# DYNAMIC VISION-BRAIN ALIGNMENT VIA LAYER-SENSITIVE TEMPORAL DYNAMICS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Continuous neural recordings provide rich temporal dynamics for visual decoding, yet conventional cross-modal alignment methods typically aggregate time-series neural signals into static feature vectors. Such static temporal aggregation neglects the temporal progression inherent in visual cognitive processing, thereby limiting the explicit separation between low-level sensory details and high-level semantics. To address this limitation, we propose Layer-Sensitive Temporal Alignment (LSTA), a progressive framework designed to model continuous temporal dynamics without static feature compression. By structuring neural recordings into a holistic representation alongside early stimulus-evoked and late residual-semantic branches, the framework effectively decouples global context from time-varying perceptual signals. LSTA couples neural response latency with visual hierarchy depth, systematically aligning early perceptual dynamics with shallow visual layers and late cognitive responses with deeper representations. Extensive experiments demonstrate the effectiveness of LSTA, achieving state-of-the-art performance in zero-shot vision-brain retrieval tasks, with Top-1 and Top-5 accuracies reaching 91.4% and 98.6% on the THINGS-EEG dataset, and 42.7% and 67.4% on the THINGS-MEG dataset, respectively.

## 1 INTRODUCTION

Understanding human visual cognitive mechanisms plays a vital role in advancing non-invasive intent decoding and neuro-inspired vision algorithms (Hassabis et al., 2017; Kriegeskorte, 2015). Achieving precise vision-brain alignment by decoding visual information directly from continuous neural activity forms the cornerstone of these applications. To capture rapid visual neural dynamics, non-invasive imaging techniques such as Magnetoencephalography (MEG) and Electroencephalography (EEG) are commonly used (He & Liu, 2008). Among them, EEG offers exceptional advantages in temporal resolution, portability, and affordability, making it highly suitable for practical visual neural decoding despite its lower signal-to-noise ratio. However, effectively overcoming this noise constraint to decode visual concepts requires robust cross-modal representation algorithms that can capture complex relationships between continuous EEG signals and visual representations.

Multimodal contrastive learning (Radford et al., 2021) has recently emerged as a pivotal framework for bridging the gap between continuous EEG signals and visual representations. Within the domain of EEG-based visual decoding, several approaches leverage text modalities as intermediary supervision to enrich the semantic capacity of learned EEG features (Du et al., 2023; Liu et al., 2026). Meanwhile, Zheng et al. (2026) incorporate ensembles of diverse visual backbones to capture multi-granularity visual features corresponding to complex EEG patterns. More recently, Tang et al. (2026) employ multi-layer vision feature fusion to integrate both low-level textures and high-level semantic details for better alignment with EEG responses.

However, despite these remarkable advances, existing frameworks routinely pool continuous neural recordings into a single static vector before alignment. This rigid temporal aggregation fails to preserve the progressive evolution of neural responses over time. Specifically, neuroscientific evidence indicates that early evoked potentials tend to emphasize low-level visual properties such as edges and textures, whereas late persistent dynamics show greater sensitivity to high-level cognitive abstractions (Cichy et al., 2014; Kietzmann et al., 2019). Averaging these distinct temporal phases into



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.

a monolithic feature vector risks representation entanglement, where fine-grained sensory patterns and abstract semantics can muddy each other during cross-modal contrastive matching.

To address this challenge, we draw inspiration from neuroscientific findings showing that visual processing unfolds along a dynamic trajectory in the visual ventral stream, transitioning from early sensory features to late abstract semantics (Cichy et al., 2014; Kietzmann et al., 2019). Inspired by this spatio-temporal progression, we introduce an inductive alignment mapping: pairing early-latency neural dynamics with shallow visual features, and associating late persistent responses with deeper semantic representations. Building upon this inductive mapping, we propose Layer-Sensitive Temporal Alignment (LSTA), a progressive framework for dynamic vision-brain alignment. Rather than squashing the temporal dimension into a static vector, LSTA explicitly preserves the temporal trajectory of continuous neural recordings. We decompose the neural dynamics into an overarching holistic representation and two functional temporal stages—early stimulus-evoked and late residual-semantic responses—to capture both global context and progressive visual features. These temporally evolving features are aligned with their corresponding depths in the visual hierarchy, establishing a structured alignment between neural response dynamics and visual feature depth. Consequently, this dynamic mapping alleviates the feature entanglement inherent in traditional static pooling, allowing shallow sensory details and deep abstract semantics to complement rather than obscure each other during cross-modal matching.

- We propose Layer-Sensitive Temporal Alignment (LSTA) to address the limitations of static temporal pooling, which collapses continuous neural recordings and incurs feature entanglement. To tackle this, LSTA explicitly structures neural dynamics into three complementary components: an overarching holistic representation for global context, alongside early stimulus-evoked and late residual-semantic branches.
- We introduce a layer-sensitive alignment mechanism that pairs neural response latency with visual feature depth, aligning early perceptual dynamics with shallow model layers and late semantic dynamics with deeper representations.
- We achieve state-of-the-art performance on benchmark THINGS-EEG and THINGS-MEG datasets, substantially outperforming existing baselines in zero-shot vision-brain retrieval tasks while offering improved representation interpretability.

## 2 Related Work

**Neural Visual Decoding** Decoding visual information from brain signals has emerged as a fundamental topic at the intersection of neuroscience, BCIs, and computer vision (Kamitani & Tong, 2005). Among various non-invasive brain recording modalities, EEG stands out due to its portability and high temporal resolution, capturing dynamic millisecond-level neural responses to visual stimuli (Makeig et al., 2002). To extract robust visual representations from noisy EEG signals, recent studies widely leverage pre-trained visual models to extract rich visual priors (Radford et al., 2021). Coupled with advances in self-supervised and multi-modal contrastive learning, brain visual decoding has achieved substantial progress, particularly in zero-shot vision-brain retrieval tasks (Li et al., 2024; Wu et al., 2025; Tang et al., 2026). Furthermore, recent works (Du et al., 2023; Liu et al., 2026) incorporate textual semantics as an auxiliary modality, facilitating effective decoupling between neural signals and visual feature spaces.

**Multi-modal Contrastive Learning** Multi-modal contrastive learning plays a pivotal role in cross-modal alignment and has achieved widespread success across diverse domain-bridging tasks, spanning vision-language pre-training (Radford et al., 2021; Jia et al., 2021; Li et al., 2022; Alayrac et al., 2022) and brain-visual cross-modal matching (Du et al., 2023; Tang et al., 2026; Liu et al., 2026). At its core, multi-modal contrastive learning pulls representations of paired cross-modal samples together while pushing unpaired samples apart in a shared embedding space (Yuan et al., 2021; Girdhar et al., 2023). These frameworks are typically optimized via InfoNCE-style noise-contrastive estimation objectives (Chen et al., 2020; He et al., 2020), with recent methodologies expanding toward self-supervised representation learning and masked contrastive paradigms (Baevski et al., 2022; Oquab et al., 2024; Tian et al., 2024). Building upon these foundational principles, recent brain–visual decoding paradigms heavily rely on contrastive frameworks to align neural activity with visual features in a shared embedding space (Wu et al., 2025; Tang et al., 2026; Liu et al., 2026; Zheng et al., 2026).

**Multi-layer Feature Fusion** Extracting and fusing multi-layer representations has long been a foundational technique to bridge low-level structural details with high-level semantic abstractions. In general computer vision, incorporating multi-scale intermediate features from deep backbones is widely adopted across self-supervised learning and generative modeling (Dosovitskiy et al., 2021; Bao et al., 2022; He et al., 2022; van den Oord et al., 2017). In terms of visual layer selection, Lin et al. (2025) categorize existing strategies into two main paradigms: similarity-based and proportion-based selection. For cross-modal alignment, multi-layer feature fusion has similarly become standard practice in vision-language (Li et al., 2022; Alayrac et al., 2022) and multi-modal architectures via layer-wise cross-attention, feature pyramids, or dynamic weighting mechanisms. Inspired by these advances, recent brain visual decoding paradigms have begun exploiting multi-layer visual features, with Zheng et al. (2026) underscoring the fusion from low-level textures to high-level semantics and Tang et al. (2026) utilizing multi-layer representations to strengthen contrastive learning.

# 3 METHOD

## 3.1 PROBLEM STATEMENT

We denote the paired brain-visual dataset as $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{V}_i)\}_{i=1}^{N}$, where $\mathbf{X}_i \in \mathbb{R}^{C\times T}$ represents $C$-channel neural recordings over duration $T$, and $\mathbf{V}_i \in \mathbb{R}^{3\times H\times W}$ is the RGB visual image. The framework maps both heterogeneous modalities into a shared embedding space, pulling matched pairs together and pushing mismatched ones apart for downstream neural decoding and retrieval tasks.

## 3.2 LAYER-SENSITIVE CROSS-MODAL ALIGNMENT

Aligning brain signals with a single visual layer is inherently insufficient, as a static or single-level representation fails to capture the multi-faceted nature of visual stimuli, which span from low-level textures to high-level semantics (Zeiler & Fergus, 2014; Oquab et al., 2024). To overcome this, we leverage pre-trained visual encoders (such as CLIP or ResNet), whose hierarchical architectures naturally embed information across varying depths—shallow layers encode fine-grained visual appearances, whereas deep layers abstract high-level semantic concepts. To implement this layer-wise feature integration, we adopt the proportion-based selection strategy summarized by Lin et al. (2025).

Concurrently, studies by Cichy et al. (2014); Kietzmann et al. (2019) reveal that the brain processes visual stimuli in a temporally progressive manner, transitioning continuously from early sensory perception to late cognitive abstraction. Motivated by this physiological evolution, as depicted Fig. 1, we propose a multi-stage temporal alignment strategy. We decompose the continuous cognitive trajectory into three functionally specialized components: holistic processing ($f_{\text{hol}}$), stimulus-driven early visual processing ($f_{\text{stim}}$), and late semantic residual processing ($f_{\text{res}}$). Specifically, given an input multi-channel neural signal $\mathbf{X} \in \mathbb{R}^{C\times T}$ across $C$ channels over a total recording duration $T$, we partition $\mathbf{X}$ along the temporal axis based on the active visual stimulus exposure duration $T_s$ ($0 < T_s < T$):

$$\mathbf{X}_{\text{hol}} = \mathbf{X}[:, 0:T], \quad \mathbf{X}_{\text{stim}} = \mathbf{X}[:, 0:T_s], \quad \mathbf{X}_{\text{res}} = \mathbf{X}[:, T_s:T], \tag{1}$$

where $\mathbf{X}_{\text{hol}}$ encompasses the complete spatio-temporal response, $\mathbf{X}_{\text{stim}}$ isolates the early sensory dynamics during active visual stimulation, and $\mathbf{X}_{\text{res}}$ captures the post-stimulus residual cognitive processing.

## 3.3 SPATIO-TEMPORAL EEG ENCODER

Cognitive processes natively evolve over time, requiring models to adequately preserve continuous temporal dynamics. To achieve this, we adopt the Inverted Transformer (iTransformer) (Liu et al., 2024), which processes the continuous temporal sequence of each channel as an entity to extract rich temporal features. Furthermore, to incorporate crucial spatial topology, a Graph Attention Network (GAT) (Veličković et al., 2018) is introduced as a residual branch before the iTransformer backbone, formulated as:

$$\mathbf{H} = \text{iTransformer}\Big(\mathbf{X} + \text{GAT}(\mathbf{X})\Big), \tag{2}$$

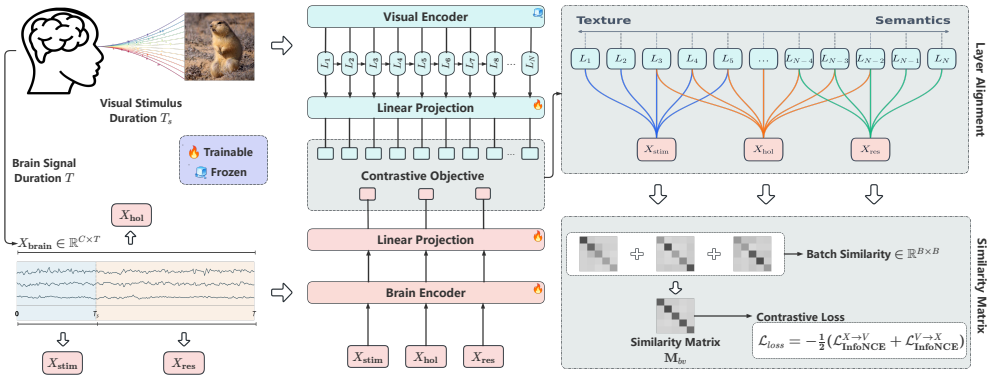



Figure 1: Overview of the proposed LSTA framework for brain-vision matching. Given visual stimuli and corresponding EEG signals, visual features ranging from low-level texture details to high-level semantic information are extracted from intermediate layers $L_1, L_2, \ldots, L_N$ of a Visual Encoder. Concurrently, the Brain Encoder processes different temporal segments of the EEG signals—stimulus-onset, holistic, and residual—to output brain feature representations. Multi-level contrastive learning is then performed between the brain embeddings and the multi-layer visual representations to establish hierarchical cross-modal alignment.

where $\mathbf{X}$ represents the input EEG signals and $\text{GAT}(\cdot)$ extracts inter-channel spatial relationships. To capture dynamic functional connectivity without relying on fixed electrode distance, the GAT module constructs an adaptive adjacency matrix optimized in an end-to-end self-learning manner. This residual integration seamlessly blends spatial domain topology into the temporal encoding pipeline.

### 3.4 Layer-Sensitive Contrastive Objective

Our method fuses hierarchical visual information across different cognitive branches. To optimize cross-modal alignment during training and fully leverage this mechanism during inference, we design a multi-layer conditioning and loss framework.

**Similarity Matrix** Given a batch of brain signal representations across three cognitive stages, denoted as $\mathbf{H}_{\text{hol}}, \mathbf{H}_{\text{stim}}, \mathbf{H}_{\text{res}} \in \mathbb{R}^{B\times d}$, and visual feature representations from the $l$-th layer of the vision backbone $\mathbf{H}_{v_l} \in \mathbb{R}^{B\times d}$ (where $l \in \{1, \ldots, K\}$), the layer-sensitive similarity matrix $\mathbf{M}_{bv} \in \mathbb{R}^{B\times B}$ is formulated as:

$$\mathbf{M}_{bv} = \sum_{l=1}^{K} \left( W_{l,\text{hol}} \mathbf{H}_{\text{hol}} \mathbf{H}_{v_l}^{\top} + W_{l,\text{stim}} \mathbf{H}_{\text{stim}} \mathbf{H}_{v_l}^{\top} + W_{l,\text{res}} \mathbf{H}_{\text{res}} \mathbf{H}_{v_l}^{\top} \right), \tag{3}$$

where $W_{l,m}$ denotes the layer-sensitive alignment weight for cognitive stage $m \in \{\text{hol}, \text{stim}, \text{res}\}$ at the $l$-th visual depth. To prevent unconstrained weights from absorbing mismatched features across non-corresponding visual depths, $W_{l,m}$ is formulated as a Gaussian prior mask centered at visual depth $\mu_m$:

$$W_{l,m} \propto \exp\left( -\frac{(l-\mu_m)^2}{2\sigma_m^2} \right) \tag{4}$$

This prior profile acts as a structural regularizer to alleviate cross-layer feature entanglement. By configuring $\mu_{\text{stim}} < \mu_{\text{hol}} < \mu_{\text{res}}$, $W_{l,m}$ regularizes the receptive window for each temporal branch, encouraging early perceptual and late semantic dynamics to remain appropriately decoupled during joint contrastive optimization.

**Similarity Loss** We adopt a symmetric contrastive learning objective that aligns matching neural-visual pairs to exhibit high similarity, while suppressing the similarity of mismatched pairs within each batch. The similarity loss is defined as

$$\mathcal{L}_{l,m} = -\frac{1}{2B}\sum_{i=1}^{B}\left(\log\frac{\exp\left(\hat{\mathbf{h}}_m^{(i)\top}\hat{\mathbf{h}}_{v_l}^{(i)}/\tau\right)}{\sum_{k=1}^{B}\exp\left(\hat{\mathbf{h}}_m^{(i)\top}\hat{\mathbf{h}}_{v_l}^{(k)}/\tau\right)} + \log\frac{\exp\left(\hat{\mathbf{h}}_m^{(i)\top}\hat{\mathbf{h}}_{v_l}^{(i)}/\tau\right)}{\sum_{k=1}^{B}\exp\left(\hat{\mathbf{h}}_m^{(k)\top}\hat{\mathbf{h}}_{v_l}^{(i)}/\tau\right)}\right), \tag{5}$$

where $\hat{\mathbf{h}}_m^{(i)}$ and $\hat{\mathbf{h}}_{v_l}^{(i)}$ represent the projected and $\ell_2$-normalized embeddings of the $i$-th sample in a shared cross-modal latent space, derived from the $m$-th temporal neural branch and the $l$-th hierarchical visual layer, respectively. Here, $\tau$ denotes a learnable temperature parameter, and $B$ represents the batch size of paired neural-visual samples. This bidirectional formulation encourages mutual alignment between neural and visual representations by pulling matched pairs together from both modal perspectives. To optimize all cross-modal alignment components jointly across all $K$ visual layers, the total loss function $\mathcal{L}_{\text{total}}$ is defined as the weighted combination of individual pairwise losses:

$$\mathcal{L}_{\text{total}} = \sum_{l=1}^{K}\left(W_{l,\text{hol}}\cdot\mathcal{L}_{l,\text{hol}} + W_{l,\text{stim}}\cdot\mathcal{L}_{l,\text{stim}} + W_{l,\text{res}}\cdot\mathcal{L}_{l,\text{res}}\right) \tag{6}$$

# 4 Experiment

## 4.1 Experiment Details

We conducted the contrastive training stage on a single NVIDIA RTX 5090 GPU for 200 epochs using AdamW with a peak learning rate of $1 \times 10^{-4}$, a cosine decay schedule, and a global batch size of 128. Models are trained and evaluated under a strict zero-shot protocol on the THINGS-EEG (Grootswagers et al., 2019; Gifford et al., 2022) and THINGS-MEG (Hebart et al., 2023) datasets by completely withholding test concepts during training. For preprocessing, we strictly follow Tang et al. (2026); Zheng et al. (2026) by extracting a 1000 ms time window starting from the stimulus onset. The boundary between the stimulus-onset branch and the residual branch is determined by the actual stimulus offset, which corresponds to 200 ms for THINGS-EEG and 500 ms for THINGS-MEG. For detailed preprocessing pipeline and dataset descriptions, please refer to Appendix A.

## 4.2 Quantitative Evaluation

We extensively evaluate our method on the 200-way zero-shot brain-visual retrieval task using both the THINGS-EEG and THINGS-MEG datasets. To comprehensively demonstrate the effectiveness of our approach, we aim to compare it with several state-of-the-art baseline models, including BraVL (Du et al., 2023),NICE and its spatial variants (NICE-SA, NICE-GA) (Song et al., 2024),

Table 1: Average Top-1/Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG and THINGS-MEG. Inter-subject evaluation uses standard LOSO without domain adaptation; ”—” indicates not reported. Detailed data are provided in Appendix B.

| **Method** | **EEG** | | | | **MEG** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Intra-subject** | | **Inter-subject** | | **Intra-subject** | | **Inter-subject** | |
| | **Top-1** | **Top-5** | **Top-1** | **Top-5** | **Top-1** | **Top-5** | **Top-1** | **Top-5** |
| BraVL | 5.8 | 17.5 | 1.8 | 7.0 | — | — | — | — |
| NICE | 16.1 | 43.6 | 6.2 | 21.4 | 12.8 | 36.0 | — | — |
| NICE-SA | 14.7 | 41.7 | 7.0 | 23.1 | 12.7 | 35.0 | — | — |
| NICE-GA | 15.6 | 42.8 | 5.9 | 21.6 | 14.3 | 42.3 | — | — |
| ATM | 28.5 | 60.4 | 11.8 | 33.7 | — | — | — | — |
| UBP | 50.9 | 79.7 | 12.4 | 33.4 | 26.7 | 55.2 | 2.2 | 10.4 |
| BVSA | 58.2 | 89.1 | 15.8 | 39.4 | 32.5 | 62.9 | 5.4 | 13.6 |
| HVF | 75.7 | 94.6 | 20.0 | 44.1 | 33.9 | 60.7 | 5.4 | 15.2 |
| HCF | 84.6 | 98.2 | 23.4 | 54.9 | — | — | — | — |
| **Ours** | **91.4** | **98.6** | **35.1** | **63.1** | **42.7** | **67.4** | **7.0** | **19.3** |

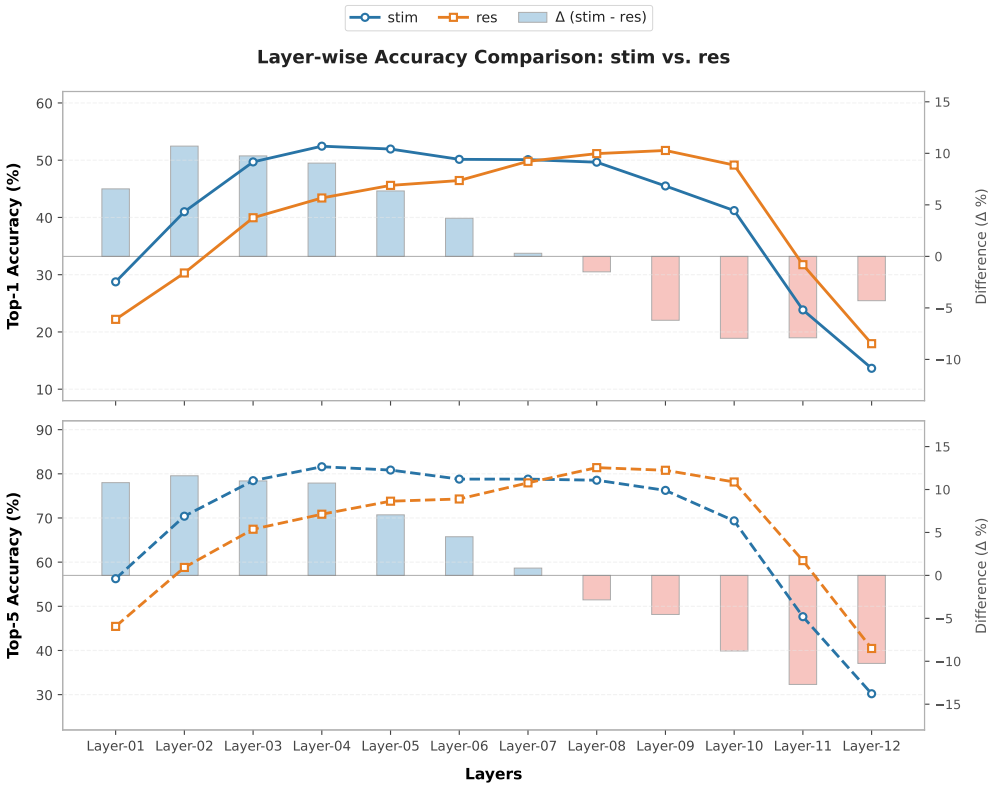
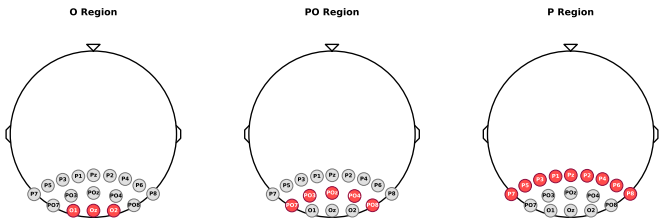



Figure 2: Layer-wise Top-1 and Top-5 classification accuracy of the stimulus branch ($f_{\text{stim}}$) and residual branch ($f_{\text{res}}$) across the 12 layers of the DINOv2 backbone on the THINGS-EEG dataset. The bar chart ($\Delta$) represents the performance differential between the two branches (stim $-$ res). Additional layer-wise evaluation results across other backbone configurations and detailed performance metrics are provided in Appendix B.

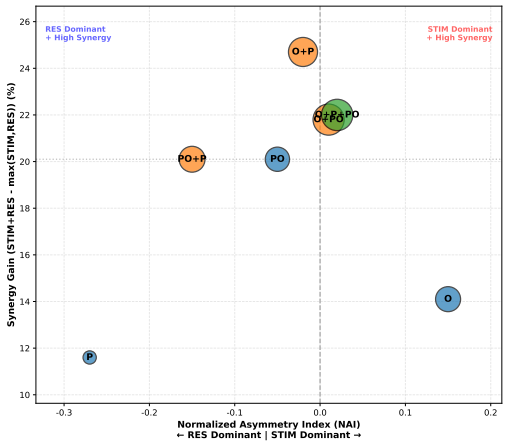
ATM (Li et al., 2024), UBP (Wu et al., 2025), BVSA (Liu et al., 2026), HVF (Zheng et al., 2026), and HCF (Tang et al., 2026). The quantitative results are summarized in Tab. 1. As observed, our proposed method consistently outperforms all existing baselines by a significant margin across different modalities and evaluation settings. Specifically, in the EEG intra-subject setting, our model achieves a remarkable Top-1 accuracy of 91.4% and a Top-5 accuracy of 98.6%, surpassing the previous best method (HCF) by 6.8 percentage points in Top-1. Furthermore, in the more challenging inter-subject EEG setting, our method reaches 35.1% in Top-1 accuracy, yielding a substantial improvement over prior works. A similar performance superiority is also observed in the MEG dataset, achieving 42.7% Top-1 accuracy in the intra-subject setting and 7.0% in the inter-subject setting, demonstrating the strong generalizability and robustness of our layer-sensitive alignment strategy across different neuroimaging modalities.

### 4.3 Layer-Sensitive Hierarchy Validation

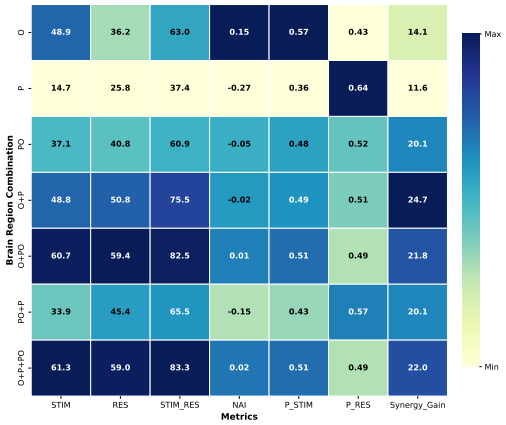
To examine whether the neural-visual alignment conforms to the established spatiotemporal hierarchy (Cichy et al., 2014; Kietzmann et al., 2019; Zeiler & Fergus, 2014; Oquab et al., 2024), we evaluate the stimulus branch ($f_{\text{stim}}$) and residual branch ($f_{\text{res}}$) across varying depths of the visual backbone, with the layer-wise performance depicted in Fig. 2. The empirical results strongly support this hierarchical progression: in shallow layers, the stimulus branch consistently outperforms the residual branch by over $10\%$ in Top-1 accuracy due to its heightened sensitivity to low-level perceptual details. As features progress toward intermediate and deep layers, the performance trend reverses, with the residual branch taking the lead and achieving a maximum advantage of $7.95\%$. This dynamic shift demonstrates that sustained residual activity effectively captures abstract, high-level semantic information, validating our decoupling design for early sensory encoding and late representation.




(a) **Brain Region Channel Topography**




(b) **Task Preference & Synergy Map**




(c) **Multi-Metric Performance Profiles**

Figure 3: Functional synergy and task preference analysis across brain region combinations. (a) Regional channel topographies for O, PO, and P regions. (b) Task Preference & Synergy Map showing asymmetry index versus synergy gain. (c) Multi-Metric Performance Profiles across single, dual, and triple region combinations.

### 4.4 Spatio-Temporal Brain Region Synergy Analysis

Prior neuroscientific evidence indicates that distinct brain regions subserve functional specialization and continuous transitions during visual perception (Kravitz et al., 2011). To evaluate whether our framework aligns with this biological mechanism, we assess the dual-branch representations ($f_{\text{stim}}$ and $f_{\text{res}}$) across individual Occipital (O), Parieto-Occipital (PO), and Parietal (P) regions, as well as their spatial combinations. As shown in Fig. 3, the empirical observations suggest a spatial-functional preference: O region tends to be more responsive to immediate visual stimuli ($f_{\text{stim}}$), whereas P region appears to lean toward post-stimulus residual semantic processing ($f_{\text{res}}$). Positioned between them, PO region functions like a transition zone, balancing both aspects and yielding notable synergy gain when integrating the dual branches ($f_{\text{stim}} + f_{\text{res}}$). Furthermore, prominent synergy gains are consistently observed across multiple region combinations, indicating that cross-region collaboration effectively captures complementary neural dynamics. Overall, these observations align well with the idea that spatial topographies may reflect the temporal evolution from sensory input to semantic persistence.

### 4.5 Temporal Dynamics of Residual Response

Previous studies typically extract a fixed 1000 ms EEG time window following stimulus onset. To ensure a fair comparison, all baseline methods evaluated in this work follow this standard setup. In our framework, we decompose the total visual evoked response into distinct functional branches: the early stimulus-evoked phase and the subsequent persistent (residual) response phase. To systematically investigate the impact of the residual response duration on visual object recognition, we conduct experiments across different time windows. Specifically, on the THINGS-EEG dataset—where

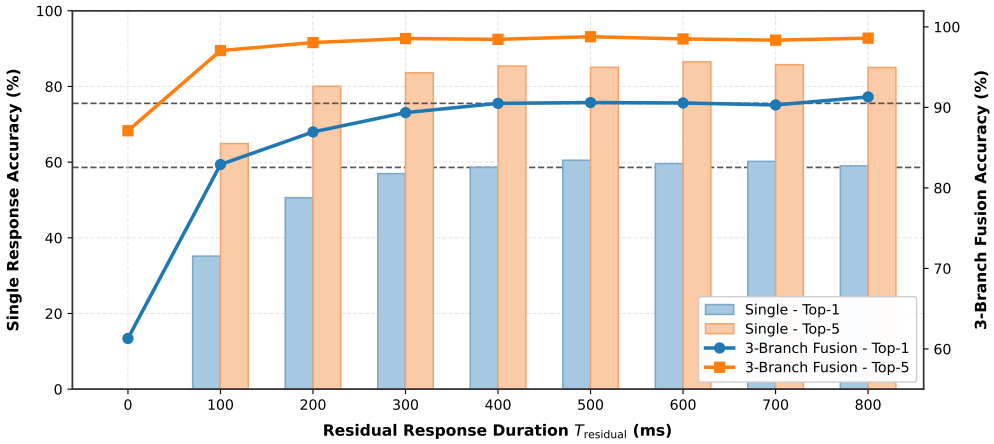



Figure 4: Impact of residual response duration $T_{\text{res}}$ on zero-shot object decoding performance on THINGS-EEG. Bars indicate single $f_{\text{res}}$ branch accuracy and lines show 3-branch fusion accuracy for Top-1 and Top-5 metrics.

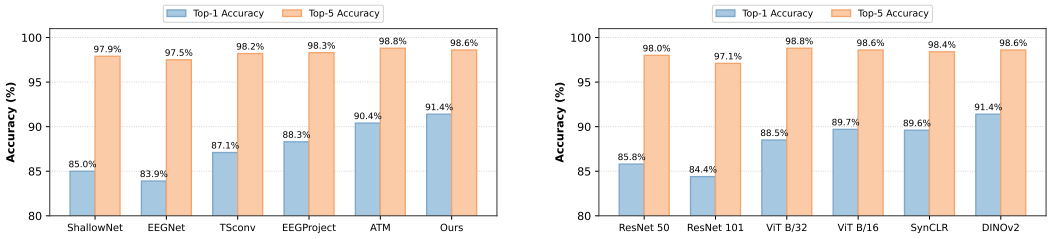
Table 2: Ablation study on different EEG signal branch components and visual layer allocation profiles on the THINGS-EEG dataset for zero-shot EEG-visual retrieval, showing performance under both Normal Gaussian prior and Reversed Gaussian prior. All numbers represent subject-wise average accuracies. For the specific implementation details of the Reversed Gaussian prior, please refer to the Appendix C.

| **Branch Components** | | | **Normal Gaussian Prior** | | **Reversed Gaussian Prior** | |
|---|---|---|---|---|---|---|
| **Holistic ($f_{\text{hol}}$)** | **Stimulus ($f_{\text{stim}}$)** | **Residual ($f_{\text{res}}$)** | **Top-1** | **Top-5** | **Top-1** | **Top-5** |
| — | ✓ | — | 61.3 | 87.1 | 58.7 | 86.6 |
| — | — | ✓ | 59.0 | 85.0 | 49.0 | 78.9 |
| ✓ | — | — | 87.4 | 98.4 | 69.2 | 92.0 |
| — | ✓ | ✓ | 83.3 | 97.0 | 73.3 | 93.9 |
| ✓ | — | ✓ | 89.6 | 98.1 | 77.0 | 95.1 |
| ✓ | ✓ | — | 90.3 | 98.9 | 72.2 | 93.0 |
| ✓ | ✓ | ✓ | 91.4 | 98.6 | 82.3 | 97.1 |

the active visual stimulus lasts for 200 ms—we vary the residual response duration $T_{\text{res}}$ from 0 ms to 800 ms. As illustrated in Fig. 4, the decoding performance improves sharply as $T_{\text{res}}$ increases up to 300 ms–400 ms, after which additional residual duration yields negligible gains and the performance plateaus across both the single $f_{\text{res}}$ branch and 3-branch fusion settings.

### 4.6 Ablation Studies

**Branch Ablation** To evaluate each functional branch and validate visual layer allocation guided by cognitive priors, we conduct ablation studies on THINGS-EEG (Tab. 2). Individual branches yield sub-optimal performance, with $f_{\text{hol}}$, $f_{\text{stim}}$, and $f_{\text{res}}$ reaching Top-1 accuracies of 87.4%, 61.3%, and 59.0% respectively. Combining $f_{\text{stim}}$ and $f_{\text{res}}$ significantly boosts Top-1 accuracy to 83.3%, confirming that early perceptual features and late semantic residuals are highly complementary. Our full model integrating all three branches achieves the best performance of 91.4% Top-1 and 98.6% Top-5 accuracy. Furthermore, reversing the cognitive prior $W_{l,m}$ to misalign early responses with deep visual layers causes consistent performance drops across all settings, with the full model Top-1 accuracy falling to 82.3%. This validates that aligning temporal latency with visual hierarchy reflects intrinsic visual cognitive progression. Detailed layer weighting strategies, including analyses on learnable variants, hyperparameter configurations, and further ablation setups, are provided in Appendix C.




(a) **EEG Encoders** (b) **Visual Backbones**

Figure 5: Performance comparisons across different architectural backbones on the zero-shot EEG-visual retrieval task using the THINGS-EEG dataset. (a) Impact of different EEG encoders. (b) Comparison of neural decoding performance across different visual backbones.

**Impact of EEG Encoders** To validate the design choice and superior feature extraction capability of our EEG encoder, we conduct an ablation study by replacing our encoder with several representative alternatives, including classic backbones (ShallowNet (Schirrmeister et al., 2017), and EEGNet (Lawhern et al., 2018)) and recent architectures (ATM (Li et al., 2024), TSConv (Song et al., 2024) and EEGProject (Wu et al., 2025)). As shown in Fig. 5-(a), substituting our encoder with other backbones leads to a noticeable degradation in performance. In particular, compared to the competitive TSConv and EEGProject, our full proposed encoder achieves the best performance with $91.4\%$ Top-1 and $98.6\%$ Top-5 accuracy. These results demonstrate that our architecture captures discriminative EEG dynamics more effectively for visual alignment.

**Impact of Visual Backbones** Pre-trained visual encoders capture rich visual semantics that fundamentally dictate downstream neural decoding quality. To systematically investigate this influence, we benchmarked a diverse set of visual backbones on the THINGS-EEG dataset, spanning different architectural inductive biases and pre-training paradigms (e.g., supervised, contrastive, and self-supervised learning). Specifically, our comparisons cover CNN-based models, such as ResNet-50 and ResNet-101 (He et al., 2016), Vision Transformers including ViT-B/16 and ViT-B/32 (Dosovitskiy et al., 2021), and self-supervised/foundation frameworks like SynCLR (Tian et al., 2024) and DINOv2 (Oquab et al., 2024). Fig. 5-(b) compares the Top-1 and Top-5 retrieval accuracies across these visual backbones.

## 5 Conclusion

To resolve the feature interference caused by standard static matching, this paper introduces LSTA, a layer-sensitive framework designed for progressive vision-brain alignment. By bridging temporal neural latency with spatial feature depth, the proposed method captures the dynamic progression of visual cognition by synchronizing early sensory dynamics and late persistent semantics with shallow and deep visual layers, respectively, in continuous EEG and MEG signals. Exceptional zero-shot retrieval performance on THINGS-EEG and THINGS-MEG benchmarks confirms the effectiveness of our design. Simultaneously, evaluations across various visual backbones empirically validate the core hypothesis that early neural responses preferentially align with shallow layers, whereas late residual dynamics correspond to deeper semantic stages.

**Limitations and future work** The current evaluations are limited to the THINGS-EEG and THINGS-MEG benchmark datasets, with the functional analysis across brain regions serving as a preliminary exploration of regional task preferences and spatial synergy gains. Building upon these baseline insights, future research will focus on tightly coupling these functional brain regions with temporal dynamics and visual hierarchy representations, extending our approach toward a unified spatial-temporal-hierarchical alignment architecture.

## LLM USAGE STATEMENT

We utilized LLMs for grammar checking and language polishing to enhance overall readability, as well as for literature retrieval to assist in discovering related work.

## REPRODUCIBILITY STATEMENT

To ensure reproducibility, we have provided the source code, experimental environment, and detailed parameter configurations in the supplementary materials.

## REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 23716–23736, 2022.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 1298–1312, 2022.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

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.

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.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Tijl Grootswagers, Amanda K. Robinson, and Thomas A. Carlson. The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188:668–679, 2019.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

Bin He and Zhongming Liu. Multimodal functional neuroimaging: integrating functional mri and eeg/meg. *IEEE reviews in biomedical engineering*, 1:23–40, 2008.

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 (CVPR)*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022.

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.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 4904–4916, 2021.

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

Tim C Kietzmann, Courtney J Spoerer, Lynn KA Sörensen, Radoslaw M Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, 2019.

Dwight J Kravitz, Kadharbatcha S Saleem, Chris I Baker, and Mortimer Mishkin. A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, 12(4):217–230, 2011.

Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

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.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 12888–12900, 2022.

Junyan Lin, Haoran Chen, Yue Fan, Yingqi Fan, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multi-layer visual feature fusion in multimodal llms: Methods, analysis, and best practices. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4156–4166, 2025.

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

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations (ICLR)*, pp. 11116–11140, 2024.

Scott Makeig, Marissa Westerfield, T-P Jung, Sonia Enghoff, Jeanne Townsend, Eric Courchesne, and Terrence J Sejnowski. Dynamic brain sources of visual evoked responses. *Science*, 295 (5555):690–694, 2002.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

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

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 (ICLR)*, pp. 47648–47665, 2024.

Jingyi Tang, Shuai Jiang, Fei Su, and Zhicheng Zhao. Aligning what eeg can see: Structural representations for brain-vision matching. *arXiv preprint arXiv:2603.07077*, 2026. URL https://arxiv.org/abs/2603.07077.

Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15887–15898, 2024.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2246–2257, 2025.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6991–7000, 2021.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pp. 818–833, 2014.

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 (ICLR)*, 2026.

# A Dataset Details

**THINGS-EEG** The THINGS-EEG dataset (Gifford et al., 2022) includes EEG data from 10 participants across 1,654 training concepts (10 images per concept, presented four times) and 200 test concepts (1 image per concept, repeated 80 times). Preprocessing follows Zheng et al. (2026) and Tang et al. (2026): raw EEG was filtered to 0.1–100 Hz (yielding 63 channels at 1,000 Hz), epoch-extracted from 0 to 1,000 ms relative to stimulus onset with a pre-stimulus 200 ms baseline correction, and downsampled to 250 Hz. By averaging across trial repetitions for each image to boost signal quality, 17 posterior parietal and occipital channels relevant to visual processing were selected for final analysis.

**THINGS-MEG** The THINGS-MEG dataset (Hebart et al., 2023) captures 271-channel MEG activity across 4 participants. Stimuli were presented for $500$ ms per trial, interleaved with a $1,000 \pm 200$ ms inter-trial interval. The data is partitioned into 1,854 training concepts (12 images $\times$ 1 presentation) and 200 test concepts (1 image $\times$ 12 repetitions), with the test concepts removed from the training set to construct a zero-shot task. In accordance with Liu et al. (2026); Zheng et al. (2026), the raw MEG data underwent 0.1–100 Hz band-pass filtering, epoch segmentation from 0 to 1,000 ms relative to onset with baseline correction, and downsampling to 200 Hz. To boost signal quality, repetitions were averaged across corresponding trials.

# B Results Details

**Per-Subject Retrieval on THINGS-EEG and THINGS-MEG** Tab. 3 and Tab. 4 evaluate the 200-way zero-shot Top-1 and Top-5 retrieval performance across individual subjects on THINGS-EEG and THINGS-MEG, respectively. Neural features are extracted from the respective datasets. For our method, DINOv2 is adopted as the visual backbone, and all reported metrics are averaged over five independent runs with different random seeds.

Table 3: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG.

| Method | Sub1 | | Sub2 | | Sub3 | | Sub4 | | Sub5 | | Sub6 | | Sub7 | | Sub8 | | Sub9 | | Sub10 | | 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 |
| 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 |
| 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 |
| BVSA | 57.0 | 85.5 | 55.5 | 89.0 | 57.5 | 92.0 | 54.5 | 88.5 | 51.0 | 81.5 | 61.5 | 92.0 | 56.5 | 85.5 | 66.0 | 92.5 | 55.0 | 89.0 | 67.0 | 95.5 | 58.2 | 89.1 |
| HVF | 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 |
| HCF | 81.9 | 98.3 | 86.0 | 98.9 | 84.3 | 97.7 | 82.2 | 98.8 | 74.5 | 94.1 | 90.4 | 98.9 | 82.8 | 97.2 | 90.9 | 99.4 | 74.7 | 98.4 | 93.1 | 99.8 | 84.6 | 98.2 |
| Ours | 95.5 | 99.5 | 90.7 | 99.0 | 90.1 | 98.9 | 86.9 | 96.9 | 87.4 | 97.9 | 95.4 | 99.3 | 90.1 | 98.7 | 94.4 | 98.6 | 89.5 | 97.7 | 94.3 | 99.0 | 91.4 | 98.6 |

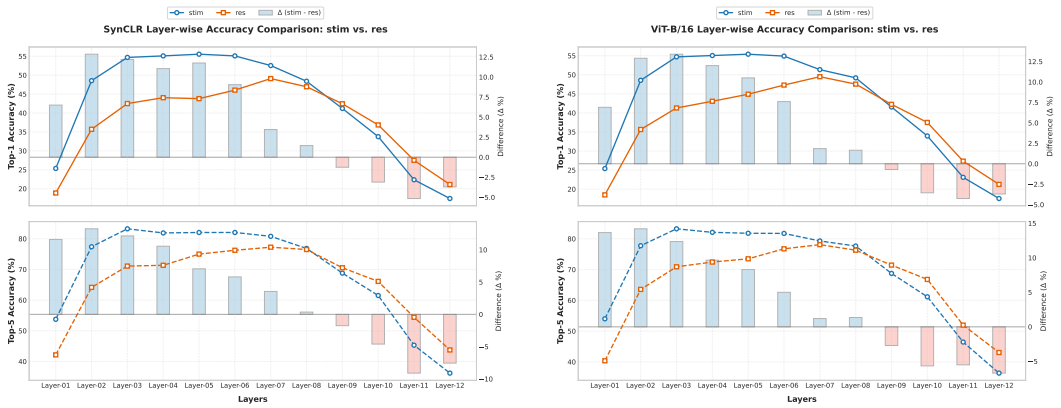
Table 4: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-MEG.

| Method | Sub1 | | Sub2 | | Sub3 | | Sub4 | | Avg | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 |
| 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 |
| UBP | 15.0 | 38.0 | 46.0 | 80.5 | 27.3 | 59.0 | 18.5 | 43.5 | 26.7 | 55.2 |
| BVSA | 20.5 | 47.0 | 52.0 | 86.0 | 32.0 | 67.0 | 25.5 | 51.5 | 32.5 | 62.9 |
| HVF | 14.0 | 31.8 | 63.8 | 91.8 | 41.0 | 78.3 | 17.0 | 41.0 | 33.9 | 60.7 |
| Ours | 28.4 | 57.0 | 67.3 | 91.4 | 44.1 | 73.8 | 31.0 | 47.5 | 42.7 | 67.4 |

**Additional Layer-wise Performance** To provide a comprehensive evaluation across different model architectures and training paradigms, we extend our layer-wise analysis to additional vision backbones, including Vision Transformers (SynCLR, ViT-B/16, ViT-B/32, and DINOv2) and Convolutional Neural Networks (ResNet-50 and ResNet-101), with the detailed classification results on the THINGS-EEG dataset presented in Fig. 6 and Fig. 7, and the exact layer-wise differentials summarized in Tab. 5. Across both ViT and ResNet families, the stimulus branch ($f_{\text{stim}}$) achieves higher Top-1 and Top-5 accuracies than the residual branch ($f_{\text{res}}$) in lower layers, yielding a positive differential ($\Delta > 0$). In contrast, in the deepest layers, the overall performance drops and $f_{\text{res}}$ overtakes $f_{\text{stim}}$, resulting in a negative differential ($\Delta < 0$).

Table 5: Layer-wise differential performance (Δ) in Top-1 and Top-5 retrieval accuracy (%) on the THINGS-EEG dataset across 12 layers. We evaluate additional vision backbones including Vision Transformers and Convolutional Neural Networks to provide a comprehensive evaluation across different model architectures and training paradigms. Note that ResNet models consist of 4 main stage layers.

| Model | Layer 1 | | Layer 2 | | Layer 3 | | Layer 4 | | Layer 5 | | Layer 6 | | Layer 7 | | Layer 8 | | Layer 9 | | Layer 10 | | Layer 11 | | Layer 12 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 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 | Top-1 | Top-5 |
| SynCLR | 6.5 | 11.6 | 12.9 | 13.3 | 12.2 | 12.2 | 11.1 | 10.6 | 11.8 | 7.1 | 9.1 | 5.8 | 3.5 | 3.6 | 1.5 | 0.4 | -1.3 | -1.8 | -3.1 | -4.6 | -5.2 | -9.1 | -3.7 | -7.6 |
| ViT-B/16 | 6.9 | 13.6 | 12.9 | 14.2 | 13.4 | 12.3 | 12.0 | 9.6 | 10.5 | 8.3 | 7.6 | 5.0 | 1.9 | 1.2 | 1.6 | 1.4 | -0.7 | -2.7 | -3.6 | -5.6 | -4.2 | -5.5 | -3.7 | -6.7 |
| ViT-B/32 | 5.7 | 9.2 | 8.1 | 13.1 | 10.0 | 10.0 | 13.3 | 9.0 | 8.4 | 6.3 | 6.1 | 5.6 | 3.9 | 4.0 | 1.6 | 1.2 | -3.2 | -2.8 | -5.9 | -8.9 | -6.8 | -10.0 | -7.8 | -12.3 |
| DINOv2 | 6.6 | 10.8 | 10.7 | 11.6 | 9.8 | 11.0 | 9.1 | 10.8 | 6.4 | 7.0 | 3.7 | 4.5 | 0.3 | 0.8 | -1.5 | -2.9 | -6.2 | -4.5 | -7.9 | -8.8 | -7.9 | -12.7 | -4.3 | -10.2 |
| ResNet-50 | 9.3 | 11.0 | 6.5 | 5.7 | 0.0 | 0.1 | -3.8 | -6.5 | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| ResNet-101 | 11.6 | 13.5 | 8.9 | 7.8 | -0.5 | 0.5 | -2.0 | -3.4 | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |



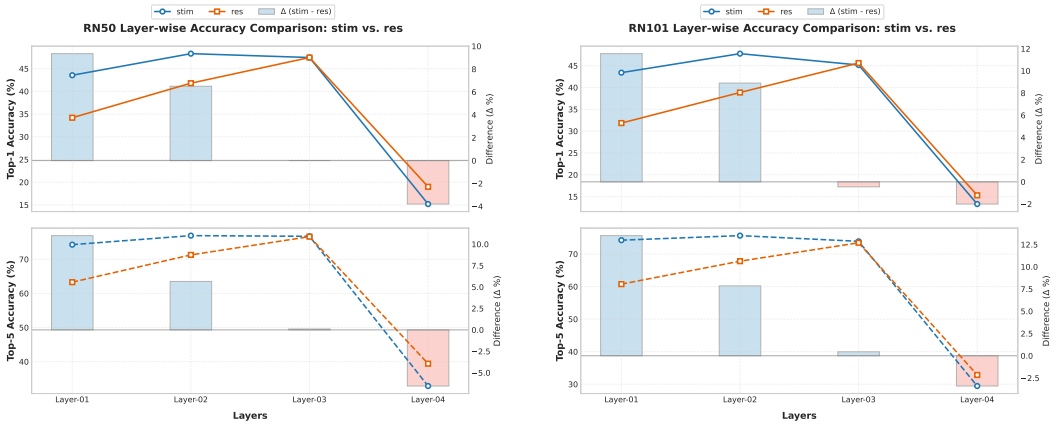

(a) **SynCLR** (b) **ViT-B/16**

(c) **ViT-B/32** (d) **DINOv2**

Figure 6: Layer-wise Top-1 and Top-5 classification accuracy of the stimulus branch ($f_{\text{stim}}$) and residual branch ($f_{\text{res}}$) across the layers of SynCLR, ViT-B/16, ViT-B/32, and DINOv2 backbones on the THINGS-EEG dataset. The bar chart (Δ) represents the performance differential between the two branches ($f_{\text{stim}} - f_{\text{res}}$).

## C Additional Ablation Studies

**Ablation Study on Weighting Strategies** To investigate the effectiveness of different weight distribution strategies, we evaluate five distinct settings: Uniform Weights, Softmax Learnable Weights (Soft Attention), Reversed Gaussian Prior, Gaussian-Guided Attention, and Normal Gaussian Prior. In detail, Uniform Weights assigns equal weights across layers; Soft Attention utilizes three independent Softmax functions to dynamically optimize layer weights for each branch; Reversed Gaussian Prior swaps $\mu_{\text{stim}}$ and $\mu_{\text{res}}$ while placing $\mu_{\text{hol}}$ at both ends of the visual layers; Gaussian-Guided Attention first determines the visual layer range via the Gaussian prior and subsequently learns detailed internal weights using Softmax; and Normal Gaussian Prior represents the standard layer-sensitive Gaussian weighting scheme.




(a) **ResNet-50** (b) **ResNet-101**

Figure 7: Layer-wise Top-1 and Top-5 classification accuracy of the stimulus branch ($f_{\text{stim}}$) and residual branch ($f_{\text{res}}$) across the layers of ResNet-50 and ResNet-101 backbones on the THINGS-EEG dataset. The bar chart ($\Delta$) represents the performance differential between the two branches ($f_{\text{stim}} - f_{\text{res}}$).

Table 6: Top-1 and Top-5 retrieval accuracy (%) for different weight distribution strategies. Note that results for the Reversed Gaussian Prior are averaged across two configurations with $\mu_{\text{hol}}$ placed at the lower and upper bounds of the visual layers, respectively.

| | **Retrieval Accuracy (%)** | |
|---|---|---|
| **Weighting Strategy** | **Top-1** | **Top-5** |
| Uniform Weights | 90.1 | 98.9 |
| Soft Attention | 89.7 | 98.6 |
| Reversed Gaussian Prior | 82.3 | 97.1 |
| Gaussian-Guided Attention | 91.2 | 98.7 |
| Normal Gaussian Prior | 91.4 | 98.6 |

As shown in Tab. 6, the Reversed Gaussian Prior leads to lower retrieval accuracy compared to other configurations, suggesting that an inverted depth prior is less aligned with cross-modal features. Similarly, learnable Soft Attention yields a slight performance drop relative to the prior-guided scheme. The converged weights under Soft Attention, illustrated in Fig. 8, show that even without explicit guidance, the learned weights display a similar overall trend to our domain priors. For instance, earlier visual layers receive higher weights for the stimulus-evoked branch, while deeper layers contribute more to the residual-semantic branch. However, because unconstrained learnable attention can absorb features across multiple depths during optimization, it may introduce unnecessary cross-layer interactions. Gaussian-Guided Attention and Normal Gaussian Prior yield virtually identical performance. This shows that once layer ranges are constrained by the domain prior, intra-layer attention refinement brings negligible gain, making either strategy a viable choice. In the main experiments, we adopt the Normal Gaussian Prior for its simplicity and consistency with our prior-guided design.

**Ablation Study on Prior Weight Profiles** To evaluate the sensitivity of the Gaussian prior parameters, we conducted ablation experiments on the center parameter $\mu_{\text{stim}}$ using the two-branch architecture ($f_{\text{stim}}$ and $f_{\text{res}}$) on the THINGS-EEG dataset, with the results illustrated in Fig. 9. In this setup, we fix $\sigma = 1.0$ and enforce $\mu_{\text{stim}} + \mu_{\text{res}} = 11$, where $\mu = 0$ corresponds to the first visual layer $L_1$. As shown in Fig. 9-(a), both Top-1 and Top-5 retrieval accuracies remain high and stable within the light blue shaded "Prior Zone", demonstrating that the model exhibits local robustness to moderate variations in $\mu_{\text{stim}}$ near the biologically plausible domain. However, as $\mu_{\text{stim}}$ deviates significantly beyond this region, performance drops rapidly. Furthermore, the performance distribution comparison in Fig. 9-(b) highlights a marked accuracy decline and larger variance in the Non-Prior Region compared to the Prior Region. These findings indicate that while the model is resilient to minor fluctuations in prior weights, severely violating the pre-established temporal-hierarchy cor-

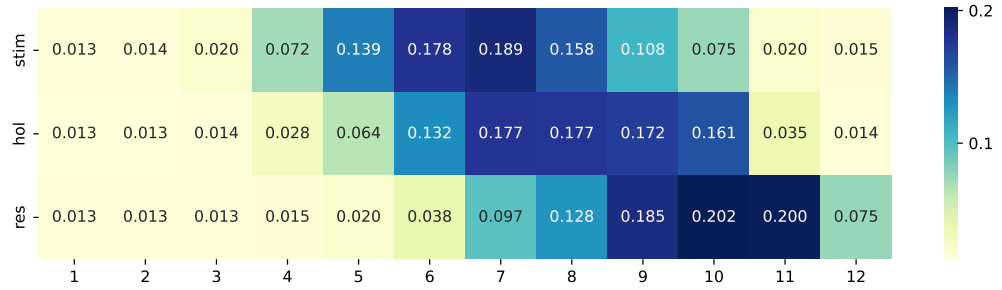
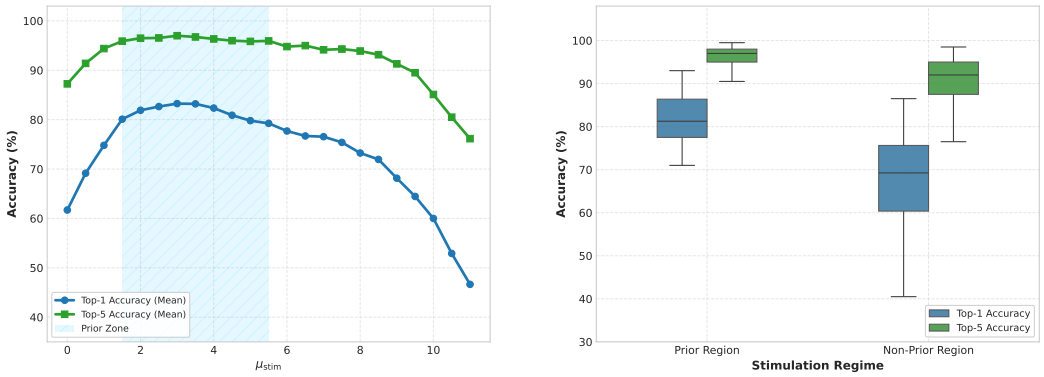



Figure 8: Layer-wise learned attention weight allocation under the soft-alignment framework for cross-modal vision-brain decoding.



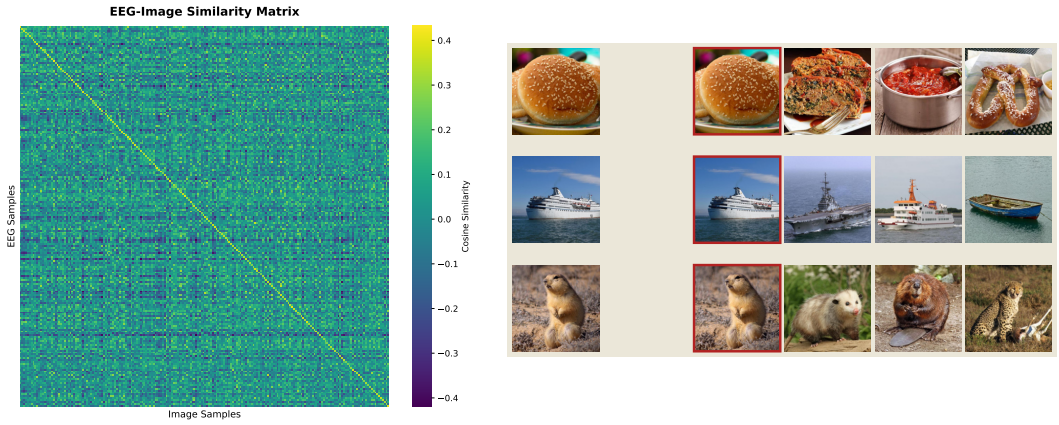

(a) (b)

Figure 9: Ablation study of the Gaussian prior center $\mu_{\text{stim}}$ in the two-branch model ($f_{\text{stim}}$ and $f_{\text{res}}$) on THINGS-EEG. Here, $\sigma = 1.0$, $\mu_{\text{stim}} + \mu_{\text{res}} = 11$, and $\mu = 0$ corresponds to the first visual layer $L_1$. (a) Retrieval accuracies versus $\mu_{\text{stim}}$. The shaded area denotes the Prior Zone. (b) Performance distributions for configurations within the Prior Region versus the Non-Prior Region .

respondence leads to substantial performance degradation, underscoring the necessity of properly aligning prior parameters with neural dynamics.

## D Qualitative Analysis

To further complement our quantitative evaluation, we provide qualitative visualizations on the unseen test split of the THINGS-EEG dataset in Fig. 10. As shown in Fig. 10-(a), the cross-modal similarity matrix exhibits a clear and distinct diagonal pattern, proving that our model effectively aligns paired neural-visual representations in the shared embedding space. Furthermore, the zero-shot image retrieval results in Fig. 10-(b) demonstrate that our model successfully hits the exact ground-truth visual stimulus as the Top-1 match (highlighted in red) alongside representative retrieved candidates, confirming the strong cross-modal decoding capacity of LSTA.




(a) **Similarity Matrix** (b) **Zero-Shot Retrieval**

Figure 10: Cross-modal alignment performance of our model on the THINGS-EEG dataset. (a) EEG-Image similarity matrix demonstrating strong diagonal alignment across test samples. (b) Qualitative image retrieval results using EEG feature queries, illustrating accurate Top-1 predictions (highlighted in red) and subsequent candidates ordered by similarity.