LEARNING BRAIN-ALIGNED VISUAL REPRESENTATION WITH ADAPTIVE TEMPORAL RECALIBRATION FOR EEG-BASED VISUAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding visual stimuli from brain activity is a central task in understanding visual perception and advancing BCI applications. Existing methods typically align EEG signals with frozen pretrained visual embeddings and process temporal activity uniformly. This paradigm suffers from two limitations: a representational mismatch between machine vision and human perception, and temporal sparsity of decodable visual information in EEG. Together, these yield poorly aligned decoders that struggle with zero-shot generalization. To address these challenges, we propose **BEAT**, a **B**rain-aligned visual **E**ncoder with **A**daptive **T**emporal recalibration. To mitigate the representational mismatch, we introduce a dual-alignment strategy, integrating multi-teacher visual distillation with cross-modal contrastive learning. Together, these objectives guide the visual encoder to learn semantically meaningful and neurally aligned representations. To address temporal sparsity, an adaptive temporal recalibration module emphasizes visually informative signal intervals while suppressing noisy ones via temporal contextualization. This lightweight calibrator could be readily applied to various backbones in a plug-and-play manner. Experiments demonstrate state-of-the-art zero-shot brain-to-image retrieval, while reconstruction results further validate the utility of the learned representations.

1 INTRODUCTION

Decoding visual content from brain activity provides an important means of understanding how the human brain represents the visual world and developing non-invasive brain–computer interfaces (Kriegeskorte & Douglas, 2018; Kay et al., 2008). Among non-invasive recording techniques, electroencephalography (EEG) is particularly attractive because of its accessibility and millisecond-level temporal resolution, offering a direct window into the rapid dynamics of human visual processing. Recent EEG-based approaches have demonstrated that neural responses elicited by visual stimuli can support zero-shot image retrieval and image reconstruction (Song et al., 2024; Guo et al., 2025).

A prevailing paradigm in this field is to map brain-derived features into the embedding space of powerful pre-trained image encoders such as CLIP (Radford et al., 2021), typically using contrastive objectives to align neural and visual representations (Khosla et al., 2020). Progress has been driven by increasingly sophisticated brain encoders (Schirrmeister et al., 2017; Li et al., 2024; Song et al., 2024; Benchetrit et al., 2024), the integration of auxiliary text descriptions (Song et al., 2025; Zhang et al., 2025), the use of blurred image priors (Wu et al., 2025), and human-aligned image models (Rajabi et al., 2025).

Despite recent advances, EEG-based visual decoding remains constrained by two fundamental challenges: **1) Representational mismatch between machine vision and human perception**. Pretrained vision encoders provide rich semantic information, but their representations are optimized for machine-vision objectives rather than neural predictability. Consequently, the visual information encoded by these models may only partially overlap with what can be reliably recovered from EEG. Existing approaches adopt predefined visual targets that ignore this discrepancy. They force EEG encoders to approximate a target space only partially supported by neural evidence, thereby resulting

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.

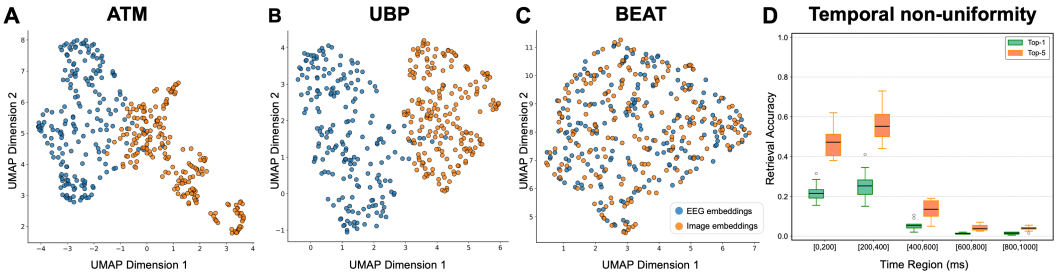


Figure 1: (A–C) Joint UMAP visualizations of EEG and image embeddings on the held-out test set. Existing methods, ATM and UBP, retain a pronounced modality gap, whereas BEAT substantially improves the distributional alignment between the two modalities. (D) Retrieval accuracy across post-stimulus intervals, showing that decodable visual information is distributed non-uniformly over time.

in pronounced cross-modal gap that limits decoding and generalization (Fig. 1A-C). **2) Temporal sparsity of decodable visual information in EEG.** EEG densely samples post-stimulus activity, but decodable visual information is not uniformly distributed over time (Fig. 1D): some intervals contain more reliable stimulus-related responses, whereas others provide weaker task-relevant evidence and are more affected by ongoing brain activity and measurement noise (Song et al., 2024; Goldenholz et al., 2009; Repovs, 2010). However, most existing EEG encoders process the entire temporal sequence uniformly, implicitly assuming that all time points are equally informative. This uniform treatment introduces temporal redundancy and may allow task-irrelevant fluctuations to interfere with representation learning.

To address these challenges, we introduce **BEAT**, a **B**rain-aligned visual **E**ncoder with **A**daptive **T**emporal recalibration. BEAT learns a neural-compatible visual target and introduces a plug-and-play temporal recalibration module. To address the representational mismatch, we propose a **Brain-aligned Visual Encoder** (BAE), which learns neural-consistent and semantically meaningful visual features via a dual-alignment strategy. Instead of directly adopting the image embedding space as the decoding target, BAE is trained with cross-modal contrastive learning to align its representations with neural signals, encouraging it to focus on visual attributes that are genuinely reflected in brain activity. In parallel, to further bridge the gap between human and machine vision, we introduce a multi-teacher visual distillation mechanism that transfers complementary high-level semantic representations from multiple frozen visual encoders to BAE, reducing reliance on the representational biases of any single encoder.

To address temporal sparsity, we introduce **Adaptive Temporal Recalibration** (ATR) module, which partitions each EEG channel into temporal patches and models their dependencies by adaptive temporal contextualization, enabling the network to amplify informative intervals and suppress less-informative ones. Its plug-and-play design allows it to be readily integrated with diverse EEG backbones.

By jointly adapting the visual target space and recalibrating temporally varying neural patterns, BEAT learns more compatible brain–vision representations for visual decoding. Our main contributions are summarized as follows:

- We propose **BEAT**, a unified brain–vision decoding framework that jointly adapts the visual target space and recalibrates temporally varying EEG signals, enabling more compatible cross-modal representation learning.
- BEAT comprises two complementary modules: a **Brain-aligned Visual Encoder** (BAE), which learns a neural-compatible yet semantically grounded visual target through joint brain–image alignment and multi-teacher visual distillation; and an **Adaptive Temporal Recalibration** (ATR) module, which models channel-wise temporal dependencies to emphasize visually informative signal intervals.
- Extensive experiments demonstrate that BEAT achieves state-of-the-art zero-shot EEG-to-image retrieval. Reconstruction experiments further validate the utility of the learned representations, while ablation studies confirm the effectiveness of both BAE and ATR.

2 Related Work

Visual Decoding from Neural Signals. Decoding visual signals from the human brain remains a formidable challenge in neuroscience, inspiring numerous studies that leverage deep learning methodologies. For example, NICE (Song et al., 2024) employs a self-supervised framework to align EEG features with CLIP image embeddings via multimodal contrastive learning, achieving notable decoding performance through its end-to-end training paradigm. Building on this, Li et al. (2024) propose ATM, an effective EEG encoder designed to extract salient representations from EEG signals and subsequently align them with image embeddings. More recently, several studies have attempted to mitigate the mismatch between human visual perception and machine feature spaces. For instance, Wu et al. (2025) introduce a data augmentation strategy that applies Gaussian blur to images, aiming to simulate the loss of fine-grained details inherent in biological vision. However, these methods remain constrained by their reliance on fixed, non-learnable, and inherently dense image representations as targets, which prevents them from effectively resolving the fundamental basis mismatch. In contrast, our approach uses the pretrained vision model as a semantic teacher to learn a visual target space that is both semantically grounded and compatible with neural representations.

Cross-modal Alignment via Contrastive Learning. Contrastive learning has achieved substantial success in learning shared representations across modalities, as exemplified by CLIP (Radford et al., 2021) and BLIP (Li et al., 2022). Building on these advances, neural decoding studies have aligned brain signals elicited by text (Lévy et al., 2025), speech (Défossez et al., 2023), and images (Song et al., 2024) with corresponding pretrained representation spaces. In visual decoding, the image encoder is typically frozen, and its embedding space serves as a predefined target for neural alignment. Although this paradigm provides strong semantic supervision, it also ties neural decoding to a representation space shaped by the original machine-vision objective. In contrast, our approach uses the pretrained vision encoder as a semantic teacher rather than the final decoding target. Neural-guided contrastive learning and visual distillation jointly construct a learnable visual space that preserves pretrained semantics while adapting to neural representations.

3 Method

3.1 Problem Statement

We aim to decode visual stimuli from EEG signals recorded while healthy participants viewed a sequence of natural images. Let $X \in \mathbb{R}^{C\times T}$ denote an EEG trial associated with an image I, where C and T are the numbers of channels and time points, respectively. Let $\{\mathcal{P}_k\}_{k=1}^{K}$ denote a set of frozen pretrained visual teachers, where the k-th teacher extracts a native feature representation $F_{I,k} \in \mathbb{R}^{d_k}$ from the image. We adopt a dual-teacher setting ($K = 2$) and jointly learn visual representations for brain decoding under neural supervision and semantic supervision from multiple teachers.

As illustrated in Fig. 2, BEAT consists of two principal components: 1) the **Brain-aligned Visual Encoder (BAE)**, which uses multiple teachers to supervise independent visual branches and aligns each branch with a single EEG representation through cross-modal contrastive learning; and 2) the **Adaptive Temporal Recalibration (ATR)** module that models within-channel temporal dependencies and recalibrates temporally varying EEG responses to emphasize visually informative signal intervals

3.2 Brain-aligned Visual Encoder

BAE combines semantic supervision from multiple frozen visual teachers with EEG-guided cross-modal alignment to learn a visual target space for brain decoding. Each teacher supervises an independent high-level visual branch, allowing that branch to acquire the semantic knowledge of its corresponding teacher. All visual branches are also aligned with a single EEG representation, adapting to neural representations while preserving visual knowledge from multiple sources.

1) Visual encoder. The visual encoder consists of a shared backbone and independent high-level visual branches. Given an image I, a shared visual backbone $\mathcal{G}$ first extracts common features. The k-th independent high-level branch $\mathcal{B}_k$ and visual projection head $\mathcal{Q}_k$ then produce a representation

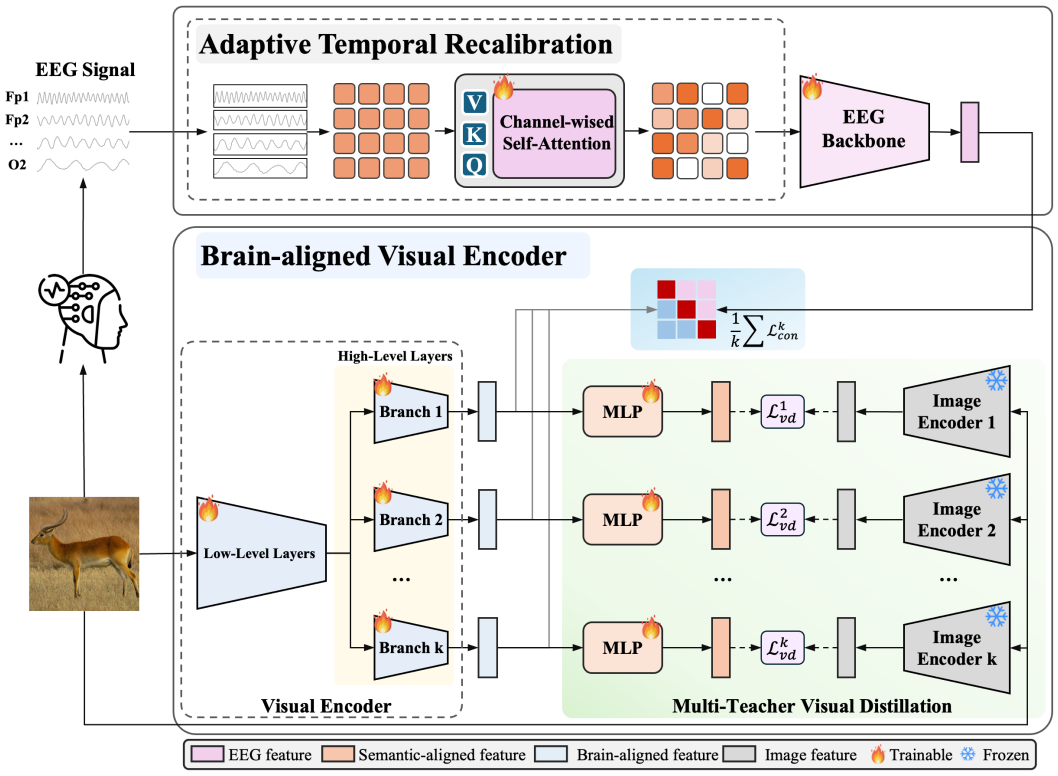


Figure 2: Overview of the proposed framework. The Adaptive Temporal Recalibration (ATR) module refines EEG representations by dynamically calibrating temporal dynamics to reduce variability across trials. The Brain-aligned Visual Encoder (BAE) uses a shared backbone and multi independent high-level visual branches, aligning visual representations with EEG representation through multi-teacher visual distillation and contrastive learning. $\mathcal{L}_{con}$ denotes the contrastive loss, and $\mathcal{L}_{vd}$ denotes the visual distillation loss.

for brain–visual alignment:

$$F_{d,k} = \mathcal{Q}_k((\mathcal{B}_k(H)) \in \mathbb{R}^d, \quad k = 1, \ldots, K. \tag{1}$$

Specifically, the visual student is a convolutional network with four residual stages. The shared backbone $\mathcal{G}$ comprises the first three residual stages, each $\mathcal{B}_k$ is a separate fourth residual stage with independent parameters, and each $\mathcal{Q}_k$ is an independent visual projection head. The shared backbone extracts common visual information, while the independent high-level branches provide separate adaptation paths for semantic supervision from different teachers. The EEG signal is processed by the ATR module $\mathcal{A}$ and the EEG encoder $\mathcal{E}$ to obtain a single neural representation shared by all visual branches:

$$F_e = \mathcal{E}\big(\mathcal{A}(X)\big) \in \mathbb{R}^d. \tag{2}$$

2) Brain–visual alignment. For a minibatch of B paired EEG–image examples, each branch jointly optimizes EEG-to-image and image-to-EEG matching using a bidirectional InfoNCE loss:

$$L_{\text{Con}} = -\frac{1}{2BK} \sum_{k=1}^{K} \sum_{i=1}^{B} \left[\log \frac{\exp(s_{ii}^{(k)})}{\sum_{j=1}^{B} \exp(s_{ij}^{(k)})} + \log \frac{\exp(s_{ii}^{(k)})}{\sum_{j=1}^{B} \exp(s_{ji}^{(k)})} \right]. \tag{3}$$

$s_{ij}^{(k)}$ is the scaled cosine similarity between the EEG (F_e) and visual representations from different branches ($F_{d,k}$), with $i = j$ indicating a matched pair. This objective supervises a single EEG representation through alignment with multiple visual branches and encourages each branch to capture visual information relevant to the paired neural responses.

Table 1: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG, with bold text highlighting the best performance. Our results are averaged over three runs with different random seeds.

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**
Intra-subject: train and test on one subject																						
NICE	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
CogCap	31.4	79.7	31.4	77.8	38.2	85.7	40.4	85.8	24.4	66.3	34.8	78.8	34.7	81.0	48.1	88.6	37.4	79.4	35.6	79.3	35.6	80.2
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
FLORA	30.0	67.0	36.0	64.0	38.5	70.5	48.0	77.0	28.0	58.0	43.5	73.0	39.0	72.0	51.5	80.0	37.5	71.0	43.0	77.5	39.5	71.0
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
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
BEAT	**85.7**	**95.7**	**81.0**	**95.7**	**76.8**	**96.5**	**73.5**	**94.0**	**75.8**	**95.0**	**88.3**	**98.2**	**80.0**	95.3	**89.0**	**98.5**	75.5	**96.2**	85.2	98.2	**81.1**	**96.3**
Inter-subject: leave one subject out for test																						
NICE	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
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
CogCap	16.3	42.3	16.2	37.9	8.8	26.8	15.4	37.6	10.1	31.7	14.0	35.4	10.7	26.9	13.9	34.2	9.0	32.4	15.3	38.6	13.0	34.4
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
FLORA	17.5	38.7	18.7	36.4	**19.8**	**37.6**	19.3	37.2	16.6	34.4	20.3	40.1	18.4	37.4	20.9	41.2	17.9	35.7	19.9	39.7	18.9	37.9
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
BEAT	**24.3**	**52.8**	**28.7**	**59.2**	16.8	35.0	**19.5**	44.2	**24.4**	**54.0**	17.7	41.7	**19.0**	**45.2**	**24.8**	**52.2**	**33.5**	**66.2**	**33.0**	**65.3**	**24.2**	**51.6**

3) Multi-teacher visual distillation. To bridge the gap between human and machine vision, we distill diverse visual features from multiple pretrained visual encoders, mitigating the representational biases of any single encoder. Specifically, we assign a frozen teacher $\mathcal{P}_k$ to each visual branch. The teacher provides a fixed target in its native feature space, while an independent student-side distillation head $\mathcal{D}_k : \mathbb{R}^d \to \mathbb{R}^{d_k}$ maps the visual representation of that branch into the corresponding teacher space:

$$\begin{aligned} F_{I,k} &= \mathrm{sg}(\mathcal{P}_k(I)), \\ \widetilde{F}_{I,k} &= \mathcal{D}_k(F_{d,k}). \end{aligned} \tag{4}$$

Here, sg denotes the stop-gradient operation. The head $\mathcal{D}_k$ consists of two linear layers with a GELU activation between them and takes the unnormalized $F_{d,k}$ as input. After applying ℓ_2 normalization to both the distillation prediction and the teacher target, we define the visual distillation loss for each branch using cosine distance:

$$L_{\mathrm{VD}} = \frac{1}{BK} \sum_{k=1}^{K} \sum_{i=1}^{B} \left[1 - \frac{\left(\widetilde{F}^i_{I,k}\right)^\top F^i_{I,k}}{\left\|\widetilde{F}^i_{I,k}\right\|_2 \left\|F^i_{I,k}\right\|_2} \right]. \tag{5}$$

Knowledge from multiple teachers is transferred to the brain-aligned visual space by jointly optimizing the student branches and the shared backbone.

4) Joint optimization and retrieval. The overall training objective is a weighted combination of the contrastive and distillation losses aggregated across branches:

$$L = \lambda_{\mathrm{Con}} L_{\mathrm{Con}} + \lambda_{\mathrm{VD}} L_{\mathrm{VD}}, \tag{6}$$

Here, λ_{Con} and λ_{VD} control the contributions of neural alignment and visual distillation, respectively. For zero-shot retrieval, we rank candidate images by the mean cosine similarity between the EEG embedding and the visual embeddings from all branches.

3.3 Adaptive Temporal Recalibration

EEG signals contain rich temporal information, but stimulus-related patterns are distributed non-uniformly and are often entangled with noise and ongoing brain activity. ATR recalibrates temporal responses independently within each channel while sharing parameters across channels. It preserves the input shape, $\mathcal{A} : \mathbb{R}^{C\times T} \to \mathbb{R}^{C\times T}$, allowing integration with different EEG backbones without modifying their input interfaces.

1) Temporal patch embedding. Given an EEG trial $X \in \mathbb{R}^{C\times T}$, ATR first applies reversible instance normalization to each channel $X^m \in \mathbb{R}^{1\times T}$. After replication padding of p_{pad} samples at the end, the normalized sequence is divided into $N = \lfloor (T + p_{\text{pad}} - P)/S \rfloor + 1$ overlapping patches of length P with stride S. The patches are projected into D-dimensional embeddings:

$$z_0^m = [\phi(\bar{X}_1^m); \ldots; \phi(\bar{X}_N^m)] + E_{\text{pos}}, \tag{7}$$

where $\bar{X}_n^m$ denotes a normalized temporal patch, $\phi : \mathbb{R}^P \to \mathbb{R}^D$ is a shared linear embedding layer, and E_{pos} denotes learnable positional embeddings.

2) Temporal contextualization. The patch sequence is processed by L Transformer encoder layers:

$$z_l^m = \mathcal{T}_l(z_{l-1}^m), \qquad l = 1, \ldots, L, \tag{8}$$

where $\mathcal{T}_l$ comprises multi-head self-attention with residual attention and a feed-forward sublayer. Attention contextualizes each patch using temporal evidence within the same channel, enabling ATR to emphasize informative intervals and suppress less-relevant activity. The contextualized tokens are flattened and mapped back to the original temporal length, followed by denormalization:

$$\begin{aligned} \tilde{X}^m &= \mathcal{R}_m^{-1}\left(\mathcal{H}\left(\text{Flatten}(z_L^m)\right)\right), \\ \tilde{X} &= \text{Stack}\left(\tilde{X}^1, \ldots, \tilde{X}^C\right) = \mathcal{A}(X) \in \mathbb{R}^{C\times T}, \end{aligned} \tag{9}$$

where $\mathcal{H} : \mathbb{R}^{ND} \to \mathbb{R}^T$ is a shared reconstruction head, and $\mathcal{R}_m^{-1}$ restores the normalization statistics of channel m. ATR thus preserves the channel organization and temporal length while recalibrating the temporal content.

4 Implementation Details

Datasets. THINGS-EEG (Gifford et al., 2022) is a large-scale EEG dataset that includes data from 10 subjects with the RSVP paradigm. The training set includes 1654 concepts with 10 images for each concept. The test set includes 200 concepts with one image per concept for each subject.

Brain-aligned Visual Encoder (BAE). By default, we use OpenCLIP (Ilharco et al., 2021) ViT-B/16 and ViT-B/32 as two frozen visual teachers, each with a native feature dimensionality of 512. The visual student uses a pretrained ResNet-50 with shared layer1–3 and two separate layer4 branches with independent parameters. Both layer4 branches are initialized with the same pretrained weights, while their visual projection heads are initialized independently and each produces a 128-dimensional representation. A single EEG encoder outputs a 128-dimensional representation that is aligned with each of the two visual representations. Each student-side distillation head consists of two linear layers with dimensions $128 \to 256 \to 512$ and intermediate GELU activation, and is trained by minimizing cosine distance in the native teacher space. The contrastive and distillation losses are each averaged over the two branches and then combined with $\lambda_{\text{Con}} = \lambda_{\text{VD}} = 1$.

Adaptive Temporal Recalibration (ATR). The ATR module uses a patch size of 10 and a stride of 5. It is implemented as a transformer encoder with 3 layers, each containing 4 attention heads. Each block employs a model dimensionality of 64 and a feed-forward dimension of 128.

EEG Encoders. To assess the generalizability of BEAT across neural backbones, we evaluate five EEG encoders: *EEGProjector* (Wu et al., 2025), *EEGNet* (Lawhern et al., 2018), *TSConv* (Song et al., 2024), *ShallowNet* and *DeepNet* (Schirrmeister et al., 2017). Unless otherwise specified, EEGProjector is used as the default encoder.

Baselines. We compare BEAT with NICE (Song et al., 2024), ATM (Li et al., 2024), CogCap (Zhang et al., 2025), UBP (Wu et al., 2025), FLORA (Li et al., 2025), NeuroBridge (Zhang et al., 2026),HVF (Zheng et al., 2026).

Evaluation metrics. We evaluate BEAT on two tasks: zero-shot image retrieval and EEG-based image reconstruction. For retrieval, we report Top-1 and Top-5 accuracy in a 200-way setting. For reconstruction, we assess both low-level image fidelity and high-level semantic similarity, with detailed metric definitions provided in the appendix.

Image reconstruction. Following Li et al. (2024), we use a two-stage reconstruction pipeline: a diffusion prior predicts CLIP image embeddings conditioned on our EEG embeddings aligned with

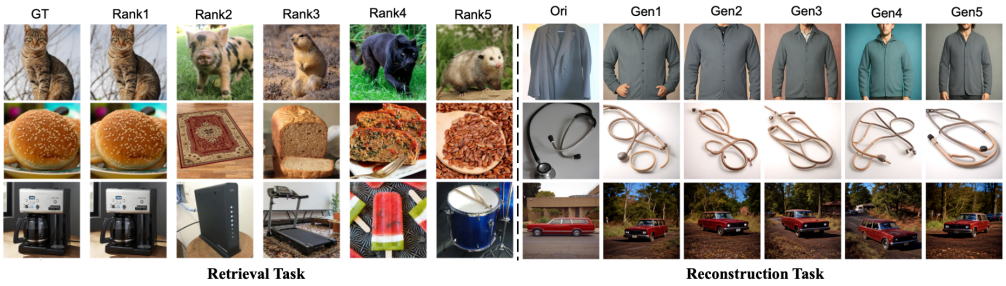


Figure 3: Examples of Top-5 retrieval results in EEG-image retrieval tasks (left panel). The comparison between the original image and the image generated using the corresponding EEG data (right panel). GT denotes ground truth; Ori denotes original image.

Table 2: Quantitative assessments of the reconstruction quality on THINGS-EEG.

Method	PixCorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	SwAV↓
ATM	0.135	0.341	0.759	0.837	0.695	0.739	0.560
CogCap	0.150	0.347	0.754	0.623	0.669	0.715	0.590
FLORA	0.101	**0.360**	0.595	0.622	0.604	0.812	0.661
HVF	**0.195**	0.336	**0.843**	0.905	0.756	0.808	0.554
BEAT	0.154	0.344	0.824	**0.914**	**0.927**	**0.924**	**0.485**

multiple visual teachers. The predicted CLIP embeddings then guide SDXL (Podell et al., 2024) through IP-Adapter (Ye et al., 2023). Further details are provided in the Appendix.

Training setup. The learning rate is set to 1×10^{-3}. The visual encoder is trained with a $0.1\times$ smaller learning rate relative to the base rate. All models are optimized using AdamW with a weight decay of 1×10^{-4}. All retrieval results are reported as the average over three runs with different random seeds to ensure statistical reliability. The experiments are performed on one H200 GPU.

5 Results

5.1 Visual Decoding Performance

Retrieval performance. Table 1 shows quantitative comparisons between our approach and baselines on the EEG datasets. Our approach outperforms previous state-of-the-art in intra-subject settings. Notably, BEAT achieves a Top-1 accuracy of 81.1% and Top-5 accuracy of 96.3% for the zero-shot retrieval task on the THINGS-EEG dataset. Although inter-subject brain decoding is challenging due to individual variability, BEAT improves the Top-1 and Top-5 accuracies by 5.2% and 5.7%. These improvements suggest that the combination of adaptive temporal recalibration and brain-aligned visual representation learning helps capture stimulus-related EEG information, leading to better cross-modal alignment and more accurate retrieval. We also present the Top-5 retrieved images corresponding to the real visual stimulus seen by subjects in the left panel of Fig. 3.

Reconstruction performance. We align the brain signals induced by visual stimuli with the embeddings of images and then use a generative model to produce images. The examples in Fig. 3 (right) illustrate semantic similarity between reconstructed images and the corresponding visual stimuli. To further evaluate the reconstruction quality in a quantitative manner, we adopt a set of representative low-level and high-level metrics and the results are reported in Table 2. BEAT surpasses other approaches in key high-level metrics, highlighting its ability to capture rich semantic information from noisy EEG signals. Additional reconstruction examples are provided in the Appendix.

5.2 Ablation Study

We conduct ablation studies on THINGS-EEG under the intra-subject setting. For each EEG backbone, the baseline directly aligns EEG representations with fixed visual embeddings from a frozen

Table 3: Average Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG across ten subjects. Each EEG backbone is evaluated with ATR, BAE, and their combination. Bold denotes the best result within each backbone.

	Baseline		+ ATR		+ BAE		BEAT	
Backbone	**Top-1**	**Top-5**	**Top-1**	**Top-5**	**Top-1**	**Top-5**	**Top-1**	**Top-5**
EEGProjector	37.85	68.28	46.95	77.30	51.62	80.93	**81.08**	**96.32**
EEGNet	22.70	50.48	34.97	67.05	47.63	78.75	**70.77**	**93.17**
TSConv	35.08	67.43	43.20	75.00	52.03	80.52	**69.10**	**91.62**
ShallowNet	26.38	58.38	34.57	66.85	45.10	76.43	**65.30**	**90.03**
DeepNet	32.17	60.87	36.90	66.55	57.50	85.10	**64.78**	**90.33**

ViT-B/16 encoder. EEGProjector is used unless otherwise specified. All results are averaged over three runs with different random seeds.

Ablation on core components. Table 3 evaluates ATR, BAE, and their combination across five EEG backbones. Both modules consistently improve retrieval performance over the baseline, and their combination achieves the highest Top-1 and Top-5 accuracy for every backbone. Specifically, adding ATR improves Top-1 accuracy by 4.73%–12.27% across the five backbones, supporting its effectiveness across different EEG architectures. Adding BAE yields larger improvements of 13.77%–25.33%, demonstrating the benefit of learning a brain-aligned visual target space over using fixed visual embeddings. Combining both modules further improves performance over either module alone. For example, with EEGProjector, Top-1 accuracy increases from 37.85% to 46.95% with ATR and 51.62% with BAE, and reaches 81.08% with both. These results support the complementary contributions of temporal recalibration and visual representation adaptation.

Ablation on different distillation settings. Table 4 compares direct alignment, single-teacher distillation, and multi-teacher distillation, with ATR retained in all settings. Direct alignment with frozen ViT-B/16 and ViT-B/32 embeddings achieves Top-1 accuracies of 46.95% and 50.08%, respectively. Introducing a trainable visual student with both contrastive alignment and single-teacher distillation raises these accuracies to 78.78% and 78.83%, highlighting the benefit of learning visual targets under joint neural and semantic supervision. Using both teachers to supervise separate student branches further improves Top-1 and Top-5 accuracy to 81.08% and 96.32%, outperforming both single-teacher settings. This further improvement supports the effectiveness of incorporating supervision from multiple visual teachers within BAE.

Additional module ablations and hyperparameter sensitivity analyses are provided in the Appendix.

Table 4: Intra-subject retrieval performance under different distillation settings.

Setting	**Configuration**	**Intra-subject**	
		Top-1 (%)	**Top-5 (%)**
Direct alignment	B16	46.95 ± 0.39	77.30 ± 0.44
	B32	50.08 ± 0.75	80.05 ± 0.48
Single-teacher distillation	B16	78.78 ± 1.50	95.60 ± 0.17
	B32	78.83 ± 1.23	95.48 ± 0.37
Multi-teacher distillation	B16+B32	**81.08 ± 0.90**	**96.32 ± 0.30**

5.3 Learning a Neural-Compatible Visual Target

Having established the effectiveness of BAE, we analyze the properties that make a visual representation suitable for neural alignment. As shown in Fig. 4A, standard pretrained models (CLIP, ResNet50) require 101-104 principal components to explain 90% of its variance, whereas BAE requires only 22–30 components across all tested dimensions (128 to 1024). Crucially, the Top-1 classification accuracy remains robust between 80.9% and 82.1%, confirming this low-dimensional structure reflects concentrated semantic distillation rather than trivial representation collapse. Since neural signals inherently possess low intrinsic dimensionality, aligning them to widely distributed visual space easily leads to cross-modal overfitting. A learned, compact visual representation may

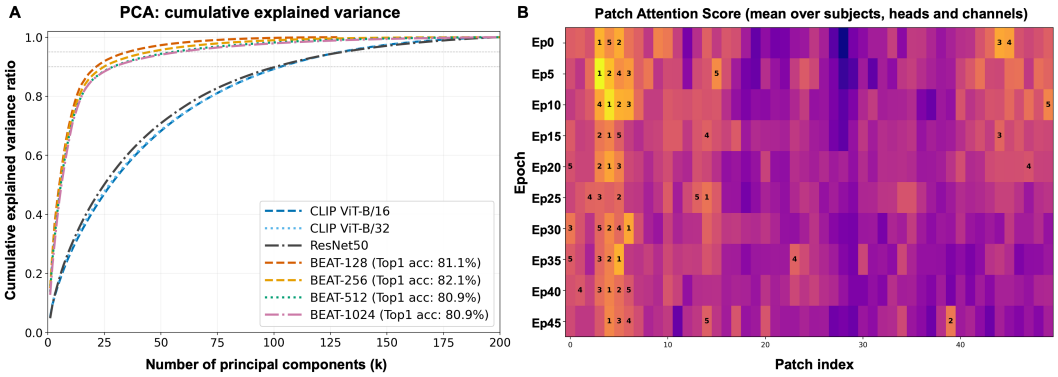


Figure 4: Analysis of the visual target space. (A) PCA spectra of CLIP, ResNet and BAE representations. (B) Temporal patch attention learned by ATR across training epochs, Numbers 1–5 indicate the ranks of the five highest-attention patches at each epoch.

therefore provide a more neurally compatible alignment target than the full pretrained representation.

Table 5: Retrieval performance under different test-time temporal masking strategies. Δ denotes the absolute change from the unmasked condition in percentage points.

Condition	**Top-1 (%)**	**ΔTop-1**	**Top-5 (%)**	**ΔTop-5**
Unmasked	46.95 ± 0.39	—	77.30 ± 0.44	—
Mask lowest-attention	40.37 ± 0.20	-6.58	73.42 ± 0.85	-3.88
Random mask	36.30 ± 0.88	-10.65	67.98 ± 0.51	-9.32
Mask highest-attention	18.17 ± 2.35	-28.78	43.42 ± 3.83	-33.88

5.4 Temporal Informativeness of Neural Signals

The window-wise retrieval analysis in Fig. 1D provides an independent estimate of temporal informativeness, suggesting that decodable visual information is concentrated primarily within the first 400 ms after stimulus onset. Consistent with this profile, the last-layer attention scores learned by ATR, averaged over subjects, attention heads, and EEG channels, are generally higher for early temporal patches and lower for later patches (Fig. 4B). To further assess whether these attention scores reflect the contribution of temporal patches to retrieval, we conducted a test-time masking analysis. After computing ATR attention scores on the test set, we masked 20% of temporal patches using three strategies: selecting patches with the lowest attention scores, selecting patches at random, or selecting patches with the highest attention scores. Implementation details are provided in the appendix. As shown in Table 5, masking the highest-attention patches produced the largest decline in retrieval performance. Masking the lowest-attention patches had the smallest effect, while random masking yielded intermediate decreases. Together, the window-wise decoding profile and masking results suggest that ATR places greater temporal emphasis on intervals with stronger standalone decodability and that patches receiving higher attention contribute more to retrieval performance.

6 Conclusion

We present BEAT, a unified framework for EEG-based visual decoding that jointly adapts visual alignment targets and recalibrates temporal neural signals. On the visual side, BAE integrates complementary semantic guidance from multiple pretrained vision teachers to learn a neural-compatible visual space, moving beyond alignment with a single fixed visual representation. On the neural side, ATR adaptively emphasizes informative temporal responses before EEG encoding. Experiments on zero-shot image retrieval and reconstruction demonstrate the effectiveness of this joint design, supported by module ablations and representation analyses.

AI USE STATEMENT

We used LLM to assist with manuscript editing, including grammar checking, language polishing and adjusting LaTeX formatting.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the proposed method and experimental setup in Sections 3 and 4. The supplementary material provides dataset descriptions and preprocessing procedures in Appendix A, detailed model architectures in Appendix B, and optimization settings, data augmentation, random seeds, evaluation protocols, and the computing environment in Appendix C.

ETHICS STATEMENT

This study uses the publicly available datasets. We did not recruit new participants or collect additional human-subject data.

REFERENCES

Yohann Benchetrit, Hubert Banville, and Jean-Remi King. Brain decoding: toward real-time reconstruction of visual perception. In *International Conference on Learning Representations*, 2024.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5 (10):1097–1107, 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.

Daniel M Goldenholz, Seppo P Ahlfors, Matti S Hämäläinen, Dahlia Sharon, Mamiko Ishitobi, Lucia M Vaina, and Steven M Stufflebeam. Mapping the signal-to-noise-ratios of cortical sources in magnetoencephalography and electroencephalography. *Human brain mapping*, 30(4):1077–1086, 2009.

Zhanqiang Guo, Jiamin Wu, Yonghao Song, Jiahui Bu, Weijian Mai, Qihao Zheng, Wanli Ouyang, and Chunfeng Song. Neuro-3d: Towards 3d visual decoding from eeg signals. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23870–23880, 2025.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Nikolaus Kriegeskorte and Pamela K Douglas. Cognitive computational neuroscience. *Nature neuroscience*, 21(9):1148–1160, 2018.

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.

Jarod Lévy, Mingfang Zhang, Svetlana Pinet, Jérémy Rapin, Hubert Banville, Stéphane d'Ascoli, and Jean-Rémi King. Brain-to-text decoding: A non-invasive approach via typing. *arXiv preprint arXiv:2502.17480*, 2025.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 102822–102864, 2024.

Dongyang Li, Haoyang Qin, Mingyang Wu, Chen Wei, and Quanying Liu. Brainflora: Uncovering brain concept representation via multimodal neural embeddings. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5577–5586, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

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 *International Conference on Machine Learning*, 2025.

Grega Repovs. Dealing with noise in eeg recording and data analysis. In *Informatica Medica Slovenica*, volume 15, pp. 18–25, 2010.

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 *International Conference on Learning Representations*, 2024.

Yonghao Song, Yijun Wang, Huiguang He, and Xiaorong Gao. Recognizing natural images from eeg with language-guided contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2025. doi: 10.1109/TNNLS.2025.3562743.

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.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

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.

Jiawen Zheng, Haonan Jia, Ming Li, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. *arXiv preprint arXiv:2602.07495*, 2026.

APPENDIX FOR LEARNING BRAIN-ALIGNED VISUAL REPRESENTATION WITH ADAPTIVE TEMPORAL RECALIBRATION FOR EEG-BASED VISUAL DECODING

Anonymous authors
Paper under double-blind review

CONTENTS

A Dataset Details **3**
A.1 THINGS-EEG . . . 3
A.2 THINGS-MEG . . . 3

B Detailed Model Architecture **4**
B.1 Overall Framework and Training Flow . . . 4
B.2 Brain-Aligned Visual Encoder . . . 4
B.3 Adaptive Temporal Recalibration . . . 5
B.4 EEG Encoders . . . 5

C Implementation Details **6**
C.1 Optimization . . . 6
C.2 Data Augmentation . . . 6
C.3 Random Seeds and Reproducibility . . . 6
C.4 Zero-Shot Retrieval Protocol . . . 6
C.5 Reconstruction Protocol and Metrics . . . 7
C.6 Computing Environment . . . 7

D Additional Ablation Studies **7**
D.1 Compatibility with Different EEG Encoders . . . 7
D.2 Hyperparameter Sensitivity . . . 8

E Temporal Decoding Analysis **9**

F Analysis of the Brain-Aligned Visual Encoder **10**
F.1 Visual Backbone and Initialization . . . 10
F.2 Embedding Dimension and Intrinsic Dimensionality . . . 10
F.3 PCA Analysis . . . 11

G Analysis of Adaptive Temporal Recalibration **12**
G.1 Attention-Guided Temporal Perturbation . . . 12

H Additional Retrieval and Reconstruction Results **12**

H.1 More Retrieval examples 12

H.2 More Reconstruction examples 14

A Dataset Details

A.1 THINGS-EEG

Things-EEG (Gifford et al., 2022) is a large-scale EEG dataset including 10 subjects recorded under a rapid serial visual presentation paradigm. The training set contains 1,654 concepts, each associated with 10 images, and each image is repeated 4 times (1,654 concepts $\times$ 10 images/concept $\times$ 4 trials/image) per subject. The test set contains 200 concepts, each associated with 1 image, and each image is repeated 80 times (200 concepts $\times$ 1 image/concept $\times$ 80 trials/image) per subject. For data preprocessing, we follow the procedure described in Song et al. (2024). The raw EEG data are band-pass filtered to [0.1, 100] Hz, recorded from 63 channels at a sampling rate of 1,000 Hz. The EEG data are epoched into trials from 0 to 1,000 ms after stimulus onset, with baseline correction using the mean over the preceding 200 ms. The epoched data are then down-sampled to 250 Hz. Trial repetitions are averaged, resulting in a total of 16,540 training samples and 200 test samples per subject. In the final data loader, repetitions are averaged before model input, producing tensors of shape $16{,}540 \times 63 \times 250$ for training and $200 \times 63 \times 250$ for testing, per subject.

A.2 THINGS-MEG

To evaluate the generalizability of our method, we followed prior work (Wu et al., 2025; Zhang et al., 2026) and conducted additional experiments on the THINGS-MEG (Hebart et al., 2023) dataset, which contains data from four human participants performing the same visual recognition task. THINGS-MEG includes data from four participants and comprises 271 channels. The experimental design uses a stimulus duration of 500 ms, followed by a blank screen lasting 1,000 ± 200 ms. The training stage consists of 1,854 concepts $\times$ 12 images $\times$ 1 repetition, and the test stage consists of 200 concepts $\times$ 1 image $\times$ 12 repetitions. During data processing, the 200 test concepts are removed from the training set to construct a zero-shot setting, following the procedure in the original study. The MEG data are then epoched into trials from 0 to 1,000 ms after stimulus onset. For preprocessing, a band-pass filter in the range [0.1, 100] Hz is applied, the data are down-sampled to 200 Hz, and baseline correction is performed. The resulting intra- and inter-subject zero-shot retrieval performance is summarized in Table S1.

Table S1: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-MEG with bold text highlighting the best performance.

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**
Intra-subject: train and test on one subject										
NICE	6.4	23.4	16.2	40.8	14.0	38.9	6.4	25.2	10.8	32.1
ATM	12.5	30.0	36.0	65.5	25.0	50.0	15.0	36.0	22.1	45.3
CogCap	13.5	33.9	31.8	71.3	24.0	53.8	14.9	35.6	22.7	48.9
UBP	15.0	38.0	46.0	80.5	27.3	59.0	**18.5**	**43.5**	26.7	55.2
FLORA	13.9	35.2	41.1	78.4	26.1	57.3	15.9	36.9	24.5	52.6
NeuroBridge	**16.5**	**41.6**	53.7	85.3	40.4	73.2	18.1	43.1	32.2	**60.8**
HVF	14.0	31.8	63.8	**91.8**	41.0	78.3	17.0	41.0	33.9	60.7
BEAT	13.3	32.7	**71.0**	91.0	**52.2**	**79.3**	16.8	39.3	**38.3**	60.6
Inter-subject: leave one subject out for test										
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
BEAT	2.0	9.5	**6.8**	**19.8**	**7.0**	**22.0**	2.3	10.3	**4.5**	**15.4**

B Detailed Model Architecture

B.1 Overall Framework and Training Flow

BEAT is formulated as a multi-teacher distillation framework with K frozen visual teachers, each supervising a corresponding student branch. For an averaged EEG trial $X \in \mathbb{R}^{63\times250}$ and its paired image I, ATR recalibrates the temporal sequence and a single EEG encoder produces $F_e \in \mathbb{R}^d$, with $d = 128$ by default. The visual student shares the ResNet-50 stem and residual stages 1–3, followed by K independent stage-4 branches and projection heads that produce $\{F_d^{(k)} \in \mathbb{R}^d\}_{k=1}^K$. All visual branches align with the same EEG embedding. A separate head maps each branch embedding to its corresponding teacher's native feature space for cosine feature distillation. ATR, the EEG encoder, the shared visual trunk, all visual branches, and all heads are jointly trainable in the default configuration.

The teacher count K specifies a configuration of the framework. In our default experimental implementation, we set $K = 2$, using frozen OpenCLIP ViT-B/16 and ViT-B/32 teachers to supervise two corresponding visual branches. The training procedure below is written for general K, while the concrete teacher choices and network dimensions describe this default implementation.

B.2 Brain-Aligned Visual Encoder

Frozen Visual Teachers

Each frozen teacher provides a target in its native feature space, whose dimension d_k may differ across teachers.For the default $K = 2$ implementation, the teachers are OpenCLIP ViT-B/16 (`laion2b_s34b_b88k`) and ViT-B/32 (`laion2b_s34b_b79k`), each with a native output dimension of 512. Images have spatial resolution 224×224 and are converted to tensors. Teacher features are cached separately and detached and ℓ_2-normalized when computing the distillation loss. Teacher parameters are excluded from optimization. Each visual branch uses its assigned teacher's native feature target.

Trainable Visual Encoder

The visual student is initialized from the ImageNet ResNet-50 checkpoint `resnet50-19c8e357.pth`. Its stem (convolution, batch normalization, ReLU, and max pooling) and residual stages 1–3 are shared. The final residual stage, `layer4`, is copied into K independently trainable branches initialized with the same pretrained weights, giving two branches in the default implementation. Each branch is globally average-pooled to 2,048 dimensions and

Algorithm 1 Training procedure of BEAT with K frozen visual teachers ($K = 2$ by default).

Require: Paired EEG trials and images (X, I); ATR $\mathcal{A}$; EEG encoder $\mathcal{E}$; shared visual backbone $\mathcal{G}$; visual branches $\mathcal{B}_k$; distillation heads $\mathcal{D}_k$; frozen visual teachers $\mathcal{P}_k$, $k = 1, \ldots, K$; loss weights λ_{Con} and λ_{VD}.

for each training minibatch (X, I) **do**
  Compute the shared EEG representation $F_e \leftarrow \mathcal{E}(\mathcal{A}(X))$.
  Extract shared visual features $H \leftarrow \mathcal{G}(I)$.
  for $k = 1, \ldots, K$ **do**
    Compute the branch representation $F_{d,k} \leftarrow \mathcal{B}_k(H)$.
    Obtain the frozen teacher target $F_{I,k} \leftarrow \text{sg}(\mathcal{P}_k(I))$.
    Predict the teacher-space feature $\widetilde{F}_{I,k} \leftarrow \mathcal{D}_k(F_{d,k})$.
  end for
  Compute L_{Con} by averaging the bidirectional InfoNCE losses between F_e and $F_{d,k}$ over all K branches.
  Compute L_{VD} by averaging the cosine distances between ℓ_2-normalized $\widetilde{F}_{I,k}$ and $F_{I,k}$ over the minibatch and all K branches.
  $L \leftarrow \lambda_{\text{Con}} L_{\text{Con}} + \lambda_{\text{VD}} L_{\text{VD}}$.
  Update $\mathcal{A}$, $\mathcal{E}$, $\mathcal{G}$, $\{\mathcal{B}_k\}_{k=1}^K$, and $\{\mathcal{D}_k\}_{k=1}^K$ by minimizing L, keeping all $\mathcal{P}_k$ frozen.
end for

passed through its own $\mathrm{Linear}(2048, 512)$–ReLU–$\mathrm{Linear}(512, d)$ head. The K projection heads are initialized independently.

VISUAL DISTILLATION AND TRAINING OBJECTIVES

Each student-side distillation head $\mathcal{D}_k : \mathbb{R}^d \to \mathbb{R}^{d_k}$ maps the unnormalized branch representation $F_{d,k}$ into the native feature space of its corresponding frozen teacher $\mathcal{P}_k$. The head consists of $\mathrm{Linear}(d, 256)$–GELU–$\mathrm{Linear}(256, d_k)$, with $d_1 = d_2 = 512$ for the default teacher pair. The teacher target and student prediction are

$$F_{I,k} = \mathrm{sg}(\mathcal{P}_k(I)), \qquad \widetilde{F}_{I,k} = \mathcal{D}_k(F_{d,k}), \tag{1}$$

where sg denotes the stop-gradient operation. Visual distillation uses the cosine distance between the ℓ_2-normalized prediction and teacher target, averaged over the B paired examples in a minibatch and all K branches:

$$L_{\mathrm{VD}} = \frac{1}{BK} \sum_{k=1}^{K} \sum_{i=1}^{B} \left[1 - \frac{(\widetilde{F}^i_{I,k})^\top F^i_{I,k}}{\|\widetilde{F}^i_{I,k}\|_2 \|F^i_{I,k}\|_2} \right]. \tag{2}$$

For brain–visual alignment, let $s^{(k)}_{ij}$ denote the scaled cosine similarity between the EEG representation of example i and the visual representation of example j from branch k. The bidirectional InfoNCE objective averages EEG-to-image and image-to-EEG matching losses across all branches:

$$L_{\mathrm{Con}} = -\frac{1}{2BK} \sum_{k=1}^{K} \sum_{i=1}^{B} \left[\log \frac{\exp(s^{(k)}_{ii})}{\sum_{j=1}^{B} \exp(s^{(k)}_{ij})} + \log \frac{\exp(s^{(k)}_{ii})}{\sum_{j=1}^{B} \exp(s^{(k)}_{ji})} \right]. \tag{3}$$

where $\lambda_{\mathrm{Con}} = \lambda_{\mathrm{VD}} = 1$ by default. By default, we use $K = 2$ and jointly optimize ATR, the EEG encoder, the shared visual backbone, the visual branches, and the distillation heads while keeping all visual teachers frozen.

B.3 ADAPTIVE TEMPORAL RECALIBRATION

TEMPORAL PATCH EMBEDDING

ATR receives a batch of shape $B \times 63 \times 250$ and processes every EEG channel as an independent univariate sequence. Replication padding of five samples is appended at the end. With patch length $P = 10$ and stride $S = 5$, unfolding yields 50 overlapping patches per channel. A shared linear layer maps each 10-sample patch to a 64-dimensional token. Learned positional embeddings are added to the 50 tokens. The Transformer weights and reconstruction head are shared across all EEG channels.

The temporal tokens are processed by three channel-independent Transformer encoder layers, each with four attention heads, model dimension 64, feed-forward dimension 128, GELU activation, and residual attention. Residual and feed-forward dropout are 0.3, and attention-probability dropout is 0. The positional encoding is learnable. Reversible instance normalization is enabled without affine parameters: each channel is normalized before patch processing and denormalized after reconstruction. A shared flattening head maps the 50×64 token features back to 250 samples per channel, with reconstruction-head dropout set to 0.

B.4 EEG ENCODERS

The default EEG projector flattens the recalibrated 63×250 signal, applies $\mathrm{Linear}(15750, d)$, then a residual GELU–$\mathrm{Linear}(d, d)$–Dropout(0.3) block and LayerNorm. The same EEG representation supervises all K visual branches. The configurable dimension d can be 128, 256, 512, or 1,024 for `EEGProjectLayer`; the alternative EEG backbones currently require $d = 128$. The default EEG encoder is `EEGProjectLayer`. We additionally evaluate EEGNet, TSConv, ShallowNet, and DeepNet. Table S2 compares the capacity of the downstream EEG encoders. Each count includes the encoder backbone and its final 128-dimensional projection head. The shared ATR module is excluded so that the table isolates differences among EEG encoders; ATR contains 0.90M parameters and can be added equally to every row when reporting the complete EEG path.

Table S2: Parameter counts of the EEG encoders. All encoders receive 63-channel inputs and produce 128-dimensional embeddings. The shared ATR module is excluded.

EEG Encoder	Parameters
EEGProjector	2.03M
EEGNet	0.19M
TSConv	0.30M
ShallowNet	0.30M
DeepNet	0.50M

C Implementation Details

C.1 Optimization

The hyperparameters used for intra-subject training are summarized in Table S3.

Table S3: Hyperparameters used for intra-subject training.

Hyperparameter	Value
Train/test batch size	256/200
Maximum training epochs	50
Optimizer	AdamW
EEG/ATR learning rate	1×10^{-3}
Visual encoder learning rate	1×10^{-4}
Heads learning rate	1×10^{-3}
Weight decay	1×10^{-4}
Learning-rate scheduler	CosineAnnealingLR
Scheduler parameters	$T_{\max} = 50$, $\eta_{\min} = 10^{-5}$
Shared EEG / visual branch dimension	$d = 128$
Loss weights $(\lambda_{\mathrm{Con}}, \lambda_{\mathrm{VD}})$	$(1, 1)$
Initial contrastive temperature	0.07 per branch; learnable logit scales

AdamW optimizes the EEG encoder (including ATR and the two learnable logit scales), the complete visual student, and the two heads in three parameter groups. Their learning rates are 10^{-3}, 10^{-4}, and 10^{-3}, respectively, with weight decay 10^{-4}.

C.2 Data Augmentation

No online EEG augmentation is applied; repeated trials are averaged and the training loader shuffles the resulting examples. The default configuration reads the preprocessed 250-Hz whitened EEG data, uses all 63 channels, and selects samples 0–249. Images are loaded at 224×224 and normalized with the OpenCLIP statistics in Section B.2. Following Wu et al. (2025), the default visual input is a fixed medium FoveaBlur image generated offline with an exponential profile, kernel size 51, and $g = 3$.

C.3 Random Seeds and Reproducibility

The experiments use seeds 0, 1, and 2. `pytorch_lightning.seed_everything` is called before model and data-loader construction. Ten subject-specific models are trained for each seed, giving 30 runs for the intra-subject main result.

C.4 Zero-Shot Retrieval Protocol

The test set comprises one averaged EEG trial and one candidate image for each of 200 concepts absent from the training concepts. Let $e_i = F_{e,i}/\|F_{e,i}\|_2$ and $z_j^{(k)} = F_{d,j}^{(k)}/\|F_{d,j}^{(k)}\|_2$. Retrieval ranks

images by the mean of the two branch cosine similarities:

$$s_{ij} = \frac{1}{2}\left(e_i^\top z_j^{(1)} + e_i^\top z_j^{(2)}\right) = e_i^\top \frac{z_j^{(1)} + z_j^{(2)}}{2}. \tag{4}$$

The mean visual embedding is not normalized again, and the learnable contrastive logit scales are not applied at retrieval. For each EEG query, all 200 candidate images are ranked; Top-k accuracy records whether the paired image occurs among the first k candidates. Chance performance is 0.5% for Top-1 and 2.5% for Top-5. Intra-subject models are trained and evaluated independently for each subject.

C.5 Reconstruction Protocol and Metrics

Two-stage image generation. Following Li et al. (2024), the first stage uses a diffusion prior with a lightweight U-Net to model CLIP image embeddings conditioned on EEG embeddings. Starting from a noisy embedding, the prior iteratively denoises it across diffusion timesteps under EEG conditioning to predict a CLIP image embedding. In the second stage, this predicted embedding is supplied to IP-Adapter (Ye et al., 2023), which conditions the pretrained SDXL (Podell et al., 2024) image generator to synthesize the reconstruction. This separates the mapping from EEG to visual embeddings from the subsequent image-generation process. Subject-wise quantitative results and additional reconstruction examples are provided in Section H.

Following Scotti et al. (2023), we evaluated reconstruction fidelity using a representative set of low- and high-level image metrics. PixCorr was computed as the mean Pearson correlation between the flattened pixels of each ground-truth image and its paired reconstruction, whereas SSIM measured their structural similarity. For SwAV, features were extracted from the average-pooling layer of a pretrained SwAV-ResNet50, and the correlation distance between paired ground-truth and reconstructed-image features was averaged across test samples; lower values indicate greater similarity. AlexNet, InceptionV3, and CLIP scores were calculated using two-way identification. Specifically, the Pearson correlation between the feature embeddings of a ground-truth image and its paired reconstruction was compared with the correlation between that image and each non-matching reconstruction. A comparison was considered correct when the paired correlation was higher. Accuracy was averaged over all pairwise comparisons and test images, with chance performance at 50%. Features were obtained from the second and fifth convolutional layers of AlexNet, the final pooling layer of InceptionV3, and the final image-embedding layer of CLIP ViT-L/14.

C.6 Computing Environment

All experiments were conducted on a single NVIDIA H200 GPU with 20 CPU cores and 200 GB of host memory. BEAT was implemented in Python 3.10 using PyTorch and PyTorch Lightning with CUDA acceleration.

D Additional Ablation Studies

D.1 Compatibility with Different EEG Encoders

We evaluate ATR and BAE with five EEG encoders: EEGProjector, EEGNet, TSConv, ShallowNet, and DeepNet. For each encoder, we compare the baseline, ATR-only, BAE-only, and combined ATR+BAE configurations under the same 200-way zero-shot retrieval protocol. The subject-wise Top-1 and Top-5 accuracies are reported in Table S4. For EEGProjector, the ATR+BAE configuration corresponds to the main BEAT experiment.

BAE alone improves both retrieval metrics over the corresponding baseline for all five encoders, showing that its benefit is not restricted to EEGProjector. Combining ATR and BAE also produces substantial improvements over every encoder baseline. The combined configuration gives the highest Top-1 and Top-5 averages for EEGProjector, EEGNet, TSConv, and ShallowNet. Overall, the results show that both modules can be applied across different EEG encoding architectures.

Table S4: Top-1 and Top-5 accuracy (%) of 200-way zero-shot retrieval on THINGS-EEG across different EEG encoders and module configurations.

Configuration	S1		S2		S3		S4		S5		S6		S7		S8		S9		S10		Average	
	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5
EEGProjector	30.83	61.00	34.67	64.83	36.67	66.83	41.33	73.83	30.33	58.33	41.83	74.50	36.00	63.67	47.50	79.33	36.50	67.17	42.83	73.33	37.85	68.28
+ ATR	39.83	68.33	43.67	74.67	49.17	79.50	49.67	83.50	36.67	67.33	52.17	80.17	41.83	72.00	57.83	87.50	46.50	76.67	52.17	83.33	46.95	77.30
+ BAE	52.83	82.83	37.50	72.50	64.17	90.17	36.83	70.67	30.50	62.33	42.33	74.17	53.83	85.33	77.67	95.17	52.67	84.83	67.83	91.33	51.62	80.93
+ ATR + BAE	85.67	95.67	81.00	95.67	76.83	96.50	73.50	94.00	75.83	95.00	88.33	98.17	80.00	95.33	89.00	98.50	75.50	96.17	85.17	98.17	81.08	96.32
EEGNet	17.50	41.17	19.17	47.33	20.50	48.17	26.83	57.33	15.50	39.50	26.33	56.00	19.50	48.17	29.67	59.83	23.67	48.67	28.33	58.67	22.70	50.48
+ ATR	33.50	63.00	33.00	66.83	36.17	67.00	32.83	68.00	27.33	57.67	40.00	71.33	35.17	63.00	41.83	74.83	27.83	63.50	42.00	75.33	34.97	67.05
+ BAE	46.00	80.00	47.33	79.50	38.17	71.00	43.67	75.83	35.00	67.00	63.00	88.17	43.17	72.83	53.50	83.67	50.83	83.33	55.67	86.17	47.63	78.75
+ ATR + BAE	76.83	95.67	71.50	92.67	70.50	93.50	59.17	90.00	59.50	88.67	75.83	95.33	73.33	92.67	76.50	95.67	66.33	90.17	78.17	97.33	70.77	93.17
TSConv	29.33	62.67	28.17	63.33	34.00	68.33	38.00	71.50	25.83	54.17	43.33	77.17	35.83	65.50	44.00	76.17	32.50	62.50	39.83	73.00	35.08	67.43
+ ATR	45.33	75.17	39.33	75.83	43.67	75.50	42.00	77.17	33.67	62.83	49.33	79.17	42.83	71.33	50.33	83.33	34.33	69.17	51.17	80.50	43.20	75.00
+ BAE	50.50	79.67	50.67	78.17	49.00	79.00	48.33	75.83	36.50	69.83	62.17	89.83	57.17	84.33	62.83	86.83	43.67	76.17	59.50	85.50	52.03	80.52
+ ATR + BAE	74.50	93.67	65.83	91.67	68.67	93.17	57.67	84.83	64.00	87.33	77.33	95.17	73.00	93.50	77.17	95.50	60.50	87.33	72.33	94.00	69.10	91.62
ShallowNet	20.50	50.67	22.50	53.50	24.50	59.17	32.50	63.17	18.33	47.50	29.00	63.83	26.17	59.67	33.17	68.17	24.83	53.67	32.33	64.50	26.38	58.38
+ ATR	30.17	66.00	32.83	69.17	37.33	67.83	31.17	67.83	25.50	53.67	39.67	70.33	34.33	61.50	44.67	77.50	30.00	60.67	40.00	74.00	34.57	66.85
+ BAE	40.83	74.50	41.33	75.17	45.50	77.00	40.50	69.17	31.17	62.67	53.67	83.83	51.67	82.17	55.50	83.17	38.33	70.83	52.50	85.83	45.10	76.43
+ ATR + BAE	71.50	91.67	60.33	89.17	63.00	91.83	53.83	83.67	56.67	82.83	70.00	91.67	72.33	93.67	73.17	95.50	58.33	83.67	73.83	96.67	65.30	90.03
DeepNet	27.33	54.67	29.00	59.17	33.83	62.00	37.83	70.50	22.83	48.83	38.17	68.17	29.50	54.83	41.67	70.67	28.33	52.50	33.17	67.33	32.17	60.87
+ ATR	35.67	64.50	34.83	66.83	41.00	70.83	41.67	70.67	24.50	51.67	42.50	71.67	34.33	62.83	47.67	79.67	25.83	54.00	41.00	72.83	36.90	66.55
+ BAE	57.67	87.17	57.83	85.00	59.17	86.33	53.33	80.83	39.33	72.17	72.50	94.17	49.33	78.00	65.67	90.17	62.50	89.33	57.67	87.83	57.50	85.10
+ ATR + BAE	73.83	95.33	63.83	91.67	68.50	94.50	57.50	86.67	55.67	87.00	63.00	90.17	72.17	92.00	69.83	92.50	52.33	79.67	71.17	93.83	64.78	90.33

D.2 Hyperparameter Sensitivity

We examine the sensitivity of BEAT to the temporal patch configuration in ATR, the relative weights of the two BAE objectives, and the shared embedding dimension of the EEG encoder and visual student. All experiments use EEGProjector under the intra-subject THINGS-EEG protocol, with all non-evaluated settings fixed to the reference configuration.

Temporal patch length and stride. ATR divides each 250-sample EEG sequence into overlapping temporal patches. We evaluate three patch-length and stride pairs, $(P, S) \in \{(10, 5), (25, 10), (50, 25)\}$. With end padding, these configurations produce 50, 24, and 10 temporal patches, respectively. As shown in Table S5, retrieval accuracy decreases as the temporal segmentation becomes coarser. Compared with the reference setting $(10, 5)$, $(25, 10)$ reduces the subject-averaged Top-1 and Top-5 accuracies by 2.90 and 1.1 percentage points, while $(50, 25)$ reduces them by 10.25 and 4.54 points.

Table S5: Sensitivity to the ATR patch length P and stride S. Patch parameters are measured in EEG samples. Results are obtained with three seeds and averaged across ten subjects.

P	S	No. Patches	Top-1 (%)	Top-5 (%)
10	5	50	81.08	96.32
25	10	24	78.18	95.20
50	25	10	70.83	91.78

Balance between neural alignment and semantic preservation. The BAE objective is $\mathcal{L} = \lambda_{\text{Con}}\mathcal{L}_{\text{Con}} + \lambda_{\text{VD}}\mathcal{L}_{\text{VD}}$. We fix $\lambda_{\text{Con}} = 1$ and vary $\lambda_{\text{VD}} \in \{0.5, 1, 2\}$. Table S6 shows that Top-1 accuracy remains within a 0.52-point range across the evaluated settings.

Table S6: Sensitivity to the relative weights of semantic distillation and neural contrastive alignment. λ_{Con} is fixed to 1. Results are averaged across ten subjects.

$\lambda_{\text{VD}}/\lambda_{\text{Con}}$	Top-1 (%)	Top-5 (%)
0.5	81.03	96.02
1.0	81.08	96.32
2.0	81.55	96.38

Shared EEG–student embedding dimension. We vary the shared embedding dimension $d \in \{128, 256, 512, 1024\}$ of the EEG encoder and the projection heads of all visual student branches.

Each setting maps the EEG and student-image representations into the same d-dimensional alignment space; the frozen teachers retain their native feature dimensions. Table S7 reports the subject-wise Top-1 and Top-5 accuracies and their averages across ten subjects.

The highest subject-averaged performance is obtained at $d = 256$, with Top-1 and Top-5 accuracies of 82.08% and 96.68%, respectively. Compared with the default $d = 128$ setting, these correspond to gains of 1.00 and 0.37 percentage points, calculated before rounding. Increasing the dimension further to 512 or 1024 does not improve either average. Across the evaluated dimensions, the subject-averaged Top-1 and Top-5 accuracies span 1.22 and 0.62 percentage points, respectively, indicating limited sensitivity over this range. The best dimension varies across subjects, so the aggregate advantage of 256 dimensions is not uniform at the individual level. These results show that a larger alignment space does not necessarily yield better retrieval performance. The default 128-dimensional representation remains a compact choice with performance close to the best evaluated setting.

Table S7: Sensitivity to the shared EEG–student embedding dimension d in 200-way zero-shot retrieval on THINGS-EEG. Accuracies are reported in percent. Average denotes the arithmetic mean across ten subjects; boldface marks the highest subject-averaged accuracy for each metric.

Subject	$d = 128$		$d = 256$		$d = 512$		$d = 1024$	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
S1	85.67	95.67	83.67	96.50	82.33	96.83	82.83	95.83
S2	81.00	95.67	82.50	96.17	79.83	96.67	82.17	96.83
S3	76.83	96.50	79.83	97.00	78.83	95.67	76.83	96.50
S4	73.50	94.00	78.17	95.00	73.33	93.17	72.33	92.50
S5	75.83	95.00	75.00	95.50	76.67	95.00	76.50	95.17
S6	88.33	98.17	89.50	98.00	89.67	98.83	86.83	97.33
S7	80.00	95.33	80.00	97.00	80.17	95.00	81.33	97.17
S8	89.00	98.50	88.50	98.50	88.00	98.33	88.33	97.33
S9	75.50	96.17	79.33	95.67	76.33	95.67	77.33	95.00
S10	85.17	98.17	84.33	97.50	84.17	97.67	84.17	97.00
Average	81.08	96.32	**82.08**	**96.68**	80.93	96.28	80.87	96.07

E Temporal Decoding Analysis

Experimental setup. We examine how EEG-to-image retrieval depends on the post-stimulus interval available to the decoder. The fixed-window experiment corresponds to Fig. 1D of the main paper, and a complementary cumulative-window experiment evaluates progressively longer observation periods. Both experiments use a baseline EEGProjector aligned directly with frozen OpenCLIP ViT-B/16 image features. The proposed BAE, dual-teacher distillation, and ATR modules are all excluded. Training uses only the bidirectional EEG–image contrastive objective.

Temporal-window construction. We use all ten THINGS-EEG participants and the preprocessing described in Section A. After averaging repeated trials, each participant has 16,540 training examples and 200 evaluation examples. All 63 EEG channels are retained. At 250 Hz, the 0–1,000 ms post-stimulus epoch contains 250 samples, with one sample every 4 ms. We construct two sets of windows:

$$\mathcal{W}_{\text{fixed}} = \big\{[200k,\ 200(k+1)) : k = 0, \ldots, 4\big\}, \tag{5}$$

$$\mathcal{W}_{\text{cumulative}} = \big\{[0,\ 200k) : k = 1, \ldots, 5\big\}, \tag{6}$$

where all boundaries are in milliseconds relative to stimulus onset. Thus, the fixed windows are 0–200, 200–400, 400–600, 600–800, and 800–1,000 ms, each containing 50 samples. The cumulative windows are 0–200, 0–400, 0–600, 0–800, and 0–1,000 ms, containing 50, 100, 150, 200, and 250 samples, respectively. For a window $[a, b)$, we select zero-based sample indices $a/4, \ldots, b/4 - 1$ from the preprocessed EEG. The same cropping is applied during training and evaluation, without temporal resampling or zero-padding.

Results. Fig. 1D of the main paper and Fig. S1 show that visual decodability varies across post-stimulus intervals. The 200–400 ms interval yields the strongest fixed-window performance. Combining early responses improves retrieval, with cumulative mean accuracy peaking at 0–600 ms and decreasing for longer windows. These observations motivate adaptive temporal recalibration to emphasize informative temporal evidence.

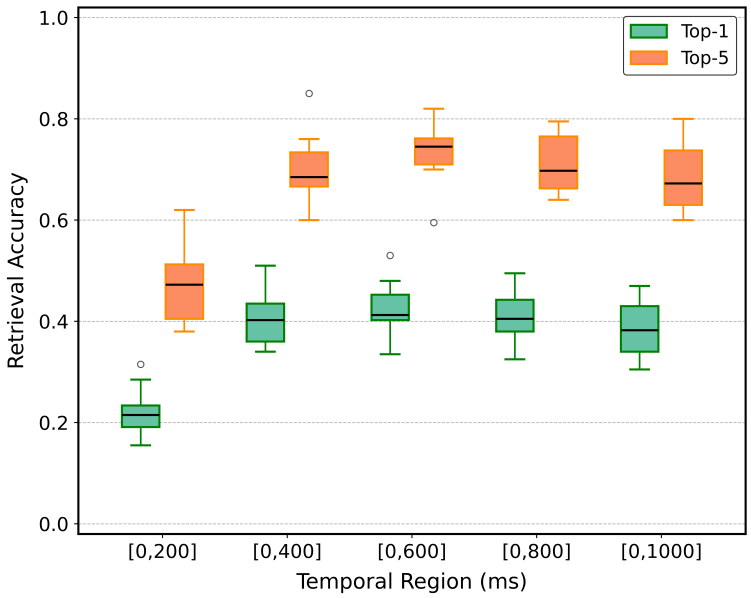


Figure S1: Retrieval performance with cumulative EEG windows. Boxplots show Top-1 and Top-5 accuracies across ten participants.

F Analysis of the Brain-Aligned Visual Encoder

F.1 Visual Backbone and Initialization

The default visual encoder is a ResNet-50 initialized with ImageNet-pretrained weights, and all backbone parameters are optimized during training. To examine the effects of model capacity and initialization, we compare ResNet-18(**resnet18-5c106cde.pth**), ResNet-34(**resnet34-333f7ec4.pth**), and ResNet-50(**resnet50-19c8e357.pth**) under random and ImageNet-pretrained initialization while keeping the remaining training configuration fixed.

Table S8: Ablation Study of different visual encoder backbones. ✓/✗ means with/without model pretrained weight.

Visual Backbone	Weight	Top-1 (%)	Top-5 (%)
ResNet-18	✗	77.60 ± 0.49	95.23 ± 0.53
ResNet-18	✓	80.05 ± 0.52	95.85 ± 0.18
ResNet-34	✗	76.87 ± 2.99	94.67 ± 0.84
ResNet-34	✓	81.83 ± 1.59	96.48 ± 0.38
ResNet-50	✗	73.55 ± 0.66	94.01 ± 0.76
ResNet-50	✓	81.08 ± 0.90	96.32 ± 0.30

F.2 Embedding Dimension and Intrinsic Dimensionality

The default model uses a 128-dimensional common EEG–image space. We additionally train variants with output dimensions of 256, 512, and 1,024 to distinguish the intrinsic geometry of the learned representation from an architectural bottleneck imposed by its nominal output dimension.

As summarized in Table S9, increasing the output dimension produces only a moderate increase in the number of active directions, while zero-shot retrieval accuracy remains consistently high.

Table S9: Retrieval performance and PCA statistics of the evaluated visual representations. Top-1 values provide performance context for the BAE variants. The PCA statistics are computed from the 200 test images using the features described in Section F.3. PR denotes the participation ratio.

Representation	**Nominal Dim.**	**Top-1 (%)**	**PCs at 90%**	**PCs at 95%**	**PR**
ViTB/16	512	–	103	130	70.99
ViTB/32	512	–	104	130	68.49
ResNet-50 backbone	2,048	–	101	131	66.30
BAE-128	128	81.08	22	36	14.39
BAE-256	256	82.08	25	45	15.24
BAE-512	512	80.93	29	55	16.45
BAE-1024	1,024	80.87	30	58	19.60

F.3 PCA Analysis

We perform PCA on the 200 held-out test images. The learned BAE features are extracted from all subjects trained with random seed 0,1,2, using the same ordered set of images for every representation. We compare the 512-dimensional frozen ViTB/16 and ViTB/32 teacher features, the 2,048-dimensional features obtained by forwarding the test images through the ResNet-50 backbone before BEAT-specific optimization, and the outputs of BAE models with nominal dimensions of 128, 256, 512, and 1,024.

PCA is fitted independently to each representation. Before fitting, every feature coordinate is standardized across the 200 images to have zero mean and unit variance. We use full singular value decomposition and retain all available principal components. Given the ordered PCA eigenvalues $\{\lambda_i\}_{i=1}^{r}$, cumulative explained variance is computed as

$$C(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{r} \lambda_i}. \tag{7}$$

We report the smallest k for which $C(k)$ reaches 90% and 95%. We additionally measure spectral concentration using the participation ratio

$$\mathrm{PR} = \frac{\left(\sum_{i=1}^{r} \lambda_i\right)^2}{\sum_{i=1}^{r} \lambda_i^2}, \tag{8}$$

where a lower value indicates that the variance is concentrated in fewer principal directions.

As shown in Table S9, the ViTB/16,ViTB/32 and ResNet-50 representations require 103,104 and 101 principal components, respectively, to explain 90% of their variance. In contrast, BAE requires only 22, 25, 29, and 30 components at nominal dimensions of 128, 256, 512, and 1,024. Their participation ratios likewise remain substantially lower than those of CLIP and ResNet-50. Together with the Top-1 retrieval accuracies of 80.87–82.08%, these results indicate that BAE learns a compact yet discriminative visual target space, rather than obtaining an apparently low-dimensional structure solely from the 128-dimensional output bottleneck.

Figure S2 provides a component-wise view of the same PCA spectra. The CLIP and ResNet-50 spectra exhibit comparatively long tails, indicating that their variance remains distributed over many directions. By contrast, the BAE spectra decay rapidly, particularly at lower nominal dimensions. As the nominal BAE dimension increases, the tail becomes gradually longer, consistent with the increase from 22 to 30 components at the 90% explained-variance threshold, while remaining substantially more concentrated than the two reference representations.

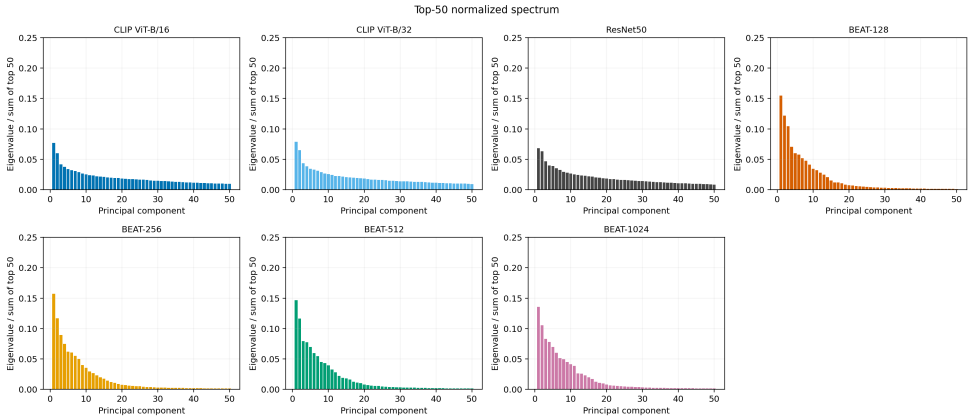


Figure S2: Normalized eigenvalue spectra of the top 50 principal components for the evaluated visual representations. Within each panel, the displayed eigenvalues are normalized by the sum of the top-50 eigenvalues.

G Analysis of Adaptive Temporal Recalibration

G.1 Attention-Guided Temporal Perturbation

We perform a test-time perturbation experiment to examine whether the attention learned by ATR identifies temporal patches that are functionally important for EEG–image retrieval. To isolate the contribution of ATR, this experiment uses EEGProjector with ATR and the fixed ViT-B/16 image features, without BAE. The model is trained and validated without masking. After training, the same best checkpoint is evaluated under four conditions: unmasked input, masking the highest-attention patches, masking the lowest-attention patches, and masking a random subset of patches. Thus, differences among the four conditions arise only from the test-time intervention rather than retraining or checkpoint selection.

For each test sample, we first perform an unmasked forward pass and extract the post-softmax self-attention matrix from the final ATR encoder layer. Let $A_{c,h,q,j}$ denote the attention assigned by query patch q to key patch j for EEG channel c and attention head h. We define the importance of patch j as its mean incoming attention,

$$s_j = \frac{1}{CHN} \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{q=1}^{N} A_{c,h,q,j}, \qquad j = 1, \ldots, N, \tag{9}$$

where $C = 63$ is the number of EEG channels, $H = 4$ is the number of attention heads, and $N = 50$ is the number of temporal patches. For every sample, we mask 20% of the temporal patches, corresponding to 10 of the 50 patch tokens.

The high- and low-attention conditions select the patches with the largest and smallest s_j, respectively. The random condition uses a deterministic sample-wise random subset with the same mask budget. Selected tokens are set to zero for all EEG channels after patch projection and positional encoding and before the Transformer encoder. A second forward pass through the same checkpoint produces the perturbed EEG representation.

H Additional Retrieval and Reconstruction Results

H.1 More Retrieval Examples

Fig. S3 presents additional qualitative retrieval examples for visual comparison. We randomly select six representative queries and show their Top-5 retrieved images, forming a set of "good cases". These examples demonstrate that the model can closely align with the target stimuli and retrieve highly relevant candidates. In all six cases, the ground-truth image appears at the Top-1 position, indicating reliable and precise retrieval performance.

The retrieved results exhibit consistent similarity to the queries at both low-level and high-level aspects. For instance, in animal-related queries, all Top-5 retrieved images depict animals, suggesting that animal-related semantics are captured in a compact and consistent manner within the learned neural embedding space. In the "bullet" example, the Top-5 retrieved images not only match the object category, but also share similar orientations and shapes, indicating that the model is sensitive to fine-grained visual attributes beyond coarse semantic categories.

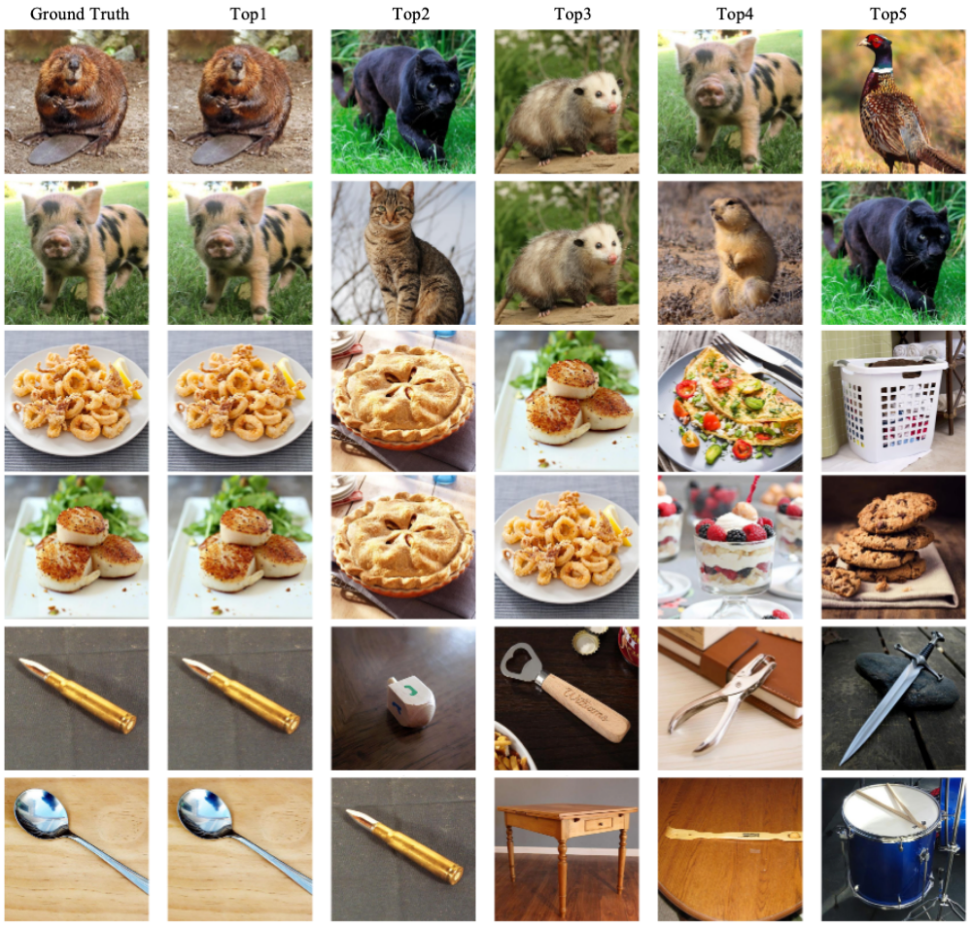


Figure S3: Good cases: Top-5 retrieval results for various stimuli.

In contrast, Fig. S4 shows six representative Top-1 failure but Top-5 success cases. In these examples, the ground-truth image can still be found within the Top-5 retrieved results, but the retrieved images exhibit only weak similarity to the query in both low-level properties (e.g., color, texture, and local structure) and high-level semantics (e.g., object category and scene type). This suggests that, under certain challenging conditions, the learned embedding space may fail to fully capture the fine-grained visual and semantic characteristics required for precise retrieval, leading to semantically plausible yet suboptimal matches.

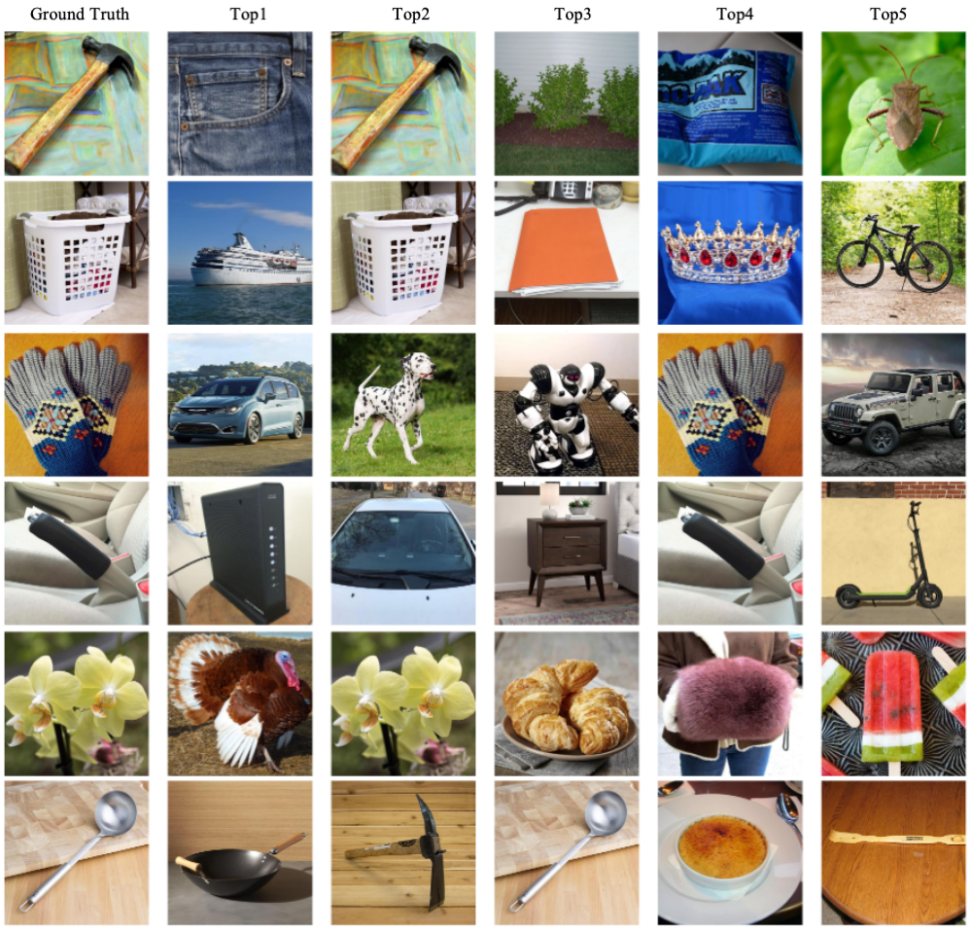


Figure S4: Top-1 failure but Top-5 success cases for various stimuli.

H.2 More Reconstruction Examples

Table S10, S11 and S12 reports reconstruction performance for all 10 subjects based on ViT-B/16, ViT-B/32 and both. We can find the multi-teacher distillation get the best score at most metrics. Performance varies across individuals and Fig. S5 presents randomly selected reconstruction examples from these subjects.

The results reveal substantial variability in reconstruction quality across individuals. For most subjects, the reconstructed images are broadly consistent with the semantics of the original stimuli, indicating that the learned mapping captures subject-specific neural representations reasonably well.

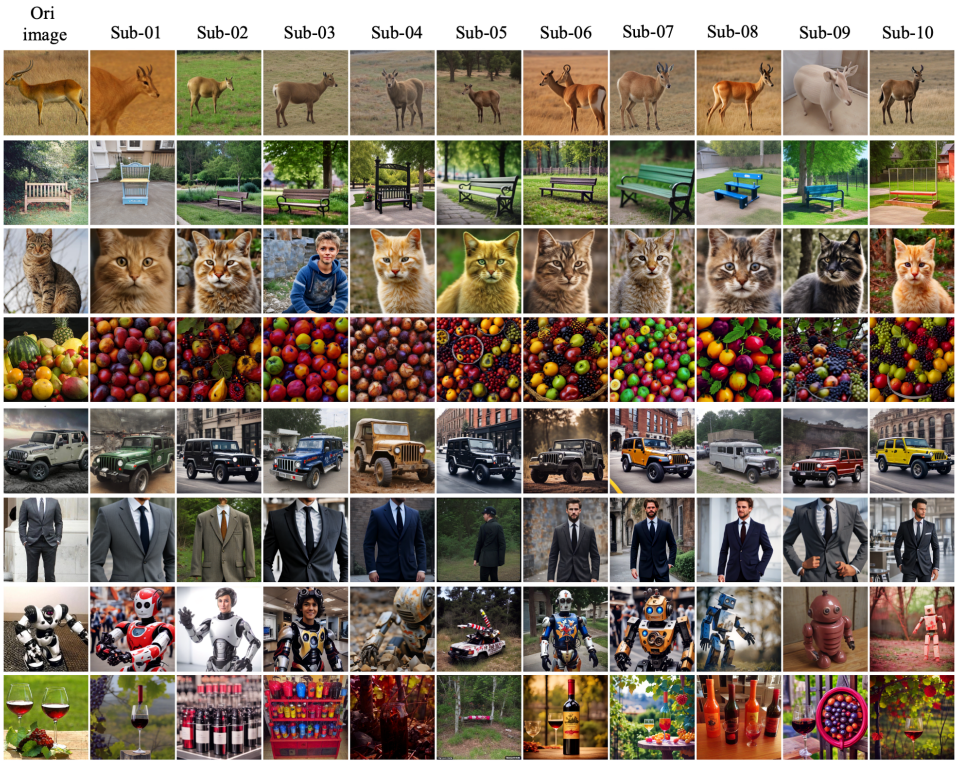


Figure S5: Details of eight examples in Image Reconstruction task on THINGS-EEG dataset across 10 subjects

Table S10: Reconstruction performance of the single-teacher, single-branch model with a ViT-B/16 teacher across all subjects on THINGS-EEG. Arrows indicate whether higher or lower values are better.

Subject	PixCorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	InceptionV3↑	CLIP↑	EffNet-B↓	SwAV↓
S1	0.1766	0.3528	0.8526	0.9341	0.9253	0.9148	0.6935	0.4837
S2	0.1739	0.3664	0.8063	0.8565	0.8644	0.8734	0.7202	0.5094
S3	0.1585	0.3442	0.8132	0.8912	0.9096	0.9113	0.7029	0.4887
S4	0.1162	0.3574	0.8053	0.8876	0.9210	0.9148	0.7173	0.5001
S5	0.1505	0.3339	0.8137	0.9134	0.9316	0.9355	0.6728	0.4693
S6	0.1463	0.3263	0.8113	0.9016	0.9003	0.9088	0.7173	0.4874
S7	0.1433	0.3354	0.7903	0.8968	0.9141	0.9141	0.6921	0.4852
S8	0.1901	0.3696	0.8542	0.9311	0.9642	0.9335	0.6595	0.4682
S9	0.1444	0.3657	0.8336	0.8987	0.9352	0.9201	0.6828	0.4723
S10	0.1710	0.3537	0.8300	0.8995	0.9164	0.9006	0.6967	0.4975
Average	0.1571	0.3505	0.8210	0.9010	0.9182	0.9127	0.6955	0.4862

Table S11: Reconstruction performance of the single-teacher, single-branch model with a ViT-B/32 teacher across all subjects on THINGS-EEG. Arrows indicate whether higher or lower values are better.

Subject	PixCorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	InceptionV3↑	CLIP↑	EffNet-B↓	SwAV↓
S1	0.1587	0.3475	0.8192	0.9203	0.9058	0.9096	0.7169	0.5132
S2	0.1370	0.3234	0.7822	0.8658	0.9005	0.8986	0.7322	0.5173
S3	0.1398	0.3416	0.8304	0.8939	0.9424	0.9112	0.6861	0.4861
S4	0.1405	0.3524	0.8215	0.9342	0.9491	0.9436	0.6811	0.4636
S5	0.1398	0.3494	0.8215	0.8742	0.9014	0.8891	0.7249	0.5139
S6	0.1552	0.3518	0.8189	0.9137	0.9502	0.9320	0.6631	0.4690
S7	0.1460	0.3571	0.7918	0.8900	0.9132	0.9007	0.7110	0.4952
S8	0.1710	0.3767	0.8590	0.9128	0.9382	0.9196	0.6750	0.4803
S9	0.1463	0.3667	0.8313	0.9094	0.9099	0.9056	0.6977	0.4856
S10	0.1489	0.3698	0.8309	0.8957	0.9015	0.9068	0.7109	0.4977
Average	0.1483	0.3536	0.8207	0.9010	0.9212	0.9117	0.6999	0.4922

Table S12: Reconstruction performance across all subjects on THINGS-EEG. Arrows indicate whether higher or lower values are better.

Subject	**PixCorr↑**	**SSIM↑**	**AlexNet(2)↑**	**AlexNet(5)↑**	**InceptionV3↑**	**CLIP↑**	**EffNet-B↓**	**SwAV↓**
S1	0.1622	0.3666	0.8362	0.9030	0.9154	0.8971	0.7102	0.4881
S2	0.1463	0.3302	0.7999	0.9173	0.9292	0.9222	0.6943	0.4871
S3	0.1535	0.3403	0.8017	0.8905	0.9272	0.9232	0.6977	0.4977
S4	0.1560	0.3323	0.8105	0.9112	0.9107	0.9231	0.7008	0.4862
S5	0.1551	0.3082	0.8246	0.9192	0.9255	0.9270	0.6951	0.4931
S6	0.1495	0.3286	0.8617	0.9308	0.9390	0.9192	0.7065	0.4903
S7	0.1507	0.3450	0.8537	0.9312	0.9592	0.9469	0.6596	0.4574
S8	0.1624	0.3720	0.8159	0.9102	0.9346	0.9301	0.6651	0.4690
S9	0.1332	0.3626	0.8152	0.8980	0.9105	0.9276	0.7125	0.5030
S10	0.1754	0.3587	0.8216	0.9291	0.9216	0.9258	0.6904	0.4757
Average	0.1544	0.3444	0.8241	0.9141	0.9273	0.9242	0.6932	0.4848

REFERENCES

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.

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.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 102822–102864, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding Natural Images from EEG for Object Recognition. In *International Conference on Learning Representations*, 2024.

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.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

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.