COGEEG: COGNITION-GUIDED HYPERBOLIC LEARNING FOR CROSS-SUBJECT EEG–VISUAL RETRIEVAL

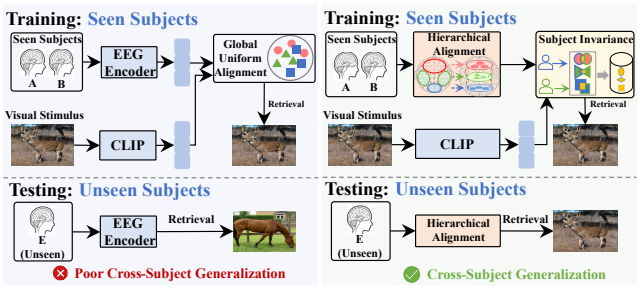
Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-subject EEG–visual retrieval requires learning from source subjects while generalizing to unseen subjects. Existing methods typically perform subject alignment on a single global representation. However, such a level-agnostic paradigm overlooks the heterogeneity across cognitive levels (*i.e.,* different levels exhibit distinct EEG–visual correspondence patterns and varying degrees of cross-subject distribution shift), thereby limiting cross-subject generalization. To address this limitation, we propose **CogEEG**, a **Cog**nition-guided hyperbolic learning framework for cross-subject **EEG**–visual retrieval from *global uniform alignment* to *cognitive-level-aware adaptive alignment*. Specifically, CogEEG builds multi-level EEG representations from brain-region priors and captures their hierarchy via level-specific hyperbolic geometry and geodesic cross-level interaction. These representations are further aligned with multi-level visual features to establish structured EEG–visual correspondences. Building on the hierarchical representations, CogEEG dynamically adjusts domain-adversarial alignment across cognitive levels based on visual reliability and subject discriminability, enabling level-aware cross-subject alignment. This alignment reduces level-specific subject shifts while improving generalization to unseen subjects. Extensive experiments show that CogEEG achieves superior performance on cross-subject EEG–visual retrieval.

1 INTRODUCTION

EEG–image retrieval aims to identify visual content from EEG signals elicited by visual stimuli [37; 6; 44; 47]. Unlike conventional EEG classification, it requires fine-grained correspondence between neural activity and visual semantics. This task provides a useful setting for studying human visual perception and cognition [11; 39; 31], with potential applications in brain–computer interfaces [21; 20; 22] and visual content decoding [3; 17; 50; 30; 10]. Recent methods exploit pretrained visual models to provide rich semantic supervision and align EEG with visual features in a shared embedding space [42; 36; 37]. However, these methods mainly focus on cross-modal alignment and often assume that the learned EEG representation transfers across subjects. In practice, EEG responses can vary substantially across individuals even for the same visual stimulus. This inter-subject variability makes EEG–visual mappings learned from source subjects generalize poorly to unseen subjects.



(a) Previous Cross-Subject Methods (b) CogEEG(Ours)

Figure 1: Motivation of CogEEG: paradigm comparison of global uniform alignment and our cognitive-level-aware adaptive alignment for cross-subject EEG–visual retrieval.

Existing cross-subject EEG studies mainly follow two directions: *subject-invariant representation learning* and *new-subject adaptation and generalization*. The former reduces distribution discrepancies across subjects to learn shared EEG features [34; 49], while the latter uses target-subject information or transfer mechanisms to improve generalization [49; 19]. Most EEG–visual retrieval methods adopt the former strategy, aligning a single global EEG representation across subjects to

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.

reduce distribution discrepancies (Fig. 1(a)). While effective to some extent, this treatment ignores the fact that EEG–visual correspondence varies across cognitive levels and implicitly assumes a similar degree of subject shift at each level. In practice, different cognitive levels exhibit distinct EEG–visual correspondence patterns [7] and are affected differently by inter-subject variability [34]. A uniform global constraint is therefore unlikely to align all levels equally well and may limit cross-subject generalization. This motivates *cognitive-level-aware adaptive alignment*, where the alignment strength is adjusted according to the characteristics of each cognitive level.

A natural basis for such alignment is the hierarchy of human visual cognition. Neuroscientific studies show that human visual processing progresses across cognitive levels, transforming low-level perceptual information into object-level and high-level semantic representations [7; 14; 5]. EEG decoding studies further suggest that visually evoked responses evolve from low-level visual features toward object and semantic information over time [2]. These findings motivate us to capture visual EEG at multiple cognitive levels rather than compress it into a single feature vector. Importantly, these levels are not independent: perceptual cues contribute to object representations, which in turn support high-level semantic understanding. Such cross-level relations are inherently hierarchical and can be difficult to preserve in flat Euclidean space [4; 35]. Hyperbolic space offers a suitable alternative because its negative curvature can represent hierarchical structures with low distortion [1; 13; 40]. Together, these observations lead to our central idea: *first establish hierarchical EEG–visual correspondences, and then adapt cross-subject alignment to each cognitive level*.

Based on this idea, we propose **CogEEG**, a **Cog**nition-guided hyperbolic learning framework for cross-subject **EEG**–visual retrieval (see Fig.1 (b)). CogEEG contains two components: Brain-Region-Guided Hierarchical Cognitive Pathway (BHCP) and Adaptive Hierarchical Domain-Adversarial Learning (AHDA). Specifically, BHCP constructs hierarchical EEG representations based on brain-region priors and models their hierarchical structures in level-specific hyperbolic spaces, leveraging negative curvature geometry to preserve hierarchical relationships and geodesic cross-level interactions to integrate cross-level information. These structured EEG representations are aligned with multi-level visual features to learn hierarchy-consistent EEG–visual correspondences. Building on the hierarchical representations, AHDA adaptively aligns cognitive levels to mitigate level-specific inter-subject distribution shifts, while employing retrieval-aware gradient projection to preserve discriminative information, thereby improving generalization to unseen subjects.

Our main contributions are summarized as follows:

- We propose **CogEEG**, which moves beyond global uniform alignment by modeling cross-subject EEG–visual retrieval with cognitive-level-aware representations and alignment.
- We develop BHCP, which combines brain-region priors with hyperbolic hierarchical modeling and cross-level interaction to establish EEG–visual correspondences from low-level perception to high-level semantics.
- We introduce AHDA to adapt alignment strength across cognitive levels using visual reliability and subject discriminability, together with retrieval-aware gradient projection to preserve retrieval-relevant semantics.

We extensively validate CogEEG on THINGS-EEG [12; 15] and THINGS-MEG [18] under both intra-subject and inter-subject settings. The results show that CogEEG improves EEG-Visual retrieval across subjects and neural modalities. Specifically, CogEEG achieves 38.5% and 14.25% Top-1 accuracy on THINGS-EEG under intra- and inter-subject settings, outperforming ATM by 10.35% and 4.35%, respectively. On THINGS-MEG, it further improves Top-1 by 3.88% and 1.12%. These results demonstrate that hierarchical cognitive modeling and adaptive cross-subject alignment consistently improve cross-subject generalization.

2 Related Work

Visual Brain Decoding. Visual brain decoding aims to recover perceived content from neural activity and supports brain–computer interfaces [24]. Early EEG-based methods incorporated external visual semantics to improve neural representations. BraVL [6] jointly modeled brain, visual, and linguistic signals, while NICE [41] aligned EEG and image representations through contrastive learning. With the emergence of vision–language models, ATM [28] mapped EEG into the CLIP [38] space

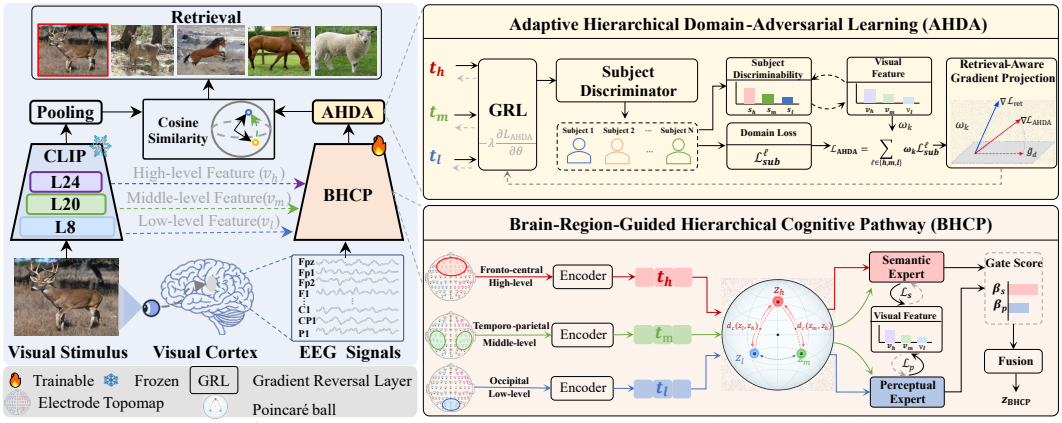


Figure 2: Overview of the CogEEG framework. First, BHCP leverages brain-region priors to learn low-, middle-, and high-level EEG features, and captures their hierarchical cognitive structure through cross-level interaction and expert fusion in hyperbolic space. Subsequently, AHDA adaptively suppresses subject-specific information based on subject discriminability and visual reliability, while retrieval-aware gradient projection preserves task-discriminative semantics. Finally, subject-invariant EEG features are aligned with hierarchical CLIP features for cross-subject retrieval.

for retrieval and reconstruction, and Cog-cap [48] further introduced multimodal experts for richer semantic supervision. These methods mainly focus on improving EEG–visual semantic alignment, but treat EEG as a single global representation. In contrast, CogEEG models EEG–visual correspondence across multiple cognitive levels and explicitly considers cross-subject variation within this hierarchy.

Hyperbolic Representation Learning Hyperbolic space provides a compact geometry for representing hierarchical structures due to its negative curvature and exponential volume growth. It has been used to model concept hierarchies [33; 8] and has recently been extended to visual and vision–language representation learning. Representative studies capture hierarchical visual concepts [26], general-to-specific image–text relations [4], compositional hierarchies [35], part-to-whole structures [25], and local-to-global visual relations [45]. Different from these works, we apply hyperbolic geometry to EEG decoding, where representations from different cognitive levels are organized and interacted within a hierarchical space.

Cross-subject Brain Decoding Cross-subject variability induces substantial distribution shifts in EEG, challenging generalizable brain decoding. Existing methods mainly follow two directions: *subject-invariant representation learning* [9; 34; 49] and *new-subject adaptation and generalization* [19; 29; 32]. The former reduces subject-specific variation through domain adaptation or representation disentanglement, while the latter improves transfer to unseen subjects through adaptation or meta-learning. Most of these methods perform alignment on a global EEG representation and therefore do not distinguish subject variation across different cognitive levels. CogEEG instead performs level-adaptive cross-subject alignment, allowing the alignment strength to vary with the visual reliability and subject discriminability of each level.

3 METHOD

Problem Formulation. Given a training set of EEG–image pairs $\mathcal{D}_{\text{tr}} = \{(\mathbf{x}_i^s, \mathbf{v}_i) \mid s \in \mathcal{S}_{\text{tr}}\}$, where $\mathbf{x}_i^s$ denotes the i-th EEG trial from source subject s and $\mathbf{v}_i$ is the corresponding image. Our goal is to learn an EEG–visual retrieval model that generalizes to unseen subjects $\mathcal{S}_{\text{te}}$, with $\mathcal{S}_{\text{tr}} \cap \mathcal{S}_{\text{te}} = \varnothing$. A learnable EEG encoder $f_\theta(\cdot)$ and a frozen visual encoder $g(\cdot)$ map EEG trials and images into a shared embedding space. During training, we optimize a symmetric contrastive objective $\mathcal{L}_{\text{ret}}$ to align paired EEG and visual representations. At inference, given an EEG trial $\mathbf{x}^s$ from an unseen subject $s \in \mathcal{S}_{\text{te}}$ and a candidate image set $\mathcal{V}_{\text{te}} = \{\mathbf{v}_1, \ldots, \mathbf{v}_M\}$, retrieval is performed by

$$\hat{\mathbf{v}} = \arg\max_{\mathbf{v}_j \in \mathcal{V}_{\text{te}}} \texttt{sim}\left(f_\theta(\mathbf{x}^s), g(\mathbf{v}_j)\right), \tag{1}$$

where $\texttt{sim}(\cdot,\cdot)$ denotes cosine similarity. The key challenge is to preserve hierarchical EEG–visual correspondences while handling inter-subject distribution shifts that vary across cognitive levels.

Overview. Figure 2 illustrates the overall framework of **CogEEG**. Following the preliminaries (§3.1), we present its two core components, BHCP and AHDA. Specifically, BHCP constructs multi-level EEG representations from brain-region priors and models their hierarchical dependencies using level-specific hyperbolic geometry and cross-level interaction. These representations are further aligned with multi-level visual features to establish structured EEG–visual correspondences. (§3.2). Based on this hierarchical representations, AHDA adjusts domain-adversarial alignment across cognitive levels based on visual reliability and subject discriminability, enabling level-aware cross-subject alignment. This alignment reduces level-specific subject shifts, improving cross-subject generalization (§3.3).

3.1 PRELIMINARIES

Poincaré Hyperbolic Geometry. Hyperbolic space has constant negative curvature and exponential volume growth, making it well suited for representing hierarchical structures with low distortion [33; 8; 43; 22]. We adopt the n-dimensional Poincar'e ball with curvature $-c$ ($c > 0$), defined as $\mathbb{B}_c^n = \{\mathbf{z} \in \mathbb{R}^n : c\|\mathbf{z}\|_2^2 < 1\}$, which has radius $1/\sqrt{c}$. Given $\mathbf{x}, \mathbf{y} \in \mathbb{B}_c^n$, Möbius addition is defined:

$$\mathbf{x} \oplus_c \mathbf{y} = \frac{\left(1 + 2c\langle \mathbf{x}, \mathbf{y}\rangle + c\|\mathbf{y}\|_2^2\right)\mathbf{x} + \left(1 - c\|\mathbf{x}\|_2^2\right)\mathbf{y}}{1 + 2c\langle \mathbf{x}, \mathbf{y}\rangle + c^2\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2}. \tag{2}$$

The geodesic distance between $\mathbf{x}, \mathbf{y} \in \mathbb{B}_c^n$ is

$$d_c(\mathbf{x}, \mathbf{y}) = \frac{2}{\sqrt{c}} \operatorname{artanh}\left(\sqrt{c}\,\|(-\mathbf{x}) \oplus_c \mathbf{y}\|_2\right), \tag{3}$$

where $\oplus_c$ denotes Möbius addition. For a scalar γ, the Möbius scalar multiplication is defined as

$$\gamma \otimes_c \mathbf{y} = \frac{1}{\sqrt{c}} \tanh\left(\gamma \operatorname{artanh}\left(\sqrt{c}\|\mathbf{y}\|_2\right)\right) \frac{\mathbf{y}}{\|\mathbf{y}\|_2}. \tag{4}$$

This operation controls the radial position of a representation in the Poincaré ball, providing a natural mechanism for modeling hierarchical depth. We use exponential and logarithmic maps at the origin to move between the Euclidean tangent space and the Poincar'e ball. For $\mathbf{v} \in T_0\mathbb{B}c^n$ and $\mathbf{x} \in \mathbb{B}c^n$,

$$\mathtt{Exp}_0^c(\mathbf{v}) = \tanh\left(\sqrt{c}\|\mathbf{v}\|_2\right) \frac{\mathbf{v}}{\sqrt{c}\|\mathbf{v}\|_2}, \ \mathtt{Log}_0^c(\mathbf{x}) = \operatorname{artanh}\left(\sqrt{c}\|\mathbf{x}\|_2\right) \frac{\mathbf{x}}{\sqrt{c}\|\mathbf{x}\|_2}. \tag{5}$$

The exponential map projects Euclidean features onto the hyperbolic manifold, while the logarithmic map returns hyperbolic representations to the tangent space.

3.2 BRAIN-REGION-GUIDED HIERARCHICAL COGNITIVE PATHWAY

Existing EEG–visual retrieval methods typically aggregate multi-channel EEG signals into a single global representation for cross-modal alignment [41; 28; 48]. However, visually evoked EEG exhibits region-dependent processing patterns: occipital responses are more sensitive to low-level visual features, temporo-parietal regions encode object and spatial information, and fronto-central regions are associated with higher-level cognitive processing [23]. Collapsing these signals into a single representation can obscure cognitive-level differences and their cross-level relations, making level-dependent subject variation difficult to characterize. To preserve this structure, we introduce **Brain-Region-Guided Hierarchical Cognitive Pathway (BHCP)**, which organizes regional EEG activity into a perceptual-to-semantic hierarchy (Fig. 2). We partition EEG channels into Occipital, Temporal–Parietal, and Frontal–Central groups, corresponding to low-, middle-, and high-level cognitive processing, respectively. Each group is encoded by an independent spatiotemporal graph convolutional network (STGCN) [46]:

$$\mathbf{f}_\ell = \mathtt{STGCN}_\ell(\mathbf{x}_\ell), \qquad \ell \in \{\text{low}(l), \text{middle}(m), \text{high}(h)\}, \tag{6}$$

where $\mathbf{x}_\ell$ denotes the regional EEG signal at level ℓ, and $\mathbf{f}_\ell$ is the corresponding spatiotemporal representation. Although regional STGCNs capture region-dependent EEG information, the resulting features $\mathbf{f}_l$, $\mathbf{f}_m$, and $\mathbf{f}_h$ remain independent Euclidean representations, leaving their perceptual-to-semantic hierarchy insufficiently modeled. Therefore, we linearly map each regional feature $\mathbf{f}_\ell$ into the tangent space and denote the resulting representation as $\mathbf{t}_\ell$. We then project $\mathbf{t}_\ell$ to a level-specific

Poincaré space with learnable curvature c_ℓ via $\mathbf{p}_\ell = \mathtt{Exp}_0^{c_\ell}(\mathbf{t}_\ell)$, allowing different cognitive levels to adapt their geometric structures.

Cross-Level Interaction. Level-specific Poincar'e spaces allow each cognitive level to retain its own geometric structure, but they also separate the representations across different manifolds. To enable cross-level interaction, we transform all levels into a common Poincar'e space whose curvature $\bar{c}$ is the mean of the level-specific curvatures. Specifically, each representation is transferred to the shared manifold through the tangent space at the origin: $\mathbf{z}_\ell = \mathtt{Exp}_0^{\bar{c}}\left(\mathtt{Log}_0^{c_\ell}(\mathbf{p}_\ell)\right)$. We then capture dependencies across cognitive levels using Poincar'e distance-based attention in the shared space. Each $\mathbf{z}_\ell$ is transformed into query ($\mathbf{q}_\ell$), key ($\mathbf{k}_\ell$), and value representations through hyperbolic linear transformations. The attention weight from level ℓ to level j is computed as

$$A_{\ell j} = \frac{\exp\left(-\bar{c}\, d_{\bar{c}}^2(\mathbf{q}_\ell, \mathbf{k}_j)\right)}{\sum_r \exp\left(-\bar{c}\, d_{\bar{c}}^2(\mathbf{q}_\ell, \mathbf{k}_r)\right)}, \tag{7}$$

where $d_{\bar{c}}(\cdot,\cdot)$ denotes the Poincaré distance. The resulting attention weights are used to aggregate value representations across levels, producing the cross-level enhanced feature $\tilde{\mathbf{z}}_\ell$.

Perceptual–Semantic Expert Construction. After cross-level interaction, the enhanced representations $\tilde{\mathbf{z}}_\ell$ contain complementary information across cognitive levels. To preserve the distinction between perceptual and semantic cues during fusion, we introduce two hyperbolic experts: a perceptual expert over the low–middle levels and a semantic expert over the middle–high levels. The perceptual expert transforms the low- and middle-level representations into $\tilde{\mathbf{z}}_l^p$ and $\tilde{\mathbf{z}}_m^p$, while the semantic expert produces $\tilde{\mathbf{z}}_m^s$ and $\tilde{\mathbf{z}}_h^s$ from the middle- and high-level representations. Both experts share the middle level, which bridges low-level perception and high-level semantics. The expert-specific representations are then aggregated in hyperbolic space using weighted Fréchet means:

$$\mathbf{e}_p = \widetilde{\mathbf{z}}_l^p \oplus_{\bar{c}} \left[1/3 \otimes_{\bar{c}} \left((-\widetilde{\mathbf{z}}_l^p) \oplus_{\bar{c}} \widetilde{\mathbf{z}}_m^p\right)\right], \quad \mathbf{e}_s = \widetilde{\mathbf{z}}_m^s \oplus_{\bar{c}} \left[2/3 \otimes_{\bar{c}} \left((-\widetilde{\mathbf{z}}_m^s) \oplus_{\bar{c}} \widetilde{\mathbf{z}}_h^s\right)\right] \tag{8}$$

where $\mathbf{e}_p$ and $\mathbf{e}_s$ denote the perceptual and semantic expert representations, respectively.

Hierarchical Visual Specialization. Although the two experts operate on different cognitive levels, the shared middle-level representation and the common retrieval objective do not explicitly enforce perceptual–semantic specialization. We therefore introduce hierarchical visual supervision to differentiate their roles. The expert representations are first mapped from the Poincaré space to the tangent space as $\mathbf{u}_{r,i} = \mathtt{Log}_0^{\bar{c}}(\mathbf{e}_{r,i})$, where $r \in \{p, s\}$ denotes the perceptual or semantic expert. We then extract low-, middle-, and high-level visual representations from different depths of a frozen CLIP encoder [38]. For each visual level ℓ, let $\mathcal{V}_\ell = \{\mathbf{v}_{\ell,j}\}_{j=1}^N$ denote the representations of the N unique images in a mini-batch. Each expert is aligned with the corresponding visual levels through an instance-discrimination objective:

$$\mathcal{L}_{\text{ins}}^{r,\ell} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}(\mathbf{u}_{r,i}, \mathbf{v}_{\ell,y_i})/\tau_v\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(\mathbf{u}_{r,i}, \mathbf{v}_{\ell,j})/\tau_v\right)}, \tag{9}$$

where B is the number of EEG samples, N is the number of unique images in the mini-batch, $y_i \in \{1, \ldots, N\}$ indexes the image paired with the i-th EEG sample, and τ_v is the temperature. The perceptual expert is supervised by low- and middle-level visual features, while the semantic expert is supervised by middle- and high-level features. This shared middle-level supervision maintains continuity between the two experts while preserving their distinct perceptual and semantic roles. The overall specialization objective is

$$\mathcal{L}_{\text{spec}} = \frac{1}{2}\left(\mathcal{L}_p + \mathcal{L}_s\right), \quad \mathcal{L}_p = \mathcal{L}_{\text{ins}}^{p,l} + \mathcal{L}_{\text{ins}}^{p,m}, \quad \mathcal{L}_s = \mathcal{L}_{\text{ins}}^{s,m} + \mathcal{L}_{\text{ins}}^{s,h}. \tag{10}$$

Geometry-Aware Expert Fusion. To combine the perceptual and semantic experts adaptively, we introduce a geometry-aware gating mechanism based on their hyperbolic relations. Specifically, we encode the geodesic distance between the two experts and their relative radial positions into a geometric relation feature α, which is then used to predict the fusion weights:

$$[\beta_p, \beta_s] = \mathtt{softmax}\left(\mathtt{Linear}(\alpha)\right), \quad \alpha = \left[\sqrt{\bar{c}}\, d_{\bar{c}}(\mathbf{e}_p, \mathbf{e}_s),\ \sqrt{\bar{c}}\left(d_{\bar{c}}(\mathbf{e}_s, \mathbf{0}) - d_{\bar{c}}(\mathbf{e}_p, \mathbf{0})\right)\right], \tag{11}$$

where $\mathtt{Linear}(\cdot)$ denotes a linear layer, β_p and β_s denote the contributions of the perceptual and semantic experts, respectively. Since $\beta_p + \beta_s = 1$, we fuse the two expert representations by geodesic interpolation, parameterized by β_s, and then map the result back to the tangent space:

$$\mathbf{z}_{\text{BHCP}} = \mathtt{Log}_0^{\bar{c}}\left(\mathbf{e}_p \oplus_{\bar{c}} \left[\beta_s \otimes_{\bar{c}} \left((-\mathbf{e}_p) \oplus_{\bar{c}} \mathbf{e}_s\right)\right]\right). \tag{12}$$

The resulting $\mathbf{z}_{\text{BHCP}}$ is used as the final EEG features for retrieval against frozen CLIP image features.

3.3 Adaptive Hierarchical Domain-Adversarial Learning

Although BHCP establishes hierarchical EEG–visual representations, subject-specific variation may persist at different cognitive levels, limiting generalization to unseen subjects. Importantly, these levels can differ both in subject dependence and in the reliability of their visual semantics. Applying the same adversarial constraint to all levels may therefore either under-suppress subject variation or distort retrieval-relevant information. To address this issue, we introduce **Adaptive Hierarchical Domain-Adversarial Learning (AHDA)**, which performs subject-adversarial alignment at each cognitive level and adapts the alignment strength according to visual reliability and subject discriminability (Fig. 2). We first estimate subject discriminability at each level using a shared subject classifier. Each representation is passed through a gradient reversal layer (GRL) and a shared feature projection, while a learnable level embedding captures level-specific characteristics:

$$\boldsymbol{\pi}_i^\ell = \texttt{softmax}\left(\texttt{D}\left(\left[\phi\left(\texttt{GRL}(\mathbf{t}_i^\ell)\right); \boldsymbol{\eta}_\ell\right]\right)\right), \tag{13}$$

where $\phi(\cdot)$ denotes the shared feature projection, $\boldsymbol{\eta}_\ell$ is the learnable embedding of cognitive level ℓ, and $\texttt{D}(\cdot)$ is the shared subject classifier. The predicted distribution $\boldsymbol{\pi}_i^\ell \in \mathbb{R}^S$ is defined over the S source subjects. The subject-classification loss at level ℓ is

$$\mathcal{L}_{\text{sub}}^\ell = -\frac{1}{B}\sum_{i=1}^{B} \log \boldsymbol{\pi}_{i,s_i}^\ell, \tag{14}$$

where s_i denotes the subject identity of the i-th EEG sample. Through the GRL, the classifier is optimized to distinguish source subjects, while the EEG encoder is encouraged to suppress subject-specific information from each cognitive level.

Reliability–Discriminability Estimation. The subject predictions above measure residual subject discriminability, but they do not indicate whether a cognitive level still preserves reliable visual semantics. We therefore introduce visual reliability as a complementary signal. Specifically, we measure the similarity margin between the matched visual reference and the hardest mismatched one: $\Delta_\ell = B^{-1}\sum_{i=1}^{B}\left[\rho_{i,y_i}^\ell - \max_{j\neq y_i}\rho_{ij}^\ell\right]$, $R_\ell = \sigma\left(\Delta_\ell/\tau_r\right)$. Here, ρ_{ij}^ℓ denotes the EEG–visual similarity, $\sigma(\cdot)$ denotes the sigmoid function, and a larger R_ℓ indicates more reliable visual semantics at level ℓ. We further quantify residual subject dependence using the complement of the normalized entropy of the subject predictions: $U_\ell = 1 + \frac{1}{B\log S}\sum_{i=1}^{B}\sum_{s=1}^{S}\boldsymbol{\pi}_{i,s}^\ell \log \boldsymbol{\pi}_{i,s}^\ell$, where larger U_ℓ indicates stronger residual subject dependence and thus a greater need for alignment.

Adaptive Level-Wise Alignment. To reduce minibatch-level fluctuations, we maintain exponential moving averages $\bar{R}_\ell$ and $\bar{U}_\ell$ for each cognitive level. These estimates are then used to adaptively weight the level-wise subject-classification losses:

$$\mathcal{L}_{\text{AHDA}} = \sum_{\ell\in\mathcal{K}} w_\ell \mathcal{L}_{\text{sub}}^\ell, \quad w_\ell = \frac{\exp\left(\bar{R}_\ell\bar{U}_\ell/\tau_w\right)}{\sum_{r\in\mathcal{K}}\exp\left(\bar{R}_r\bar{U}_r/\tau_w\right)}, \tag{15}$$

where $\mathcal{L}_{\text{AHDA}}$ denotes the overall hierarchical domain-adversarial objective. A larger w_ℓ is assigned to levels that preserve reliable visual semantics while remaining strongly subject-discriminative, thereby concentrating cross-subject alignment where it is most needed.

Retrieval-Preserving Gradient Projection. Adaptive hierarchical weighting controls the alignment strength at each cognitive level, but the adversarial update can still conflict with the retrieval objective. To prevent cross-subject alignment from degrading retrieval-relevant information, we regulate the adversarial gradient according to its agreement with the retrieval gradient. We define the retrieval and adversarial gradients as $\mathbf{g}_r = \nabla_{\mathbf{t}_\ell}\mathcal{L}_{\text{ret}}$ and $\mathbf{g}_d = \nabla_{\mathbf{t}_\ell}(\mathcal{L}_{\text{AHDA}})$, respectively. When $\langle \mathbf{g}_d, \mathbf{g}_r\rangle < 0$, the conflicting component of the adversarial gradient is removed:

$$\widetilde{\mathbf{g}}_d = \mathbf{g}_d - (\min(0, \langle \mathbf{g}_d, \mathbf{g}_r\rangle)/\|\mathbf{g}_r\|_2^2)\mathbf{g}_r. \tag{16}$$

The projected gradient satisfies $\langle \tilde{g}_d, g_r\rangle \geq 0$, eliminating first-order conflicts with the retrieval objective (see Appendix C.1 for proof). Thus, AHDA reduces subject-specific variation while preserving retrieval-relevant information.

Training Objective. Finally, CogEEG is jointly optimized by

$$\mathcal{L} = \mathcal{L}_{\text{ret}} + \lambda_{\text{spec}}\mathcal{L}_{\text{spec}} + \lambda_d\mathcal{L}_{\text{AHDA}}, \tag{17}$$

where λ_{spec} and λ_d balance the expert-specialization and domain-adversarial objectives, respectively.

Table 1: Top-1 (T-1) and Top-5 (T-5) accuracy (%) for 200-way zero-shot brain-to-image retrieval on THINGS-EEG under intra-subject and inter-subject settings. **Bold** and underline denote the best and second-best results, respectively. * denotes reproduced results.

Method	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
Intra-subject: train and test on one subject																						
BraVL [6]	6.1	17.9	4.9	14.9	5.6	17.4	5.0	15.1	4.0	13.4	6.0	18.2	6.5	20.4	8.8	23.7	4.3	14.0	7.0	19.7	5.8	17.5
NICE [41]	13.2	39.5	13.5	40.3	14.5	42.7	20.6	52.7	10.1	31.5	16.5	44.0	17.0	42.1	22.9	56.1	15.4	41.6	17.4	45.8	16.1	43.6
EEGNet* [27]	24.5	59.0	21.5	55.0	25.5	57.5	19.5	48.0	20.5	53.0	27.0	67.5	21.5	52.5	29.0	66.5	22.0	54.0	23.0	56.0	23.40	56.90
ATM* [28]	**34.0**	**65.5**	25.5	**65.0**	23.5	59.5	25.0	56.5	23.0	52.0	31.0	67.5	26.0	57.5	36.0	69.0	27.0	58.0	30.5	62.0	28.15	61.25
CogEEG (Ours)	**34.0**	63.5	**35.0**	**65.0**	**41.5**	**73.0**	**41.0**	**72.0**	**26.0**	**60.5**	**37.5**	**72.0**	**38.0**	**68.5**	**49.0**	**80.0**	**39.5**	**66.0**	**43.5**	**77.0**	**38.50**	**69.75**
Inter-subject: leave one subject out for test																						
BraVL [6]	2.3	8.0	1.5	6.3	1.9	6.7	2.1	8.1	2.2	7.6	1.6	6.4	2.3	8.5	1.8	7.0	1.4	5.9	1.7	6.7	1.9	7.1
NICE [41]	7.6	22.8	5.9	20.5	6.0	22.3	6.3	20.7	4.4	18.3	5.6	22.2	5.6	19.7	6.3	22.0	5.7	17.6	8.4	28.3	6.2	21.4
NICE-G [41]	5.9	21.4	6.4	22.7	5.5	20.1	6.1	21.0	4.7	19.5	6.2	22.5	5.9	19.1	7.3	25.3	6.2	18.3	6.2	26.3	6.0	21.60
EEGNet* [27]	15.0	30.5	8.0	22.5	5.0	14.5	7.0	16.0	3.5	17.0	10.5	25.0	6.0	22.5	6.5	18.5	6.5	13.0	11.0	34.0	7.90	21.35
ATM* [28]	12.0	29.5	9.0	25.5	7.5	16.5	9.0	21.5	10.0	23.5	12.0	31.0	8.5	**24.0**	11.0	29.5	6.5	19.0	13.5	36.0	9.90	25.60
CogEEG (Ours)	**18.0**	**41.5**	**19.5**	**33.5**	**10.5**	**28.5**	**14.5**	**36.5**	**10.5**	**30.0**	**13.5**	**37.5**	**11.0**	22.0	**13.0**	**35.0**	**13.5**	**31.0**	**18.5**	**42.5**	**14.25**	**33.80**

4 EXPERIMENTS

Datasets and Evaluation Metrics. The THINGS-EEG dataset [12; 15] is a large-scale RSVP benchmark with ten participants, 1,654 training concepts, 200 unseen test concepts, and 82,160 trials per subject. In addition, we evaluate CogEEG on the THINGS-MEG dataset [18]. The evaluation metrics are Top-1 and Top-5 accuracy. More details in Appendix A and B.

Implementation Details. *i) Network Architecture.* CogEEG combines a global ATMS encoder with BHCP, which extracts three-level EEG features using STGCNs, models them in learnable-curvature Poincaré spaces, and performs hyperbolic cross-level fusion. The resulting 1024-D feature is fused with the global representation, while AHDA conducts level-wise subject-adversarial alignment. *ii) Training and Testing.* CogEEG is trained for 40 epochs with a batch size of 512 using AdamW (10^{-4} learning rate, cosine decay to 10^{-5}, and 0.01 weight decay). The expert-specialization and domain-adversarial weights are both 0.01. At inference, EEG features are matched against 200 frozen image embeddings for Top-1/Top-5 retrieval. *iii) Reproducibility.* CogEEG is implemented in PyTorch and evaluated on five NVIDIA RTX 3090 GPUs. More details in Appendix B.

4.1 COMPARE WITH THE STATE-OF-THE-ART

Results on Things-EEG Datasets. Table 1 reports 200-way zero-shot EEG–image retrieval results under both intra-subject and inter-subject settings. We compare CogEEG with representative methods, including BraVL [6], NICE [41], EEGNet [27], and ATM [28]. In the intra-subject setting, each subject is trained and evaluated independently, whereas the inter-subject setting follows a leave-one-subject-out protocol to evaluate generalization to unseen subjects. Under the intra-subject setting, CogEEG *vs.* ATM achieves **+10.35**% Top-1. More importantly, under the challenging inter-subject setting, CogEEG *vs.* ATM achieves **+4.35**% Top-1. The stronger Top-1 gains indicate improved discrimination of the correct visual target. Overall, CogEEG benefits EEG–visual alignment by modeling hierarchical cognitive structure and improves cross-subject generalization by suppressing subject-specific variations while preserving retrieval-relevant semantics. More details in Appendix E.

Results on THINGS-MEG. Table 2 further illustrates the effectiveness of CogEEG on MEG signals. In terms of Top-1 accuracy, CogEEG achieves 24.50% under the intra-subject setting, exceeding ATM by 3.88%. In particular, the improvement reaches 10.5% on Subject 2 (43.0% *vs.* 32.5%). Under the more challenging inter-subject setting, CogEEG obtains 5.00% Top-1 accuracy, compared with 3.88% for ATM. These gains are notable considering that THINGS-MEG contains dense high-dimensional neural recordings but only four subjects. The results suggest that BHCP effectively organizes complex neural responses through hierarchical representation learning, while AHDA further reduces subject-dependent shifts when training subjects are limited. Thus, the advantages of CogEEG are not restricted to EEG signals but extend consistently to MEG-based visual retrieval.

Table 2: Top-1 (T-1) and Top-5 (T-5) accuracy (%) for brain-to-image retrieval on THINGS-MEG.

Method	S1		S2		S3		S4		Avg.	
	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
Intra-subject										
EEGNet* [27]	7.5	20.0	12.5	41.5	12.0	28.5	2.0	9.5	8.50	24.88
NICE [41]	8.7	30.5	21.8	56.6	16.5	49.7	10.3	32.3	14.30	42.30
ATM* [28]	10.0	29.5	32.5	68.5	25.5	56.5	14.5	32.0	20.62	46.62
CogEEG	**13.5**	**35.0**	**43.0**	**79.0**	**25.5**	**58.0**	**16.0**	**39.0**	**24.50**	**52.75**
Inter-subject										
EEGNet* [27]	4.0	10.0	3.0	10.0	3.5	11.5	3.0	4.5	3.38	9.00
ATM* [28]	3.0	7.5	3.0	10.5	6.0	10.0	3.5	9.5	3.88	9.38
CogEEG	**3.5**	**11.0**	**5.0**	**14.0**	**7.0**	**14.0**	**4.5**	**9.5**	**5.00**	**12.13**

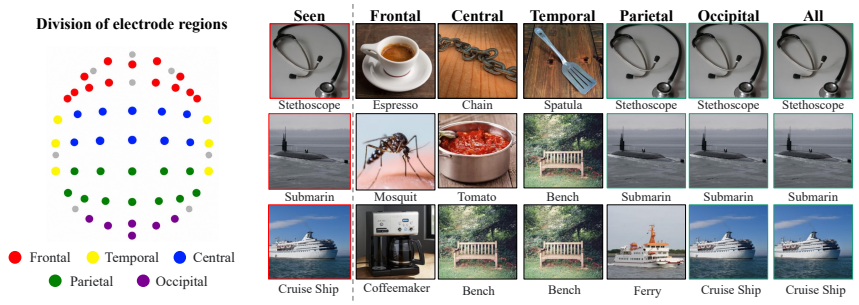


Figure 3: Regional EEG retrieval analysis on THINGS-EEG.

Table 3: Regional EEG retrieval analysis on THINGS-EEG under the intra-subject setting.

Method	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
Frontal	1.5	3.0	1.5	4.5	3.0	7.0	3.0	8.0	1.5	5.5	2.5	4.5	2.0	4.0	2.0	7.5	1.5	4.5	1.5	5.5	2.00	5.40
Temporal	8.0	19.0	12.0	24.0	6.5	22.0	13.0	26.0	5.0	14.5	8.0	23.5	6.5	19.5	9.5	27.0	4.0	15.5	11.5	29.0	8.40	22.00
Central	2.0	7.0	2.5	5.0	2.0	6.5	3.0	7.5	3.0	6.0	4.5	11.0	3.5	6.0	5.0	19.5	3.5	7.0	2.0	7.5	3.10	8.30
Parietal	11.0	26.5	13.5	31.5	15.0	40.5	20.0	46.5	15.5	35.0	19.0	39.5	13.5	32.5	31.5	54.0	15.0	38.5	20.5	50.0	17.45	39.45
Occipital	32.5	**63.5**	20.5	51.0	33.5	64.5	37.0	64.5	20.0	51.0	32.0	65.0	32.0	63.5	35.5	71.0	32.0	60.0	43.0	71.0	31.80	62.50
All	**34.0**	**63.5**	**35.0**	**65.0**	**41.5**	**73.0**	**41.0**	**72.0**	**26.0**	**60.5**	**37.5**	**72.0**	**38.0**	**68.5**	**49.0**	**80.0**	**39.5**	**66.0**	**43.5**	**77.0**	**38.50**	**69.75**

4.2 Motivation Analysis

Regional Contribution Analysis. We examine the contribution of different EEG regions to visual retrieval under the intra-subject setting. In Table 3, the occipital region achieves the best single-region performance, exceeding the parietal, temporal, central, and frontal regions by 14.35%, 23.40%, 28.70%, and 29.80% in Top-1 accuracy, respectively. The parietal region also consistently outperforms the remaining regions, showing that visual information is unevenly distributed across EEG regions. Notably, using all channels improves Top-1 accuracy by **6.70**% over the occipital-only setting, indicating that non-occipital regions provide complementary information. The retrieval examples in Fig. 3 show a similar pattern: occipital and parietal signals recover more relevant visual content, while full-channel EEG gives the most consistent results. These observations support our use of region-specific modeling in CogEEG, rather than encoding all channels as a homogeneous representation.

Hierarchical EEG–Visual Correspondence. Fig. 4 examines how visual features at different semantic levels correspond to EEG activity. We use frozen low-, middle-, and high-level visual features to predict EEG responses from predefined channel groups and temporal windows, and measure their contributions with Shapley values. The resulting patterns differ across EEG levels: low- and middle-level EEG representations show clear preference for their matched visual features, whereas the high-level correspondence is less pronounced. This level-specific structure supports modeling EEG–visual relations hierarchically, as done in BHCP, rather than collapsing all information into a single global embedding.

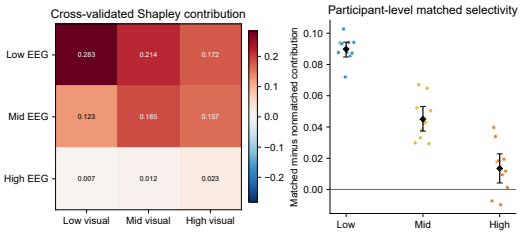


Figure 4: Motivation analysis of hierarchical EEG–visual correspondence via Shapley contribution and matched-level selectivity.

Cross-subject Alignment Varies Across Cognitive Levels. Table 4 examines domain-adversarial alignment at different cognitive levels. Aligning the low level improves Top-1 accuracy by **4**%, whereas further aligning the middle and high levels yields smaller and non-monotonic gains. This shows that cross-subject alignment does not benefit all cognitive levels equally, supporting the need for level-adaptive alignment in AHDA.

4.3 Ablation Study

Effectiveness of BHCP and AHDA. We ablate the two main components of CogEEG in Table 5. Removing both BHCP and AHDA recovers the ATM baseline. Adding BHCP alone improves Top-1 accuracy by **4.10**%, showing the benefit of organizing EEG features across cognitive levels and modeling their cross-level dependencies. Replacing BHCP with AHDA yields a comparable gain of **4.05**%, indicating that level-adaptive subject alignment is also effective when applied to the baseline representation. Combining both components gives the best result, outperforming ATM by **4.35**%

Table 4: Analysis of inter-subject variability across cognitive levels on THINGS-EEG.

f_l'	f_m'	f_h'	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
			T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
Baseline			12.0	29.5	9.0	25.5	7.5	16.5	9.0	21.5	10.0	23.5	12.0	31.0	8.5	24.0	11.0	29.5	6.5	19.0	13.5	36.0	9.90	25.60
✓			16.5	40.5	17.5	34.0	8.5	21.5	14.0	**38.0**	10.5	31.5	11.0	36.5	**11.5**	**33.5**	**14.5**	**36.5**	**13.5**	**32.5**	**21.5**	**44.5**	13.90	34.90
✓	✓		17.5	**42.0**	18.0	**39.0**	10.0	26.5	**14.5**	36.5	**11.0**	**32.0**	**13.5**	34.5	11.0	31.5	13.0	**36.5**	**13.5**	**32.5**	18.5	43.0	14.05	**35.40**
✓	✓	✓	**18.0**	41.5	**19.5**	33.5	**10.5**	**28.5**	**14.5**	36.5	10.5	30.0	**13.5**	**37.5**	11.0	22.0	13.0	35.0	**13.5**	31.0	18.5	42.5	**14.25**	33.80

Table 5: Ablation study of BHCP and AHDA on THINGS-EEG under the inter-subject setting.

BHCP	AHDA	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
		T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
✗	✗	12.0	29.5	9.0	25.5	7.5	16.5	9.0	21.5	10.0	23.5	12.0	31.0	8.5	24.0	11.0	29.5	6.5	19.0	13.5	36.0	9.90	25.60
✓	✗	13.5	37.5	**20.0**	**38.0**	8.0	20.0	16.0	**38.0**	**11.5**	28.0	12.5	35.5	**11.0**	31.5	**13.0**	**35.5**	12.5	30.0	**22.0**	46.5	14.00	**34.05**
✗	✓	15.0	36.0	19.0	36.5	9.0	20.5	**16.5**	36.5	9.5	29.0	13.0	36.0	9.0	**32.0**	**13.0**	34.5	**14.5**	**32.0**	21.0	**47.0**	13.95	34.00
✓	✓	**18.0**	**41.5**	19.5	33.5	**10.5**	**28.5**	14.5	36.5	10.5	**30.0**	**13.5**	**37.5**	**11.0**	22.0	**13.0**	35.0	13.5	31.0	18.5	42.5	**14.25**	33.80

Table 6: Ablation study of BHCP fusion strategies on THINGS-EEG under the intra-subject setting.

Method	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
Baseline	25.6	60.4	22.0	54.5	25.0	62.4	31.4	60.9	12.9	43.0	21.3	51.1	30.5	61.5	38.8	72.0	30.4	51.5	29.1	63.5	26.7	58.1
Sum	33.5	61.5	33.5	62.0	41.0	72.5	40.0	**73.5**	**27.5**	**61.0**	34.5	68.5	37.0	65.5	48.5	**81.0**	36.5	64.0	**43.5**	74.5	37.55	68.40
Concat	33.0	62.0	33.5	64.5	40.0	70.0	**42.0**	72.0	22.5	58.0	**38.0**	**73.5**	**41.0**	68.0	44.5	75.5	38.0	64.0	**43.5**	**77.0**	37.60	68.45
Ours(Eq. 7)	**34.0**	**63.5**	**35.0**	**65.0**	**41.5**	**73.0**	41.0	72.0	26.0	60.5	37.5	72.0	38.0	**68.5**	**49.0**	80.0	**39.5**	**66.0**	**43.5**	**77.0**	**38.50**	**69.75**

Table 7: Ablation study of AHDA alignment strategies on THINGS-EEG under the inter-subject.

Method	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
Baseline	12.0	29.5	9.0	25.5	7.5	16.5	9.0	21.5	10.0	23.5	12.0	31.0	8.5	24.0	11.0	29.5	6.5	19.0	13.5	36.0	9.90	25.60
Uniform	16.5	40.0	18.5	**36.0**	10.0	24.5	14.0	**40.5**	**10.5**	**32.0**	12.5	36.0	**11.5**	**35.0**	**13.5**	**36.5**	**14.0**	30.0	**18.5**	**43.5**	13.95	**35.40**
Ours(Eq. 15)	**18.0**	**41.5**	**19.5**	33.5	**10.5**	**28.5**	**14.5**	36.5	**10.5**	30.0	**13.5**	**37.5**	11.0	22.0	13.0	35.0	13.5	**31.0**	**18.5**	42.5	**14.25**	33.80

in Top-1 accuracy. Overall, BHCP improves the structure of EEG–visual representations, while AHDA reduces level-dependent subject variation; their combination provides complementary gains for cross-subject retrieval.

Effectiveness of the Fusion Strategy. We compare the proposed hyperbolic cross-level interaction with two simple fusion strategies, i.e., summation (*Sum*) and direct concatenation (*Concat*), while keeping the remaining settings unchanged. As shown in Table 6, our fusion strategy improves Top-1 accuracy by **0.95**% over Sum and by **0.90**% over Concat. These results demonstrate that explicitly modeling cross-level dependencies provides more effective integration of hierarchical EEG representations than simple feature aggregation.

Effectiveness of Adaptive Hierarchical Alignment. Table 7 compares Baseline, uniform level-wise average alignment (*Uniform*), and our adaptive alignment strategy. Uniform alignment improves the average Top-1 accuracy by **4.05**% over the baseline, confirming the benefit of reducing subject-specific variations across cognitive levels. Our adaptive strategy further improves Top-1 accuracy by **0.30**% over Uniform and by **4.35**% points over the baseline. These results suggest that different cognitive levels benefit from different alignment strengths, and adapting them according to visual reliability and residual subject discriminability enables more targeted cross-subject alignment.

Effect of Representation Geometry. We compare Euclidean and hyperbolic representations under the same EEG–visual retrieval setting. In Table 8, hyperbolic geometry improves Top-1 accuracy by **0.55**%, while achieving comparable Top-5 performance. The gain, though moderate, indicates that modeling EEG representations in hyperbolic space is beneficial for fine-grained retrieval and is consistent with the hierarchical design of BHCP.

Table 8: Retrieval Results on THINGS-EEG.

Method	T-1 (%)	T-5 (%)
Euclidean	37.95	**70.05**
Hyperbolic	**38.50**	69.75

5 Conclusion

We present **CogEEG**, a hierarchical hyperbolic framework for cross-subject EEG–visual retrieval. Instead of aligning a single global EEG representation across subjects, CogEEG models EEG–visual correspondence at multiple cognitive levels and performs level-adaptive cross-subject alignment. BHCP combines functional brain-region priors with hyperbolic cross-level modeling to capture hierarchical visual representations, while AHDA adjusts subject alignment according to visual reliability and subject discriminability, with gradient protection to preserve retrieval-relevant semantics. Extensive experiments show that these designs consistently improve cross-subject retrieval and enable CogEEG to achieve state-of-the-art performance.

AI USE STATEMENT

We used generative AI tools for language proofreading and improving the clarity of manuscript writing. We did not use generative AI tools for generating experimental results, conducting data analysis, or making scientific claims. We also used generative AI tools for brainstorming research directions and checking grammatical errors. All AI-assisted content was carefully reviewed and verified by the authors. We take full responsibility for the final content of this manuscript.

REFERENCES

[1] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019. 2

[2] Erika W Contini, Susan G Wardle, and Thomas A Carlson. Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *neuropsychologia*, 105:165–176, 2017. 2

[3] Sicheng Dai, Hongwang Xiao, Shan Yu, and Qiwei Ye. Autoregressive visual decoding from eeg signals. *arXiv preprint arXiv:2602.22555*, 2026. 1

[4] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731. PMLR, 2023. 2, 3

[5] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. 2

[6] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023. 1, 2, 7, 17

[7] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991. 2

[8] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018. 3, 4

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 3

[10] Jianxiong Gao, Yichang Liu, Baofeng Yang, Jianfeng Feng, and Yanwei Fu. Cinebrain: A large-scale multi-modal audiovisual brain dataset for brain-conditioned video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36224–36234, 2026. 1

[11] Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254:119121, 2022. 1

[12] Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *Journal of Vision*, 23(9):4579–4579, 2023. 2, 7, 14

[13] Siddhant Gole, Akash Pal, Amit More, S Divakar Bhat, Subhasis Chaudhuri, and Biplab Banerjee. Hyperbolic prototype learning with uncertainty-aware consistency for continual test-time segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34694–34703, 2026. 2

[14] Kalanit Grill-Spector and Rafael Malach. The human visual cortex. *Annu. Rev. Neurosci.*, 27 (1):649–677, 2004. 2

[15] Tijl Grootswagers, Ivy Zhou, Amanda K Robinson, Martin N Hebart, and Thomas A Carlson. Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1):3, 2022. 2, 7, 14

[16] Matthias Guggenmos, Philipp Sterzer, and Radoslaw Martin Cichy. Multivariate pattern analysis for meg: A comparison of dissimilarity measures. *Neuroimage*, 173:434–447, 2018. 14

[17] Zhanqiang Guo, Jiamin Wu, Yonghao Song, Jiahui Bu, Weijian Mai, Qihao Zheng, Wanli Ouyang, and Chunfeng Song. Neuro-3d: Towards 3d visual decoding from eeg signals. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23870–23880. IEEE, 2025. 1

[18] Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023. 2, 7, 14

[19] Shuai Huang, Huan Luo, Haodong Jing, Qixian Zhang, Litao Chang, Yating Feng, Xiao Lin, Chendong Qin, Han Chen, Shuwen Jia, et al. Need: Cross-subject and cross-task generalization for video and image reconstruction from eeg signals. *Advances in Neural Information Processing Systems*, 38:173134–173173, 2026. 1, 3

[20] Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. Eeg-tcnet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. In *2020 IEEE international conference on systems, man, and cybernetics (SMC)*, pp. 2958–2965. IEEE, 2020. 1

[21] Ji-Hoon Jeong, Jeong-Hyun Cho, Byeong-Hoo Lee, and Seong-Whan Lee. Real-time deep neurolinguistic learning enhances noninvasive neural language decoding for brain–machine interaction. *IEEE Transactions on Cybernetics*, 53(12):7469–7482, 2022. 1

[22] Sangmin Jo, Wootaek Jeong, Da-Woon Heo, Yoohwan Hwang, and Heung-Il Suk. Hyfi: Hyperbolic feature interpolation for brain-vision alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5575–5583, 2026. 1, 4

[23] Hamid Karimi-Rouzbahani. Three-stage processing of category and variation information by entangled interactive mechanisms of peri-occipital and peri-frontal cortices. *Scientific reports*, 8(1):12213, 2018. 4

[24] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. 2

[25] Hayeon Kim, Ji Ha Jang, Junghun James Kim, and Se Young Chun. Uncertainty-guided compositional alignment with part-to-whole semantic representativeness in hyperbolic vision-language models. *arXiv preprint arXiv:2603.22042*, 2026. 3

[26] Hyeongjun Kwon, Jinhyun Jang, Jin Kim, Kwonyoung Kim, and Kwanghoon Sohn. Improving visual recognition with hyperbolical visual hierarchy mapping. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17364–17374. IEEE, 2024. 3

[27] 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. 7, 17

[28] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024. 2, 4, 7, 17

[29] Shumeng Li, Jintao Guo, Jian Zhang, Yulin Zhou, Luyang Cao, and Yinghuan Shi. Duala: Dual-level alignment of subjects and stimuli for cross-subject fmri decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42722–42731, 2026. 3

[30] Dongjun Liu, Weichen Dai, Jingsheng Qian, Honggang Liu, Hangjie Yi, and Wanzeng Kong. Linguistic priors for visual decoupling: Towards symmetric vision-brain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7869–7878, 2026. 1

[31] Minxu Liu, Donghai Guan, Chuhang Zheng, Chunwei Tian, Jie Wen, and Qi Zhu. Vieeg: Hierarchical visual neural representation for eeg brain decoding. *arXiv preprint arXiv:2505.12408*, 2025. 1

[32] Mu Nan, Muquan Yu, Weijian Mai, Jacob S Prince, Hossein Adeli, Rui Zhang, Jiahang Cao, Benjamin Becker, John A Pyles, Margaret M Henderson, et al. Meta-learning in-context enables training-free cross subject brain decoding. *arXiv preprint arXiv:2604.08537*, 2026. 3

[33] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 3, 4

[34] Ozan Özdenizci, Ye Wang, Toshiaki Koike-Akino, and Deniz Erdoğmuş. Learning invariant representations from eeg via adversarial inference. *IEEE access*, 8:27074–27085, 2020. 1, 2, 3

[35] Avik Pal, Max Van Spengler, Guido D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 87371–87399, 2025. 2, 3

[36] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, and Mubarak Shah. Generative adversarial networks conditioned by brain signals. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3430–3438. IEEE, 2017. 1

[37] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11): 3833–3849, 2020. 1

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021. 2, 5

[39] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017. 1

[40] Xiangbo Shu, Chengjian Li, Zheng Yin, et al. Hierarchical text semantics-driven multi-granularity human motion generation. *Acta Electronica Sinica*, 54(1):451–465, 2026. 2

[41] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. In *International conference on learning representations*, volume 2024, pp. 47648–47665, 2024. 2, 4, 7, 17

[42] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4503–4511. IEEE, 2017. 1

[43] Hoang M Truong, Hai Nguyen-Truong, and Dang Huynh. Semantic alignment in hyperbolic space for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2605.08874*, 2026. 4

[44] Yujie Wei, Chenglong Ma, Jianxiong Gao, Chenhui Wang, Shiwei Zhang, Biao Gong, Shuai Tan, Hangjie Yuan, and Hongming Shan. Bridging brain and semantics: A hierarchical framework for semantically enhanced fmri-to-video reconstruction. *arXiv preprint arXiv:2605.14569*, 2026. 1

[45] Suhan Woo, Seongwon Lee, Jinwoo Jang, and Euntai Kim. Hypevpr: Exploring hyperbolic space for perspective to equirectangular visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12183–12192, 2026. 3

[46] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 4

[47] Chenglong Yu, Shuai Shen, Xiangsheng Li, and Yang Li. D^2-FOSA: Dual-diffusion guided eeg-to-image reconstruction with frequency-oriented semantic alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26698–26710, 2026. 1

[48] Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14486–14493, 2025. 3, 4

[49] Li-Ming Zhao, Xu Yan, and Bao-Liang Lu. Plug-and-play domain adaptation for cross-subject eeg-based emotion recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 863–870, 2021. 1, 3

[50] Jiahao Zhou, Chenghao Xu, Wei Wang, Erkun Yang, and Cheng Deng. Eegit: Teaching vision transformers to understand the eeg signal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40438–40447, 2026. 1

Overview. The supplementary includes the following sections:

- Datasets for experiments (§A).
- More Implementation Details (§B).
- Additional Method Analysis(§C)
- Baseline Descriptions (§D).
- Qualitative Experiments(§E).
- Ablation Study(§F).

A DATASETS FOR EXPERIMENTS

A.1 THINGS-EEG DATASET.

We conduct experiments on the THINGS-EEG dataset [12; 15], which contains EEG responses from 10 participants collected under an RSVP paradigm. Each participant has 16,540 training images with four repetitions and 200 test images with 80 repetitions, yielding 82,160 trials in total. EEG signals were originally recorded with 64 channels (see Fig.5) at 1,000 Hz. We use 63 scalp EEG channels from the raw recordings, apply a 0.1–100 Hz band-pass filter, and downsample the signals to 250 Hz. Epochs are extracted from -200 to 1,000 ms relative to stimulus onset and baseline-corrected using the -200 to 0 ms interval, after which the 0–1,000 ms post-stimulus segment is retained. Multivariate noise normalization [16] is estimated from the training data only and applied separately for each participant. During training, individual EEG trials are used directly. At test time, the 80 repetitions of each image are grouped in sets of four and averaged, producing 20 averaged samples per image to improve the signal-to-noise ratio. The same evaluation protocol is used for all reproduced methods. For methods with public code, we report results reproduced under our preprocessing pipeline; otherwise, we report the results from the original papers.

A.2 THINGS-MEG DATASET.

To further assess the cross-modal generalizability of BHCP, we conduct additional experiments on the THINGS-MEG dataset [18], which contains 271-channel MEG recordings from four participants collected across 12 sessions. The original training split comprises 1,854 concepts with 12 images per concept and one presentation per image, while the test split contains 200 concepts with one image per concept and 12 repetitions. To construct a zero-shot setting consistent with THINGS-EEG, we remove the 200 test concepts from the training pool, ensuring no semantic overlap between training and test sets. Each stimulus was presented for 500 ms with an inter-stimulus interval of $1{,}000 \pm 200$ ms. MEG epochs are extracted from -100 to 1,300 ms relative to stimulus onset and baseline-corrected using the -100 to 0 ms pre-stimulus interval. The signals are band-pass filtered between 0.1 and 40 Hz and downsampled to 200 Hz, after which the 0–1,000 ms interval is retained for analysis, yielding 201 time points when both endpoints are included. Given the small cohort size ($n = 4$), we use THINGS-MEG primarily to assess cross-modal generalization and do not draw population-level statistical conclusions. For image retrieval, the exact presented stimulus image is used as the target rather than another exemplar from the same concept.

To adapt BHCP to MEG, we retain its three hierarchical local branches together with the global encoder and redefine the branch inputs according to the MEG sensor layout. Based on CTF channel-name prefixes, the 271 sensors are partitioned into three coarse spatial groups: posterior (MLO, MRO, MZO), temporo-parietal (MLP, MRP, MZP, MLT, MRT), and fronto-central (MLF, MRF, MZF, MLC, MRC, MZC). Sensor indices follow the order of channel names in the preprocessed data, and we verify that the resulting groups are mutually exclusive and collectively cover all 271 sensors. Each group is processed by an independent STGCN with its own learnable within-group adjacency matrix, followed by the corresponding Poincaré projection and cross-level fusion. The global encoder is kept unchanged except for adapting its input dimensions to the 271-sensor, 201-time-point MEG representation.

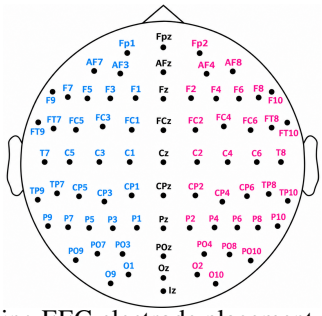


Figure 5: The corresponding EEG electrode placement of THINGS-EEG dataset.

B More Implementation Details

B.1 Evaluation metric implementation

Retrieval Accuracy. For EEG–image retrieval, we replace category text embeddings with image embeddings and retrieve the Top-K images most relevant to each EEG query. We report Top-K retrieval accuracy to evaluate EEG–visual alignment. Since contrastive learning is sensitive to batch size, we additionally compare retrieval performance under different batch sizes, including 16 and 1,024, to examine the robustness of different methods to contrastive training configurations.

B.2 Method Implementation Details

Network Architecture. Our framework combines a global ATMS EEG encoder with BHCP for hierarchical EEG representation learning. Each input EEG sample contains 63 channels and 250 time points. BHCP employs three brain-region STGCN branches with a hidden dimension of 64, a regional feature dimension of 256, and a dropout rate of 0.1. The regional features are mapped to learnable-curvature Poincaré balls, with curvature magnitudes initialized to 1.0, and subsequently interact through distance-based attention in a common-curvature space. Perceptual and semantic experts process the low–middle and middle–high representations, respectively, followed by geometry-dependent gating and Fréchet fusion. The resulting representation is projected to 1024 dimensions and combined with the global EEG representation through addition and layer normalization. Frozen multilevel visual features provide training-only supervision for expert specialization. AHDA adopts a shared level-conditioned subject discriminator with projection and hidden dimensions of 128 and a level-embedding dimension of 16.

Training and Testing. We evaluate CogEEG under intra-subject and inter-subject settings, as summarized in Table 9. In the intra-subject setting, AHDA is disabled since each model is trained on a single subject and no cross-subject alignment is required. In the inter-subject setting, we adopt a ten-fold leave-one-subject-out protocol and employ the complete CogEEG framework with both BHCP and AHDA. The held-out subject is strictly excluded from training, hyperparameter tuning, and checkpoint selection. For both settings, models are trained for 40 epochs with a batch size of 512 using AdamW, an initial learning rate of 10^{-4}, weight decay of 0.01, and cosine decay to 10^{-5}. The expert-specialization and domain-adversarial loss weights are both set to 0.01, with a specialization temperature of 0.07. The specialization weight is linearly warmed up over the first 10 epochs. For AHDA, the discriminator is warmed up for 5 epochs, followed by a 10-epoch ramp of the GRL coefficient to 1.0, and adaptive hierarchy weights use an EMA momentum of 0.9. Conflicting auxiliary gradients are projected, with their norms capped at 3% and 10% of the retrieval gradient for expert supervision and domain adaptation, respectively. During inference, EEG representations are matched against 200 frozen candidate image embeddings, and Top-1/Top-5 accuracies are averaged across subjects or leave-one-subject-out folds.

C Additional Method Analysis

C.1 Property of Retrieval-Aware Gradient Projection

As described in Sec. 3.3, adaptive hierarchical domain-adversarial alignment reduces subject-specific variations across cognitive levels. However, its gradient may conflict with the retrieval objective

Table 9: Training and evaluation protocols for intra-subject and inter-subject settings.

Setting	Intra-subject	Inter-subject
Training subjects	Single subject	Nine source subjects
Test subject	Same subject	One unseen subject
Protocol	Subject-specific	10-fold leave-one-subject-out
BHCP	✓	✓
AHDA	–	✓
Evaluation	Top-1 / Top-5	Top-1 / Top-5
Final reporting	Average over 10 subjects	Average over 10 folds

and potentially suppress retrieval-relevant information. To alleviate this conflict, we project the adversarial gradient according to its agreement with the retrieval gradient.

Let

$$\mathbf{g}_r = \nabla_{\mathbf{t}^\ell} \mathcal{L}_{\text{ret}}, \qquad \mathbf{g}_d = \nabla_{\mathbf{t}^\ell} \mathcal{L}_{\text{AHDA}} \tag{18}$$

denote the retrieval and adversarial gradients at cognitive level ℓ, respectively. Following Eq. (16), the projected adversarial gradient is defined as

$$\tilde{\mathbf{g}}_d = \mathbf{g}_d - \frac{\min\left(0, \langle \mathbf{g}_d, \mathbf{g}_r \rangle\right)}{\|\mathbf{g}_r\|_2^2} \mathbf{g}_r. \tag{19}$$

Proposition 1 (First-Order Retrieval Compatibility). For $\mathbf{g}_r \neq \mathbf{0}$, the projected adversarial gradient $\tilde{\mathbf{g}}_d$ defined in Eq. 19 satisfies

$$\langle \tilde{\mathbf{g}}_d, \mathbf{g}_r \rangle \geq 0. \tag{20}$$

Therefore, the projected adversarial gradient contains no component that conflicts with the retrieval gradient to first order.

Proof. We consider two cases.

When $\langle \mathbf{g}_d, \mathbf{g}_r \rangle \geq 0$, we have

$$\min\left(0, \langle \mathbf{g}_d, \mathbf{g}_r \rangle\right) = 0, \tag{21}$$

and hence $\tilde{\mathbf{g}}_d = \mathbf{g}_d$. Therefore,

$$\langle \tilde{\mathbf{g}}_d, \mathbf{g}_r \rangle = \langle \mathbf{g}_d, \mathbf{g}_r \rangle \geq 0. \tag{22}$$

When $\langle \mathbf{g}_d, \mathbf{g}_r \rangle < 0$, the projected gradient becomes

$$\tilde{\mathbf{g}}_d = \mathbf{g}_d - \frac{\langle \mathbf{g}_d, \mathbf{g}_r \rangle}{\|\mathbf{g}_r\|_2^2} \mathbf{g}_r. \tag{23}$$

Taking its inner product with $\mathbf{g}_r$ gives

$$\langle \tilde{\mathbf{g}}_d, \mathbf{g}_r \rangle = \langle \mathbf{g}_d, \mathbf{g}_r \rangle - \frac{\langle \mathbf{g}_d, \mathbf{g}_r \rangle}{\|\mathbf{g}_r\|_2^2} \langle \mathbf{g}_r, \mathbf{g}_r \rangle \tag{24}$$

$$= \langle \mathbf{g}_d, \mathbf{g}_r \rangle - \langle \mathbf{g}_d, \mathbf{g}_r \rangle \tag{25}$$

$$= 0. \tag{26}$$

Thus, in both cases,

$$\langle \tilde{\mathbf{g}}_d, \mathbf{g}_r \rangle \geq 0. \tag{27}$$

This completes the proof.

Interpretation. Consider an adversarial update $\Delta \mathbf{t}^\ell = -\eta \tilde{\mathbf{g}}_d$ with learning rate $\eta > 0$. Using a first-order Taylor approximation, the corresponding change in the retrieval objective is

$$\mathcal{L}_{\text{ret}}(\mathbf{t}^\ell + \Delta \mathbf{t}^\ell) \approx \mathcal{L}_{\text{ret}}(\mathbf{t}^\ell) - \eta \langle \mathbf{g}_r, \tilde{\mathbf{g}}_d \rangle. \tag{28}$$

Since Proposition 1 guarantees $\langle \mathbf{g}_r, \tilde{\mathbf{g}}_d \rangle \geq 0$, the projected adversarial update does not increase the retrieval objective to first order. Therefore, the proposed projection removes conflicting adversarial components while retaining updates compatible with retrieval learning.

Table 10: Ablation study of AHDA alignment strategies on THINGS-EEG under the inter-subject setting.

Method	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
Baseline	12.0	29.5	9.0	25.5	7.5	16.5	9.0	21.5	10.0	23.5	12.0	31.0	8.5	24.0	11.0	29.5	6.5	19.0	13.5	36.0	9.90	25.60
w/o Eq.16	16.0	40.0	18.0	**33.5**	**11.0**	28.0	**14.5**	**37.0**	**10.5**	**32.0**	13.0	36.0	**12.0**	**30.5**	**13.0**	**35.5**	13.0	**33.0**	**19.0**	**43.0**	14.00	**34.85**
Ours	**18.0**	**41.5**	**19.5**	**33.5**	10.5	**28.5**	**14.5**	36.5	**10.5**	30.0	**13.5**	**37.5**	11.0	22.0	**13.0**	35.0	**13.5**	31.0	18.5	42.5	**14.25**	33.80

D Baseline Descriptions.

We compare CogEEG with representative EEG–visual decoding methods:

- **EEGNet** [27] introduces a compact convolutional neural network that captures temporal and spatial EEG patterns using depthwise and separable convolutions, providing an efficient backbone for EEG representation learning.
- **BraVL** [6] jointly models brain, visual, and linguistic representations through multimodal learning, enabling neural decoding by exploiting complementary cross-modal information.
- **NICE** [41] adopts a contrastive EEG–image alignment framework with spatial feature extraction, learning discriminative EEG representations for zero-shot visual retrieval.
- **ATM** [28] introduces the Adaptive Thinking Mapper, which combines temporal–spatial EEG modeling with positional information to align EEG representations with pretrained visual embeddings.

E Qualitative Experiments

E.1 Qualitative Comparison of Retrieval Results

Figure 6 compares qualitative retrieval results across different methods. Existing approaches show inconsistent performance across visual categories: BraVL [6] correctly retrieves the spoon but fails on most other stimuli; NICE [41] succeeds on the off-road vehicle and dog; EEGNet [27] performs well on the hoodie, and dog; and ATM correctly retrieves the underwater scene and plate. In contrast, CogEEG consistently retrieves the exact stimulus across all shown examples, covering objects, apparel, animals, and natural scenes. These results indicate that CogEEG provides more stable EEG–visual alignment across diverse semantic categories, which is consistent with the benefit of hierarchical representation learning and adaptive cross-subject alignment.

E.2 More Retrieval Results

Figure 7 shows that CogEEG can recover the exact stimulus for several visually distinctive categories, such as the warship, antelope, wooden tool, concrete pipes, and metal rods, all of which are ranked at Top-1. For harder cases, the ground-truth image is still retrieved within the top candidates, e.g., the bassoon at Top-2, the basketball at Top-3, the bench at Top-2, and the bananas at Top-5. These examples indicate that CogEEG is effective at identifying distinctive visual concepts, while retrieval becomes more challenging when EEG responses provide weaker or less discriminative cues, causing the correct image to be displaced by visually or semantically related distractors.

F Ablation Study

F.1 Effectiveness of AHDA and Gradient Projection.

Table 10 evaluates the contribution of adaptive hierarchical alignment and retrieval-aware gradient projection. Compared with the baseline, AHDA without gradient projection improves the average Top-1 accuracy by 4.10%, demonstrating the effectiveness of level-adaptive cross-subject alignment. With gradient projection, the full model further improves Top-1 by 0.25%, achieving an overall gain of 4.35% over the baseline. This suggests that resolving conflicts between adversarial alignment and retrieval optimization helps preserve retrieval-discriminative information.

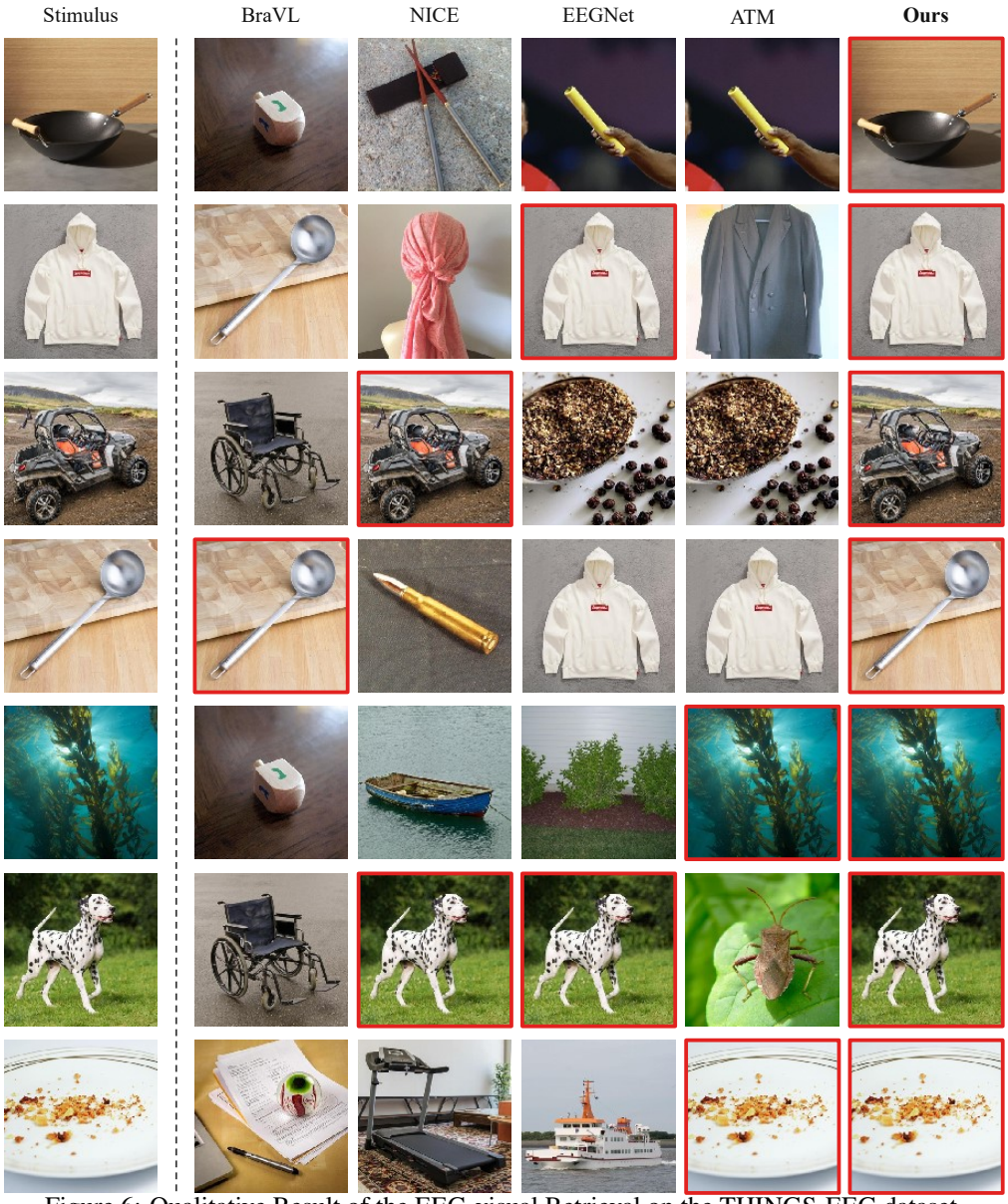


Figure 6: Qualitative Result of the EEG-visual Retrieval on the THINGS-EEG dataset.

G Discussion

G.1 Limitation.

Despite its effectiveness, CogEEG still has several limitations. First, BHCP relies on predefined brain-region priors to construct low-, middle-, and high-level EEG representations. Such a fixed partition may not fully reflect subject-specific functional organization or the dynamic nature of visual processing. Second, although learnable-curvature Poincaré spaces provide an effective geometry for hierarchical modeling, the current framework adopts a fixed three-level cognitive hierarchy, which may be insufficient for capturing finer-grained neural representations. In addition, CogEEG introduces extra computation through hyperbolic operations, expert fusion, and level-wise domain-adversarial learning. Finally, our experiments mainly focus on THINGS-EEG and THINGS-MEG under 200-way zero-shot retrieval, and the generalization of CogEEG to more diverse neural recording conditions and open-world visual concepts remains to be further investigated.

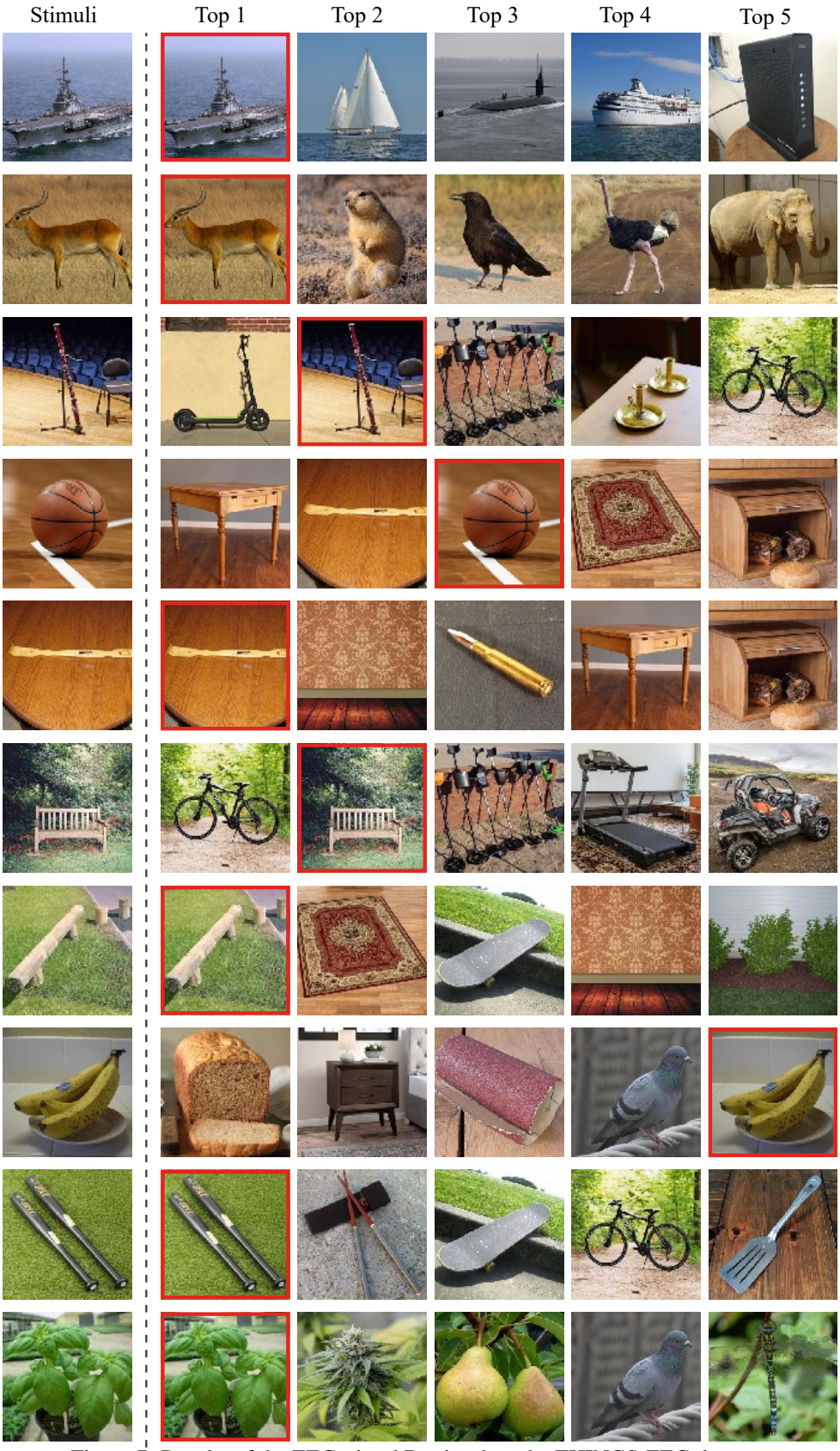


Figure 7: Results of the EEG-visual Retrieval on the THINGS-EEG dataset.

G.2 Future Work.

These limitations also suggest several promising directions. First, adaptive brain-region grouping could be explored to learn subject- or stimulus-dependent functional organization instead of relying on fixed anatomical partitions. Second, dynamically learning the number and structure of cognitive levels may enable CogEEG to capture finer visual processing hierarchies. Another interesting direction is to incorporate temporal cognitive dynamics into the current spatial hierarchy, allowing the model to jointly characterize where and when visual information emerges in EEG signals. Moreover, more efficient hyperbolic interaction and domain-alignment mechanisms could reduce the additional training cost. Finally, extending CogEEG to open-set retrieval, neural visual reconstruction, and broader EEG/MEG datasets would provide a more comprehensive evaluation of its cross-subject and cross-modality generalization.

G.3 Broader Impact

CogEEG advances cross-subject EEG–visual retrieval by improving the transferability of neural representations across individuals, which may benefit brain–computer interfaces and neural decoding systems that require reduced subject-specific calibration. More broadly, hierarchical modeling of neural responses may contribute to understanding how visual information is represented across cognitive levels. At the same time, neural signals can contain sensitive information about an individual's perceptual and cognitive states. Therefore, practical deployment should consider informed consent, data privacy, secure storage, and safeguards against unauthorized neural decoding. We view CogEEG primarily as a research framework and encourage its use under appropriate ethical and privacy protections.