruction, commitment, and center-alignment losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_S\,\mathcal{L}_{\text{com}}^S + \lambda_R\,\mathcal{L}_{\text{com}}^R + \lambda_{\text{SCA}}\,\mathcal{L}_{\text{SCA}}, \tag{8}$$

where $\lambda_S, \lambda_R, \lambda_{\text{SCA}}$ are weighting coefficients.

3.2 Single-modality encoder transferring

After pretraining, the tokenizer and all codebooks are frozen; only the EEG encoder and the classification head are optimized. For each EEG token i, the frozen tokenizer, applied to the unmasked EEG input, produces a reference representation $\widetilde{\mathbf{z}}_i^T$ in the tangent space at $\mathbf{o}$, which serves as the teacher target. We map it to $\mathbb{L}_c^d$ and assign it to the nearest entry in the frozen shared codebook:

$$k_i^E = \underset{1\le j\le K_S}{\arg\min}\, d_{\mathcal{L}}^c\left(\exp_{\mathbf{o}}^c\left(\widetilde{\mathbf{z}}_i^T\right), \mathbf{c}_j^S\right). \tag{9}$$

Although k_i^E is assigned to an EEG token, the shared codebook was learned from paired EEG and eye-signal observations. The resulting index therefore provides a target determined by multimodal pretraining rather than by EEG observations alone.

During this stage, a set of positions Ω is randomly masked with ratio γ, and the selected EEG tokens are replaced by a learnable mask token. The masked sequence $\widetilde{\mathbf{X}}^E$ is processed by the EEG encoder F_ϕ, which produces a Euclidean representation $\mathbf{z}_i^E$ for each masked position from the remaining temporal and cross-channel context. Each representation is lifted to the Lorentz manifold as $\mathbf{p}_i = \exp_{\mathbf{o}}^c(\widetilde{\mathbf{z}}_i^E)$, where $\widetilde{\mathbf{z}}_i^E = (0, \mathbf{z}_i^E)$.

A hyperbolic classifier (with parameters θ) predicts the shared-code category using Lorentz decision hyperplanes. Its parameterization is given in Appendix B.6. The encoder is trained with a masked code prediction (MCP) loss, which combines categorical prediction term and geodesic alignment term:

$$\mathcal{L}_{\text{MCP}} = -\frac{1}{|\Omega|}\sum_{i\in\Omega}\log p_{\phi,\theta}\left(k_i^E \mid \widetilde{\mathbf{X}}_i^E\right) + \frac{1}{|\Omega|}\sum_{i\in\Omega}\left[d_{\mathcal{L}}^c\left(\mathbf{p}_i, \text{sg}\left(\mathbf{c}_{k_i^E}^S\right)\right)\right]^2, \tag{10}$$

where $p_{\phi,\theta}$ is the predicted distribution over shared-code indices. The first term predicts the teacher index k_i^E, whereas the second brings the masked-token representation $\mathbf{p}_i$ closer to the corresponding shared code on the Lorentz manifold.

EEG-Only Downstream Classification. After pretraining with masked code prediction, we remove the multimodal tokenizer, codebooks, and token-prediction module, and retain the pretrained EEG encoder F_ϕ as the downstream backbone. Given an unmasked EEG input $\mathbf{X}^E$, the encoder produces token representations, we average these token representations to obtain a sample-level feature, which is passed to a task-specific linear classifier for auditory attention decoding. The entire downstream model is trained using categorical cross-entropy.

4 Experiments

4.1 Baseline And Dataset Description.

We compare against four groups of baseline methods. Specifically, we include task-specific AAD models, such as DARNet (Yan et al., 2024), DBPNet (Ni et al., 2024), and S2M (Wang et al., 2026);

foundation models pretrained on large-scale EEG datasets, including LaBraM (Jiang et al., 2024) and CBraMod (Wang et al., 2025a); the hyperbolic EEG representation model HEEGNet (Li et al., 2026), which explores hyperbolic representations for capturing hierarchical and domain-invariant EEG features; and multimodal models that integrate EEG with complementary modalities, including CM-SSL (Fan et al., 2026), E2MO (Yin et al., 2026), and EEGMOCE (Zhou et al., 2026).

AVGCAAD Dataset The dataset is a public audio-visual dataset for AAD, containing multimodal physiological and behavioral recordings (Rotaru et al., 2024a). During the experiment, subjects were instructed to attend to one of two competing speech streams, with the attended speaker positioned at $\pm 90^{\circ}$. The dataset includes 64-channel EEG and 4-channel EOG recordings sampled at 8,192 Hz. Each trial lasted 10 minutes, during which the attended direction was switched after the first 5 minutes. We excluded subjects without public release permission and those with incomplete channel recordings, resulting in a final evaluation set of 12 subjects.

EARAAD Dataset The dataset is an EEG-based AAD dataset designed to evaluate selective auditory attention decoding across different EEG modalities (Fuglsang et al., 2020). The dataset contains simultaneously recorded scalp EEG, around-ear EEG, and in-ear EEG signals, including 29 scalp EEG channels with 3 EOG channels, 19 around-ear EEG channels, and 12 in-ear EEG channels. Fifteen normal-hearing participants were recorded while attending to one of two competing speech streams, with each participant completing six 10-minute trials.

4.2 Cross-subject scenario Comparison with State-of-the-Art Methods

Table 1 compares our method with task-specific AAD architectures and recent EEG representation-learning methods under 1s and 2s decision windows. Our method achieves the best performance across all four dataset-window settings, demonstrating consistent improvements over existing approaches. On AVGCAAD, our method surpasses EEGMOCE (Zhou et al., 2026) by 0.65% and 0.32% for the 1s and 2s windows, respectively. On EARAAD, it improves over LaBraM (Jiang et al., 2024) by 0.72% and 0.46%. Overall, our method achieves consistently superior performance compared with existing state-of-the-art methods across all datasets and decision-window settings.

The comparison with HEEGNet (Li et al., 2026) is particularly informative because both methods exploit hyperbolic representations. Our method consistently outperforms HEEGNet across the four settings, indicating that the improvements do not originate solely from the use of hyperbolic geometry. Instead, our approach further models shared cross-modal physiological structures, separates modality-specific variations through residual refinements, and transfers the learned shared representations to an EEG-only encoder. Our method also consistently improves over CM-SSL (Fan et al., 2026), achieving gains of 2.16%–4.48% across the four settings. This suggests that explicitly organizing shared and modality-specific information through a structured representation hierarchy is more effective than generic cross-modal self-supervision for these AAD benchmarks.

Table 1: Mean $\pm$ standard deviation of AAD decoding accuracy (%) achieved by the proposed model in comparision with various baseline models.

Type	Methods	AVGCAAD		EARAAD	
		1 s	2 s	1 s	2 s
AAD Task	DARNet (Yan et al., 2024)	52.94 ± 7.66	55.53 ± 8.02	51.47 ± 5.23	52.84 ± 14.86
	DBPNet (Ni et al., 2024)	51.50 ± 5.77	52.70 ± 4.66	52.64 ± 9.67	53.35 ± 16.46
	S^2M (Wang et al., 2026)	52.90 ± 3.20	53.81 ± 4.48	57.18 ± 14.32	57.99 ± 11.39
Foundation Model	LaBraM (Jiang et al., 2024)	51.66 ± 5.11	51.19 ± 4.05	$\underline{57.72} \pm 12.17$	$\underline{58.47} \pm 14.56$
	CBraMod (Wang et al., 2025a)	50.88 ± 2.32	51.89 ± 3.24	52.04 ± 10.28	53.99 ± 12.95
Hyperbolic	HEEGNet (Li et al., 2026)	52.82 ± 4.71	53.24 ± 5.03	55.42 ± 18.11	55.99 ± 17.61
Multimodal	CM-SSL (Fan et al., 2026)	53.10 ± 2.33	53.86 ± 1.99	53.95 ± 5.14	55.55 ± 10.72
	E2MO (Yin et al., 2026)	53.03 ± 2.62	54.30 ± 2.73	56.35 ± 8.29	57.12 ± 12.95
	EEGMOCE (Zhou et al., 2026)	$\underline{54.62} \pm 4.61$	$\underline{55.70} \pm 2.77$	55.84 ± 15.56	56.35 ± 19.32
	Ours	$\mathbf{55.26} \pm 2.96$	$\mathbf{56.02} \pm 2.49$	$\mathbf{58.43} \pm 13.54$	$\mathbf{58.93} \pm 12.69$

Bold indicates the best performance, and underline indicates the second-best performance.

4.3 Cross-Dataset scenario Comparison with State-of-the-Art Methods

To evaluate cross-dataset generalization, we conducted bidirectional transfer experiments between AVGCAAD and EARAAD. For each direction, the last subject in the source dataset was used for validation, while the remaining subjects were used for training. For baselines requiring fixed input channels, we retained only the channels shared by both datasets. In contrast, our method supports varying channel numbers and directly handles channel differences between training and testing.

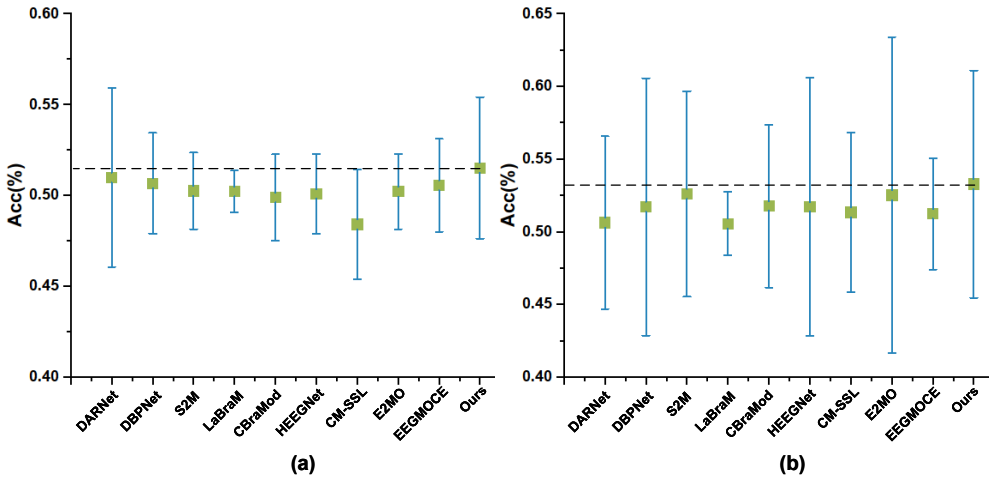

Figure 2: Performance (accuracy and error values) comparison under cross-dataset scenario. (a) Models are trained on EARAAD and evaluated on AVGCAAD. (b) Models are trained on AVGCAAD and evaluated on EARAAD.

As shown in Fig. 2, performance varied across transfer directions, highlighting the difficulty of cross-dataset recognition. Most methods performed better from AVGCAAD to EARAAD than in the reverse direction, indicating an asymmetric domain shift. Our method achieved the highest mean accuracy in both directions, demonstrating consistent cross-dataset performance. Its channel-count-agnostic design also avoids restricting the datasets to shared channels, making it suitable for heterogeneous recording settings.

4.4 Hyperbolic Geometry Analysis of Physiological Representations

To investigate whether physiological signals exhibit geometric properties suitable for hyperbolic representation learning, we measure the relative hyperbolicity δ_{rel} (Khrulkov et al., 2020; Zhou et al., 2026) of EEG, EYE/Ear-EEG, and multimodal (MULTI) representations on AVGCAAD and EARAAD. To separate the geometry of the signals themselves from that imposed by a learned encoder, we compute δ_{rel} on both the raw signals and the features produced by EEGNet (Lawhern et al., 2018), purely Euclidean CNN that we adopt as a standard geometry-agnostic encoder, so that any change between raw and encoded features reflects Euclidean encoding rather than a hyperbolic inductive bias. Lower δ_{rel} values indicate stronger tree-like structures, with values below 0.30 considered compatible with hyperbolic representations.

As shown in Table 2, all raw representations achieve low δ_{rel} values across both datasets, indicating that unimodal and multimodal physiological signals exhibit hyperbolic-compatible structures. Notably, multimodal representations maintain similar geometric characteristics after fusion, suggesting that cross-modal integration does not remove the underlying tree-like structure. We further analyze the impact of Euclidean feature extraction. Compared with raw signals, EEGNet-encoded representations generally show increased δ_{rel} values, indicating that Euclidean encoding weakens the original geometric characteristics of physiological signals. This effect is observed across both unimodal and multimodal representations. Overall, these findings suggest that multimodal physiological signals naturally contain hyperbolic-compatible structures, motivating geometry-aware representation learning approaches that explicitly model such characteristics.

Table 2: Relative hyperbolicity (δ_{rel}) of physiological representations under raw and Euclidean-encoded (EEGNet) settings. Values are reported as mean ± standard deviation across subjects. Lower δ_{rel} indicates a more tree-like geometry and thus better suitability for hyperbolic embedding.

	AVGCAAD (n=12)			EARAAD (n=15)		
Representation	EEG	EYE	MULTI	EEG	Ear-EEG	MULTI
Raw δ_{rel}	0.126±0.012	0.205±0.010	0.137±0.014	0.168±0.016	0.179±0.006	0.154±0.011
EEGNet δ_{rel}	0.245±0.030	0.274±0.027	0.298±0.024	0.206±0.008	0.164±0.013	0.281±0.048

4.5 Ablation Study

To evaluate the contribution of each component, we conduct ablation studies by removing key designs from our framework. The results on AVGCAAD and EARAAD are shown in Fig. 3. The full model achieves the highest accuracy across all evaluation settings, confirming the effectiveness of the proposed design. Replacing the hyperbolic representation with a Euclidean one consistently reduces performance, showing that hyperbolic representations provide a more suitable space for modeling multimodal physiological signals. Removing the residual codebook also leads to performance degradation, demonstrating the importance of decomposing representations into shared and modality-specific components. In addition, removing the auxiliary geometric loss further decreases accuracy, indicating that explicit geometric constraints help improve the learned representations.

Among all components, removing the masked transfer strategy causes the largest performance drop, particularly on AVGCAAD. This result highlights the importance of transferring multimodal knowledge to the EEG-only encoder through masked prediction. By recovering shared-code targets from contextual EEG information, the masked transfer strategy enables the encoder to effectively utilize multimodal supervision during EEG-only inference. Overall, the ablation results verify that each component contributes to the final performance, with their combination enabling effective multimodal knowledge transfer.

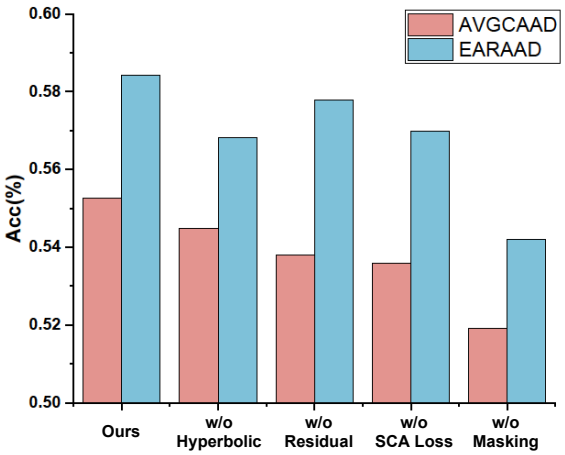

Figure 3: Ablation study results.

5 Conclusion

This work addresses the gap between multimodal training and EEG-only inference in auditory attention decoding. We propose a framework that combines shared and modality-specific codes to capture complementary physiological information, then transfers the shared vocabulary to an EEG encoder through masked code prediction. Experiments on AVGCAAD and EARAAD demonstrate consistently strong decoding performance, with ablations supporting the contribution of the main components. The current evaluation focuses on two public datasets and EOG or ear-EEG as auxiliary signals. Future work will assess generalization to broader physiological modalities, acquisition systems, and real-world listening environments.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability and clarity of the manuscript and to review the presentation and apparent consistency of mathematical expressions and equations. We did not use generative AI tools for other research tasks requiring disclosure, including developing the research methodology, designing experiments, implementing the proposed method, analyzing or interpreting results, or generating or processing data. Any AI-assisted edits and suggestions concerning mathematical expressions were reviewed by the authors, who independently checked the equations against the derivations and verified the final text. We take full responsibility for the content of this work, including all claims, equations, and results.

REFERENCES

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Martin G Bleichner, Bojana Mirkovic, and Stefan Debener. Identifying auditory attention with ear-eeg: ceegrid versus high-density cap-eeg comparison. *Journal of neural engineering*, 13(6): 066004, 2016.

András Búzás, András Makai, Géza I Groma, Zsolt Dancsházy, István Szendi, Laszlo B Kish, Ana Raquel Santa-Maria, and András Dér. Hierarchical organization of human physical activity. *Scientific Reports*, 14(1):5981, 2024.

Jiang Chang, Zhixin Zhang, Yuhua Qian, and Pan Lin. Multi-scale hyperbolic contrastive learning for cross-subject eeg emotion recognition. *IEEE Transactions on Affective Computing*, 16(3): 1716–1731, 2025.

Kai Cui, Jia Li, Yu Liu, Xuesong Zhang, Zhenzhen Hu, and Meng Wang. Physiosync: Temporal and cross-modal contrastive learning inspired by physiological synchronization for eeg-based emotion recognition. *IEEE Transactions on Computational Social Systems*, 13(2):1341–1354, 2026. doi: 10.1109/TCSS.2025.3602913.

Xiaomao Fan, Shuangdu Li, Jikui Liu, Jinzhou Cao, Jingyan Jiang, and Yumeng Liu. Cross-modality self-supervision for unimodal physiological signals representation learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6631–6635, 2026. doi: 10.1109/ICASSP55912.2026.11461644.

Søren A. Fuglsang, Jonatan Märcher-Rørsted, Torsten Dau, and Jens Hjortkjær. Selective auditory attention in normal-hearing and hearing-impaired listeners, January 2020. URL `https://doi.org/10.5281/zenodo.3618205`.

Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

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, Simon L Kappel, and Preben Kidmose. A direct comparison of simultaneously recorded scalp, around-ear and in-ear eeg for neural selective auditory attention decoding to speech. *Scientific Reports*, 15(1):41655, 2025.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2019.

Wei-Bang Jiang, Liming Zhao, and Bao-liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, pp. 16405–16426, 2024.

Wei-Bang Jiang, Xi Fu, Yi Ding, and Cuntai Guan. Towards robust multimodal physiological foundation models: Handling arbitrary missing modalities. *arXiv preprint arXiv:2504.19596*, 2025.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ECG words and sentences via pre-training ECG language model. In *The Thirteenth International Conference on Learning Representations*, 2025.

Netiwit Kaongoen, Jaehoon Choi, Jin Woo Choi, Haram Kwon, Chaeeun Hwang, Guebin Hwang, Byung Hyung Kim, and Sungho Jo. The future of wearable eeg: a review of ear-eeg technology and its applications. *Journal of Neural Engineering*, 20(5):051002, oct 2023. doi: 10.1088/1741-2552/acfcda.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6417–6427, 2020. doi: 10.1109/CVPR42600.2020.00645.

Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. Hyperbolic geometry of complex networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 82(3):036106, 2010.

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

Shanglin Li, Chu Shiwen, Okan Koç, Yi Ding, Qibin Zhao, Motoaki Kawanabe, and Ziheng Chen. Heegnet: Hyperbolic embeddings for eeg. In *International Conference on Learning Representations*, volume 2026, pp. 74317–74346, 2026.

Huan Liu, Tianyu Lou, Yuzhe Zhang, Yixiao Wu, Yang Xiao, Christian S. Jensen, and Dalin Zhang. Eeg-based multimodal emotion recognition: A machine learning perspective. *IEEE Transactions on Instrumentation and Measurement*, 73:1–29, 2024. doi: 10.1109/TIM.2024.3369130.

Mutian Liu, Banghua Yang, Lin Meng, Yonghuai Zhang, Shouwei Gao, Peng Zan, and Xinxing Xia. Sta-net: Spatial–temporal alignment network for hybrid eeg-fnirs decoding. *Information Fusion*, 119:103023, 2025. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.103023.

Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. Codebrain: Bridging decoupled tokenizer and multi-scale architecture for eeg foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 83526–83574, 2026.

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

Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, 132 (9):3484–3508, 2024.

Sahar Hassanzadeh Mostafaei, Jafar Tanha, and Amir Sharafkhaneh. A novel deep learning model based on transformer and cross modality attention for classification of sleep stages. *Journal of Biomedical Informatics*, 157:104689, 2024. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2024.104689.

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 *IJCAI*, 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, 2015.

Saurav Pahuja, Gabriel Ivucic, Felix Putze, Siqi Cai, Haizhou Li, and Tanja Schultz. Enhancing subject-independent eeg-based auditory attention decoding with wgan and pearson correlation coefficient. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3715–3720, 2023. doi: 10.1109/SMC53992.2023.10394150.

Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10023–10044, 2022. doi: 10.1109/TPAMI.2021.3136921.

Rajasekhar Pillalamarri and Udhayakumar Shanmugam. A review on eeg-based multimodal learning for emotion recognition. *Artificial Intelligence Review*, 58(5):131, 2025.

Jathurshan Pradeepkumar, Xihao Piao, Zheng Chen, and Jimeng Sun. Tokenizing single-channel eeg with time-frequency motif learning. In *International Conference on Learning Representations*, volume 2026, pp. 36611–36637, 2026.

Caitlin N Price and Deborah Moncrieff. Defining the role of attention in hierarchical auditory processing. *Audiology research*, 11(1):112–128, 2021.

Iustina Rotaru, Simon Geirnaert, Alexander Bertrand, and Tom Francart. Audiovisual, gaze-controlled auditory attention decoding dataset ku leuven (av-gc-aad), April 2024a. URL `https://doi.org/10.5281/zenodo.11058711`.

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, 2024b.

Jian Shen, Huajian Liang, Ruirui Ma, Yihong Xu, Kexin Zhu, Haoran Gao, Kechen Hou, Yanan Zhang, Xiaowei Zhang, and Bin Hu. Emotion recognition using multimodal physiological signals through regional to global fusion with a spatial-temporal semantic alignment mechanism. *Information Fusion*, 132:104224, 2026. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2026.104224.

Ryohei Shimizu, YUSUKE Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=Ec85b0tUwbA`.

Ivo V Stuldreher, Nattapong Thammasan, Jan BF van Erp, and Anne-Marie Brouwer. Physiological synchrony in eeg, electrodermal activity and heart rate reflects shared selective auditory attention. *Journal of neural engineering*, 17(4):046028, 2020.

Jamiyan Sukhbaatar, Satoshi Imamura, Ibuki Inoue, Shoya Murakami, Kazi Mahmudul Hassan, Seungwoo Han, Ingon Chanpornpakdi, and Toshihisa Tanaka. Singlem: Single-channel large eeg model. *IEEE Journal of Biomedical and Health Informatics*, pp. 1–14, 2026. doi: 10.1109/JBHI.2026.3732223.

Sai Sun, Hongbo Yu, Rongjun Yu, and Shuo Wang. Functional connectivity between the amygdala and prefrontal cortex underlies processing of emotion ambiguity. *Translational psychiatry*, 13(1):334, 2023.

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

Jiaqi Wang, Zhengyu Ma, Xiongri Shen, Chenlin Zhou, Leilei Zhao, Han Zhang, Yi Zhong, Siqi Cai, Zhenxi Song, and Zhiguo Zhang. S ̃2 m-former: Spiking symmetric mixing branchformer for brain auditory attention detection. *Advances in Neural Information Processing Systems*, 38:126593–126620, 2026.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *International conference on learning representations*, volume 2025, pp. 75310–75346, 2025a.

Jiyao Wang, Xiao Yang, Qingyong Hu, Jack Tang, Can Liu, Dengbo He, Yuntao Wang, Yingcong Chen, and Kaishun Wu. Physdrive: A multimodal remote physiological measurement dataset for in-vehicle driver monitoring. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025b. doi: 10.52202/085713-1230.

Johanna Wilroth, Bo Bernhardsson, Frida Heskebeck, Martin A Skoglund, Carolina Bergeling, and Emina Alickovic. Improving eeg-based decoding of the locus of auditory attention through domain adaptation. *Journal of Neural Engineering*, 20(6):066022, dec 2023. doi: 10.1088/1741-2552/ad0e7b.

Xiran Xu, Boda Xiao, Bo Wang, Yujie Yan, Xihong Wu, Heping Cheng, and Jing Chen. Utilizing eyeblink information to improve eeg-based auditory attention decoding. *Biomedical Signal Processing and Control*, 123:110555, 2026. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2026.110555.

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. *Advances in Neural Information Processing Systems*, 37:31688–31707, 2024.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

Hao-Long Yin, Jian-Ming Zhang, Ren-Jie Dai, Wei-Long Zheng, Qinyu Lv, Zhenghui Yi, and Bao-Liang Lu. A multimodal eeg-eye movement model for automatic depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2218–2226, 2026.

Liangfei Zhang, Yifei Qian, Ognjen Arandjelović, Tianyi Zhu, and Hongjiang Xiao. Multimodal latent emotion recognition from micro-expression and physiological signal. *Pattern Recognition*, 169:111963, 2026.

Runhe Zhou, Shanglin Li, Guanxiang Huang, Xinliang Zhou, Qibin Zhao, Motoaki Kawanabe, Yi Ding, and Cuntai Guan. EEG-based multimodal learning via hyperbolic mixture-of-curvature experts. In *Forty-third International Conference on Machine Learning*, 2026.

A Lorentz Operations Used in Hyperbolic Residual Codebook (HRC)

This appendix summarizes the standard hyperbolic operations used by the Hyperbolic Residual Codebook (HRC). These definitions are included for completeness and reproducibility, and do not constitute the main methodological contribution.

A.1 Lorentz manifold and inner product

We use the Lorentz model with constant negative curvature $-c$ and curvature parameter $c > 0$. For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{d+1}$, their Lorentz inner product is defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} = -x_0 y_0 + \sum_{\ell=1}^{d} x_\ell y_\ell. \tag{11}$$

The d-dimensional Lorentz manifold is

$$\mathbb{H}_c^d = \left\{ \mathbf{x} \in \mathbb{R}^{d+1} \;\middle|\; \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = -\frac{1}{c}, \quad x_0 > 0 \right\}. \tag{12}$$

Its canonical origin is $\mathbf{o} = (1/\sqrt{c}, 0, \ldots, 0)$. The tangent space at $\mathbf{o}$ is

$$T_{\mathbf{o}}\mathbb{L}_c^d = \left\{ \mathbf{v} \in \mathbb{R}^{d+1} \;\middle|\; \langle \mathbf{o}, \mathbf{v} \rangle_{\mathcal{L}} = 0 \right\}. \tag{13}$$

Therefore, augmenting a Euclidean embedding $\mathbf{z} \in \mathbb{R}^d$ as $\widetilde{\mathbf{z}} = (0, \mathbf{z})$ places it in $T_{\mathbf{o}}\mathbb{L}_c^d$. Tangent vectors at $\mathbf{o}$ are spacelike, so the Lorentz norm $\|\mathbf{v}\|_{\mathcal{L}} = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle_{\mathcal{L}}}$ is real-valued and well-defined.

A.2 Exponential map

The exponential map converts a tangent vector into a point on the Lorentz manifold. For $\mathbf{v} \in T_{\mathbf{o}}\mathbb{L}_c^d$, it is defined as

$$\exp_{\mathbf{o}}^c(\mathbf{v}) = \cosh\left(\sqrt{c}\,\|\mathbf{v}\|_{\mathcal{L}}\right)\mathbf{o} + \frac{\sinh\left(\sqrt{c}\,\|\mathbf{v}\|_{\mathcal{L}}\right)}{\sqrt{c}\,\|\mathbf{v}\|_{\mathcal{L}}}\,\mathbf{v}. \tag{14}$$

When $\|\mathbf{v}\|_{\mathcal{L}} = 0$, the second coefficient is evaluated by continuity and $\exp_{\mathbf{o}}^c(\mathbf{0}) = \mathbf{o}$. In HRC, this map transforms Euclidean token embeddings and residuals into valid Lorentz points, so that their distances to hyperbolic code vectors can be evaluated.

A.3 Lorentz distance and logarithmic map

For $\mathbf{x}, \mathbf{y} \in \mathbb{L}_c^d$, their geodesic distance is

$$d_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{c}} \operatorname{arcosh}\left(-c\,\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}\right). \tag{15}$$

HRC uses this distance for nearest-neighbor assignment because it measures the geodesic separation between a token and a code vector on the Lorentz manifold.

The logarithmic map is the inverse of the exponential map: it maps a Lorentz point back to the tangent space at the origin. Let

$$\alpha = -c\,\langle \mathbf{o}, \mathbf{x} \rangle_{\mathcal{L}}. \tag{16}$$

For $\mathbf{x} \in \mathbb{L}_c^d$ we have $\alpha \geq 1$, so that both $\operatorname{arcosh}(\alpha)$ and $\sqrt{\alpha^2 - 1}$ are well-defined. For $\mathbf{x} \neq \mathbf{o}$, the logarithmic map is

$$\log_{\mathbf{o}}^c(\mathbf{x}) = \frac{\operatorname{arcosh}(\alpha)}{\sqrt{\alpha^2 - 1}}\,(\mathbf{x} - \alpha\mathbf{o}). \tag{17}$$

We define $\log_{\mathbf{o}}^c(\mathbf{o}) = \mathbf{0}$ by continuity. HRC uses the logarithmic map to express selected hyperbolic codes in a common tangent space. This operation makes the Euclidean residual computation well-defined and avoids directly subtracting two manifold points.

A.4 Weighted Fréchet Mean

Given a collection of points $\{\mathbf{p}_i\}_{i=1}^M \subset \mathbb{L}_c^d$ with nonnegative weights $\{\eta_i\}_{i=1}^M$ satisfying $\sum_{i=1}^M \eta_i > 0$, their weighted Fréchet mean is defined as

$$\mathrm{wFM}_{\boldsymbol{\eta}}\left(\{\mathbf{p}_i\}_{i=1}^M\right) = \operatorname*{arg\,min}_{\mathbf{q} \in \mathbb{H}_c^d} \sum_{i=1}^M \eta_i d_{\mathcal{L}}^2(\mathbf{q}, \mathbf{p}_i). \tag{18}$$

Multiplying all weights by the same positive constant does not change the minimizer. We therefore normalize them such that $\sum_i \eta_i = 1$ in implementation. Uniform weights, $\eta_i = 1/M$, recover the standard Fréchet mean.

Unlike an arithmetic mean in the ambient space, the Fréchet mean minimizes intrinsic geodesic distances and remains on the Lorentz manifold. Because hyperbolic space is a Hadamard manifold,

the squared-distance objective is geodesically convex and admits a unique Fréchet mean for a finite collection of points with positive total weight.

The corresponding weighted Fréchet variance is

$$\nu_{\boldsymbol{\eta}}^2 = \sum_{i=1}^{M} \eta_i d_{\mathcal{L}}^2 \left(\boldsymbol{\mu}, \mathbf{p}_i\right), \qquad \boldsymbol{\mu} = \mathrm{wFM}_{\boldsymbol{\eta}} \left(\{\mathbf{p}_i\}_{i=1}^{M}\right). \tag{19}$$

Although HRC uses the Fréchet mean to construct geometric anchors, the Fréchet variance is not explicitly included in the alignment objective; it is reported here only for completeness.

B Reconstruction and Optimization Details

This appendix describes the optimization and reconstruction operations used to train HRC. These operations support end-to-end learning but are not themselves the principal contribution of the proposed quantization mechanism.

B.1 Token rearrangement and decoding

The straight-through token representations are rearranged according to the temporal and channel positions from which they were generated. For a batch of B samples, this operation produces

$$\widehat{\mathbf{Z}} \in \mathbb{R}^{B \times L \times d}, \tag{20}$$

where L is the total number of tokens and d is the token dimension. A Transformer decoder G_ψ maps these quantized tokens back to the input space:

$$\widehat{\mathbf{X}} = G_\psi \left(\widehat{\mathbf{Z}}\right) \in \mathbb{R}^{B \times N_t \times C \times P}. \tag{21}$$

Here, N_t denotes the number of temporal segments, C is the total number of physiological channels, and P denotes the number of samples within each segment. The reconstructed tensor is subsequently divided into modality-specific components $\widehat{\mathbf{X}}^E$ and $\widehat{\mathbf{X}}^Y$ according to the original channel partition.

B.2 Modality-wise reconstruction objective

Because the two physiological modalities may contain different numbers of channels, we normalize their reconstruction errors independently. For modality $m \in \{E, Y\}$, the reconstruction loss is

$$\mathcal{L}_{\mathrm{rec}}^m = \frac{\left\| \mathbf{X}^m - \widehat{\mathbf{X}}^m \right\|_1}{B\, N_t\, C_m\, P}, \qquad m \in \{E, Y\}, \tag{22}$$

where C_m is the number of channels associated with modality m. Normalizing by the number of reconstructed values prevents modalities with more channels from dominating solely because of their dimensionality.

The two reconstruction terms are combined as

$$\mathcal{L}_{\mathrm{rec}} = \lambda_E \mathcal{L}_{\mathrm{rec}}^E + \lambda_Y \mathcal{L}_{\mathrm{rec}}^Y, \tag{23}$$

where λ_E and λ_Y control the relative contributions of the two modalities.

B.3 Masked EEG Input Construction

Let $\mathbf{a}_i^E$ denote the projected EEG embedding at patch–channel position i. We sample a mask set Ω according to the masking ratio γ and construct the encoder input as

$$\widetilde{\mathbf{a}}_i^E = \begin{cases} \mathbf{e}_{\text{mask}}, & i \in \Omega, \\ \mathbf{a}_i^E, & i \notin \Omega, \end{cases} \tag{24}$$

where $\mathbf{e}_{\text{mask}}$ is a learnable embedding shared across masked positions. We denote the resulting masked sequence by $\widetilde{\mathbf{X}}^E = \{\widetilde{\mathbf{a}}_i^E\}$, which is processed by the EEG encoder:

$$\left\{\mathbf{h}_i^E\right\} = F_\phi\left(\widetilde{\mathbf{X}}^E\right). \tag{25}$$

The prediction objective is evaluated only at positions in Ω. In implementation, at least one valid position is masked in every sample to avoid an empty prediction set.

B.4 HMLR-Based Hyperbolic Prediction Head

The prediction head is implemented using an HMLR layer rather than a Euclidean affine transformation. We denote this parameterized mapping by

$$\text{HMLR}_\theta^c : \mathbb{R}^{d_h} \to \mathbb{L}_c^d, \qquad \widehat{\mathbf{u}}_i = \text{HMLR}_\theta^c\left(\mathbf{h}_i^E\right), \tag{26}$$

where d_h is the output dimension of the EEG encoder and d is the spatial dimension of the shared Lorentz codebook. The output satisfies the Lorentz constraint

$$\langle \widehat{\mathbf{u}}_i, \widehat{\mathbf{u}}_i \rangle_{\mathcal{L}} = -\frac{1}{c}, \qquad \widehat{u}_{i,0} > 0. \tag{27}$$

Unlike a Euclidean prediction head, HMLR does not treat the output space as an unconstrained Euclidean vector space. It performs the trainable transformation while respecting the geometry of $\mathbb{L}_c^d$, producing valid Lorentz points that can be directly compared with the frozen codebook through $d_{\mathcal{L}}$. The exact HMLR parameterization is given in Appendix B.6.

B.5 Masked Prediction Losses

The code probability at masked position i is

$$p_{\phi,\theta}\left(k \mid \widetilde{\mathbf{X}}^E, i\right) = \frac{\exp(\ell_{i,k})}{\sum_{j=1}^{K_S} \exp(\ell_{i,j})}. \tag{28}$$

The categorical masked-prediction loss is

$$\mathcal{L}_{\text{mask}} = -\frac{1}{|\Omega|} \sum_{i \in \Omega} \log p_{\phi,\theta}\left(k_i^\star \mid \widetilde{\mathbf{X}}^E, i\right), \tag{29}$$

and the geometric alignment term is

$$\mathcal{L}_{\text{geo}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} d_{\mathcal{L}}^2\left(\widehat{\mathbf{u}}_i, \text{sg}\left(\mathbf{c}_{k_i^\star}^S\right)\right). \tag{30}$$

Accordingly, $\mathcal{L}_{\text{MCP}} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{geo}}$, which is written in expanded form in Eq. equation 10.

B.6 Lorentz Multinomial Logistic Regression

Lorentz multinomial logistic regression performs classification by measuring the signed hyperbolic distance between an input point and a class-specific decision hyperplane. Let

$$\mathbf{p} = (p_t, \mathbf{p}_s) \in \mathbb{L}_c^d, \tag{31}$$

where p_t and $\mathbf{p}_s$ denote its time-like and spatial components, respectively. Each class $k \in \{1, \dots, K_S\}$ is associated with trainable parameters $a_k \in \mathbb{R}$ and $\mathbf{z}_k \in \mathbb{R}^d$. Its signed hyperplane distance is

$$v_{\mathbf{z}_k, a_k}(\mathbf{p}) = \frac{1}{\sqrt{c}} \operatorname{sign}(\alpha_k)\, \beta_k \left| \operatorname{arsinh}\left(\frac{\sqrt{c}\, \alpha_k}{\beta_k} \right) \right|, \tag{32}$$

where

$$\begin{aligned} \alpha_k &= \cosh(\sqrt{c} a_k)\, \langle \mathbf{z}_k, \mathbf{p}_s \rangle - \sinh(\sqrt{c} a_k) \|\mathbf{z}_k\|_2 p_t, \\ \beta_k &= \sqrt{\left\| \cosh(\sqrt{c} a_k) \mathbf{z}_k \right\|_2^2 - \left(\sinh(\sqrt{c} a_k) \|\mathbf{z}_k\|_2 \right)^2}. \end{aligned} \tag{33}$$

The corresponding logit is

$$\ell_{i,k} = \frac{v_{\mathbf{z}_k, a_k}(\widehat{\mathbf{u}}_i)}{\tau}. \tag{34}$$

The sign of $v_{\mathbf{z}_k, a_k}(\widehat{\mathbf{u}}_i)$ indicates on which side of the class hyperplane the input lies, whereas its magnitude measures the hyperbolic distance to that hyperplane. HMLR therefore defines class boundaries intrinsically in Lorentz space instead of applying a Euclidean affine classifier to hyperbolic representations.

C δ-HYPERBOLICITY

To quantify how well a set of representations is described by hyperbolic geometry, we adopt Gromov's δ-hyperbolicity, which Khrulkov et al. (2020) introduced as a measure of the degree of tree-like structure inherent in embeddings. Intuitively, δ reports how close the metric relations among points are to those of a tree: the metric relations between any four points coincide with those in a tree up to an additive constant δ, and a smaller $\delta \geq 0$ indicates a more strongly hyperbolic (tree-like) structure.

The measure is defined through the Gromov product. For the Lorentz model $\mathbb{L}_c^d$ with geodesic distance $d_{\mathcal{L}}$, the Gromov product of two points $\mathbf{y}, \mathbf{z} \in \mathbb{L}_c^d$ with respect to a base point $\mathbf{x} \in \mathbb{L}_c^d$ is

$$(\mathbf{y}, \mathbf{z})_{\mathbf{x}} = \frac{1}{2} \left(d_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) + d_{\mathcal{L}}(\mathbf{x}, \mathbf{z}) - d_{\mathcal{L}}(\mathbf{y}, \mathbf{z}) \right). \tag{35}$$

Geometrically, a larger value indicates that $\mathbf{x}$ lies further from the geodesic connecting $\mathbf{y}$ and $\mathbf{z}$.

The space is said to be δ-hyperbolic, for some $\delta \geq 0$, if it satisfies the four-point condition: for any $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w} \in \mathbb{L}_c^d$,

$$(\mathbf{x}, \mathbf{z})_{\mathbf{w}} \geq \min \left\{ (\mathbf{x}, \mathbf{y})_{\mathbf{w}}, (\mathbf{y}, \mathbf{z})_{\mathbf{w}} \right\} - \delta. \tag{36}$$

Trees satisfy this condition exactly with $\delta = 0$, so δ can be read as the deviation from perfect tree geometry; the smallest δ for which Eq. equation 36 holds characterizes the hyperbolicity of the point set.

Because the raw δ scales with the diameter of the data and is therefore not directly comparable across modalities or datasets, we report the diameter-normalized value (Khrulkov et al., 2020)

$$\delta_{\mathrm{rel}}(X) = \frac{2\, \delta(X)}{\operatorname{diam}(X)} \in [0, 1], \qquad \operatorname{diam}(X) = \max_{\mathbf{x}, \mathbf{y} \in X} d_{\mathcal{L}}(\mathbf{x}, \mathbf{y}), \tag{37}$$

where X denotes the set of representations under analysis. Values of δ_{rel} close to zero indicate strong hierarchical structure that admits a low-distortion hyperbolic embedding. We use δ_{rel} to verify that the target modalities carry intrinsic tree-like structure, which motivates embedding them in hyperbolic rather than Euclidean space.

D Baselines and Evaluation Protocol

Baselines. We compare against four families of methods. *(i) Task-specific AAD models*—DARNet (Yan et al., 2024), DBPNet (Ni et al., 2024), and S2M (Wang et al., 2026)—are supervised networks designed directly for auditory attention decoding and represent the conventional, non-pretrained approach; DARNet builds spatiotemporal representations through a dual attention refinement module, while DBPNet fuses temporal- and frequency-domain branches. *(ii) Foundation models* pretrained on large-scale EEG corpora, namely LaBraM (Jiang et al., 2024) and CBraMod (Wang et al., 2025a), are adapted to our task and fine-tuned under the same protocol as our model. *(iii) The hyperbolic EEG representation model* HEEGNet (Li et al., 2026) likewise embeds EEG features in hyperbolic space to capture hierarchical and domain-invariant structure, and thus isolates the contribution of our codebook-guided design from the choice of geometry alone. *(iv) Multimodal models* that integrate EEG with complementary modalities—CM-SSL (Fan et al., 2026), E2MO (Yin et al., 2026), and EEGMOCE (Zhou et al., 2026)—exploit both EEG and the auxiliary modality (eye-tracking on AVGCAAD, Ear-EEG on EARAAD); comparing against them tests whether our EEG-only model can match methods that retain the second modality at inference. All baselines are evaluated under the same cross-subject and cross-dataset protocols described below: for the foundation models we fine-tune the released pretrained encoders, and the remaining models are trained following their original configurations, so that differences reflect the representation rather than the evaluation pipeline.

Cross-dataset channel handling. Because our model tokenizes each (channel, patch) pair independently, the Transformer operates on a sequence of channel–patch tokens and is inherently agnostic to the absolute number of input channels. Our model therefore requires no channel alignment for cross-dataset transfer: it is pretrained on the full channel set of one dataset and applied directly to the other, even though the two corpora differ in channel count, without any change to its architecture. The baseline methods that assume a fixed input channel dimension cannot accommodate this mismatch. For these methods we instead restrict both datasets to the EEG channels they have in common, which yields a shared set of 27 channels, and we train and evaluate them on this identical 27-channel subset so that a single fixed-width architecture applies to both datasets. In all cases, sequence length is reconciled by linear resampling to the common patch length, per-channel normalization is applied to the selected channels, and non-finite values are repaired (finite per-channel mean, or zero for a fully invalid channel) before tokenization, giving a single reproducible pipeline for cross-dataset evaluation.

E Implementation Details and Hyperparameters

To support reproducibility, we report the complete set of architectural and optimization hyperparameters used in our three training stages: (i) tokenizer pretraining, which learns the hyperbolic residual codebook; (ii) encoder pretraining via masked code prediction (MCP); and (iii) downstream fine-tuning for auditory attention decoding. Input, patching, and backbone settings that are shared across stages are listed in Table 3, and the stage-specific settings are listed in Tables 4, 5, and 6. All experiments use mixed precision, gradient-norm clipping, and AdamW; we adopt a leave-one-subject-out protocol with the seed fixed to 42.

Table 3: Shared input, patching, and Transformer backbone configuration.

Hyperparameter	Value
EEG channels (AVGCAAD / EARAAD)	64 / 29
EYE(Ear-EEG) channels (AVGCAAD / EARAAD)	4 / 34
Number of patches N_t	8
Patch size t_{patch}	16
Input sequence length ($N_t \times t_{\text{patch}}$)	128
Max sequence length	1024
Sampling stride	128
Per-channel normalization	Yes
Patch (temporal) embedding	Linear $16 \to 128$ + GELU
Positional encoding	Sinusoidal
Embedding dimension D_{embed}	128

Table 4: Stage 1: Tokenizer pretraining (hyperbolic residual codebook).

Hyperparameter	Value
Transformer encoder / decoder	
Encoder layers	10
Encoder attention heads	8
Decoder layers	2
Decoder attention heads	4
Feed-forward dimension ($4 \times D_{\text{embed}}$)	512
Dropout	0.1
Activation	GELU
Codebook / quantizer	
Geometry	Hyperbolic (Lorentz)
Curvature c	1.0
Shared codebook size	256
Residual (private) codebook size	128
Code dimension	128
Commitment weight β	0.25
EMA decay ρ	0.95
k-means initialization	Yes (10 iters)
Max Lorentz norm	2.0
Dead-code threshold / interval	1.0 / 100
Loss weights	
EEG / EYE reconstruction weight	1.0 / 1.0
λ_{SCA} weight	0.1
λ_S and λ_R weight	0.25
Optimization	
Optimizer	AdamW
Learning rate	3×10^{-4}
Weight decay	1×10^{-5}
LR scheduler	Cosine
Warmup epochs	3
Minimum LR	1×10^{-5}
Batch size	128
Gradient accumulation	2
Gradient clipping (norm)	1.0
Precision	16-mixed
Max / min epochs	100 / 50
Early-stopping patience	25

Table 5: Stage 2: Encoder pretraining via masked code prediction (MCP).

Hyperparameter	Value
Encoder architecture	
Input patch dimension	16
Embedding dimension E_{embed}	256
Attention heads	8
Transformer layers	6
Feed-forward dimension ($4 \times E_{\text{embed}}$)	1024
Dropout	0.3
Stochastic depth	0.0
Activation	GELU
Mask-token init. std.	0.02
MCP objective	
Mask ratio γ	0.4
Prediction head	HMLR (Lorentz MLR)
Prediction temperature τ	0.2
Commitment weight β	1.0
Distillation weight	0.1
Optimization	
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	0.1
LR scheduler	Cosine
Warmup ratio	0.1
Batch size	128
Gradient accumulation	2
Gradient clipping (norm)	1.0
Precision	16-mixed
Max epochs	100
Early-stopping patience	10

Table 6: Stage 3: Downstream fine-tuning for auditory attention decoding.

Hyperparameter	Value
Fine-tuning mode	Linear probing (backbone frozen)
Classifier	Fully connect layer
Hyperbolic dimension	128
Curvature c	1.0
Classifier temperature	1.0
Pooling	Centroid
Optimizer	AdamW
Head learning rate	1×10^{-4}
Weight decay	1×10^{-4}
LR scheduler	Cosine
Batch size	32
Epochs	100
Early-stopping patience	10