BRIDGING THE HIERARCHIES: NEURO-INSPIRED HIERARCHY-AWARE ALIGNMENT FOR VISUAL NEURAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual neural decoding aims to retrieve perceived visual stimuli from neural signals and is an important research direction in brain–computer interfaces. Although EEG- and MEG-based methods have made promising progress, they insufficiently account for the neuroscientific priors that visual processing is hierarchically organized and information propagates progressively along cortical pathways. Two key limitations remain: **I) incomplete inter-hierarchy correspondence**, as existing methods associate neural and visual representations at a selected semantic level, leaving correspondence between the two representational hierarchies as a whole insufficiently captured; and **II) insufficient modeling of intra-hierarchy propagation**, as existing alignment strategies typically treat semantic levels in isolation, without modeling progressive information propagation across successive levels within each hierarchy. We propose **Neuro-Inspired Hierarchy-Aware Alignment (NIA)** to jointly address these limitations. NIA integrates a Hierarchy-Partitioned Channel Transformer (HPCT), Hierarchy-Matched Contrastive Learning (HMCL), and a Parallel Dual-Hierarchy Cascade (PDHC) to organize neural representations into a hierarchy, establish structured correspondence between the neural and visual representational hierarchies, and model progressive shallow-to-deep propagation within both hierarchies. Experiments on THINGS-EEG2 and THINGS-MEG demonstrate state-of-the-art zero-shot brain-to-image retrieval. On THINGS-EEG2, NIA improves Top-1/Top-5 accuracy by 9.5/1.3 and 9.9/10.0 percentage points under the intra- and inter-subject settings, respectively, validating the effectiveness of inter-hierarchy correspondence and intra-hierarchy progressive information propagation.

1 INTRODUCTION

Brain–computer interfaces (BCIs) aim to establish direct communication pathways between the human brain and external devices, allowing the decoding of human intentions, perceptions, and cognitive states (Vidal et al., 1973; Farwell & Donchin, 1988; Wolpaw et al., 1991; Hochberg et al., 2006; McFarland & Wolpaw, 2011; Nicolas-Alonso & Gomez-Gil, 2012). Within this research area, visual neural decoding aims to retrieve perceived visual stimuli based on the corresponding neural signals (Kamitani & Tong, 2005; Kay et al., 2008; Miyawaki et al., 2008), which holds promise for practical applications such as assistive communication for patients with motor impairments (Lee et al., 2020; Chaudhary et al., 2016), attention monitoring in complex visual tasks (Yao et al., 2025) and brain-guided human-computer interaction (Zhao et al., 2025; Shen et al., 2019).

Existing visual neural decoding methods have made promising progress, typically by aligning neural signals with the output of a selected layer of the visual encoder, such as final-layer (Song et al., 2024; Li et al., 2024a) or intermediate-layer (Du et al., 2026) representations. Yet, a central neuroscientific prior suggests a broader perspective: visual processing in the brain is hierarchically organized, with neural representations at different cortical stages reflecting different levels of abstraction (DiCarlo et al., 2012; Felleman & Van Essen, 1991). A comparable representational hierarchy is observed in visual encoders, whose different layers capture visual representations at different semantic levels (Zeiler & Fergus, 2014; Yosinski et al., 2014). Thus, although the selected layer varies across methods, using its output as the sole alignment target restricts neural–visual association to a selected

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

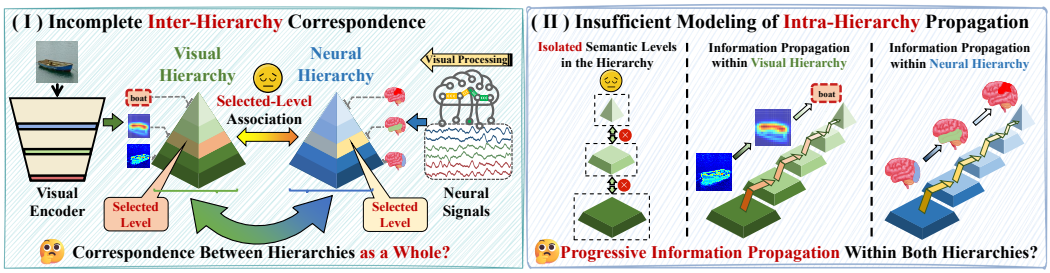


Figure 1: **Motivation of NIA. I) Incomplete inter-hierarchy correspondence:** Association at a selected semantic level leaves correspondence between the neural and visual representational hierarchies as a whole insufficiently captured. **II) Insufficient modeling of intra-hierarchy propagation:** Treating semantic levels in isolation during alignment leaves progressive information propagation across successive levels unmodeled.

semantic level, leaving correspondence between the neural and visual representational hierarchies as a whole insufficiently captured. This raises the first pivotal problem: ***I)*** *how to establish inter-hierarchy semantic correspondence between the neural and visual representational hierarchies as a whole?*

Beyond hierarchical organization, a further neuroscientific prior is that visual information propagates progressively along cortical pathways, with successive stages building upon information from preceding stages (Cichy et al., 2014; DiCarlo et al., 2012). Visual encoders exhibit a comparable progression, with deeper representations built upon information passed from shallower layers (Zeiler & Fergus, 2014; Yosinski et al., 2014). When different semantic levels are considered during neural–visual alignment, they are typically treated in isolation, without explicitly modeling progressive information propagation across successive levels within each representational hierarchy. BrainHIVE (Zheng et al., 2026) also incorporates multiple semantic levels by fusing final representations with different semantic biases from diverse visual encoders, but these representations are directly combined rather than connected through explicit progressive propagation across semantic levels. This leaves a second pivotal problem: ***II)*** *how to model intra-hierarchy progressive information propagation during neural–visual alignment?*

To address the above problems, we propose a **Neuro-Inspired Hierarchy-Aware Alignment (NIA)** framework that establishes correspondence between the neural and visual representational hierarchies while modeling progressive information propagation within each. NIA comprises three key modules: a Hierarchy-Partitioned Channel Transformer (HPCT), Hierarchy-Matched Contrastive Learning (HMCL), and a Parallel Dual-Hierarchy Cascade (PDHC). For problem **I)**, HPCT treats sensor channels as tokens and captures inter-channel dependencies, motivated by the observation that channel-wise response patterns reflect spatially distributed neural activity related to visual processing (Michel & Murray, 2012). It further partitions its attention heads into groups associated with different semantic levels, thereby organizing the neural representations into a hierarchy. HMCL then relates this neural hierarchy to the visual hierarchy formed by different layers of a pretrained visual encoder through joint contrastive supervision over their constituent semantic levels. Together, HPCT and HMCL extend neural–visual correspondence from a selected semantic level to the two representational hierarchies. For problem **II)**, inspired by the neuroscientific prior that visual information propagates progressively along cortical pathways (Cichy et al., 2014; DiCarlo et al., 2012), PDHC applies parallel gated shallow-to-deep cascades to the two hierarchies before alignment. This enables information to propagate across successive semantic levels, incorporating the progressive organization of both hierarchies into their structured cross-modal correspondence.

Overall, our main contributions are summarized as follows:

- **Framework.** Inspired by neuroscientific priors on hierarchical organization and progressive information propagation, we propose NIA to jointly address inter-hierarchy correspondence and intra-hierarchy progressive information propagation in visual neural decoding. NIA establishes structured correspondence between the neural and visual representational hierarchies as a whole while modeling progressive information propagation within each.
- **Technical Highlights.** We develop HPCT to organize neural representations into a hierarchy through channel tokenization and attention-head grouping, and HMCL to establish inter-hierarchy

semantic correspondence through joint contrastive supervision across their constituent semantic levels. PDHC further applies parallel gated shallow-to-deep cascades within both hierarchies before alignment, incorporating intra-hierarchy progressive information propagation into neural–visual correspondence learning.

- **Experimental Validation.** Extensive experiments on THINGS-EEG2 (Gifford et al., 2022) and THINGS-MEG (Hebart et al., 2023) demonstrate accurate and robust zero-shot brain-to-image retrieval across neural modalities and subject settings. Ablation studies further validate the effectiveness of the proposed designs.

2 Related Work

2.1 Visual Neural Decoding

Visual neural decoding aims to retrieve perceived stimuli from neural signals recorded by fMRI (Ogawa et al., 1990), EEG (Nunez & Srinivasan, 2006), or MEG (Hämäläinen et al., 1993). Most existing methods establish cross-modal alignment between neural responses and visual representations. Ye et al. (2024) formulated the task as self-supervised cross-modal retrieval, while NICE (Song et al., 2024) modeled the temporal and spatial structure of EEG signals. ATM (Li et al., 2024a) projected neural signals into pretrained visual embedding spaces, and subsequent methods improved alignment using multimodal semantic cues (Zhang et al., 2025), semantic decoupling (Chen et al., 2024a), uncertainty-aware visual priors (Wu et al., 2025), cognitive prior augmentation (Zhang et al., 2026), or adaptive visual supervision (Wu et al., 2026).

Recent studies have further explored the role of visual representational hierarchy. Du et al. (2026) showed that intermediate visual representations can provide more suitable alignment targets than final-layer representations, while BrainHIVE (Zheng et al., 2026) integrated hierarchical representations from multiple visual encoders into a fused visual token.

Overall, existing methods primarily align neural representations with a selected visual level or a fused visual target, leaving inter-hierarchy correspondence and intra-hierarchy propagation insufficiently addressed. In contrast, NIA establishes structured correspondence between the neural and visual representational hierarchies while modeling progressive information propagation within each.

2.2 Semantic Hierarchies in Biological Vision and Visual Encoders

Human visual processing progresses from early perceptual analysis to increasingly abstract object representations along cortical pathways (Felleman & Van Essen, 1991; DiCarlo et al., 2012), with distinct processing stages unfolding across cortical regions and temporal windows (Cichy et al., 2014). Deep visual encoders exhibit a similar depth-wise organization: shallow layers capture perceptual patterns, whereas deeper layers encode more abstract and category-related semantics (Zeiler & Fergus, 2014; Yosinski et al., 2014). Studies comparing DNN representations with neural responses further reveal correspondences between network depth and the spatial and temporal organization of biological vision (Güçlü & Van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Cichy et al., 2016). ReAlnet (Lu et al., 2026) also shows that multi-layer alignment with EEG responses can improve the brain similarity of vision models. These findings motivate jointly aligning the neural and visual representational hierarchies while preserving their progressive organization.

3 Method

3.1 Problem Definition

We study zero-shot visual neural decoding and formulate it as a cross-modal representation alignment problem between neural signals and visual stimuli. Let $\mathcal{D} = \{(x_N^{(n)}, x_I^{(n)})\}_{n=1}^{N}$ denote a paired dataset of N samples, where $x_N^{(n)} \in \mathbb{R}^{C\times T}$ represents a neural signal with C channels and T time points, and $x_I^{(n)} \in \mathbb{R}^{H_I\times W_I\times 3}$ denotes the corresponding visual stimulus. The dataset is divided into training and test sets with disjoint visual concepts, ensuring a zero-shot retrieval setting. The model is trained to align neural and visual representations in a shared latent space, and inference retrieves the corresponding stimulus from a candidate image pool using cross-modal similarity.

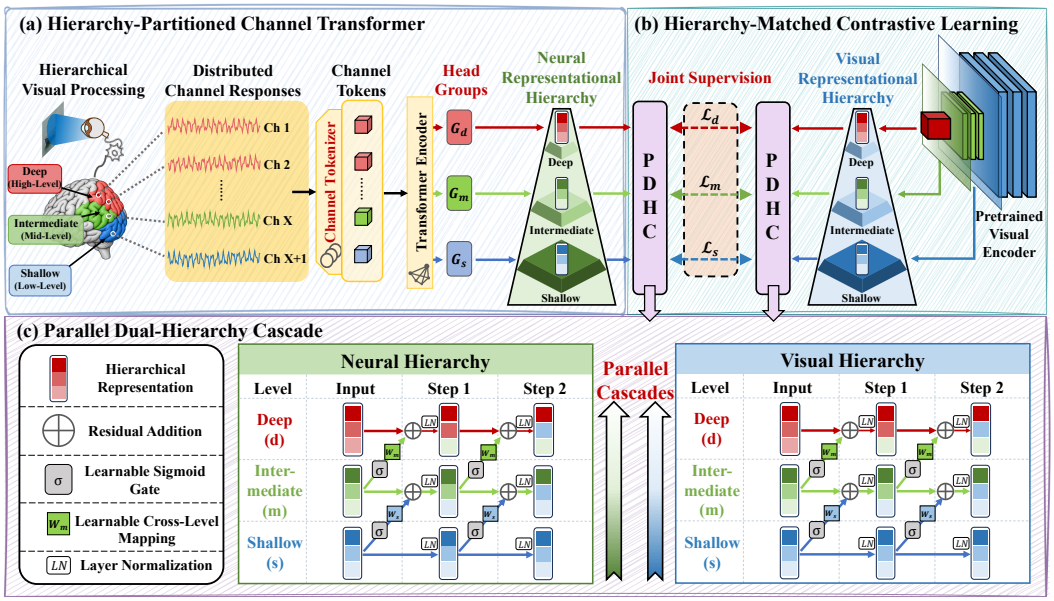


Figure 2: **Overview of the proposed Neuro-Inspired Hierarchy-Aware Alignment (NIA) framework.** (a) HPCT tokenizes EEG/MEG channels and groups attention heads to organize neural representations into a hierarchy. (b) HMCL establishes structured correspondence between the neural and visual representational hierarchies through joint contrastive supervision across corresponding semantic levels. (c) PDHC performs gated shallow-to-deep information propagation within both hierarchies before contrastive alignment.

3.2 Hierarchy-Partitioned Channel Transformer

Existing neural encoders typically learn a unified neural embedding for alignment with a selected visual representation (Song et al., 2024; Li et al., 2024a). To support correspondence with the visual hierarchy as a whole, we introduce HPCT, which combines channel-tokenized encoding with semantic-level head grouping to organize neural representations into a hierarchy.

Channel-wise response patterns can reflect spatially distributed neural activity related to visual processing (Michel & Murray, 2012). Motivated by this observation, HPCT treats sensor channels as tokens to capture dependencies among their response patterns. Given a neural signal $x_N \in \mathbb{R}^{C\times T}$, the time series of the c-th channel is mapped into a d_t-dimensional token through a shared linear projection:

$$e_c = W_{\text{tok}} x_{N,c} + b_{\text{tok}}, \tag{1}$$

where $W_{\text{tok}} \in \mathbb{R}^{d_t\times T}$ and $b_{\text{tok}} \in \mathbb{R}^{d_t}$. Stacking the channel tokens gives $E = [e_1, e_2, \ldots, e_C]^\top \in \mathbb{R}^{C\times d_t}$. We then prepend a learnable class token $e_{\text{cls}} \in \mathbb{R}^{1\times d_t}$ and add channel positional embeddings $P_{\text{ch}} \in \mathbb{R}^{(C+1)\times d_t}$, yielding $X_0 = [e_{\text{cls}}; E] + P_{\text{ch}}$. Before self-attention, X_0 is mapped into a d_a-dimensional attention space as $\bar{X}_0 = X_0 W_{\text{in}}$. Shared QKV projections then allow channel representations to incorporate information from other channels while retaining channel identity through the positional embeddings.

To organize the encoded neural information into separate semantic branches, we partition the H attention heads into three disjoint groups, $\mathcal{G}_s$, $\mathcal{G}_m$, and $\mathcal{G}_d$, associated with shallow, intermediate, and deep semantic levels, respectively. All groups operate on the same projected channel-token sequence. For each level $k \in \{s, m, d\}$, the outputs within $\mathcal{G}_k$ are concatenated and mapped back to the attention dimension through an independent output projection:

$$X^{(k)} = \bar{X}_0 + \text{Concat}_{h\in\mathcal{G}_k}(O_h) W_O^{(k)}. \tag{2}$$

No concatenation is performed across head groups, allowing the three branches to share the channel-token context while learning within disjoint attention subspaces. The class-token output of each

branch is subsequently mapped into the shared d-dimensional contrastive space:

$$z_N^{(k)} = p_k\left(X_{\text{cls}}^{(k)}\right), \qquad k \in \{s, m, d\}, \tag{3}$$

where $p_k : \mathbb{R}^{d_a} \to \mathbb{R}^d$ denotes the level-specific branch and contrastive projections.

This design combines shared channel context with separate representational capacity across semantic levels. HPCT thus provides structurally separated neural representations, whose semantic organization is subsequently guided by their correspondence with different visual encoder depths through HMCL.

3.3 Hierarchy-Matched Contrastive Learning

Using the output of a selected final or intermediate visual layer as the sole alignment target restricts neural–visual association to a particular semantic level (Song et al., 2024; Li et al., 2024a; Du et al., 2026). To establish structured correspondence between the neural and visual representational hierarchies, we introduce HMCL, which uses representations from different depths of a frozen pretrained visual encoder as level-specific targets to jointly align the corresponding neural and visual representations.

Visual representational hierarchy. Let E_ϕ denote a pretrained visual encoder with L layers. We select three representative depths, l_s, l_m, and l_d, where $1 \le l_s < l_m < l_d \le L$, to obtain shallow, intermediate, and deep visual representations. For each selected layer, a level-specific pooling and projection head $h_k(\cdot)$ maps its output into a d-dimensional embedding:

$$z_I^{(k)} = h_k\left(E_\phi^{(l_k)}(x_I)\right), \qquad k \in \{s, m, d\}. \tag{4}$$

Here, $E_\phi^{(l_k)}(x_I)$ denotes the output of the l_k-th layer. The resulting representations form an ordered visual hierarchy, providing supervision for the semantic organization of the neural hierarchy organized by HPCT.

Joint contrastive supervision. Given the neural and visual representations $\{z_N^{(k)}, z_I^{(k)}\}_{k\in\{s,m,d\}}$, we first apply PDHC to propagate information within each hierarchy, as detailed in the following subsection. HMCL then jointly optimizes neural–visual matching at the three semantic levels using the propagated representations. We denote their ℓ_2-normalized forms by $\tilde{z}_{N,i}^{(k)}$ and $\tilde{z}_{I,i}^{(k)}$, respectively. At each level k, we adopt a symmetric contrastive objective (Oord et al., 2018; Radford et al., 2021):

$$\mathcal{L}_{\text{con}}^{(k)} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(\text{sim}(\tilde{z}_{N,i}^{(k)}, \tilde{z}_{I,i}^{(k)})/\tau_k)}{\sum_{j=1}^{B}\exp(\text{sim}(\tilde{z}_{N,i}^{(k)}, \tilde{z}_{I,j}^{(k)})/\tau_k)} + \log\frac{\exp(\text{sim}(\tilde{z}_{I,i}^{(k)}, \tilde{z}_{N,i}^{(k)})/\tau_k)}{\sum_{j=1}^{B}\exp(\text{sim}(\tilde{z}_{I,i}^{(k)}, \tilde{z}_{N,j}^{(k)})/\tau_k)}\right]. \tag{5}$$

Here, B is the mini-batch size, τ_k is the learnable temperature at level k, and $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. The two terms encourage neural-to-visual and visual-to-neural matching, respectively. With $K = 3$, the overall HMCL objective averages the contrastive losses across the three semantic levels and includes a temperature-prior regularizer with $\tau_0 = 0.07$ and $\lambda_{\text{temp}} = 0.01$:

$$\mathcal{L}_{\text{HMCL}} = \frac{1}{K}\sum_{k\in\{s,m,d\}}\left[\mathcal{L}_{\text{con}}^{(k)} + \lambda_{\text{temp}}(\tau_k - \tau_0)^2\right], \tag{6}$$

Together, HPCT provides structurally separated neural representations, while HMCL guides their semantic organization through correspondence with different visual encoder depths. They jointly address problem **I)** by establishing structured correspondence between the neural and visual representational hierarchies as a whole.

3.4 Parallel Dual-Hierarchy Cascade

To address problem **II)**, we introduce PDHC to model progressive information propagation across successive semantic levels within each hierarchy before contrastive alignment. Inspired by visual information propagation along cortical pathways (DiCarlo et al., 2012; Felleman & Van Essen, 1991;

Cichy et al., 2014), PDHC applies parallel gated shallow-to-deep cascades to the neural and visual representations. The two cascades share the same structure but have separate learnable parameters.

Let $q \in \{N, I\}$ index the neural or visual hierarchy. We initialize $z_q^{(k,0)} = z_q^{(k)}$ for $k \in \{s, m, d\}$ and perform two propagation steps. At each step $t \in \{0, 1\}$, the level-specific representations are updated simultaneously:

$$\begin{aligned} z_q^{(s,t+1)} &= \mathrm{LN}_s\left(z_q^{(s,t)}\right), \\ z_q^{(m,t+1)} &= \mathrm{LN}_m\left(z_q^{(m,t)} + g_{q,s} W_{q,s} z_q^{(s,t)}\right), \\ z_q^{(d,t+1)} &= \mathrm{LN}_d\left(z_q^{(d,t)} + g_{q,m} W_{q,m} z_q^{(m,t)}\right). \end{aligned} \tag{7}$$

Here, $\mathrm{LN}_k(\cdot)$ denotes the level-specific LayerNorm, $W_{q,s}, W_{q,m} \in \mathbb{R}^{d \times d}$ are learnable mappings between adjacent levels, and the propagated information is controlled by learnable sigmoid gates $g_{q,s} = \sigma(a_{q,s})$ and $g_{q,m} = \sigma(a_{q,m})$. Thus, gated information is propagated between adjacent levels in the shallow-to-deep direction, with all level-specific representations normalized after each propagation step.

Through these gated residual updates, two propagation steps allow information to accumulate progressively across the hierarchy. The resulting representations $z_q^{(k,2)}$ are ℓ_2-normalized and used to compute the contrastive losses in Eq. (5) and the overall HMCL objective.

4 Experiments

4.1 Experimental Setup

Datasets and Processing. We conduct experiments on THINGS-EEG2 (Gifford et al., 2022) and THINGS-MEG (Hebart et al., 2023), both based on the THINGS (Hebart et al., 2019) image database. THINGS-EEG2 contains EEG recordings from 10 participants, with 1,654 training and 200 test concepts; each training concept has 10 images with 4 repetitions, and each test concept one image repeated 80 times. For THINGS-MEG, we exclude the 200 test concepts from the original 1,854 concepts under a concept-disjoint zero-shot protocol, yielding 1,654 training and 200 test concepts for 4 participants; each training concept has 12 images, and each test concept one image repeated 12 times. Following prior work (Wu et al., 2025; Zhang et al., 2026; Du et al., 2026), we apply modality-specific preprocessing and repeated-trial averaging; full details are provided in Appendix A.

Retrieval Protocol. During inference, candidate images are ranked by the equally weighted average of cosine similarities across the three semantic levels, computed using the neural and visual representations after PDHC.

4.2 Comparison with Baselines

For a comprehensive evaluation, we compare the proposed method with representative approaches in EEG/MEG-based visual decoding. The compared methods include **BraVL** (Du et al., 2023), the **NICE** series—comprising **NICE**, **NICE-SA**, **NICE-GA** (Song et al., 2024), and **NICE-HA** (Rajabi et al., 2025)—**ATM-S** (Li et al., 2024a), **Neural-MCRL** (Li et al., 2024b), **VE-SDN** (Chen et al., 2024a), **CognitionCapturer** (Zhang et al., 2025), **UBP** (Wu et al., 2025), **NeuroCLIP** (Wang et al., 2025), **ATS** (Wu et al., 2026), **NeuroBridge** (Zhang et al., 2026), **Du et al.** (Du et al., 2026), and **BrainHIVE** (Zheng et al., 2026). Baseline results are taken from their respective papers, while our results are averaged over five random seeds. Controlled reproductions under matched evaluation settings are provided in Appendix E.1.

As shown in Table 1, NIA achieves the highest average retrieval accuracy among the compared methods on THINGS-EEG2, with 92.1%/99.0% Top-1/Top-5 accuracy in the intra-subject setting and 31.7%/59.4% in the inter-subject setting. Table 2 further shows that NIA achieves the highest average accuracy on THINGS-MEG, reaching 50.6%/75.5% and 6.2%/14.2% under the intra- and inter-subject settings, respectively. These results support the effectiveness of hierarchy-aware neural–visual alignment across neural modalities and subject settings. Additional neural–visual encoder combination results are provided in Appendix D.

Table 1: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG2 (Gifford et al., 2022).

Method	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Avg	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Inter-subject: leave one subject out for test																						
BraVL	2.3	8.0	1.5	6.3	1.4	5.9	1.7	6.7	1.5	5.6	1.8	7.2	2.1	8.1	2.2	7.6	1.6	6.4	2.3	8.5	1.8	7.0
NICE	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-SA	7.0	22.6	6.6	23.2	7.5	23.7	5.4	21.4	6.4	22.2	7.5	22.5	3.8	19.1	8.5	24.4	7.4	22.3	9.8	29.6	7.0	23.1
NICE-GA	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	4.8	18.3	6.2	26.3	5.9	21.6
NICE-HA	10.3	33.7	12.8	32.0	11.9	30.7	9.9	28.6	7.4	22.1	16.7	38.2	8.5	28.8	12.4	31.6	10.0	27.1	14.1	33.1	11.4	30.6
ATM-S	10.5	26.8	7.1	24.8	11.9	33.8	14.7	39.4	7.0	23.9	11.1	35.8	16.1	43.5	15.0	40.3	4.9	22.7	20.5	46.5	11.8	33.7
Neural-MCRL	13.0	31.5	12.0	30.5	14.5	35.5	12.5	35.5	11.5	29.0	13.5	35.5	14.0	36.0	18.5	38.5	13.5	32.5	17.0	39.0	14.0	34.3
NeuroCLIP	17.2	47.3	31.8	60.2	10.9	24.3	19.1	48.2	14.3	32.6	11.2	32.1	15.8	41.3	10.2	29.2	9.9	27.4	29.8	60.0	17.0	40.3
ATS	17.0	37.5	21.5	49.5	7.5	20.0	13.5	41.5	7.0	27.0	14.5	37.0	10.0	33.5	13.0	30.5	9.5	28.5	26.0	52.5	14.0	35.8
UBP	11.5	29.7	15.5	40.0	9.8	27.0	13.0	32.3	8.8	33.8	11.7	31.0	10.2	23.8	12.2	32.2	15.5	40.5	16.0	43.5	12.4	33.4
NeuroBridge	23.2	52.4	21.2	49.3	13.2	36.5	17.0	45.3	14.5	37.7	25.0	55.0	15.3	45.1	20.1	44.9	13.7	36.5	27.2	56.3	19.0	45.9
BrainHIVE	24.0	50.0	18.5	50.5	15.5	36.5	15.5	43.0	18.5	43.5	17.5	40.5	18.5	39.5	21.5	43.5	17.5	43.5	28.0	58.5	19.5	44.9
Du et al.	23.5	53.2	30.6	60.4	10.0	28.1	19.5	48.3	18.1	45.2	22.7	49.8	18.6	46.0	17.3	46.1	23.0	**54.8**	34.4	62.0	21.8	49.4
Ours	**38.1**	**67.8**	**42.0**	**72.7**	**23.9**	**51.0**	**29.7**	**57.7**	**22.7**	**48.8**	**26.8**	53.1	**28.9**	**55.4**	**33.7**	**61.1**	**24.0**	51.0	**47.0**	**75.4**	**31.7**	**59.4**
Intra-subject: train and test on one subject																						
BraVL	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	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
NICE-SA	13.3	40.2	12.1	36.1	15.3	39.6	15.9	49.0	9.8	34.4	14.2	42.4	17.9	43.6	18.2	50.2	14.4	38.7	16.0	42.8	14.7	41.7
NICE-GA	15.2	40.1	13.9	40.1	14.7	42.7	17.6	48.9	9.0	29.7	16.4	44.4	14.9	43.1	20.3	52.1	14.1	39.7	19.6	46.7	15.6	42.8
ATM-S	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	34.4	51.5	29.1	63.5	28.5	60.4
VE-SDN	32.6	63.7	34.4	69.9	38.7	73.5	39.8	72.0	29.4	58.6	34.5	68.8	34.5	68.3	49.3	79.8	39.0	69.6	39.8	75.3	37.2	69.9
CogCapturer	27.2	59.5	28.7	57.0	37.2	66.1	37.7	63.2	21.8	47.8	31.6	58.1	32.8	59.6	47.6	73.5	33.4	57.7	35.1	63.6	33.3	60.6
Neural-MCRL	27.5	64.0	28.5	61.5	37.0	69.0	35.0	66.0	22.5	51.5	31.5	61.0	31.5	62.5	42.0	74.5	30.5	59.5	37.5	71.0	32.4	64.1
UBP	41.2	70.5	51.2	80.9	51.2	82.0	51.1	76.9	42.2	72.8	57.5	83.5	49.0	79.9	58.6	85.8	45.1	76.2	61.5	88.2	50.9	79.7
NeuroCLIP	60.3	84.2	64.5	94.1	63.7	93.2	62.8	91.3	55.0	83.7	68.6	93.6	70.6	93.3	69.6	90.3	48.2	86.5	69.1	92.1	63.2	90.3
ATS	53.0	79.0	62.0	87.5	61.5	89.0	57.0	86.5	55.0	84.0	68.0	90.5	53.0	84.0	66.5	91.0	58.5	86.0	67.5	89.0	60.2	86.7
NeuroBridge	50.0	77.6	63.2	90.6	61.6	91.1	61.4	90.0	54.8	85.0	69.7	92.9	62.7	88.8	71.2	95.1	64.0	91.0	73.6	97.1	63.2	89.9
BrainHIVE	64.3	88.8	76.3	95.3	74.0	95.0	67.0	91.8	68.0	91.5	81.5	96.3	76.8	96.8	84.8	98.5	76.8	95.8	87.3	99.3	75.7	94.6
Du et al.	75.0	94.3	87.5	98.9	83.2	98.2	79.5	96.1	74.6	96.4	89.9	**99.3**	78.5	97.3	86.9	99.4	81.3	97.8	89.3	99.1	82.6	97.7
Ours	**93.9**	**99.3**	**91.1**	**99.1**	**94.6**	**99.5**	**88.0**	**98.2**	**87.5**	**97.1**	**93.7**	98.7	**91.0**	**99.5**	**94.1**	**100.0**	**90.7**	**98.8**	**97.1**	**99.6**	**92.1**	**99.0**

4.3 Ablation Study

Unless otherwise specified, ablation and diagnostic experiments are conducted under the intra-subject protocol with a fixed random seed (2026), whereas the main comparison results are averaged over five random seeds.

Table 2: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-MEG (Hebart et al., 2023).

Method	Subject 1		Subject 2		Subject 3		Subject 4		Avg	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Inter-subject: leave one subject out for test										
NeuroCLIP	2.1	6.0	2.7	16.3	2.9	12.4	2.2	8.9	2.5	10.9
ATS	2.0	6.0	6.0	14.0	4.5	15.0	1.0	9.5	3.4	11.2
UBP	2.0	5.7	1.5	17.2	2.7	10.5	2.5	8.0	2.2	10.4
NeuroBridge	4.3	**13.1**	3.6	15.6	3.0	11.2	**2.5**	**11.3**	3.4	12.8
Du et al.	1.3	6.9	6.6	**18.5**	5.4	18.5	1.5	7.9	3.7	13.0
Ours	**4.7**	10.3	**7.8**	**18.5**	**9.4**	**21.6**	**2.9**	6.2	**6.2**	**14.2**
Intra-subject: train and test on one subject										
NICE	9.6	27.8	18.5	47.8	14.2	41.6	9.0	26.6	12.8	36.0
NICE-SA	9.8	27.8	18.6	46.4	10.5	38.4	11.7	27.2	12.7	35.0
NICE-GA	8.7	30.5	21.8	56.6	16.5	49.7	10.3	32.3	14.3	42.3
NICE-HA	12.6	35.7	28.3	62.5	22.2	48.8	14.3	36.7	19.3	45.9
UBP	15.0	38.0	46.0	80.5	27.3	59.0	18.5	43.5	26.7	55.2
NeuroCLIP	15.6	31.7	67.1	91.6	35.3	74.1	12.3	28.7	32.6	56.5
ATS	18.0	43.0	56.0	89.0	36.0	67.0	19.5	50.0	32.4	62.3
NeuroBridge	16.5	41.6	53.7	85.3	40.4	73.2	18.1	43.1	32.2	60.8
BrainHIVE	14.0	31.8	63.8	91.8	41.0	78.3	17.0	41.0	33.9	60.7
Du et al.	25.5	54.5	**81.9**	**97.4**	**56.0**	**87.5**	28.6	58.3	48.0	74.4
Ours	**29.4**	**55.9**	80.1	96.5	55.6	80.9	**37.4**	**68.5**	**50.6**	**75.5**

Table 3: Component ablation on average retrieval accuracy (%).

Encoder	HMCL	PDHC	Top-1	Top-5
HPCT	✓	✓	**92.35**	**99.20**
ATM	✓	✓	89.20	98.45
EEGNet	✓	✓	78.80	95.70
EEGProject	✓	✓	79.90	96.50
TSConv	✓	✓	82.10	96.15
HPCT	✓	×	89.30	98.90
HPCT	Shared 60% target	✓	88.05	98.25
HPCT	60% target only	–	84.70	96.75

Table 4: Effect of the number of levels on EEG inter-subject retrieval accuracy (%).

Setting	Top-1	Top-5
2 levels	29.60	59.00
3 levels	**33.35**	**61.20**
4 levels	31.30	58.80
5 levels	31.35	59.60

Component Ablation. Table 3 reports the ablation results of the key modules in NIA. The full framework achieves the best performance, with 92.35% Top-1 and 99.20% Top-5 accuracy. Replacing HPCT with other neural encoders reduces retrieval accuracy, supporting its channel-tokenized design for learning inter-channel dependencies and organizing neural representations across semantic levels. Removing PDHC reduces Top-1 accuracy to 89.30%, indicating the benefit of progressive information propagation before alignment.

To disentangle neural hierarchy organization from hierarchy-matched visual supervision, we compare two controls using the same visual representation at 60% encoder depth. The *60% depth only* variant performs a single neural–visual alignment without PDHC or multi-level similarity fusion. The *Shared 60% depth* variant retains HPCT's three level-specific neural representations, level-wise contrastive supervision, PDHC, and similarity fusion, but uses the same 60%-depth visual representation as the target for all three levels. It improves Top-1/Top-5 accuracy by 3.35/1.50 percentage points over

the single-level variant, showing the combined benefit of neural hierarchy organization, progressive intra-hierarchy propagation, and multi-level similarity fusion under a fixed visual target. Replacing the shared target with distinct shallow, intermediate, and deep visual representations further improves accuracy by 4.30/0.95 points, supporting hierarchy-matched supervision from different visual depths.

Effect of Neural and Visual Encoders. Figure 3 evaluates different combinations of neural and visual encoders on THINGS-EEG2. For neural encoders, we compare the proposed **HPCT** with **EEGNet** (Lawhern et al., 2018), **TSConv** (Song et al., 2024), **ATM** (Li et al., 2024a), and **EEGProject** (Wu et al., 2025). For visual encoders, we compare CNN-based ResNet backbones including **RN50** and **RN101** (He et al., 2016), transformer-based ViT backbones including **ViT-B/16**, **ViT-L/14**, and **ViT-H/14** (Dosovitskiy et al., 2020), **DINOv2** (Oquab et al., 2023), and **InternViT** (Chen et al., 2024b). Overall, HPCT consistently outperforms other neural encoders across different visual encoder choices, achieving an average Top-1 accuracy of 83.7%. This advantage is consistent with HPCT's channel-tokenized design, which captures inter-channel dependencies to organize spatially distributed neural responses for hierarchy-aware alignment. On the visual side, InternViT achieves the highest average accuracy among the evaluated encoders, and its combination with HPCT obtains the best Top-1 accuracy of 92.4%.

Effect of Semantic Hierarchy Granularity. As shown in Table 4, three levels achieve the best performance, with 33.35% Top-1 and 61.20% Top-5 accuracy. All settings evenly partition the visual encoder depth. The results suggest that the shallow–intermediate–deep partition effectively separates semantic information while maintaining clear correspondence between the two hierarchies. Two levels provide less complete semantic separation, whereas four or five yield no further gains. Thus, NIA benefits from an appropriately structured hierarchy with clear and discriminative semantic correspondences, rather than simply increasing the number of levels.

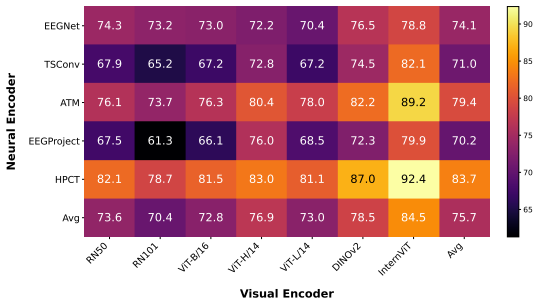


Figure 3: Neural–visual encoder combination analysis on THINGS-EEG2.

Table 5: Inference-time level selection on Top-1 and Top-5 accuracy (%).

Shallow	Inter.	Deep	Top-1	Top-5
✓	–	–	77.90	96.60
–	✓	–	86.80	98.35
–	–	✓	67.30	91.25
✓	✓	–	90.45	98.90
✓	–	✓	87.30	98.30
–	✓	✓	86.75	98.10
✓	✓	✓	**92.35**	**99.20**

Contribution of Different Semantic Levels. Table 5 analyzes the contribution of different semantic levels during inference. Among individual levels, the intermediate level performs best, while the shallow and deep levels also provide non-trivial retrieval accuracy. Aggregating all three levels achieves the highest accuracy, with 92.35% Top-1 and 99.20% Top-5, exceeding the best individual level by 5.55 percentage points in Top-1 accuracy. These results support the complementarity of the matching cues distributed across the learned hierarchies.

Effect of Semantic-Level Head Grouping. To isolate the contribution of head grouping in HPCT, we construct a parameter-matched shared-representation control. In this variant, three independent level-specific projection heads receive the same full neural representation without disjoint attention-head grouping. The remaining framework and supervision are unchanged, allowing us to examine whether separate head groups provide benefits beyond multiple supervised projections. For this analysis, Corr. denotes the Pearson correlation between the vectorized level-specific retrieval-score matrices, averaged across level pairs and subjects.

Table 6 shows that partitioning improves Top-1 accuracy by 1.70 percentage points under matched parameter counts. The partitioned model also exhibits lower inter-branch similarity correlation, suggesting more differentiated matching patterns across semantic levels. These results support allo-

cating separate representational capacity within HPCT, complementing the level-specific supervision provided by HMCL.

Effect of Information Propagation Direction. Table 7 compares four propagation directions under matched model capacity and multi-level supervision. Forward propagation achieves the best Top-1 and Top-5 accuracy, outperforming middle-out, middle-in, and reverse propagation by 2.30, 8.30, and 15.95 percentage points in Top-1 accuracy, respectively. These results demonstrate that propagation direction materially affects retrieval and support the shallow-to-deep organization adopted by PDHC, showing that progressively accumulating information across ordered semantic levels is a key contributor to its effectiveness.

Table 6: Parameter-matched head-grouping comparison on retrieval accuracy (%) and inter-branch similarity correlation.

Setting	Top-1	Top-5	Corr.
Shared	90.65	98.70	0.8314
Partitioned	**92.35**	**99.20**	0.7751

Table 7: Effect of PDHC propagation direction on retrieval accuracy (%).

Direction	Flow	Top-1	Top-5
Forward	$s \to m \to d$	**92.35**	**99.20**
Middle-out	$m \to s, d$	90.05	98.90
Middle-in	$s, d \to m$	84.05	96.90
Reverse	$d \to m \to s$	76.40	94.40

4.4 Spatial and Temporal Attention across Semantic Levels

To compare spatial and temporal differentiation across neural levels, we analyze normalized CLS attention under channel and temporal tokenization. Channel attention is aggregated over parietal (P), parieto-occipital (PO), and occipital (O) groups across all ten THINGS-EEG2 subjects; the temporal variant uses 50 non-overlapping five-sample tokens containing all 17 channels.

Table 8: Spatial and temporal attention distributions (%) across semantic levels. Parentheses indicate the attention assigned to the highest-weight temporal patch.

(a) Spatial attention distributions.

Region	Shallow	Intermediate	Deep	Δ	Consistency
Parietal	22.56	22.82	30.67	$+8.11$	10/10 increased
Parieto-occipital	34.88	36.27	38.36	$+3.48$	8/10 increased
Occipital	42.56	40.92	30.97	-11.59	9/10 decreased

(b) Temporal attention distributions.

Level	Top time patch	0–50	50–100	100–150	150–200	200–250	Center
Shallow	025–030 (14.52)	59.30	34.63	4.12	1.31	0.64	50.0
Intermediate	025–030 (19.79)	62.04	30.88	4.55	1.64	0.88	51.1
Deep	025–030 (17.61)	55.93	33.92	6.16	2.45	1.54	55.8

Table 8(a) shows clear spatial differentiation: from the shallow to the deep level, occipital attention decreases by 11.59 percentage points, while parietal and parieto-occipital attention increase by 8.11 and 3.48 points, respectively. The parietal increase and occipital decrease occur in 10/10 and 9/10 subjects. In contrast, Table 8(b) shows broadly shared temporal attention, with all levels peaking at samples 25–30 and concentrating most attention within the first 100 samples.

Together, these results indicate that hierarchy-related differences are more clearly expressed in spatially distributed channel-wise response patterns than in temporally segmented patterns. This finding supports HPCT's use of sensor channels as tokens to capture inter-channel dependencies and organize the resulting representations into a neural representational hierarchy.

5 Conclusion

We presented NIA, which extends visual neural decoding from alignment at a selected semantic level to structured correspondence between the neural and visual representational hierarchies. HPCT organizes neural representations into a hierarchy, HMCL establishes structured correspondence between the neural and visual hierarchies, and PDHC models progressive shallow-to-deep information propagation within both hierarchies before alignment. Experiments on THINGS-EEG2 and THINGS-MEG demonstrate accurate and robust zero-shot brain-to-image retrieval across EEG and MEG under both intra- and inter-subject settings.

Reproducibility Statement

We describe the model formulation, data preprocessing, and evaluation protocols in the main text. Implementation and training details are provided in Appendix A, and complete subject-wise results across five random seeds are reported in Appendix C.

AI use statement

Generative AI tools were used to assist with language editing, literature search, and the organization and presentation of the manuscript. All technical claims, citations, implementation details, experimental results, and interpretations were independently verified by the authors. No data or experimental results were generated using AI tools. The authors made all final research decisions and take full responsibility for the content of this work.

References

Ujwal Chaudhary, Niels Birbaumer, and Ander Ramos-Murguialday. Brain–computer interfaces for communication and rehabilitation. *Nature reviews neurology*, 12(9):513–525, 2016.

Hongzhou Chen, Lianghua He, Yihang Liu, and Longzhen Yang. Visual neural decoding via improved visual-eeg semantic consistency. *arXiv preprint arXiv:2408.06788*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455–462, 2014.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):27755, 2016.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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.

Yang Du, Siyuan Dai, Yonghao Song, Paul M Thompson, Haoteng Tang, and Liang Zhan. Deep models, shallow alignment: Uncovering the granularity mismatch in neural decoding. *arXiv preprint arXiv:2601.21948*, 2026.

Lawrence Ashley Farwell and Emanuel Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6):510–523, 1988.

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.

Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Umut Güçlü and Marcel AJ Van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2):413, 1993.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10):e0223792, 2019.

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.

Leigh R Hochberg, Mijail D Serruya, Gerhard M Friehs, Jon A Mukand, Maryam Saleh, Abraham H Caplan, Almut Branner, David Chen, Richard D Penn, and John P Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099):164–171, 2006.

Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685, 2005.

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.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

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.

Seo-Hyun Lee, Minji Lee, and Seong-Whan Lee. Neural decoding of imagined speech and visual imagery as intuitive paradigms for bci communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(12):2647–2659, 2020.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *Advances in Neural Information Processing Systems*, 2024a.

Yueyang Li, Zijian Kang, Shengyu Gong, Wenhao Dong, Weiming Zeng, Hongjie Yan, Wai Ting Siok, and Nizhuan Wang. Neural-mcrl: Neural multimodal contrastive representation learning for eeg-based visual decoding. *arXiv preprint arXiv:2412.17337*, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Zitong Lu, Yile Wang, and Julie D Golomb. Achieving more human brain-like vision via human eeg representational alignment. *Communications Biology*, 2026.

Dennis J McFarland and Jonathan R Wolpaw. Brain-computer interfaces for communication and control. *Communications of the ACM*, 54(5):60–66, 2011.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Christoph M Michel and Micah M Murray. Towards the utilization of eeg as a brain imaging tool. *Neuroimage*, 61(2):371–385, 2012.

Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.

Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *sensors*, 12(2):1211–1279, 2012.

Paul L Nunez and Ramesh Srinivasan. *Electric fields of the brain: the neurophysics of EEG*. Oxford university press, 2006.

Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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, 2021.

Nona Rajabi, Antônio H Ribeiro, Miguel Vasco, Farzaneh Taleb, Mårten Björkman, and Danica Kragic. Human-aligned image models improve visual decoding from the brain. In *Forty-second International Conference on Machine Learning*, 2025.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *The Twelfth International Conference on Learning Representations*, 2024.

Jacques J Vidal et al. Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering*, 2(1):157–180, 1973.

Jiyuan Wang, Li Zhang, Haipeng Lin, Qile Liu, Gan Huang, Ziyu Li, Zhen Liang, and Xia Wu. Neuroclip: Brain-inspired prompt tuning for eeg-to-image multimodal contrastive learning. *arXiv preprint arXiv:2511.09250*, 2025.

Jonathan R Wolpaw, Dennis J McFarland, Gregory W Neat, and Catherine A Forneris. An eeg-based brain-computer interface for cursor control. *Electroencephalography and clinical neurophysiology*, 78(3):252–259, 1991.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2246–2257, 2025.

Lukun Wu, Jie Li, Ziqi Ren, Kaifan Zhang, and Xinbo Gao. Shrinking the teacher: An adaptive teaching paradigm for asymmetric eeg-vision alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17859–17867, 2026.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Yuanyuan Yao, Wout De Swaef, Simon Geirnaert, and Alexander Bertrand. Eeg-based decoding of selective visual attention in superimposed videos. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Zesheng Ye, Lina Yao, Yu Zhang, and Sylvia Gustin. Self-supervised cross-modal visual retrieval from brain activities. *Pattern Recognition*, 145:109915, 2024.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

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.

Wenjiang Zhang, Sifeng Wang, Yuwei Su, Xinyu Li, Chen Zhang, and Suyu Zhong. Neurobridge: Bio-inspired self-supervised eeg-to-image decoding via cognitive priors and bidirectional semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18028–18036, 2026.

Lei Zhao, Yao Liu, Jing'ao Gao, Peng Ding, Fan Wang, Anmin Gong, Wenya Nan, Yunfa Fu, and Tianwen Li. Visual imagery-based brain-computer interaction paradigms and neural encoding and decoding. *IEEE Transactions on Human-Machine Systems*, 2025.

Jiawen Zheng, Haonan Jia, Ming Li, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. In *The Fourteenth International Conference on Learning Representations*, 2026.

Appendix Overview

Section	Content
A	**Implementation Details**
B	**Electrode Channel Distribution and Visual Processing Hierarchy**
B.1	Neuroscientific Basis of Hierarchical Visual Processing
B.2	Channel Distribution and Visual Processing Hierarchy
C	**Additional Stability Analysis**
C.1	EEG Intra-Subject Stability
C.2	EEG Inter-Subject Stability
C.3	MEG Intra-Subject Stability
C.4	MEG Inter-Subject Stability
D	**Fully Expanded Neural and Visual Encoder Combination Analysis**
E	**Additional Experimental Analyses**
E.1	Controlled Baseline Reproductions
E.2	Multi-Seed Component Ablations
E.3	Hierarchy Configuration Analysis
E.4	Tokenization and Temporal Diagnostics
E.5	Complementarity Across Visual Depths
E.6	Computational Efficiency
E.7	Effect of EEG Augmentation
E.8	UMAP Visualization of Hierarchical Representations
F	**More Retrieval Visualization**
G	**Limitations**

A Implementation Details

All experiments are implemented in PyTorch (Paszke et al., 2019) and conducted on NVIDIA GeForce RTX 3090 GPUs. We evaluate NIA under four settings: EEG intra-subject, EEG inter-subject, MEG intra-subject, and MEG inter-subject.

For EEG preprocessing, the signals are downsampled from 1,000 Hz to 250 Hz. Epochs are extracted from -200 to $1,000$ ms relative to stimulus onset, with the interval from -200 to 0 ms used for baseline correction. The model takes the post-stimulus interval from 0 to $1,000$ ms, containing 250 time points. Multivariate noise normalization is estimated exclusively from the training data and then applied to both the training and test data. For each stimulus, its four training repetitions or 80 test repetitions are averaged to obtain a single neural sample. For MEG, we use the provided signals sampled at 200 Hz and retain the post-stimulus interval from 0 to $1,000$ ms, containing 200 time points. No additional baseline correction or amplitude normalization is applied. The 12 repetitions of each MEG test image are averaged for retrieval evaluation.

For the visual hierarchy, following (Du et al., 2026), we adopt InternViT-6B-448px-V1-5 (Chen et al., 2024b) as the frozen visual foundation model. We extract hidden states from the 15th, 30th, and 45th layers of InternViT, corresponding approximately to one-third, two-thirds, and the final depth of the visual encoder. For each selected layer, the CLS token is first ℓ_2-normalized and then mapped into a 512-dimensional contrastive embedding space using an independent trainable linear projection.

For the neural hierarchy, we use the proposed Hierarchy-Partitioned Channel Transformer (HPCT). Following prior work (Zhang et al., 2026), we use 17 vision-related channels in the EEG intra-subject setting and all 63 channels in the EEG inter-subject setting. For MEG, we select 84 vision-related channels following (Du et al., 2026). Each neural channel is treated as a token, and its complete post-stimulus time series is mapped to a 252-dimensional token embedding through a shared linear projection. A learnable 252-dimensional CLS token is prepended, and learnable channel positional embeddings are added to the resulting token sequence.

HPCT employs a hierarchy-partitioned Transformer encoder to process the channel-token sequence. The encoder uses 36 attention heads, which are evenly divided into three non-overlapping groups corresponding to the shallow, intermediate, and deep semantic levels. All groups operate on the shared channel-token representations, while each group uses an independent output projection to produce its level-specific representation. The three group outputs are processed separately and are not concatenated across semantic levels. Their corresponding CLS representations are subsequently mapped into the shared 512-dimensional contrastive embedding space through independent projection heads.

PDHC applies parallel gated shallow-to-deep propagation to the neural and visual representational hierarchies before contrastive alignment. The two hierarchies use the same cascade structure with separate learnable parameters for the neural and visual cascades. We perform two propagation steps, allowing information from the shallow level to reach the deep level through the intermediate level. Each adjacent level pair is equipped with a learnable linear transformation and a sigmoid gate. At each step, all three level-specific representations are normalized by level-specific LayerNorm, while the intermediate and deep representations additionally incorporate gated, linearly transformed information from the preceding level through residual addition. The gate logits are initialized to -3, corresponding to an initial gate value of $\sigma(-3) \approx 0.047$. All PDHC parameters and gates are optimized jointly with the other trainable components of NIA.

For multi-level contrastive learning, we independently compute a symmetric neural-to-visual and visual-to-neural contrastive loss at each semantic level and average the three regularized level-specific losses, following Eq. (6). The level-specific temperatures τ_k are learnable. They are initialized to 0.07, constrained to the interval $[0.03, 0.20]$ through a sigmoid parameterization, and regularized toward their initial values.

We train the model using AdamW (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 10^{-4}. The batch sizes are 1,024 for EEG and 64 for MEG. The learning rate and number of epochs are set to 8×10^{-4} and 200 for EEG, and 4×10^{-4} and 100 for MEG. We use a constant learning rate without a learning-rate scheduler and do not apply neural-signal augmentation. The main results are averaged over five random seeds, from 2024 to 2028. Unless otherwise specified, ablation and diagnostic experiments use seed 2026.

During evaluation, the neural and visual representations at each semantic level are independently ℓ_2-normalized, and retrieval is performed using cosine similarity over the 200-way candidate pool. The final retrieval score is obtained by equally averaging the shallow-, intermediate-, and deep-level similarity matrices rather than concatenating their representations. We report Top-1 and Top-5 accuracy for each semantic level and for their equally weighted fusion.

B Electrode Channel Distribution and Visual Processing Hierarchy

B.1 Neuroscientific Basis of Hierarchical Visual Processing

Visual object recognition is widely understood as a hierarchical process rather than a single-stage mapping from retinal input to object identity. Along the visual pathway, information is progressively transformed from low-level visual features, such as local edges, orientations, contrast, and texture patterns, into increasingly abstract object-level and semantic representations. Classical studies of the primate visual system describe the visual cortex as a distributed hierarchy of interconnected areas, where information flows from early visual regions toward higher-level occipito-temporal and inferotemporal regions (Felleman & Van Essen, 1991; DiCarlo et al., 2012). This hierarchical organization provides the neuroscientific basis for assuming that visually evoked neural responses may contain information at multiple semantic levels.

From a functional perspective, early visual areas are more closely related to low-level perceptual properties, while higher-level ventral visual regions are associated with more invariant object representations. As visual processing proceeds, neural representations become less tied to local image details and more related to object identity and category-level structure. This view is also consistent with computational models of the visual brain, where different layers of deep neural networks have

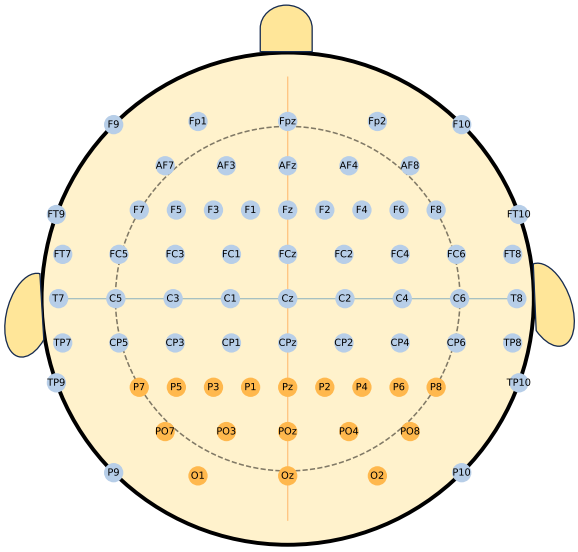


Figure 4: **Distribution of EEG electrodes used in the intra-subject setting.** The highlighted posterior electrodes correspond to the 17 vision-related channels used for EEG visual decoding, covering occipital, parieto-occipital, and parietal scalp regions. These channels provide a compact sensor-space view of visually evoked responses, while the remaining electrodes illustrate the full 10–10 layout.

been shown to correspond to different stages of visual processing and to capture representations with increasing semantic abstraction (Yamins et al., 2014; Cichy et al., 2016; Güçlü & Van Gerven, 2015).

Temporal evidence from M/EEG further supports this hierarchical view. Because EEG and MEG provide millisecond-level temporal resolution, they can reveal the rapid evolution of visual representations after stimulus onset. Prior studies have shown that human object recognition unfolds within the first few hundred milliseconds, with earlier responses carrying more image-specific visual information and later responses reflecting more abstract category- or object-level information (Cichy et al., 2014; 2016). Therefore, visually evoked EEG/MEG signals should not be regarded as containing only a single semantic level. Instead, they may reflect a mixture of low-level, intermediate-level, and high-level visual information distributed over time and across sensors.

These findings motivate both inter-hierarchy correspondence and intra-hierarchy progressive information propagation in NIA. Correspondence across multiple semantic levels accounts for the hierarchical organization of neural and visual representations, while shallow-to-deep propagation incorporates their progressive dependencies into alignment.

B.2 Channel Distribution and Visual Processing Hierarchy

EEG and MEG sensor-space channel distribution. The spatial distribution of EEG and MEG channels provides an important sensor-space basis for analyzing visually evoked neural responses. In THINGS-EEG2, EEG signals were recorded using a 64-channel EASYCAP system following the standard 10-10 electrode layout (Gifford et al., 2022). Following prior preprocessing protocols, the intra-subject EEG setting focuses on 17 vision-related electrodes over occipital, parieto-occipital, and parietal scalp regions:

$$\mathrm{O1}, \mathrm{Oz}, \mathrm{O2}, \mathrm{PO7}, \mathrm{PO3}, \mathrm{POz}, \mathrm{PO4}, \mathrm{PO8}, \mathrm{P7}, \mathrm{P5}, \mathrm{P3}, \mathrm{P1}, \mathrm{Pz}, \mathrm{P2}, \mathrm{P4}, \mathrm{P6}, \mathrm{P8}.$$

These electrodes can be roughly grouped according to their scalp locations. The O-family electrodes are located over the occipital region and are most directly related to early visual responses. The PO-family electrodes are located over parieto-occipital regions and can reflect responses from posterior visual areas as well as interactions between occipital and parietal regions. The P-family electrodes are located over parietal scalp regions and can capture more distributed posterior activity, including

visual-spatial and higher-level object-related responses. Therefore, this 17-channel subset provides a compact posterior sensor set for visual decoding.

However, EEG channels should not be interpreted as direct measurements of isolated cortical areas. Due to volume conduction, scalp EEG signals are spatially blurred mixtures of neural activity generated by distributed cortical sources. A single electrode may receive contributions from multiple visual and non-visual cortical regions, and activity from the same cortical source may spread to multiple electrodes. Therefore, the relationship between EEG channels and brain areas is not one-to-one. Instead, channel-wise EEG patterns should be interpreted as distributed sensor-space observations that reflect mixed projections of visually evoked neural activity.

For THINGS-MEG, the original recordings were acquired using a whole-head MEG system with dense sensor coverage (Hebart et al., 2023). MEG provides high temporal resolution measurements over a large sensor array, and the released THINGS-MEG data contains hundreds of sensor channels covering occipital, temporal, parietal, frontal, and central regions. In contrast to EEG, MEG is less affected by skull conductivity, but MEG sensors also measure extracranial magnetic fields generated by distributed cortical currents. Thus, MEG channels also cannot be directly equated with individual cortical regions without source localization. Nevertheless, posterior MEG sensors, especially occipital, temporal, and parietal sensor groups, are particularly relevant for visual object processing and provide useful sensor-space information for visual neural decoding.

Relation to hierarchical visual processing. The posterior distribution of visual EEG/MEG sensors is consistent with the hierarchical organization of human visual processing. Visual object recognition unfolds through multiple processing stages, transforming low-level visual features into increasingly abstract object-level and semantic representations (Gifford et al., 2022). Early visual responses are primarily associated with occipital visual areas, while later and more abstract object-related representations involve higher-level ventral and lateral occipito-temporal regions. THINGS-data also reports functionally selective visual regions and large-scale spatial and temporal patterns related to object processing, including early visual areas and category-selective regions (Hebart et al., 2023).

From the sensor-space perspective, this hierarchy implies that posterior EEG/MEG responses may contain mixed information from multiple stages of the visual pathway. Occipital sensors are more closely related to early visual responses, parieto-occipital sensors may reflect intermediate posterior visual activity, and temporal/parietal sensor patterns may carry information related to higher-level object representations. At the same time, because EEG and MEG sensors measure mixtures of distributed neural sources, hierarchical information is not isolated within a single channel or a single sensor group. Instead, it is expressed as spatially distributed response patterns across multiple channels.

This provides a motivation for channel-based neural encoding. Rather than treating each channel independently or focusing only on temporal dynamics, our channel-tokenized neural encoder treats channels as tokens and uses self-attention to capture inter-channel dependencies. In this way, the model can learn how distributed posterior sensor patterns jointly reflect visually evoked neural activity. Since these sensor-space patterns may integrate signals from early, intermediate, and higher-level stages of visual processing, modeling inter-channel dependencies provides a useful basis for extracting hierarchy-relevant neural representations. HPCT organizes the encoded channel information into a neural representational hierarchy, whose semantic correspondence with the visual hierarchy is learned through HMCL.

C Additional Stability Analysis

To further evaluate the robustness of NIA, we report complete seed-level and subject-level results under four settings: EEG intra-subject, EEG inter-subject, MEG intra-subject, and MEG inter-subject. For each setting, we provide the subject-wise Top-1/Top-5 accuracy under five random seeds, together with the mean and standard deviation computed across the five seeds. The standard deviation is computed using the five seed-level results for each subject and for the overall average.

C.1 EEG INTRA-SUBJECT STABILITY

Table 9 reports the complete EEG intra-subject results across five random seeds. NIA achieves an overall average of 92.14% Top-1 and 99.03% Top-5 accuracy. The standard deviation of the overall average is only 0.25 for Top-1 and 0.12 for Top-5, indicating that the intra-subject EEG performance is highly stable across random initializations.

Table 9: Complete seed-level stability results on THINGS-EEG2 under the intra-subject setting. Each entry reports Top-1 / Top-5 accuracy (%). Mean and Std are computed across the five random seeds.

Seed	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	Overall Avg
2024	94.0 / 99.5	92.0 / 100.0	93.5 / 99.0	88.5 / 98.5	89.5 / 97.5	94.0 / 98.0	91.5 / 99.5	93.0 / 100.0	90.5 / 99.0	96.0 / 99.5	92.25 / 99.05
2025	93.5 / 99.0	90.5 / 99.0	94.5 / 100.0	87.5 / 98.0	86.5 / 97.0	94.0 / 99.0	90.5 / 100.0	94.0 / 100.0	90.5 / 99.0	98.0 / 100.0	91.95 / 99.10
2026	94.5 / 99.5	92.5 / 99.0	95.0 / 99.5	88.0 / 98.5	87.5 / 98.0	93.0 / 98.5	90.5 / 99.5	94.5 / 100.0	91.0 / 100.0	97.0 / 99.5	92.35 / 99.20
2027	94.5 / 99.0	90.5 / 98.0	95.5 / 100.0	88.0 / 97.5	88.0 / 97.5	93.5 / 99.0	91.0 / 99.0	95.0 / 100.0	91.0 / 99.0	97.0 / 99.5	92.40 / 98.85
2028	93.5 / 99.5	90.0 / 100.0	94.0 / 98.5	88.0 / 98.5	87.0 / 96.5	93.5 / 99.5	90.5 / 100.0	93.5 / 100.0	90.0 / 97.5	97.5 / 99.5	91.75 / 98.95
Mean	94.00 / 99.30	91.10 / 99.20	94.50 / 99.40	88.00 / 98.20	87.70 / 97.30	93.60 / 98.80	90.80 / 99.60	94.00 / 100.00	90.60 / 98.90	97.10 / 99.60	92.14 / 99.03
Std	0.45 / 0.24	0.97 / 0.75	0.71 / 0.58	0.32 / 0.40	1.03 / 0.51	0.37 / 0.51	0.40 / 0.37	0.71 / 0.00	0.37 / 0.80	0.66 / 0.20	0.25 / 0.12

The results show that NIA is highly stable in the EEG intra-subject setting. Across the five random seeds, the overall Top-1 accuracy varies only from 91.75% to 92.40%, and Top-5 accuracy remains close to 99%. This suggests that the reported intra-subject performance is not caused by a favorable random initialization.

C.2 EEG INTER-SUBJECT STABILITY

Table 10 reports the complete EEG inter-subject results across five random seeds. NIA achieves an overall average of 31.68% Top-1 and 59.40% Top-5 accuracy. The overall seed-level standard deviations are 0.98 and 0.96 for Top-1 and Top-5, respectively. Compared with the intra-subject setting, the inter-subject setting shows larger variation across subjects, reflecting the difficulty of cross-subject generalization.

Table 10: Complete seed-level stability results on THINGS-EEG2 under the inter-subject setting. Each entry reports Top-1 / Top-5 accuracy (%). Mean and Std are computed across the five random seeds.

Seed	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	Overall Avg
2024	34.5 / 64.5	43.0 / 76.5	22.5 / 48.0	30.0 / 57.5	21.5 / 50.5	23.5 / 52.0	29.5 / 54.5	32.0 / 59.0	22.0 / 49.5	44.5 / 77.0	30.30 / 58.90
2025	38.0 / 74.0	40.5 / 72.0	23.0 / 51.0	28.5 / 57.0	24.5 / 45.0	26.0 / 54.0	27.0 / 49.5	35.0 / 61.0	26.0 / 49.5	45.5 / 73.5	31.40 / 58.65
2026	41.5 / 72.5	43.5 / 73.0	28.0 / 51.5	30.0 / 63.5	24.0 / 51.0	31.5 / 57.0	31.5 / 62.5	33.0 / 59.5	24.0 / 51.5	46.5 / 70.0	33.35 / 61.20
2027	37.5 / 68.5	42.0 / 72.5	24.0 / 51.5	28.5 / 54.5	22.5 / 48.0	25.5 / 48.5	29.0 / 56.5	35.0 / 63.5	23.5 / 51.0	50.0 / 81.0	31.75 / 59.55
2028	39.0 / 59.5	41.0 / 69.5	22.0 / 53.0	31.5 / 56.0	21.0 / 49.5	27.5 / 54.0	27.5 / 54.0	33.5 / 62.5	24.5 / 53.5	48.5 / 75.5	31.60 / 58.70
Mean	38.10 / 67.80	42.00 / 72.70	23.90 / 51.00	29.70 / 57.70	22.70 / 48.80	26.80 / 53.10	28.90 / 55.40	33.70 / 61.10	24.00 / 51.00	47.00 / 75.40	31.68 / 59.40
Std	2.27 / 5.31	1.14 / 2.25	2.15 / 1.64	1.12 / 3.08	1.36 / 2.16	2.68 / 2.80	1.59 / 4.22	1.17 / 1.71	1.30 / 1.48	2.00 / 3.65	0.98 / 0.96

The EEG inter-subject results remain stable across seeds, while the subject-level differences are more pronounced. For example, S10 achieves much higher performance than several other held-out subjects, indicating that inter-subject retrieval difficulty is strongly affected by subject-specific neural response patterns.

C.3 MEG INTRA-SUBJECT STABILITY

Table 11 reports the complete MEG intra-subject results across five random seeds. NIA obtains an overall average of 50.62% Top-1 and 75.45% Top-5 accuracy, which is consistent with the rounded result reported in the main comparison table. The overall seed-level standard deviations are 0.62 for Top-1 and 0.70 for Top-5, showing stable performance across random initializations.

Table 11: Complete seed-level stability results on THINGS-MEG under the intra-subject setting. Each entry reports Top-1 / Top-5 accuracy (%). Mean and Std are computed across the five random seeds.

Seed	S01	S02	S03	S04	Overall Avg
2024	32.0 / 59.5	80.0 / 95.5	55.0 / 82.5	38.0 / 66.5	51.25 / 76.00
2025	28.0 / 58.0	81.0 / 97.0	57.0 / 81.0	38.5 / 69.5	51.12 / 76.38
2026	30.0 / 56.5	80.0 / 97.0	53.0 / 79.0	37.0 / 69.0	50.00 / 75.38
2027	26.5 / 54.0	79.5 / 96.5	57.0 / 81.0	36.0 / 69.0	49.75 / 75.12
2028	30.5 / 51.5	80.0 / 96.5	56.0 / 81.0	37.5 / 68.5	51.00 / 74.38
Mean	29.40 / 55.90	80.10 / 96.50	55.60 / 80.90	37.40 / 68.50	50.62 / 75.45
Std	1.93 / 2.85	0.49 / 0.55	1.50 / 1.11	0.86 / 1.05	0.62 / 0.70

C.4 MEG Inter-Subject Stability

Table 12 reports the complete MEG inter-subject results across five random seeds. NIA achieves an overall average of 6.20% Top-1 and 14.15% Top-5 accuracy, matching the rounded result reported in the main comparison table. Although this setting is challenging due to cross-subject generalization and the limited number of MEG subjects, the overall seed-level standard deviations remain small, with 0.39 for Top-1 and 0.46 for Top-5.

Table 12: Complete seed-level stability results on THINGS-MEG under the inter-subject setting. Each entry reports Top-1 / Top-5 accuracy (%). Mean and Std are computed across the five random seeds.

Seed	S01	S02	S03	S04	Overall Avg
2024	5.5 / 10.5	8.5 / 20.5	10.5 / 21.0	2.5 / 6.0	6.75 / 14.50
2025	5.0 / 10.0	7.5 / 18.5	8.5 / 23.0	3.0 / 5.5	6.00 / 14.25
2026	4.5 / 8.0	8.0 / 19.5	10.5 / 24.5	3.0 / 6.0	6.50 / 14.50
2027	3.5 / 12.5	7.0 / 15.5	8.5 / 22.5	3.5 / 6.5	5.62 / 14.25
2028	5.0 / 10.5	8.0 / 18.5	9.0 / 17.0	2.5 / 7.0	6.12 / 13.25
Mean	4.70 / 10.30	7.80 / 18.50	9.40 / 21.60	2.90 / 6.20	6.20 / 14.15
Std	0.68 / 1.44	0.51 / 1.67	0.92 / 2.56	0.37 / 0.51	0.39 / 0.46

D Fully Expanded Neural and Visual Encoder Combination Analysis

To further analyze whether the performance improvement comes from the proposed neural encoder or only from the visual backbone, we evaluate different combinations of neural and visual encoders on THINGS-EEG2 under the intra-subject setting. We report subject-wise Top-1 and Top-5 accuracy for each combination.

Tables 13–17 provide a fully expanded comparison of neural and visual encoder combinations. The results show that HPCT consistently achieves the best average performance across all visual backbones. Even without InternViT, HPCT remains superior to other neural encoders when paired with RN50, RN101, ViT-B/16, ViT-H/14, ViT-L/14, and DINOv2. For example, with DINOv2, HPCT achieves 86.95% Top-1 accuracy, outperforming ATM by 4.75 percentage points and TSConv by 12.45 percentage points. Similarly, with RN50, HPCT reaches 82.10% Top-1 accuracy, compared with 76.10% for ATM and 67.90% for TSConv.

These results indicate that the gain of NIA does not solely come from using InternViT as a stronger visual backbone. InternViT indeed provides the strongest visual representations and further improves performance across neural encoders. However, the consistent advantage of HPCT across diverse visual backbones demonstrates that the proposed channel-tokenized neural encoder provides a robust neural representation learning advantage. Therefore, the overall improvement comes from both high-quality multi-level visual features and the hierarchy-aware neural encoding and alignment design.

Table 13: Fully expanded results of **HPCT** with different visual backbones on THINGS-EEG2 under the intra-subject setting. Each entry reports Top-1 / Top-5 accuracy (%).

Visual Backbone	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	Avg
RN50	87.0 / 98.5	84.0 / 98.0	83.5 / 96.5	79.5 / 94.5	73.5 / 92.5	84.5 / 97.0	79.5 / 98.0	82.0 / 96.5	78.5 / 96.0	89.0 / 98.5	82.10 / 96.60
RN101	81.5 / 98.0	81.0 / 95.5	80.0 / 97.0	75.5 / 94.5	68.5 / 92.0	84.0 / 99.0	76.5 / 95.0	80.0 / 98.5	71.5 / 94.5	88.5 / 98.0	78.70 / 96.20
ViT-B/16	84.5 / 98.5	83.0 / 95.5	80.0 / 96.0	78.0 / 96.5	72.0 / 93.5	85.5 / 99.5	82.0 / 97.0	82.5 / 98.5	79.5 / 96.5	88.5 / 99.0	81.55 / 97.05
ViT-H/14	84.0 / 96.5	83.0 / 97.0	85.0 / 98.5	79.0 / 97.5	71.0 / 92.0	88.0 / 99.0	80.0 / 95.5	90.0 / 99.0	81.0 / 96.5	89.5 / 98.5	83.05 / 97.00
ViT-L/14	82.5 / 96.0	80.0 / 95.5	79.5 / 97.5	74.0 / 92.5	73.0 / 94.0	85.5 / 99.0	81.0 / 96.5	84.5 / 95.5	84.0 / 97.5	87.0 / 97.0	81.10 / 96.10
DINOv2	90.0 / 99.0	85.5 / 97.5	86.5 / 99.0	82.0 / 96.5	81.0 / 94.0	90.0 / 99.5	89.5 / 98.0	87.5 / 98.5	85.5 / 96.5	92.0 / 99.5	86.95 / 97.80
InternViT	**94.5 / 99.5**	**92.5 / 99.0**	**95.0 / 99.5**	**88.0 / 98.5**	**87.5 / 98.0**	**93.0 / 98.5**	**90.5 / 99.5**	**94.5 / 100.0**	**91.0 / 100.0**	**97.0 / 99.5**	**92.35 / 99.20**

Table 14: Fully expanded results of **ATM** with different visual backbones on THINGS-EEG2 under the intra-subject setting. Each entry reports Top-1 / Top-5 accuracy (%).

Visual Backbone	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	Avg
RN50	79.0 / 96.5	77.5 / 93.5	73.5 / 93.5	71.0 / 92.0	69.0 / 90.5	83.5 / 97.5	73.5 / 94.5	80.0 / 97.5	72.5 / 91.5	81.5 / 97.0	76.10 / 94.40
RN101	74.0 / 95.5	71.0 / 94.0	71.0 / 92.5	69.0 / 91.5	61.5 / 87.5	84.5 / 97.5	71.5 / 92.5	83.0 / 94.5	71.5 / 93.5	80.0 / 98.0	73.70 / 93.70
ViT-B/16	80.0 / 96.5	79.0 / 91.5	76.0 / 95.0	69.5 / 96.0	65.5 / 88.5	83.0 / 96.5	74.0 / 95.0	80.5 / 97.5	73.0 / 95.5	83.0 / 98.0	76.35 / 95.00
ViT-H/14	80.5 / 96.5	79.0 / 98.0	78.0 / 97.5	77.0 / 97.5	69.0 / 92.0	87.0 / 98.5	79.0 / 94.5	87.5 / 99.0	77.0 / 95.0	90.0 / 98.0	80.40 / 96.65
ViT-L/14	80.0 / 94.0	77.0 / 94.0	73.0 / 96.0	71.0 / 95.5	71.5 / 93.0	82.5 / 98.0	81.0 / 96.5	82.0 / 96.5	76.5 / 94.5	85.0 / 96.0	77.95 / 95.40
DINOv2	86.5 / 97.5	79.5 / 95.0	74.5 / 97.5	77.5 / 96.0	79.0 / 92.0	88.0 / 99.0	81.0 / 99.5	85.0 / 96.0	81.0 / 97.5	90.0 / 99.0	82.20 / 96.90
InternViT	93.0 / 99.5	87.5 / 99.0	87.5 / 98.5	86.0 / 98.0	86.0 / 96.5	92.0 / 99.5	88.0 / 99.0	94.5 / 100.0	82.5 / 95.5	95.0 / 99.0	89.20 / 98.45

Table 15: Fully expanded results of **EEGNet** with different visual backbones on THINGS-EEG2 under the intra-subject setting. Each entry reports Top-1 / Top-5 accuracy (%).

Visual Backbone	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	Avg
RN50	70.0 / 93.0	73.0 / 93.5	73.5 / 92.5	71.0 / 92.5	70.0 / 90.5	79.5 / 96.0	73.0 / 94.0	78.0 / 96.5	72.5 / 90.0	83.0 / 97.5	74.35 / 93.60
RN101	69.0 / 90.0	73.0 / 91.0	70.5 / 95.5	69.5 / 93.5	68.0 / 89.0	80.0 / 96.0	73.5 / 95.0	79.5 / 96.5	70.0 / 95.0	79.5 / 96.5	73.25 / 93.80
ViT-B/16	66.5 / 89.0	72.0 / 92.0	70.5 / 92.0	68.5 / 90.5	65.0 / 90.0	80.0 / 96.5	74.0 / 94.5	77.0 / 97.0	76.0 / 93.5	81.0 / 97.5	73.05 / 93.25
ViT-H/14	63.5 / 89.0	67.5 / 93.5	73.0 / 93.0	69.5 / 94.0	65.0 / 87.0	80.0 / 95.5	72.0 / 91.5	79.5 / 97.0	71.0 / 91.5	81.0 / 95.5	72.20 / 92.75
ViT-L/14	62.0 / 86.5	68.0 / 91.0	68.5 / 88.5	65.0 / 88.5	63.0 / 89.5	81.0 / 96.0	70.5 / 91.5	77.0 / 94.5	72.5 / 92.5	76.5 / 95.0	70.40 / 91.35
DINOv2	73.0 / 91.0	74.0 / 96.0	75.0 / 94.0	69.5 / 95.5	68.5 / 91.5	83.0 / 96.0	80.5 / 95.5	80.0 / 97.0	78.5 / 96.0	83.0 / 98.0	76.50 / 95.05
InternViT	78.0 / 93.5	76.0 / 94.5	78.0 / 95.0	68.5 / 93.5	75.0 / 94.0	87.0 / 97.5	78.5 / 97.5	86.0 / 99.5	77.5 / 96.0	83.5 / 96.0	78.80 / 95.70

Table 16: Fully expanded results of **EEGProject** with different visual backbones on THINGS-EEG2 under the intra-subject setting. Each entry reports Top-1 / Top-5 accuracy (%).

Visual Backbone	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	Avg
RN50	66.5 / 88.0	70.0 / 91.0	65.0 / 88.5	63.5 / 90.0	60.5 / 86.5	75.5 / 93.5	63.0 / 87.0	74.0 / 96.0	63.0 / 88.0	74.5 / 93.0	67.55 / 90.15
RN101	59.0 / 84.0	62.0 / 87.5	60.0 / 83.0	57.5 / 87.0	50.0 / 79.0	68.0 / 91.0	56.0 / 82.0	70.5 / 92.5	58.0 / 82.0	72.0 / 94.5	61.30 / 86.25
ViT-B/16	64.0 / 89.0	68.5 / 92.0	64.5 / 90.0	63.5 / 88.5	54.5 / 83.5	74.5 / 96.0	62.5 / 89.0	72.5 / 96.0	62.0 / 87.0	74.5 / 94.5	66.10 / 90.55
ViT-H/14	73.0 / 90.5	72.5 / 93.5	75.5 / 95.5	78.5 / 97.5	63.0 / 87.0	84.0 / 97.0	76.5 / 93.5	84.5 / 97.5	68.0 / 93.5	84.5 / 97.5	76.00 / 94.30
ViT-L/14	67.5 / 90.0	68.5 / 92.0	69.0 / 88.5	66.5 / 91.0	55.5 / 81.0	80.5 / 97.0	62.5 / 91.0	77.0 / 95.5	62.0 / 86.0	76.5 / 95.0	68.55 / 90.70
DINOv2	74.0 / 92.0	69.5 / 94.5	70.0 / 92.5	69.0 / 95.0	61.0 / 86.5	76.0 / 97.5	72.5 / 92.5	80.0 / 96.5	68.0 / 92.0	83.0 / 97.5	72.30 / 93.65
InternViT	76.5 / 94.0	78.0 / 98.0	81.0 / 97.0	76.0 / 97.0	71.5 / 92.5	86.5 / 97.5	82.5 / 97.0	86.5 / 99.5	71.5 / 93.5	89.0 / 99.0	79.90 / 96.50

Table 17: Fully expanded results of **TSConv** with different visual backbones on THINGS-EEG2 under the intra-subject setting. Each entry reports Top-1 / Top-5 accuracy (%).

Visual Backbone	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	Avg
RN50	68.5 / 93.5	64.5 / 86.0	67.5 / 89.0	65.0 / 87.5	60.0 / 86.0	79.0 / 93.0	62.0 / 86.0	75.5 / 96.5	62.0 / 88.5	75.0 / 95.0	67.90 / 90.10
RN101	66.5 / 92.5	61.0 / 90.5	66.5 / 92.0	60.0 / 85.5	55.5 / 87.0	77.5 / 94.0	59.5 / 85.0	75.5 / 94.0	58.0 / 85.5	71.5 / 94.5	65.15 / 90.05
ViT-B/16	67.5 / 95.0	62.0 / 86.5	65.0 / 91.0	63.5 / 89.0	55.5 / 82.0	79.0 / 96.5	63.0 / 91.0	75.5 / 94.5	64.5 / 89.5	76.0 / 96.5	67.15 / 91.15
ViT-H/14	70.5 / 88.0	64.5 / 89.0	74.0 / 93.5	73.5 / 94.0	62.5 / 87.0	83.5 / 98.0	68.5 / 91.5	82.5 / 98.0	66.0 / 91.5	82.0 / 97.0	72.75 / 92.75
ViT-L/14	66.0 / 91.5	59.0 / 86.5	69.5 / 91.5	65.0 / 92.0	55.5 / 87.0	76.5 / 95.5	66.0 / 90.5	77.0 / 94.5	62.0 / 90.5	75.5 / 95.0	67.20 / 91.45
DINOv2	75.5 / 94.0	68.0 / 89.5	72.0 / 95.0	70.0 / 93.5	66.0 / 91.0	85.0 / 97.0	73.5 / 93.5	83.0 / 95.5	70.0 / 92.0	82.0 / 99.0	74.50 / 94.00
InternViT	81.0 / 98.0	77.0 / 93.5	83.0 / 97.0	78.0 / 96.5	69.5 / 92.5	92.5 / 98.5	82.5 / 96.5	91.5 / 98.5	78.0 / 92.5	88.0 / 98.0	82.10 / 96.15

E Additional Experimental Analyses

E.1 Controlled Baseline Reproductions

The baseline values in the main comparison tables are taken from their respective papers, whereas NIA is reported over five seeds. For controlled comparison, we reproduce Du et al. and BrainHIVE from their papers and official repositories, matching the split, preprocessing, trial averaging, training data, 200-way candidate pool, seeds, and evaluation metrics while retaining method-specific optimization. Du et al. uses matched InternViT features, whereas BrainHIVE retains its original multi-feature design.

Table 18: Controlled comparison on THINGS-EEG2 under the intra-subject 200-way retrieval setting. Each entry reports the five-seed mean ± standard deviation (%).

Method	Visual setting	Top-1	Top-5
NIA (Ours)	InternViT V1-5	92.14 ± 0.25	99.03 ± 0.12
NIA (Ours)	InternViT V2-5	91.40 ± 0.32	98.72 ± 0.12
Du et al.	InternViT V1-5	76.16 ± 0.48	95.13 ± 0.18
Du et al.	InternViT V2-5	84.57 ± 0.44	97.68 ± 0.29
BrainHIVE	RN50 + ViT-B/32 + VAE	68.23 ± 0.41	91.18 ± 0.42

E.2 Multi-Seed Component Ablations

We rerun the central-component ablation on THINGS-EEG2 Subjects 1–2 over three seeds and on all four THINGS-MEG subjects over three seeds. Both evaluations follow the intra-subject protocol.

Table 19: Multi-seed component ablation under the intra-subject protocol. EEG results average Subjects 1–2; MEG results average all four subjects. Each entry reports three-seed mean ± standard deviation (%).

Dataset	Neural encoder	HMCL	PDHC	Top-1	Top-5
THINGS-EEG2	HPCT (Ours)	✓	✓	$\mathbf{93.00 \pm 0.50}$	$\mathbf{99.17 \pm 0.63}$
	ATM	✓	✓	89.58 ± 1.15	99.00 ± 0.25
	EEGNet	✓	✓	76.58 ± 0.38	94.17 ± 0.52
	EEGProject	✓	✓	77.75 ± 1.09	95.83 ± 0.76
	TSConv	✓	✓	78.83 ± 1.26	96.33 ± 0.63
	HPCT	✓	×	90.58 ± 0.95	99.17 ± 0.14
	HPCT (60% depth only)	–	–	86.25 ± 1.15	97.92 ± 0.38
THINGS-MEG	HPCT (Ours)	✓	✓	$\mathbf{50.79 \pm 0.69}$	75.92 ± 0.50
	ATM	✓	✓	40.79 ± 0.26	68.50 ± 1.15
	HPCT	✓	×	49.17 ± 1.22	$\mathbf{76.08 \pm 1.17}$
	HPCT (60% depth only)	–	–	42.62 ± 4.77	68.33 ± 7.00

E.3 Hierarchy Configuration Analysis

We evaluate the number of hierarchy levels on THINGS-EEG2 Subjects 1–2 under both the intra- and inter-subject protocols.

We additionally test two non-uniform four-level settings to examine the effect of visual-layer placement.

Table 20: Effect of the number of hierarchy levels on THINGS-EEG2 Subjects 1–2. Each entry reports three-seed mean ± standard deviation for Top-1 / Top-5 accuracy (%).

Levels	Intra-subject	Inter-subject
2	86.25 ± 0.66 / 98.50 ± 0.00	38.50 ± 2.22 / 67.67 ± 1.91
3	$\mathbf{93.00 \pm 0.50}$ / 99.17 ± 0.63	$\mathbf{40.33 \pm 1.94}$ / $\mathbf{71.25 \pm 1.30}$
4	92.42 ± 0.29 / $\mathbf{99.42 \pm 0.14}$	40.08 ± 1.53 / 68.75 ± 0.90
5	92.03 ± 0.52 / 99.08 ± 0.14	39.25 ± 0.66 / 69.25 ± 1.15

Table 21: Visual-layer position ablation on THINGS-EEG2 Subjects 1–2 under the intra-subject protocol. Results are three-seed means (%).

Layer positions	Top-1	Top-5
25% / 50% / 75% / 100% (uniform four-level)	92.42	**99.42**
20% / 33% / 50% / 100%	90.75	98.67
20% / 40% / 60% / 100%	92.58	99.08
Three-level configuration	**93.00**	99.17

E.4 Tokenization and Temporal Diagnostics

Keeping all other components fixed, we compare channel-wise tokenization with 50 non-overlapping temporal tokens, each containing five consecutive points from all 17 channels. We also replace only HPCT self-attention with TokenMLP.

Table 22: Tokenization and token-mixing diagnostics on THINGS-EEG2 under the intra-subject setting (%). Channel-token and temporal-patch results are three-seed means; the TokenMLP row is an additional diagnostic run.

Variant	Top-1	Top-5
Channel tokens (HPCT)	**92.18**	**99.12**
Temporal patch tokens	90.05	98.72
TokenMLP in place of self-attention	14.00	31.00

To examine whether different hierarchy levels develop distinct temporal preferences, we replace the shared tokenizer with three independent tokenizer projection layers while keeping the channel-wise self-attention and remaining architecture unchanged.

Table 23: Normalized temporal distributions of tokenizer projection weights (%). Intervals denote consecutive sample-index ranges in the 250-sample neural waveform.

Tokenizer	0–50	50–100	100–150	150–200	200–250
Shared	22.96	20.84	18.90	18.73	18.56
Shallow	23.83	20.83	18.64	18.44	18.26
Intermediate	22.22	20.54	19.19	19.03	19.02
Deep	21.56	20.89	19.34	19.23	18.97

E.5 Complementarity Across Visual Depths

We report five-seed averages for each individual visual depth and their similarity aggregation on THINGS-MEG.

Table 24: Subject-wise retrieval across visual depths on THINGS-MEG. Each entry reports five-seed average Top-1 / Top-5 accuracy (%). Gain is measured against the best single depth separately for each metric.

Subject	Shallow	Intermediate	Deep	Aggregated	Gain
S01	14.90 / 43.10	25.00 / 52.70	19.70 / 47.50	**29.40 / 55.90**	+4.40 / +3.20
S02	56.80 / 88.50	72.10 / 94.70	74.10 / 95.10	**80.10 / 96.50**	+6.00 / +1.40
S03	33.70 / 66.90	45.20 / 74.90	46.40 / 77.60	**55.60 / 80.90**	+9.20 / +3.30
S04	22.60 / 58.40	30.60 / 63.40	24.90 / 58.30	**37.40 / 68.50**	+6.80 / +5.10
Mean	32.00 / 64.22	43.23 / 71.42	41.27 / 69.62	**50.62 / 75.45**	+6.60 / +3.25

E.6 Computational Efficiency

We measure computational cost on one NVIDIA GeForce RTX 3090 GPU using Python 3.12.2, PyTorch 2.6.0, CUDA 12.6, and cuDNN 9.5.1. Frozen visual features are precomputed and cached following prior work; therefore, Table 25 reports the trainable NIA and retrieval stages and excludes this one-time preprocessing cost.

Table 25: Computational efficiency of NIA across EEG and MEG evaluation settings.

Setting	Trainable params	Peak memory	Train time / epoch	Total inference	Latency / sample	Throughput
EEG intra-subject	7.78M	842 MB	5.21 s	0.332 s	1.66 ms	602.8 samples/s
EEG inter-subject	7.78M	844 MB	29.50 s	0.325 s	1.62 ms	616.3 samples/s
MEG intra-subject	7.83M	5,103 MB	44.05 s	0.360 s	1.80 ms	556.2 samples/s
MEG inter-subject	7.83M	4,990 MB	119.18 s	0.402 s	2.01 ms	498.2 samples/s

E.7 Effect of EEG Augmentation

Although neural-signal augmentation has been used to improve neural–visual alignment (Zhang et al., 2026), NIA does not employ EEG-side or image-side augmentation during training. We additionally evaluate four common EEG augmentation strategies: additive noise, temporal smoothing, channel dropout, and temporal shifting.

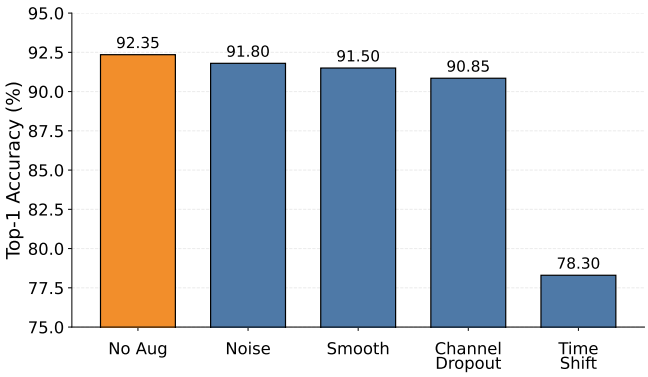
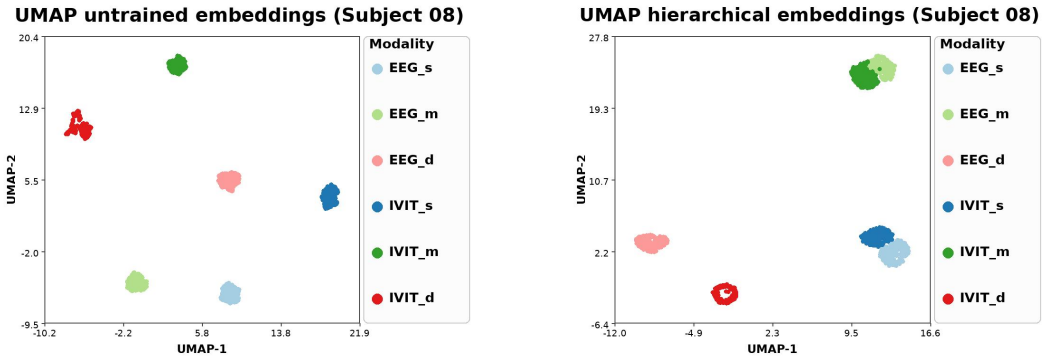


Figure 5: Top-1 retrieval accuracy (%) under different EEG augmentation strategies on THINGS-EEG2.

As shown in Figure 5, the setting without augmentation achieves the highest Top-1 accuracy of 92.35%, outperforming all evaluated augmentation strategies. These results indicate that the performance improvement of NIA does not depend on handcrafted neural-signal augmentation and further support the effectiveness of its hierarchy-aware alignment design.

E.8 UMAP Visualization of Hierarchical Representations

We use UMAP (McInnes et al., 2018) to qualitatively examine the neural and visual representation spaces before and after hierarchy-aware alignment. Before alignment refers to embeddings extracted prior to optimizing NIA's trainable alignment modules, whereas after alignment refers to embeddings produced by the fully trained model. For each condition, neural and visual embeddings from all three semantic levels are jointly projected using the same UMAP settings.



(a) Before hierarchy-aware alignment (b) After hierarchy-aware alignment

Figure 6: UMAP visualization of neural and visual representations before and after hierarchy-aware alignment.

As shown in Figure 6, neural and visual representations are more separated before alignment. After applying NIA, representations associated with corresponding semantic levels become closer, particularly at the shallow and intermediate levels. This visualization provides qualitative evidence that NIA improves neural–visual correspondence across the constituent levels of the two representational hierarchies.

F Retrieval Visualization

Figures 7–10 provide qualitative examples of zero-shot brain-to-image retrieval. Each row shows the ground-truth image and the top-5 retrieved candidates ranked by neural–visual similarity. Overall, the proposed method can consistently retrieve the correct image at the Top-1 position across diverse object categories, including vehicles, animals, food, plants, and household objects. This indicates that the learned neural representations preserve discriminative visual information and can establish reliable associations with the corresponding image representations.

Beyond exact Top-1 retrieval, the remaining top-ranked candidates are often semantically or visually related to the ground truth. For example, when the target image belongs to categories such as boats, food, animals, or household items, the retrieved alternatives frequently share similar object semantics, shapes, textures, or scene context. These results suggest that the model does not merely rely on low-level visual similarity but captures meaningful semantic cues from neural responses. At the same time, some failure-like candidates in the lower ranks show that visually similar textures or category-level associations can still introduce ambiguity, especially when different objects share similar colors, shapes, or contextual backgrounds.

Overall, these qualitative results are consistent with the quantitative retrieval performance and further support the effectiveness of hierarchy-aware neural–visual alignment. By integrating shallow, intermediate, and deep semantic cues, NIA produces a retrieval space in which the correct image is ranked highly while semantically related candidates remain close, demonstrating both discriminability and semantic consistency in zero-shot brain-to-image retrieval.

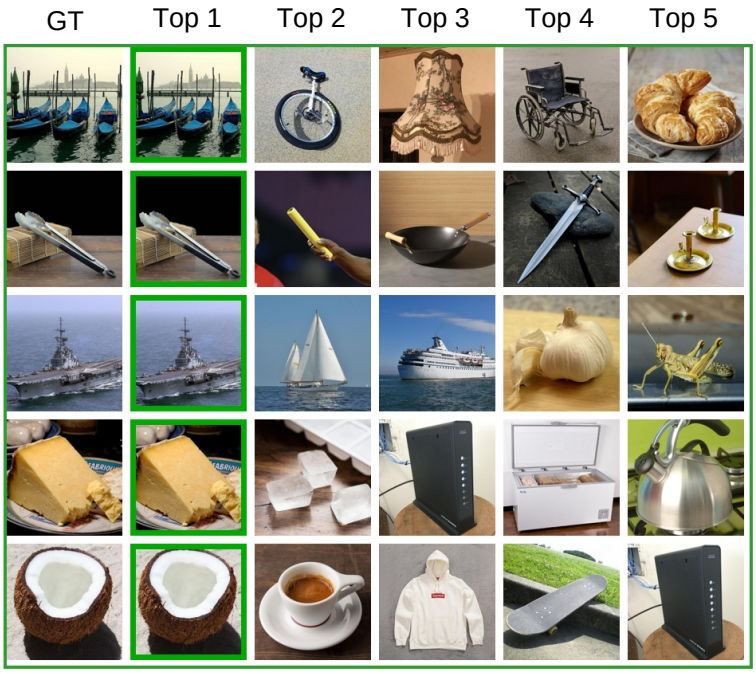


Figure 7: Additional visualization of zero-shot brain-to-image retrieval results. Each row shows the ground-truth image and the Top-5 retrieved candidates.

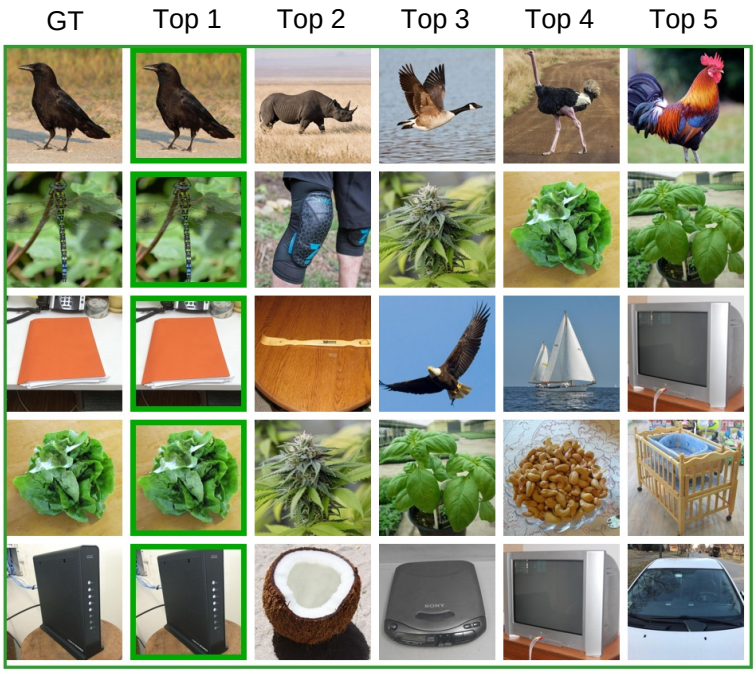


Figure 8: Additional visualization of zero-shot brain-to-image retrieval results. Each row shows the ground-truth image and the Top-5 retrieved candidates.

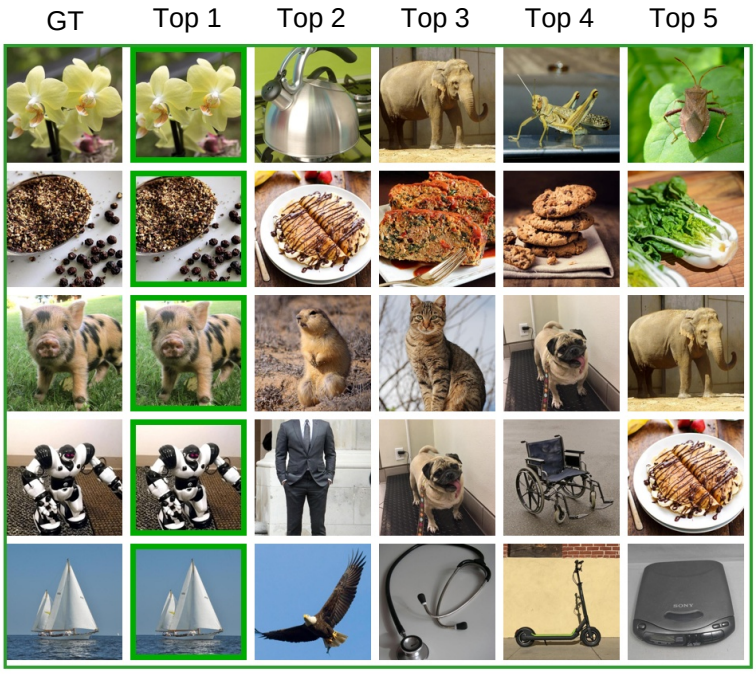


Figure 9: Additional visualization of zero-shot brain-to-image retrieval results. Each row shows the ground-truth image and the Top-5 retrieved candidates.

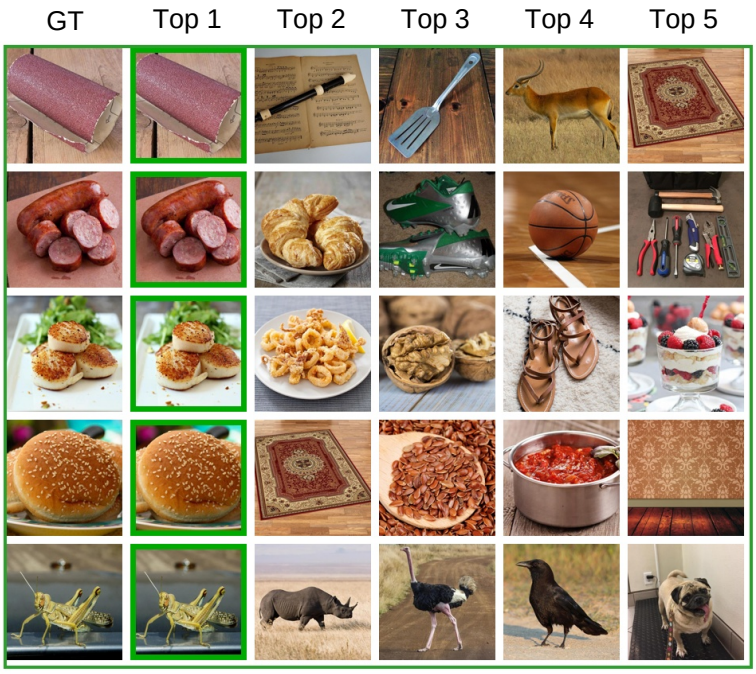


Figure 10: Additional visualization of zero-shot brain-to-image retrieval results. Each row shows the ground-truth image and the Top-5 retrieved candidates.

G LIMITATIONS

Despite its effectiveness, our method still has limitations. The correspondence between neural responses and visual representations can vary across subjects, modalities, and stimuli, which may affect the robustness of inter-hierarchy correspondence. Future work could explore more adaptive and subject-aware alignment mechanisms to better capture such variations and improve the generalization of hierarchy-aware neural–visual alignment.