Bridging the Vision-Brain Gap with an Uncertainty-Aware Blur Prior

Haitao Wu[1] Qing Li[1] Changqing Zhang[1*] Zhen He[2*] Xiaomin Ying[2*]
[1]College of Intelligence and Computing, Tianjin University
[2]Beijing Institute of Basic Medical Sciences
{wuhaitao, liqing0315, zhangchangqing}@tju.edu.cn
hezhen.bio@gmail.com, yingxmbio@foxmail.com

Abstract

*Can our brain signals faithfully reflect the original visual stimuli, even including high-frequency details? Although human perceptual and cognitive capacities enable us to process and remember visual information, these abilities are constrained by several factors, such as limited attentional resources and the finite capacity of visual memory. When visual stimuli are processed by human visual system into brain signals, some information is inevitably lost, leading to a discrepancy known as the **System GAP**. Additionally, perceptual and cognitive dynamics, along with technical noise in signal acquisition, degrade the fidelity of brain signals relative to the visual stimuli, known as the **Random GAP**. When encoded brain representations are directly aligned with the corresponding pretrained image features, the System GAP and Random GAP between paired data challenge the model, requiring it to bridge these gaps. However, in the context of limited paired data, these gaps are difficult for the model to learn, leading to overfitting and poor generalization to new data. To address these GAPs, we propose a simple yet effective approach called the **Uncertainty-aware Blur Prior (UBP)**. It estimates the uncertainty within the paired data, reflecting the mismatch between brain signals and visual stimuli. Based on this uncertainty, UBP dynamically blurs the high-frequency details of the original images, reducing the impact of the mismatch and improving alignment. Our method achieves a top-1 accuracy of **50.9%** and a top-5 accuracy of **79.7%** on the zero-shot brain-to-image retrieval task, surpassing previous state-of-the-art methods by margins of **13.7%** and **9.8%**, respectively. Code is available at GitHub.*

1. Introduction

The human brain is one of the most complex things known in the universe, and extensive studies have been devoted to unraveling its structure and function over the past several decades [26, 27, 35, 40, 47, 49, 53]. Vision, as the primary sense for humans to perceive the world, involves approximately one-third of the cortical surface. Consequently, the brain plays a crucial role in visual perception and cognition [38, 62, 63, 65]. To understand the mechanisms between human vision and brain activity, various brain imaging techniques such as Electroencephalogram (EEG), Magnetoencephalography (MEG) and Functional magnetic resonance imaging (fMRI), are utilized to measure brain responses to visual stimuli. EEG is a low-cost, portable method for measuring brain activity by detecting voltage changes caused by neuronal signals, offering high temporal resolution. However, it suffers from a low signal-to-noise ratio due to weak signals being influenced by the skull, external interference, and biological noise. MEG offers high temporal resolution, but is limited by its high cost. In contrast, fMRI provides high spatial resolution by detecting changes in blood oxygen levels, but its temporal resolution is limited due to the slower hemodynamic response.

Recently, various methods for decoding brain signals have been proposed [5, 10, 12, 15, 36, 56, 59, 61]. These methods aim to retrieve and reconstruct the original visual stimuli by aligning the representations of brain signals with the visual stimuli. However, they fail to account for the GAPs between brain signals and visual stimuli. Previous studies [6, 7, 13, 50, 71] on human perceptual and cognitive capacities have shown that the amount of visual information human can process and remember at any given moment is limited and varies across individuals due to constrained attentional resources [8, 17, 58], limitations in eye movements and scanning [32, 72], and the limited capacity of visual working memory [44]. When the digital image modality is transformed into the brain signal modality through human visual perception and cognitive processes, some information is unavoidably lost, which is called the **System GAP** between human and machine. A key factor contributing to this information loss is the structure of the human eye. As shown in Fig. 1, when an individual observes an object, the

*Corresponding authors.

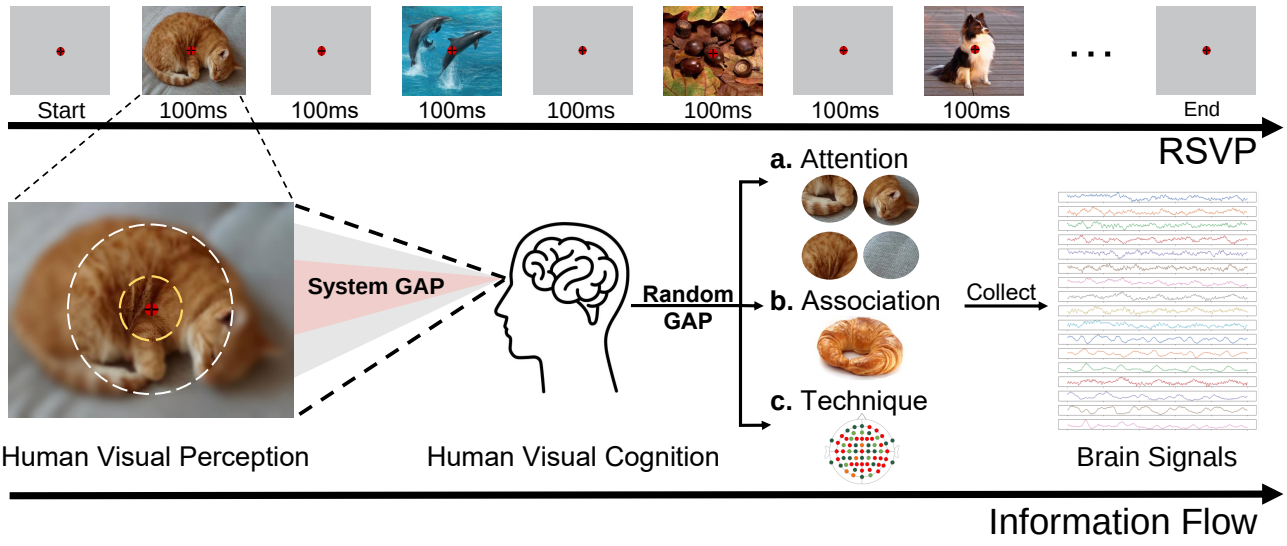


Figure 1. Overview of the information flow during Rapid Serial Visual Presentation (RSVP) and the GAPs in human visual perception and cognition. The top panel illustrates the RSVP paradigm, where a sequence of images is rapidly presented for 100ms each, with a fixation point in the center. The bottom panel highlights the GAPs in the visual processing pipeline: **System Gap**, which represents the loss of high-frequency details during the transition from raw visual stimuli to visual perception, and **Random Gap**, which arises due to (a) dynamic perceptual processes (e.g., shifts in visual attention), (b) dynamic cognitive processes (e.g., associating with similar objects or concepts), and (c) low-level technical noise in signal collection.

resolution of the visual field is not uniform and gradually decreases from the fovea toward the periphery [14].

Human perception and cognition are inherently dynamic, even when viewing or considering the same image or problem, leading to variability in brain signals responding to same stimuli. As illustrated in Fig. 1, perception can shift as attention is directed toward different parts of an image, while cognition may dynamically associate with related objects or concepts. Additionally, signal acquisition is impacted by technical noise, such as poor electrode-skin contact or instability in signal channels. These factors contribute to variability in brain signals, as shown in Fig. 2(a), and weaken the information relative to the original visual stimuli, reducing the signal-to-noise ratio. Consequently, even for two completely different stimuli, the corresponding brain signals are difficult to differentiate due to limited information and excessive noise, as shown in Fig. 2(b). We refer to this variability-induced information mismatch as **Random GAP**, attributed to its stochastic nature, which makes it exceptionally challenging to quantify. As illustrated in Fig. 2(c)(d), it demonstrates variability both across trials and among different subjects.

The most advanced visual neural decoding methods [10, 15, 36, 59] align encoded representations of brain signals with the pretrained embedding of corresponding visual stimuli by contrastive learning [11]. However, when we directly align them, the System GAP and Random GAP may prompt the model to bridge the GAPs. Limited by the scarcity of paired data, the gaps become difficult for the model to learn, leading to overfitting on the training set and poor generalization to new data. To address this issue, we aim to mitigate the impact of the GAPs and improve the alignment by introducing priors, thereby preventing the model from overfitting to these gaps. Our main contributions are summarized as follows:

1. We propose the existence of **System GAP** and **Random GAP** between visual stimuli and brain signals as shown in Fig. 1. The System GAP arises from the inability of brain signals to faithfully reflect visual stimuli, while the Random GAP arises from three factors: the dynamics of perception, the dynamics of cognition, and low-level technical noise in signal collection. These GAPs contribute to the reduction of the fidelity of brain signals in relation to the original visual stimuli.
2. We experimentally analyzed the impact of the two types of gaps. Based on observations and experimental analysis, we propose a simple and effective method called **U**ncertainty-aware **B**lur **P**rior (**UBP**). Our method achieves a top-1 accuracy of **50.9%** and a top-5 accuracy of **79.7%** on the zero-shot brain-to-image retrieval task, surpassing previous state-of-the-art methods by margins of **13.7%** and **9.8%**, respectively.

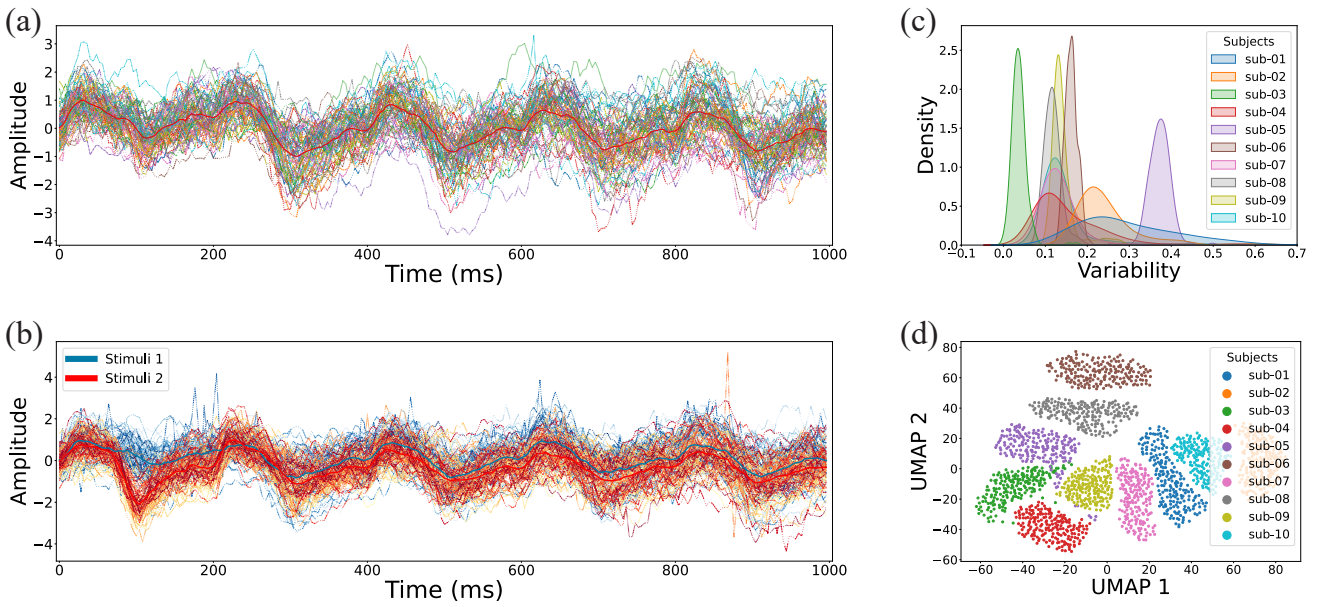


Figure 2. Illustration of brain signals. (a) EEG signals recorded over 80 trials of the same stimulus for Subject 1. The red line indicates the mean across all trials. (b) EEG signals from 80 trials of two stimuli for Subject 1. Cool colors represent Stimulus 1, warm colors represent Stimulus 2. The blue and red lines show the means for Stimulus 1 and Stimulus 2, respectively. (c) Density distribution of EEG signal variability across 10 subjects. Variability is negatively correlated with task performance and see Tab. 4 for further details. (d) UMAP projection of EEG signals from 10 subjects, showing distinct clustering patterns.

2. Related Works

2.1. Neural Decoding

Neural decoding refers to the process of interpreting neural signals (e.g., EEG, MEG, fMRI) to infer state of human perception and cognition. In recent years, significant progress has been made in this field, particularly in applications such as motor imagery decoding [2], visual decoding [3, 4, 18, 20, 43, 51, 56, 57, 61, 73, 74], text decoding [16], emotion decoding [37], inter-subject decoding[66, 80], and diagnosis of neurological disorders [64]. Visual decoding includes two primary tasks: brain-to-image retrieval and reconstruction. Several methods [10, 15, 36, 59] have been proposed for visual decoding, notably aligning the encoded representations of EEG/MEG signals with the Contrastive Vision-Language Pre-training (CLIP) [52] embedding space. However, they did not account for the GAPs between brain signals and visual stimuli, leading to overfitting on the training set and poor generalization to new data.

2.2. Multi-modal Contrastive Learning

Contrastive representation learning has attained remarkable achievements in multiple domains, including vision [11], language [19], and graph [78]. Building on the success of these works, multi-modal contrastive representation learning (MMCL) has emerged, focusing on aligning inputs from multiple modalities within a shared representation space. These models are typically pretrained on large-scale paired datasets using a contrastive loss function. Recent vision-language contrastive pre-training models, such as CLIP [52] and ALIGN [30] have demonstrated remarkable zero-shot retrieval and classification performance, along with robust generalization to a wide range of downstream tasks [60, 69]. Inspired by the success of these vision-language models, contrastive representation learning across diverse modalities has garnered increasing attention [22, 34, 46]. However, in real-world settings, for certain modality pairs like audio-visual [24] and 3D-language [75], it is challenging to obtain paired data that match precisely. This constraint restricts the generalization capabilities of the pretraining models. Fortunately, several methods have been proposed to address this issue and provide theoretical analyses [39, 70, 75]. The visual-neural data for neural decoding also suffers from poor matching. Consequently, rough alignment will inevitably lead to a reduction in generalization performance.

2.3. Uncertainty Quantification

Uncertainty quantification is essential for informing critical decisions, and several methods have emerged [1]. One notable example is out-of-distribution (OOD) or anomaly detection [9, 42, 76]. Uncertainty modeling improves the accuracy of distinguishing between normal and anomalous instances. Anomalies in the training data can lead to overfitting, causing the model to fail in generalizing to unseen

data. Consequently, it is crucial to detect abnormal data during training and evaluation. One prominent application is semi-supervised learning [77]. Several methods [54, 68, 79] leverage uncertainty to identify the incorrect pseudo-labels in unlabeled data, preventing error accumulation during model training. [45] enhances response reliability by estimating the uncertainty of LLMs. In our task, the random gap leads to a low SNR in brain signals. Consequently, it is necessary to quantify the uncertainty and dynamically mitigate the gap.

3. Visual Neural Decoding

3.1. Notation

In this paper, we begin by introducing the basic notation for visual neural decoding. We use paired data (x_v, x_b), where $x_v \in \mathbb{R}^{d_V}$ represents an image from the visual domain, and $x_b \in \mathbb{R}^{d_B}$ represents the corresponding brain signal. $\mathcal{X}_V$ is used to denote the set of all visual data from distribution $\mathcal{P}_V$, and $\mathcal{X}_B$ is employed to denote the set of all brain data from distribution $\mathcal{P}_B$. Their joint multi-modal distribution is $\mathcal{P}_M$.

3.2. Vision-Brain Contrastive Learning

The goal of vision-brain contrastive learning is to map brain data $\mathcal{X}_B$ to a k-dimensional latent space $\mathcal{H} \in \mathbb{R}^k$ that aligns with the representation of visual data $\mathcal{X}_V$. This is achieved by using a frozen visual encoder $f_V : \mathcal{X}_V \to \mathcal{H}$ to obtain visual embeddings and training a brain encoder $f_B : \mathcal{X}_B \to \mathcal{H}$ with parameters θ to map brain data into the shared latent space $\mathcal{H}$. Given the effectiveness of pre-trained vision-language models (VLMs) in providing rich visual features, f_V is taken from the vision branch of a pre-trained VLM, such as CLIP [52].

For multi-modal positive and negative pairs, we define an image-brain pair drawn from the paired vision-brain data, i.e., $(x_v, x_b) \sim \mathcal{P}_M$, as positive pairs, and draw independent samples from each domain, $x_v^- \sim \mathcal{P}_V$, $x_b^- \sim \mathcal{P}_B$, and treat (x_v, x_b^-), (x_v^-, x_b), and (x_v^-, x_b^-) as negative pairs, assuming that the samples in these pairs are independent of each other. Given positive and negative pairs (x_v, x_b, x_v^-, x_b^-), the corresponding encoders map them to (h_v, h_b, h_v^-, h_b^-). The learning objective is the symmetric cross-entropy (SCE) loss [67], computed as follows:

$$\mathcal{L}_{\text{SCE}}(f_B) = -\mathbb{E}_{x_v,x_b} \log \frac{\exp\left(f_V(x_v)^\top f_B(x_b)/\tau\right)}{\mathbb{E}_{x_b^-} \exp\left(f_V(x_v)^\top f_B(x_b^-)/\tau\right)} - \mathbb{E}_{x_v,x_b} \log \frac{\exp\left(f_V(x_v)^\top f_B(x_b)/\tau\right)}{\mathbb{E}_{x_v^-} \exp\left(f_V(x_v^-)^\top f_B(x_b)/\tau\right)}. \quad (1)$$

4. Method

Our method consists of Blur Prior and Uncertainty-aware components, addressing the System GAP and Random GAP, respectively. The algorithmic flow of our framework is illustrated in Algorithm 1 and the details are as follows.

Algorithm 1 Uncertainty-aware Blur Prior Framework

Input: Multimodal training dataset $\mathcal{P}_M$
Model: Brain encoder f_B with random parameters θ, pretrained vision encoder f_V with parameters ϕ, temperature parameter τ, learning rate η
Output: Trained model f_B
for each iteration **do**
  Obtain training sample (x_v, x_b) from dataset $\mathcal{P}_M$
  Obtain $\tilde{x}_v$ by Eq. (4) with blur radius r
  $h_b = f_B(x_b)$; $h_v = f_V(\tilde{x}_v)$
  Compute loss $\mathcal{L}$ by Eq. (1)
  Update r for sample (x_v, x_b) by Eq. (10)
  Update model parameters $\theta \leftarrow \theta - \eta \nabla \mathcal{L}$
end for
return trained model f_B

4.1. Blur Prior

Due to the existence of the System GAP between the human visual system and the original visual stimuli, a discrepancy in information arises, particularly in the loss of high-frequency details. Aligning brain signals with the images may cause the model to overfit to the high-frequency details in the images. To mitigate the System GAP, we propose a simple prior, which applies Gaussian blur to the original images, making the image modality better aligned with the brain signal modality.

Based on the characteristics of the experimental paradigm, where the focal point is concentrated on the red dot in the center of the image, we synthesized images of the macular and peripheral regions of the human eye to simulate the decrease in resolution and reduce high-frequency details. Concretely, a uniformly blurred image is generated first:

$$x_{\text{blur}}(i,j) = \sum_{m=-k}^{k} \sum_{n=-k}^{k} x(i-m, j-n) \cdot G(m,n), \quad (2)$$

where $r = 2k + 1$ denotes the radius of the Gaussian kernel, and $x(i-m, j-n)$ represents the pixel value in the original image x, while $G(m,n)$ denotes the corresponding weights provided by the Gaussian kernel. The Gaussian kernel $G(m,n)$ is defined as:

$$G(m,n) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{m^2+n^2}{2\sigma^2}\right), \quad (3)$$

where σ is the standard deviation, which controls the intensity of the blur. The fovea blur image is blended with the original image and the uniformly blurred image as:

$$\tilde{x}_v = \alpha \cdot x + (1-\alpha) \cdot x_{\text{blur}}, \quad (4)$$

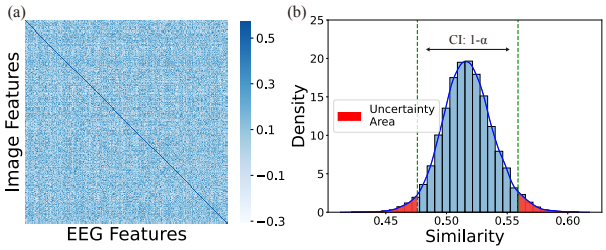


Figure 3. Semantic similarity visualization. (a) Semantic similarity matrix between image features and EEG features. The diagonal represents the similarity between corresponding pairs of features from the two modalities. (b) Density distribution of similarity scores from the diagonal of the matrix. The green dashed lines denote the confidence interval at a significance level of $1-\alpha$, indicating the range of similarity scores that are statistically significant. The red areas represent the Uncertainty Area, indicating scores outside the confidence interval.

where α is the blending factor, represented as a matrix with values between 0 and 1. For a foveated effect, we define a function of distance from the fovea as:

$$\alpha(i,j) = \exp\left(-\frac{\lambda \cdot d(i,j)}{L}\right), \tag{5}$$

where $d(i,j)$ denotes the Euclidean (2-norm) distance between pixel (i,j) and the fovea, and L denotes the maximum possible distance within the image. The parameter λ controls the rate of decay, moderating how quickly the weight $\alpha(i,j)$ decreases as the distance increases. In our setting, the level of blurriness of the image depends on the radius of the Gaussian kernel r, with other factors being fixed.

4.2. Uncertainty Quantification

The mismatch between brain signals and the original image stimuli can also be attributed to Random GAP, including dynamics of perception and cognition, along with technical noise, as shown in Fig. 1. Due to the complexity of perception and cognitive processes, which are difficult to disentangle, it is challenging to quantify the contribution of each factor to the Random GAP. Fortunately, it is found that the distribution of the semantic similarity between brain signals and images is approximately Gaussian, as shown in Fig. 3. Motivated by the above observations, we can identify the extreme outliers pairs based on the confidence interval in statistical inference. For each sample (x_v, x_b), the uncertainty is estimated based on the interval in which it falls. Subsequently, the corresponding r is adjusted to introduce varying levels of blurring, thereby dynamically mitigating the information discrepancy between brain signals and visual stimuli. Overall, a greater discrepancy corresponds to a more significant degree of blur, while a smaller discrepancy results in less blur.

Specifically, the semantic similarity matrix $\mathbf{M}$ is first computed between pair data:

$$\mathbf{M} = \mathbf{h_b} \cdot \mathbf{h_v}^\top \cdot \text{softplus}(\tau), \tag{6}$$

where $\mathbf{h_b}$ and $\mathbf{h_v}$ donate the brain signal and visual stimuli representations, respectively. τ is a learned scalar parameter, and softplus$(\cdot)$ is a smooth, non-linear activation function applied to τ to ensure positivity. The similarities can be calculated as:

$$\mathbf{S} = \text{diag}(\mathbf{M}), \tag{7}$$

where $\mathbf{S}$ denotes the vector of similarity scores, $\text{diag}(\cdot)$ denotes the diagonal of a matrix. It is assumed that the similarity scores approximately follow a normal distribution $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$. The mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ are calculated as follows:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_i, \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{S}_i - \hat{\mu})^2. \tag{8}$$

The confidence interval for the similarity scores with confidence level $1-\alpha$ is given by:

$$\left[\hat{\mu} - z_{\alpha/2}\cdot\hat{\sigma}, \hat{\mu} + z_{\alpha/2}\cdot\hat{\sigma}\right], \tag{9}$$

where $z_{\alpha/2}$ represents the critical value from the standard normal distribution corresponding to a two-sided confidence level of $1-\alpha$. For the similarity s, the corresponding degree of blur is defined as follows:

$$r(s) = \begin{cases} r_0 - c, & \text{if } s < \hat{\mu} - z_{\alpha/2}\cdot\hat{\sigma}, \\ r_0 + c, & \text{if } s > \hat{\mu} + z_{\alpha/2}\cdot\hat{\sigma}, \\ r_0, & \text{if } \hat{\mu} - z_{\alpha/2}\cdot\hat{\sigma} \le s \le \hat{\mu} + z_{\alpha/2}\cdot\hat{\sigma}, \end{cases} \tag{10}$$

where r_0 is the baseline blur radius, and c is a constant that controls the change in blur radius when s is outside this interval.

5. Experiments and Results

5.1. Datasets and Implementation Details

THINGS-EEG [21] is a large scale EEG dataset including 10 subjects with the Rapid Serial Visual Presentation (RSVP) paradigm [23, 29, 31]. The training set includes 1654 concepts with each concept 10 images, and each image repeats 4 times per subject. The test set includes 200 concepts with each concept 1 image, and each image repeats 80 times per subject. For data preprocessing, we follow the method detailed in [59]. Repetitions are averaged for the purpose of high SNR, resulting in a total of 16540 training samples and 200 test samples per subject.

THINGS-MEG [25] involves four participants and consists of 271 channels. It consists of 1854 concepts $\times$ 12 images

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

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																						
BraVL [15]	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 [59]	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 [59]	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 [59]	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 [36]	25.6	60.4	22.0	54.5	25.0	62.4	31.4	60.9	12.9	43.0	21.3	51.1	30.5	61.5	38.8	72.0	34.4	51.5	29.1	63.5	28.5	60.4
VE-SDN [10]	32.6	63.7	34.4	69.9	38.7	73.5	39.8	72.0	29.4	58.6	34.5	68.8	34.5	68.3	49.3	79.8	39.0	69.6	39.8	75.3	37.2	69.9
UBP (Ours)	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**
Inter-subject: leave one subject out for test																						
BraVL	2.3	8.0	1.5	6.3	1.4	5.9	1.7	6.7	1.5	5.6	1.8	7.2	2.1	8.1	2.2	7.6	1.6	6.4	2.3	8.5	1.8	7.0
NICE	7.6	22.8	5.9	20.5	6.0	22.3	6.3	20.7	4.4	18.3	5.6	22.2	5.6	19.7	6.3	22.0	5.7	17.6	8.4	28.3	6.2	21.4
NICE-SA	7.0	22.6	6.6	23.2	7.5	23.7	5.4	21.4	6.4	22.2	7.5	22.5	3.8	19.1	8.5	24.4	7.4	22.3	9.8	29.6	7.0	23.1
NICE-GA	5.9	21.4	6.4	22.7	5.5	20.1	6.1	21.0	4.7	19.5	6.2	22.5	5.9	19.1	7.3	25.3	4.8	18.3	6.2	26.3	5.9	21.6
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
UBP (Ours)	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**

$\times$ 1 repetition in the training set and 200 concepts $\times$ 1 image $\times$ 12 repetitions in the test set. We follow the same setting described in [59]. Repetitions of the same stimulus are averaged to ensure the SNR.

Brain Encoders. We employ a simple yet effective encoder named EEGProject, consisting of two linear layers with residual connection and a normalization layer. The detailed model architecture is provided in the appendix. To further assess the generalizability of our method, we have conducted experiments with additional architectures, including Shallownet [55], Deepnet [55], EEGnet [33], and TSconv [59].

Vision Encoders. Our research employs the visual branches of CLIP models, specifically using pretrained weights from OpenCLIP [28]. These weights are derived from training multiple models across a diverse range of data sources and computational resources. In the experiments, we utilize several weights, including RN50, RN101, ViT-B/16, ViT-B/32, ViT-L/14, ViT-H/14, ViT-g/14, and ViT-bigG/14. Unless otherwise stated, RN50 is employed as the default model.

More details on data preprocessing, hyperparameter settings, and hardware configurations are provided in the appendix.

5.2. Comparison with Baselines

Baselines. We compare our approach with recent neural decoding methods. Du et al. [15] propose BraVL, a model based on Mixture of Experts (MoE) that uses multimodal learning of brain-visual-linguistic features. Song et al. [59] present a self-supervised framework for learning image representations from EEG signals, called NICE, incorporating two plug-and-play spatial modules with self-attention and graph attention. Li et al. [36] propose a EEG encoder called the Adaptive Thinking Mapper (ATM), which incorporates position encoding and temporospatial encoding.

Table 2. Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-MEG

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	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(Ours)	15.0	38.0	46.0	80.5	27.3	59.0	18.5	43.5	26.7	55.2
Inter-subject: leave one subject out for test										
UBP(Ours)	2.0	5.7	1.5	17.2	2.7	10.5	2.5	8.0	2.2	10.4

Chen et al. [10] construct a joint semantic space and propose a Visual-EEG Semantic Decouple Framework, called VE-SDN, which explicitly extracts semantic features from both modalities to enable optimal alignment.

Comparison. Tab. 1 and Tab. 2 show quantitative comparisons between our approach and baselines on EEG and MEG test set. Our approach significantly outperforms previous sate-of-the-art in terms of both intra-subject and inter-subject settings. Notably, UBP achieves a top-1 accuracy of 50.9% and top-5 accuracy of 79.7% for the zero-shot brain-to-image retrieval task on the THINGS-EEG dataset, and a top-1 accuracy of 26.7% and top-5 accuracy of 55.2% on the THINGS-MEG dataset.

5.3. Effectiveness of Blur Prior

To illustrate that the Blur Prior does not serve as an effective data augmentation technique but rather functions as a method to bridge the System GAP, visual stimuli processed using various techniques are presented in Fig. 4, with the corresponding performance shown in Tab. 3. It can be observed from the Tab. 3 that image transformations that corrupt the high-frequency details can significantly improve the retrieval performance. Conversely, transformations that merely alter geometric properties or color distri-

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

Method	Corrupt	Dynamic	Intra-subject top-1	Intra-subject top-5	Inter-subject top-1	Inter-subject top-5
Vanilla	✗	✗	42.1	74.5	8.5	26.6
Flip	✗	✗	40.8	73.8	8.6	25.9
Crop	✗	✗	41.6	74.0	9.6	27.2
Grayscale	✗	✗	38.8	72.4	9.1	27.0
Color jitter	✗	✗	41.3	76.2	8.5	25.7
Noise	✓	✗	47.7	78.8	10.0	30.5
Low-Res	✓	✗	48.1	78.4	10.8	31.9
Uniform blur	✓	✗	49.3	80.3	11.2	31.1
Fovea blur	✓	✗	50.2	79.1	12.3	31.7
UBP	✓	✓	50.9	79.7	12.4	33.4

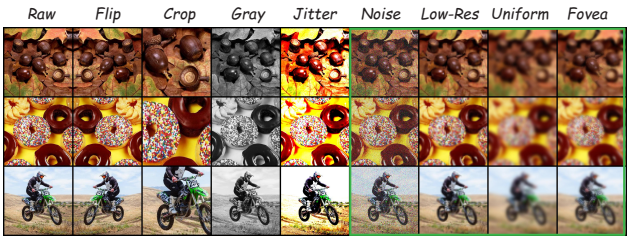


Figure 4. Illustration of various stimuli augmentations and corruptions applied to the visual stimuli. The augmentations (Flip, Crop, Grayscale, Color jitter) modify geometric properties or color distributions, while the corruptions (Gaussian noise, Low resolution, Uniform blur, Fovea blur) degrade image quality by introducing noise, lowering resolution, or simulating optical distortions.

butions do not yield comparable enhancements in performance. This further supports our motivation that reducing the information mismatch between visual stimuli and brain signals enables the model to better mitigate the overfitting issue arising from the System GAP. Moreover, the proposed Fovea Blur method, drawing inspiration from the human visual system, outperforms other corruptive transformations in terms of performance. Additionally, when the Random GAP is taken into account, the dynamic blurring method UBP further improves the retrieval performance.

5.4. Sensitivity Analysis of Various Blur Radius

To investigate the effect of varying degrees of blur on mitigating System GAP, we conducted experiments by applying a range of uniform blur radius, from 0 to 41, to the images. The results summarized in Fig. 5 show that as the blur level increases, both top-1 and top-5 accuracy improve, peaking at a blur radius of 11. As the blur level continues to increase, model performance begins to decline, which aligns with our expectation that an appropriate level of blur can reduce the mismatch between visual stimuli and brain signals. Excessive blur, such as a blur radius of 41, leads to a loss of information beyond the optimal level, increasing the

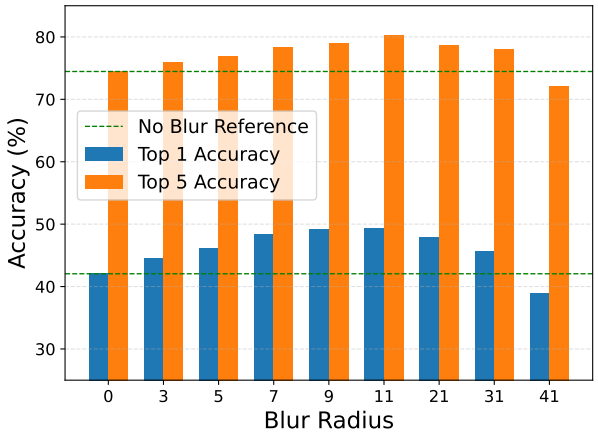


Figure 5. Comparison of Top-1 and Top-5 accuracy (%) at various blur radius, with reference accuracy for no-blur conditions.

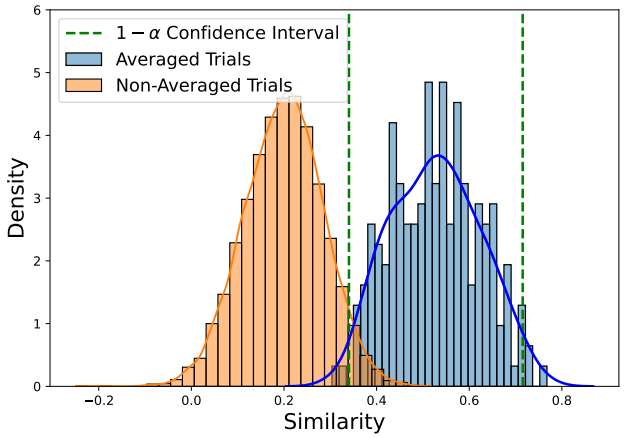


Figure 6. Distribution of similarity scores for averaged and non-averaged EEG trials. The dashed green lines denote the $1-\alpha$ confidence interval for the averaged trials. Our method effectively distinguishes the two types of trials, with non-averaged samples approximately treated as those with a large Random GAP.

information mismatch and resulting in worse performance compared to no blur.

5.5. Effectiveness of Uncertainty Quantification

Due to the unavailability of mismatch labels, direct evaluation of uncertainty quantification is challenging. However, non-averaged EEG signals, with their low signal-to-noise ratio, can serve as proxies for anomalous samples. As shown in Fig. 6, our method effectively distinguishes these anomalous samples based on the confidence intervals of the similarity distribution, thereby preventing the impact of extreme samples on generalization performance.

Figure 7. Top-1 accuracy improvement (%) of UBP across various brain and vision encoder combinations on the THINGS-EEG dataset.

5.6. Ablation Study on Various Encoders

To demonstrate that UBP is not architecture-specific, we conducted comprehensive experiments and trained thousands of models across five brain encoder architectures and eight image encoder architectures, where UBP consistently achieved robust performance improvements. Fig. 7 illustrate the improvements in top-1 accuracy of UBP on the THINGS-EEG. Detailed results, including the top-1 and top-5 accuracy for both baseline and UBP, are provided in the appendix.

5.7. Robustness to Intra-subject Variability

As shown in Tab. 4, we report the Pearson and Spearman correlations between intra-subject variability and zero-shot retrieval accuracy. Methods such as NICE-SA and VE-SDN exhibit strong negative Pearson correlations (e.g., -0.783 and -0.687 for top-1), indicating substantial sensitivity to intra-subject variability. Similarly, Vanilla presents a strong negative Pearson correlation (-0.761), underscoring its vulnerability to variability. In contrast, UBP demonstrates improved robustness, with a less negative Pearson correlation (-0.481) compared to Vanilla. UBP demonstrates more stable performance when handling subjects with high variability, with its accuracy not significantly degrading.

6. Conclusion

In this work, we propose the Uncertainty-aware Blur Prior (UBP) to address the System GAP and Random GAP in brain-to-image retrieval tasks. UBP leverages uncertainty estimation and biological priors to robustly retrieve natural images from multiple brain modalities. Extensive experiments demonstrate that UBP outperforms previous state-of-the-art methods, achieving a 13.7% improvement in Top-1 accuracy and a 9.8% improvement in Top-5 accuracy on the THINGS-EEG dataset, along with a 12.4% improvement in Top-1 accuracy and a 12.9% improvement in Top-5 accuracy on the THINGS-MEG dataset. Beyond brain-to-image retrieval, UBP holds potential for applications in stimuli reconstruction and broader multimodal learning contexts. To the best of our knowledge, we believe this is the first effort to incorporate uncertainty awareness and priors into visual neural decoding, offering new perspectives for brain-computer interfaces. Moreover, UBP provides valuable insights for other multimodal tasks, where similar challenges may arise.

Table 4. Pearson and Spearman correlation coefficients between each subject's mean variability value and the corresponding Top-1 accuracy for different methods.

	Pearson		**Spearmanr**	
Method	top-1	top-5	top-1	top-5
BraVL	-0.419	-0.451	-0.394	-0.406
NICE	-0.681	-0.705	-0.564	-0.588
NICE-SA	-0.783	-0.539	-0.745	-0.418
NICE-GA	-0.611	-0.709	-0.382	-0.450
ATM-S	-0.643	-0.608	-0.624	-0.697
VE-SDN	-0.687	-0.810	-0.787	-0.758
Vanilla	-0.761	-0.721	-0.636	-0.690
UBP	-0.481	-0.649	-0.345	-0.515
↑ Improvement	0.280	0.072	0.291	0.175

Limitations. Despite its effectiveness in reducing mismatches between brain signals and visual stimuli, UBP has certain limitations. While it uses a blur prior to approximate the loss of high-frequency details, this approach may not fully capture the complexity of the human visual system. Exploring advanced learnable methods may better bridge the GAPs and enhance generalization. Additionally, uncertainty quantification may fall short due to the complexity of the Random GAP, which is influenced by perceptual and cognitive dynamics, as well as technical noise. Future research could investigate advanced uncertainty quantification methods to enhance reliability and robustness.

Acknowledgements. This work is partially supported by the National Key R&DProgram of China (2022YFF1202400; X.Y.) and the National Natural Science Foundation of China (62376193). The authors appreciate the valuable feedback from anonymous reviewers.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021. 3

[2] Tyson Aflalo, Spencer Kellis, Christian Klaes, Brian Lee, Ying Shi, Kelsie Pejsa, Kathleen Shanfield, Stephanie Hayes-Jackson, Mindy Aisen, Christi Heck, et al. Decoding motor imagery from the posterior parietal cortex of a tetraplegic human. *Science*, 348(6237):906–910, 2015. 3

[3] Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*, 2023. 3

[4] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[5] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*, 2023. 1

[6] Ned Block. Perceptual consciousness overflows cognitive access. *Trends in cognitive sciences*, 15(12):567–575, 2011. 1

[7] Timothy J Buschman, Markus Siegel, Jefferson E Roy, and Earl K Miller. Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences*, 108(27):11252–11255, 2011. 1

[8] Patrick Cavanagh and George A Alvarez. Tracking multiple targets with multifocal attention. *Trends in cognitive sciences*, 9(7):349–354, 2005. 1

[9] Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in neural information processing systems*, 33: 1356–1367, 2020. 3

[10] Hongzhou Chen, Lianghua He, Yihang Liu, and Longzhen Yang. Visual neural decoding via improved visual-eeg semantic consistency. *arXiv preprint arXiv:2408.06788*, 2024. 1, 2, 3, 6

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3

[12] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720, 2023. 1

[13] Michael A Cohen, Daniel C Dennett, and Nancy Kanwisher. What is the bandwidth of perceptual experience? *Trends in cognitive sciences*, 20(5):324–335, 2016. 1

[14] Christine A Curcio, Kenneth R Sloan, Robert E Kalina, and Anita E Hendrickson. Human photoreceptor topography. *Journal of comparative neurology*, 292(4):497–523, 1990. 2

[15] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023. 1, 2, 3, 6

[16] Yiqun Duan, Charles Chau, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. Dewave: Discrete encoding of eeg waves for eeg to text translation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[17] Paul E Dux and Réne Marois. How humans search for targets through time: A review of data and theory from the attentional blink. *Attention, perception & psychophysics*, 71 (8):1683, 2009. 1

[18] Tao Fang, Qian Zheng, and Gang Pan. Alleviating the semantic gap for generalized fmri-to-image reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[19] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, 2021. 3

[20] Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254: 119121, 2022. 3

[21] 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. 5, 1

[22] 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*, pages 15180–15190, 2023. 3

[23] 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. 5, 1

[24] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 3

[25] 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. 5, 1

[26] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968. 1

[27] David H Hubel, Torsten N Wiesel, et al. Receptive fields of single neurones in the cat's striate cortex. *J physiol*, 148(3): 574–591, 1959. 1

[28] 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, 2021. If you use this software, please cite it as below. 6, 1

[29] Helene Intraub. Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 7(3):604, 1981. 5, 1

[30] 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 *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[31] Christian Keysers, D-K Xiao, Peter Földiák, and David I Perrett. The speed of sight. *Journal of cognitive neuroscience*, 13(1):90–101, 2001. 5, 1

[32] Eileen Kowler. Eye movements: The past 25 years. *Vision research*, 51(13):1457–1483, 2011. 1

[33] 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. 6, 1

[34] Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, and Mike Zheng Shou. Vit-lens: Towards omni-modal representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26647–26657, 2024. 3

[35] Jerome Y Lettvin, Humberto R Maturana, Warren S McCulloch, and Walter H Pitts. What the frog's eye tells the frog's brain. *Proceedings of the IRE*, 47(11):1940–1951, 1959. 1

[36] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *Advances in Neural Information Processing Systems*, 2024. 1, 2, 3, 6

[37] Xiang Li, Yazhou Zhang, Prayag Tiwari, Dawei Song, Bin Hu, Meihong Yang, Zhigang Zhao, Neeraj Kumar, and Pekka Marttinen. Eeg based emotion recognition: A tutorial and review. *ACM Computing Surveys*, 55(4):1–57, 2022. 3

[38] Liang Liang, Alex Fratzl, Glenn Goldey, Rohan N Ramesh, Arthur U Sugden, Josh L Morgan, Chinfei Chen, and Mark L Andermann. A fine-scale functional logic to convergence from retina to thalamus. *Cell*, 173(6):1343–1355, 2018. 1

[39] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 3

[40] Margaret S Livingstone and David H Hubel. Anatomy and physiology of a color system in the primate visual cortex. *Journal of Neuroscience*, 4(1):309–356, 1984. 1

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[42] Fan Lu, Kai Zhu, Wei Zhai, Kecheng Zheng, and Yang Cao. Uncertainty-aware optimal transport for semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3291, 2023. 3

[43] Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5899–5908, 2023. 3

[44] Steven J Luck and Edward K Vogel. Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in cognitive sciences*, 17(8):391–400, 2013. 1

[45] Huan Ma, Jingdong Chen, Guangyu Wang, and Changqing Zhang. Estimating llm uncertainty with logits. *arXiv preprint arXiv:2502.00290*, 2025. 4

[46] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer, 2022. 3

[47] Ian Nauhaus, Kristina J Nielsen, Anita A Disney, and Edward M Callaway. Orthogonal micro-organization of orientation and spatial frequency in primate primary visual cortex. *Nature neuroscience*, 15(12):1683–1690, 2012. 1

[48] Marc R Nuwer, Giancarlo Comi, Ronald Emerson, Anders Fuglsang-Frederiksen, Jean-Michel Guérit, Hermann Hinrichs, Akio Ikeda, Fransisco Jose C Luccas, and Peter Rappelsburger. Ifcn standards for digital recording of clinical eeg. *Electroencephalography and clinical Neurophysiology*, 106(3):259–261, 1998. 1

[49] Michael I Posner, Charles R Snyder, and Brian J Davidson. Attention and the detection of signals. *Journal of experimental psychology: General*, 109(2):160, 1980. 1

[50] Zenon Pylyshyn. Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. *Behavioral and brain sciences*, 22(3):341–365, 1999. 1

[51] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 233–243, 2024. 3

[52] 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*, pages 8748–8763. PMLR, 2021. 3, 4

[53] Marcus E Raichle, Ann Mary MacLeod, Abraham Z Snyder, William J Powers, Debra A Gusnard, and Gordon L Shulman. A default mode of brain function. *Proceedings of the national academy of sciences*, 98(2):676–682, 2001. 1

[54] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An

uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*. 4

[55] 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. 6, 1

[56] 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, 2024. 1, 3

[57] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024. 3

[58] Daniel J Simons and Daniel T Levin. Change blindness. *Trends in cognitive sciences*, 1(7):261–267, 1997. 1

[59] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3, 5, 6

[60] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024. 3

[61] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023. 1, 3

[62] Doris Y Tsao, Winrich A Freiwald, Roger BH Tootell, and Margaret S Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, 2006. 1

[63] David C Van Essen, Charles H Anderson, and Daniel J Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science*, 255(5043):419–423, 1992. 1

[64] Mário L Vicchietti, Fernando M Ramos, Luiz E Betting, and Andriana SLO Campanharo. Computational methods of eeg signals analysis for alzheimer's disease classification. *Scientific Reports*, 13(1):8184, 2023. 3

[65] Gang Wang, Keiji Tanaka, and Manabu Tanifuji. Optical imaging of functional organization in the monkey inferotemporal cortex. *Science*, 272(5268):1665–1668, 1996. 1

[66] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11333–11342, 2024. 3

[67] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 322–330, 2019. 4

[68] Zhenyu Wang, Ya-Li Li, Ye Guo, and Shengjin Wang. Combating noise: semi-supervised learning by region uncertainty quantification. *Advances in Neural Information Processing Systems*, 34:9534–9545, 2021. 4

[69] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 3

[70] Zehan Wang, Yang Zhao, Haifeng Huang, Jiageng Liu, Aoxiong Yin, Li Tang, Linjun Li, Yongqi Wang, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations. *Advances in Neural Information Processing Systems*, 36:22099–22114, 2023. 3

[71] David Whitney and Allison Yamanashi Leib. Ensemble perception. *Annual review of psychology*, 69(1):105–129, 2018. 1

[72] Jeremy M Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1:202–238, 1994. 1

[73] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8226–8235, 2024. 3

[74] Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[75] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1179–1189, 2023. 3

[76] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022. 3

[77] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2022. 4

[78] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020. 3

[79] Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho. Uncertainty aware semi-supervised learning on graph data. *Advances in Neural Information Processing Systems*, 33: 12827–12836, 2020. 4

[80] Qiongyi Zhou, Changde Du, Shengpei Wang, and Huiguang He. Clip-mused: Clip-guided multi-subject visual neu-

ral information semantic decoding. *arXiv preprint arXiv:2402.08994*, 2024. 3

Bridging the Vision-Brain Gap with an Uncertainty-Aware Blur Prior

Supplementary Material

A. Experimental details

A.1. Datasets details

THINGS-EEG [21] is a large scale EEG dataset included 10 subjects with the Rapid Serial Visual Presentation (RSVP) paradigm [23, 29, 31]. The EEG data are collected using 64-channel EASYCAP equipment with the standard 10-10 system [48]. The training set includes 1654 concepts with each concept 10 images, and each image repeats 4 times (1654 concepts $\times$ 10 images/concept $\times$ 4 trials/image) per subject. The test set includes 200 concepts with each concept 1 image, and each image repeats 80 times (200 concepts $\times$ 1 image/concept $\times$ 80 trials/image) per subject.

For data preprocessing, we follow the method detailed in [59]. Raw EEG data filtered to [0.1, 100] Hz has 63 channels and a sample rate of 1000 Hz. EEG data is epoched into trials ranging from 0 to 1000 ms after stimuli onset with baseline correction using the prior 200 ms average. EEG data is down-sampled to 250 Hz and 17 channels are selected overlying occipital and parietal cortex related to visual[1]. For the purpose of high Signal-to-Noise Ratio (SNR), EEG repetitions are averaged, resulting in total of 16540 training samples and 200 test samples per subject. Additionally, we store EEG data in float16 format to enable faster reading speeds and reduce storage requirements.

THINGS-MEG [25] dataset involves four participants and is characterized by 271 channels. The experimental design incorporates a relatively long stimulus duration of 500 ms, followed by a blank screen with a duration of 1000 ± 200 ms. It consists of 1854 concepts$\times$12 images$\times$1 repetitions in the training stage and 200 concepts$\times$1 image$\times$12 repetitions in the test stage.

We follow the settings described in [59]. During the data processing phase, 200 test concepts are discarded from the training set to construct the zero-shot task, mirroring the procedures in that study. Subsequently, the MEG data are epoched into trials covering the period from 0 to 1000 ms after the stimuli onset. For preprocessing, a band-pass filter within the range of [0.1, 100] Hz is utilized, and baseline correction is carried out after down-sampling the data to 200 Hz. Additionally, we average all MEG repetitions of one image to ensure the signal-to-noise ratio. Additionally, we store EEG data in float16 format to enable faster reading speeds and reduce storage requirements.

[1] P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, O2

A.2. Implementation details

Environment. Our method is implemented with Python 3.8.19, CUDA 12.0, and PyTorch 2.4.1. The required libraries are specified in the `requirements.txt` file provided in the repository. The experiments are performed on a machine equipped with an Intel Xeon Platinum 8352V CPU, four V100 GPUs, and 256 GB of RAM.

Training Configuration. We use a batch size of 1024 and train the model for 50 epochs. The learning rate is set to 1e-4 for intra-subject setting and 1e-5 for inter-subject setting. Gradient updates are performed using the AdamW[41] optimizer with weight decay set to 1e-4. Early stopping is employed to monitor training loss and validation performance, concluding the training process to mitigate overfitting when improvements stabilize. Notably, we use the softplus function instead of the exponential function to ensure that temperature parameter τ remains positive and continuous, as softplus offers a smoother and more stable transition, avoiding the numerical instability of the exponential function. For all above experiments, the hyperparameter r_0 is set to 0.25 and c is set to 10.

Architectures. We use EEGProject as the brain encoder, detailed as follows:

```
Input: Brain Data
       ↓
Linear Layer (Input_dim → Proj_dim)
       ↓
ResidualAdd
 |   +----------------------+
 |   | GELU                 |
 |   | Linear Layer         |
 |   | Dropout (Rate: 0.3)  |
 |   +----------------------+
 +---- ↓
LayerNorm
       ↓
Output: Latent
```

We provided the number of parameters and embedding dimension within different EEG encoders [33, 55, 59], in Tab. 21. Compared to other models, EEGProject achieves its performance through a simple yet effective architecture, while remaining lightweight with 5.154M parameters, especially in comparison to the vision branch.

We also provide the parameter counts for various CLIP vision branch models [28] to offer a comprehensive comparison across architectures in Tab. 22.

B. Results details

B.1. Retrieval Case Analysis

We present the top-5 retrieval results on THINGS-EEG dataset, including both good cases and bad cases, as shown in Fig. 8 and Fig. 9, respectively. Good cases demonstrate the model's capability to effectively align with the target stimuli and retrieve relevant results. An intriguing retrieval result is that the model not only retrieves items with similar materials but also demonstrates **associations with the orientation and quantity of objects**. These observations warrant further investigation in future studies.

Figure 8. Good Cases: Top-5 Retrieval Results for Various Stimuli.

In contrast, bad cases reveal limitations in distinguishing fine-grained features or addressing semantic inconsistencies. It is challenging to distinguish highly similar stimuli due to the limited information contained in brain signals.

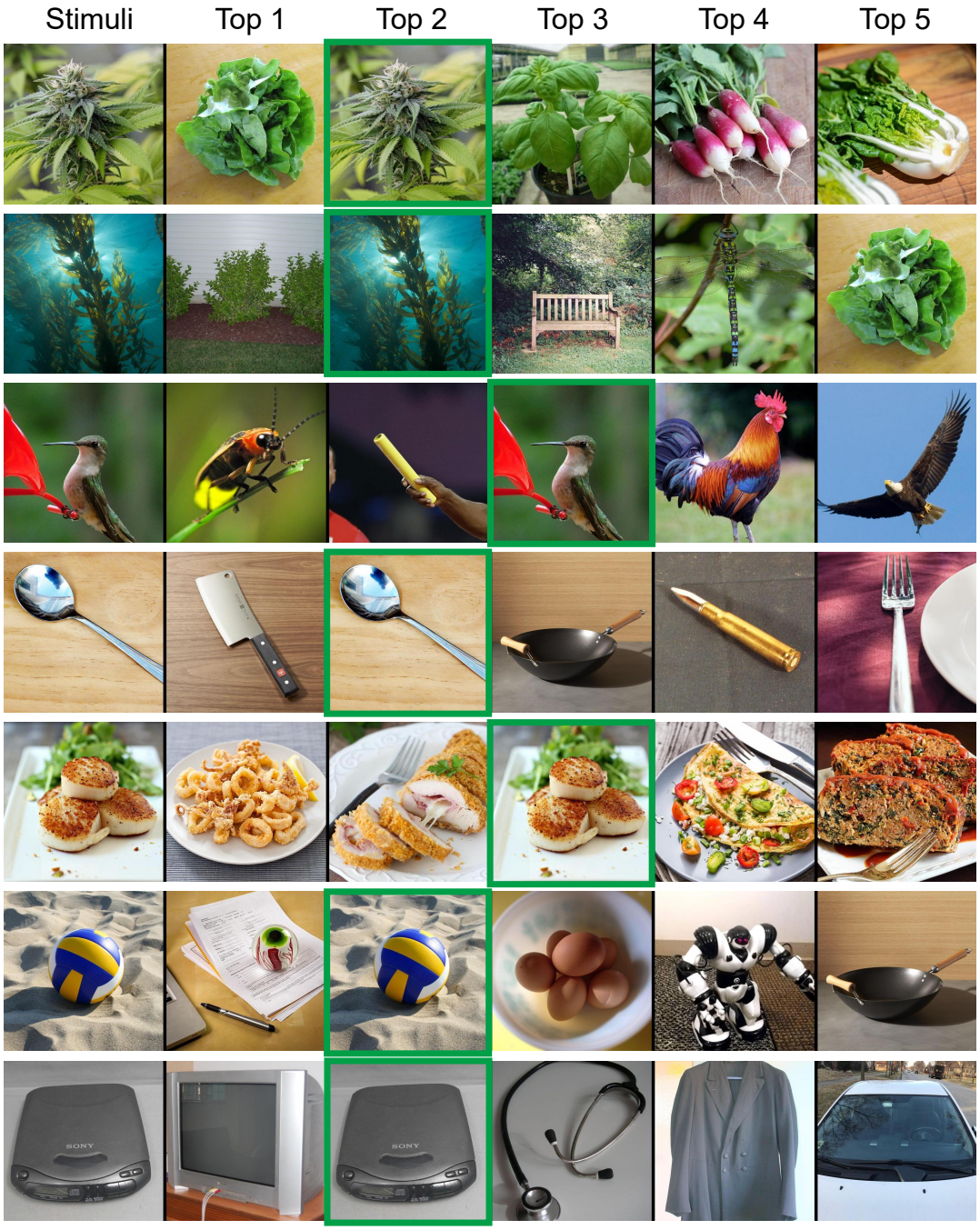
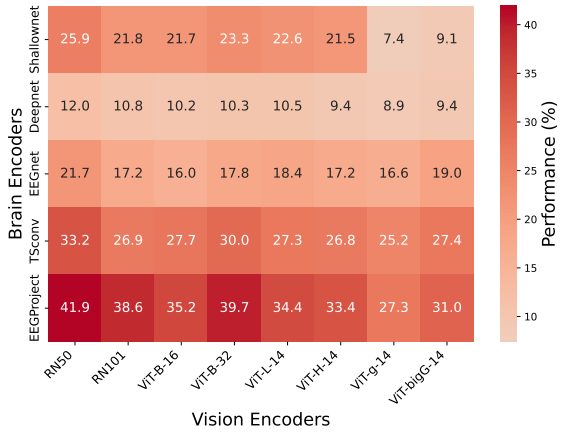
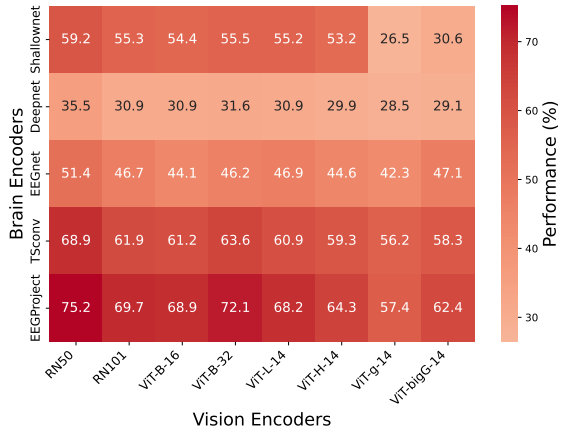


Figure 9. Bad Cases: Top-5 Retrieval Results for Various Stimuli.

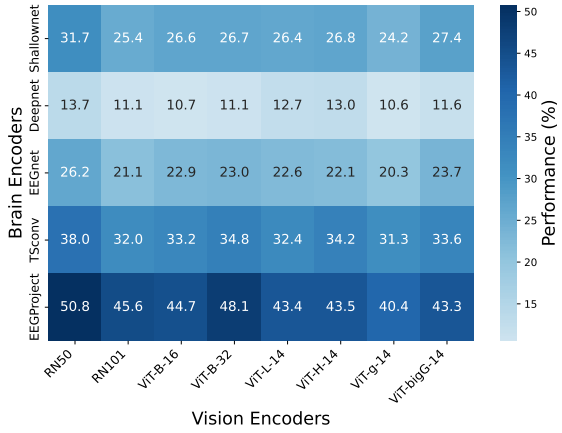
B.2. THINGS-EEG Results



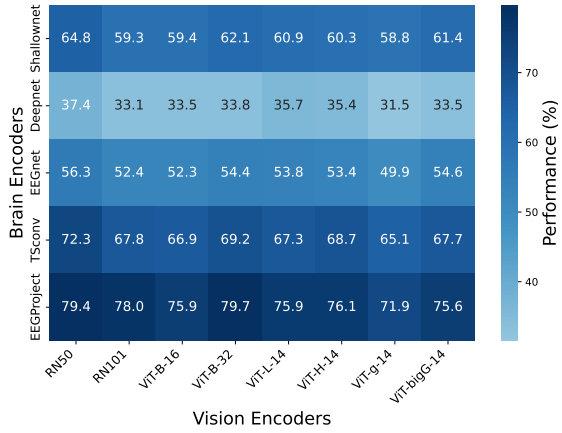
(a) **Top-1** accuracy (%) of **Vanilla** on the THINGS-EEG dataset.



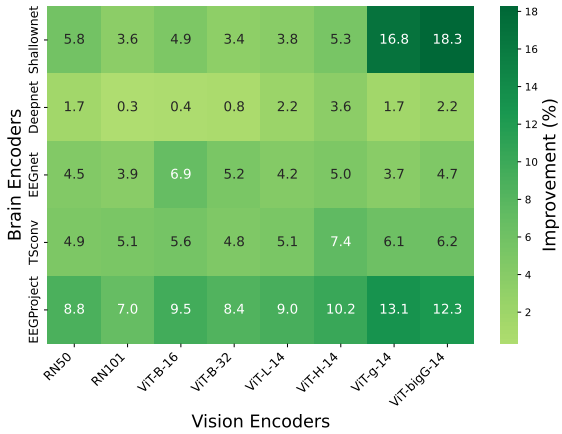
(b) **Top-5** accuracy (%) of **Vanilla** on the THINGS-EEG dataset.



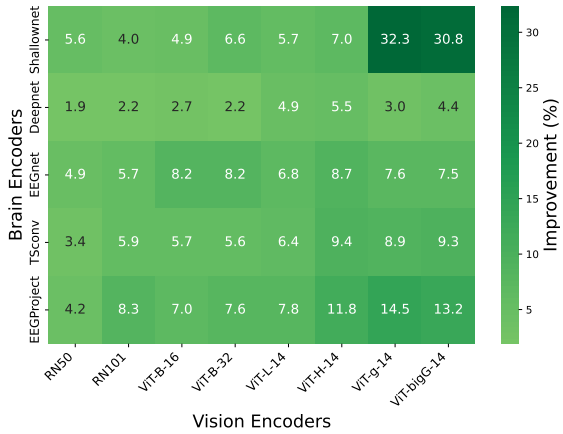
(c) **Top-1** accuracy (%) of **UBP** on the THINGS-EEG dataset.



(d) **Top-5** accuracy (%) of **UBP** on the THINGS-EEG dataset.



(e) **Top-1** accuracy **improvement** (%) on the THINGS-EEG dataset.



(f) **Top-5** accuracy **improvement** (%) on the THINGS-EEG dataset.

Figure 10. Results on the THINGS-EEG dataset.

Table 5. Top-1 and Top-5 Accuracy (%) on THINGS-EEG with CLIP **RN50** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Avg	
Backbone	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
ShallowNet	18.0	48.3	24.0	54.8	23.0	58.8	30.0	63.2	20.3	45.7	27.8	62.5	26.0	60.5	32.8	69.7	24.5	59.3	33.0	69.2	25.9	59.2
w/ UBP	24.5	57.5	25.0	60.5	31.2	64.5	35.2	66.8	20.7	49.0	37.3	70.0	33.7	65.3	42.2	78.0	27.0	63.7	40.0	72.7	31.7	64.8
DeepNet	7.8	27.5	11.0	35.8	11.2	32.0	14.5	40.0	8.2	24.0	12.2	38.5	10.5	33.2	16.3	39.0	15.5	40.5	12.5	44.5	12.0	35.5
w/ UBP	10.5	33.2	10.8	35.3	11.7	37.0	17.8	44.2	6.5	23.3	17.0	40.5	14.2	36.2	14.5	42.0	16.0	36.7	17.5	45.7	13.7	37.4
EEGNet	11.0	38.3	18.5	51.7	19.5	51.3	24.0	56.2	16.3	41.8	24.5	57.2	19.5	43.8	30.0	62.5	24.2	50.0	29.5	61.0	21.7	51.4
w/ UBP	18.5	44.0	23.2	53.2	25.7	58.2	33.3	62.0	20.0	45.8	30.2	62.7	22.5	53.2	33.7	66.2	25.0	52.0	29.5	65.5	26.2	56.3
TSConv	29.3	61.3	27.0	63.2	32.0	73.0	35.0	72.5	25.2	59.0	38.0	74.3	31.0	66.5	43.0	76.8	31.5	66.5	40.0	75.5	33.2	68.9
w/ UBP	38.0	69.5	28.7	68.5	38.8	75.7	41.3	74.3	29.7	61.0	42.0	76.7	36.2	70.8	48.5	79.7	33.2	67.3	44.0	79.0	38.0	72.3
EEGProject	30.8	64.2	39.5	71.5	42.2	78.5	42.8	76.5	33.8	69.2	45.5	79.3	40.7	71.8	49.8	80.5	41.7	74.5	52.5	85.8	41.9	75.2
w/ UBP	41.2	71.2	50.7	82.2	50.2	81.7	49.3	77.8	43.3	70.8	58.7	82.5	49.8	79.0	57.8	84.3	44.5	76.7	62.3	87.7	50.8	79.4

Table 6. Top-1 and Top-5 Accuracy (%) on THINGS-EEG with CLIP **RN101** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Avg	
Backbone	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
ShallowNet	17.8	45.7	18.2	53.0	20.7	53.5	28.0	61.0	15.2	43.0	21.8	58.0	20.3	54.8	29.7	65.3	21.0	54.2	25.2	64.2	21.8	55.3
w/ UBP	23.5	55.3	19.5	56.0	26.0	57.7	29.7	63.7	15.5	42.5	29.7	66.0	22.5	56.5	32.8	68.3	22.0	59.5	32.5	67.5	25.4	59.3
DeepNet	8.7	23.5	11.3	29.5	5.7	23.3	15.0	38.0	7.2	21.5	10.2	31.0	10.0	28.7	9.5	33.2	12.5	35.8	17.2	44.5	10.8	30.9
w/ UBP	8.8	27.8	10.0	30.3	8.2	30.5	14.3	39.5	8.5	22.0	12.2	32.8	8.8	30.0	11.2	34.5	13.3	37.7	15.5	46.3	11.1	33.1
EEGNet	12.5	32.8	14.5	42.7	15.3	43.8	22.5	54.7	13.2	40.3	16.5	51.5	16.0	42.7	20.0	56.7	17.7	45.7	24.2	56.2	17.2	46.7
w/ UBP	18.0	43.8	18.8	52.0	19.5	51.7	22.3	55.8	18.3	46.3	26.0	60.5	18.2	43.5	21.5	58.5	21.7	50.0	26.7	62.3	21.1	52.4
TSConv	24.8	56.5	24.8	58.5	23.3	60.5	33.5	67.8	19.0	48.5	27.5	68.2	25.0	58.0	34.3	71.5	25.5	60.5	31.7	69.0	26.9	61.9
w/ UBP	30.8	65.0	24.2	65.2	29.5	65.2	34.0	69.5	24.2	55.0	40.0	77.0	26.5	62.3	41.2	77.8	29.7	66.2	39.5	74.8	32.0	67.8
EEGProject	29.0	59.3	36.0	66.3	40.7	72.5	42.0	73.5	30.5	58.8	39.7	72.0	39.2	70.5	41.0	77.5	37.3	67.0	50.5	79.5	38.6	69.7
w/ UBP	35.2	69.7	45.5	80.8	46.5	78.7	47.7	77.7	40.7	71.3	49.5	83.5	42.7	75.0	55.0	81.0	39.2	75.0	53.5	87.3	45.6	78.0

Table 7. Top-1 and Top-5 Accuracy (%) on THINGS-EEG with CLIP **ViT-B-16** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Avg	
Backbone	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
ShallowNet	15.0	41.0	16.0	48.3	25.3	58.2	28.5	61.8	14.5	41.8	21.2	57.7	21.8	54.8	28.2	63.5	21.5	52.3	25.2	65.0	21.7	54.4
w/ UBP	22.3	51.8	23.5	53.0	24.0	61.8	32.3	65.0	16.8	45.0	28.2	64.0	27.2	60.0	34.7	71.0	23.0	55.3	34.3	67.0	26.6	59.4
DeepNet	9.0	22.5	6.7	29.3	9.0	30.5	13.5	36.8	4.7	17.8	10.0	31.0	10.0	31.2	9.8	33.2	15.3	37.5	14.2	39.0	10.2	30.9
w/ UBP	10.5	26.0	8.7	32.8	9.0	32.5	15.0	42.7	5.5	23.0	12.3	34.0	9.3	31.2	12.0	35.0	13.7	39.5	10.5	38.8	10.7	33.5
EEGNet	12.0	36.0	10.7	40.2	19.8	47.2	19.2	45.5	9.8	33.8	20.5	49.0	14.2	38.8	19.0	52.0	16.3	44.0	18.8	54.2	16.0	44.1
w/ UBP	18.8	41.5	18.5	48.5	23.5	55.0	24.7	57.0	16.5	44.0	27.0	57.2	20.7	49.8	29.3	58.5	23.0	50.2	27.5	61.2	22.9	52.3
TSConv	19.7	49.0	23.5	56.7	27.0	63.7	31.5	67.0	18.8	49.3	34.5	65.3	27.2	60.7	36.5	72.8	27.8	59.5	30.2	67.5	27.7	61.2
w/ UBP	28.7	63.0	28.0	63.0	31.2	66.8	37.0	72.2	22.5	54.5	41.0	74.8	33.7	63.0	42.2	76.7	31.7	63.5	36.3	71.5	33.2	66.9
EEGProject	29.5	59.5	32.8	68.0	36.5	71.8	35.3	68.0	28.2	54.0	37.2	73.3	34.3	70.0	36.8	75.0	37.3	69.2	44.0	79.7	35.2	68.9
w/ UBP	37.3	62.7	46.5	77.5	46.5	76.2	42.8	77.2	34.7	66.8	46.8	79.2	40.3	74.5	51.7	81.2	44.8	75.7	55.5	87.5	44.7	75.9

Table 8. Top-1 and Top-5 Accuracy (%) on THINGS-EEG with CLIP **ViT-B-32** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Avg	
Backbone	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
ShallowNet	15.8	45.0	17.8	49.0	27.3	57.0	28.7	65.5	17.5	41.0	25.2	60.5	20.3	54.7	30.7	63.2	22.7	56.0	27.3	63.0	23.3	55.5
w/ UBP	20.0	53.2	22.0	55.0	26.5	63.0	35.3	69.2	19.5	48.7	27.3	65.7	27.7	62.7	37.0	73.5	22.5	60.8	29.7	69.3	26.7	62.1
DeepNet	8.8	23.8	10.2	30.5	9.8	29.3	12.0	39.0	5.3	19.7	10.7	32.8	11.2	32.0	9.0	31.8	13.5	37.3	12.8	39.7	10.3	31.6
w/ UBP	8.7	29.0	10.2	34.3	9.3	34.3	12.5	41.2	8.5	21.0	11.2	33.5	14.0	33.0	12.0	32.3	12.0	37.0	12.8	42.0	11.1	33.8
EEGNet	14.3	36.0	14.2	43.0	18.3	46.3	24.0	53.0	12.8	36.5	20.0	50.2	15.5	43.8	20.3	53.5	18.8	43.8	20.2	56.0	17.8	46.2
w/ UBP	18.5	40.5	21.7	54.7	20.7	56.0	25.0	58.2	19.2	45.3	30.2	60.5	20.3	48.3	27.0	63.5	22.0	54.0	25.2	63.5	23.0	54.4
TSConv	19.8	55.8	22.7	55.8	33.8	65.0	34.5	72.5	20.0	52.5	33.7	68.3	25.3	60.5	40.3	74.5	33.7	62.5	36.5	69.0	30.0	63.6
w/ UBP	30.8	63.5	27.5	64.0	37.5	70.8	36.8	72.3	25.8	57.7	43.5	76.0	34.8	68.0	41.5	78.7	34.0	69.0	36.2	72.5	34.8	69.2
EEGProject	30.7	59.5	39.2	70.5	43.0	79.5	40.3	73.5	31.2	65.0	42.0	74.3	36.8	68.8	45.7	75.0	41.2	71.3	46.5	83.2	39.7	72.1
w/ UBP	37.5	70.0	46.5	80.5	52.8	85.5	47.2	80.8	37.5	70.7	54.5	81.7	43.8	79.5	58.8	81.5	46.5	77.0	56.0	89.7	48.1	79.7

Table 9. Top-1 and Top-5 Accuracy (%) on THINGS-EEG with CLIP **ViT-L-14** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Avg	
Backbone	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
ShallowNet	16.3	45.3	17.8	51.2	19.7	57.5	27.0	60.2	17.5	45.5	26.3	60.8	20.0	48.3	30.5	63.2	21.8	54.5	29.5	65.7	22.6	55.2
w/ UBP	15.7	47.8	23.3	52.0	24.5	61.2	32.2	66.7	17.2	50.5	33.2	69.2	25.7	61.5	35.5	71.0	23.5	58.8	33.2	70.2	26.4	60.9
DeepNet	8.3	23.0	9.0	29.0	9.0	29.7	12.8	34.7	7.3	20.5	11.5	33.3	8.3	31.7	13.0	36.3	13.5	36.5	12.0	33.7	10.5	30.9
w/ UBP	10.2	28.7	11.8	33.5	12.0	34.3	16.5	41.2	6.7	23.7	14.5	39.2	13.2	37.5	12.5	40.5	15.3	37.5	13.8	41.0	12.7	35.7
EEGNet	12.8	35.8	13.3	42.0	16.3	47.8	20.8	50.0	14.2	34.5	23.8	52.7	14.3	41.8	24.2	58.8	17.5	47.7	27.0	58.5	18.4	46.9
w/ UBP	14.3	40.5	20.7	45.5	21.7	54.0	29.0	59.5	18.0	45.7	27.5	61.3	19.7	53.0	26.3	61.3	21.8	54.0	27.5	63.0	22.6	53.8
TSConv	21.7	55.7	24.0	56.5	25.2	61.0	31.2	65.5	20.0	47.2	31.2	69.0	25.7	58.2	37.7	70.0	21.8	56.0	34.0	70.0	27.3	60.9
w/ UBP	26.5	60.0	26.0	60.5	32.3	66.5	36.3	72.0	24.5	56.0	39.0	75.2	28.2	65.3	42.8	78.5	29.0	63.5	39.0	76.0	32.4	67.3
EEGProject	26.7	55.3	33.5	66.5	35.2	73.3	32.0	70.0	27.5	51.7	36.2	72.2	33.5	72.8	44.5	75.5	29.5	67.0	45.3	77.7	34.4	68.2
w/ UBP	31.2	66.2	39.0	73.8	43.3	77.5	43.8	76.2	31.0	69.5	51.5	83.0	41.0	74.5	56.2	80.5	42.5	74.0	54.8	84.3	43.4	75.9

Table 10. Top-1 and Top-5 Accuracy (%) on THINGS-EEG with CLIP **ViT-H-14** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Avg	
Backbone	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
ShallowNet	15.2	42.2	16.8	47.0	25.0	56.5	25.5	59.3	11.5	43.0	23.2	55.0	18.0	51.0	27.8	62.7	22.7	55.5	29.5	60.0	21.5	53.2
w/ UBP	21.0	51.0	21.8	50.3	24.2	62.7	32.0	64.3	17.8	45.7	30.5	69.5	24.2	60.0	31.7	71.8	26.2	59.5	38.3	67.8	26.8	60.3
DeepNet	8.0	24.8	9.0	27.2	11.5	31.0	9.0	30.3	4.5	20.7	9.5	27.8	9.5	30.0	8.8	32.0	11.2	34.0	13.3	41.0	9.4	29.9
w/ UBP	9.3	28.7	14.3	36.5	14.0	37.3	14.0	36.5	8.5	22.2	13.8	33.8	11.7	36.5	14.5	38.0	14.0	38.5	15.7	45.8	13.0	35.4
EEGNet	9.3	30.8	17.3	43.5	18.8	46.0	18.5	47.3	12.5	36.2	21.2	49.0	14.0	40.5	21.2	51.0	17.7	48.5	21.0	53.5	17.2	44.6
w/ UBP	15.0	42.5	19.7	49.0	22.0	54.0	26.0	58.5	20.5	44.0	22.5	58.0	21.2	49.2	26.0	60.0	21.5	54.5	27.0	63.8	22.1	53.4
TSConv	20.5	50.0	22.0	55.8	27.2	60.5	29.5	61.8	19.0	49.8	30.5	61.5	22.7	57.8	35.0	70.0	27.0	59.0	34.7	66.7	26.8	59.3
w/ UBP	28.7	63.3	29.7	65.0	32.8	69.5	37.0	71.0	28.7	59.0	39.5	76.2	33.2	65.3	40.0	77.8	29.7	63.7	43.0	76.5	34.2	68.7
EEGProject	19.7	54.5	31.0	62.0	39.0	70.0	35.0	62.0	26.3	54.0	38.2	68.5	29.3	61.8	39.5	72.2	34.8	62.2	40.7	76.0	33.4	64.3
w/ UBP	32.0	64.0	45.0	75.7	45.8	81.5	42.5	77.5	34.7	66.5	47.0	81.2	40.0	73.5	51.2	82.5	45.7	72.3	51.2	86.5	43.5	76.1

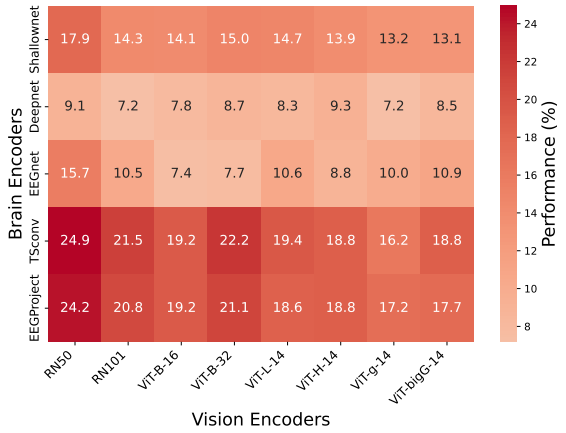
Table 11. Top-1 and Top-5 Accuracy (%) on THINGS-EEG with CLIP **ViT-g-14** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Avg	
Backbone	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
ShallowNet	6.5	23.5	4.7	21.0	6.2	28.8	9.8	31.0	5.0	17.0	7.3	25.5	6.8	24.7	11.3	33.5	5.7	25.0	10.8	34.5	7.4	26.5
w/ UBP	19.5	51.5	16.8	50.2	26.5	61.0	29.0	66.0	17.5	45.0	29.0	62.7	22.5	58.7	27.5	67.5	23.7	58.2	30.2	67.0	24.2	58.8
DeepNet	6.5	23.7	6.8	24.8	10.8	31.5	12.2	32.3	6.5	17.0	9.5	26.2	9.0	29.0	10.5	32.2	9.8	30.8	7.0	37.7	8.9	28.5
w/ UBP	9.5	26.0	8.8	26.5	11.0	33.2	12.8	33.2	6.0	21.5	12.0	33.7	11.0	36.0	9.5	33.7	11.0	32.8	14.2	38.5	10.6	31.5
EEGNet	8.3	29.3	13.3	39.2	19.3	47.0	20.3	48.7	11.5	35.0	19.5	44.0	15.0	36.8	20.0	52.2	16.5	40.5	22.0	50.7	16.6	42.3
w/ UBP	13.3	37.5	16.0	42.7	22.7	52.2	25.0	55.8	14.0	44.2	23.0	52.5	18.8	47.5	25.2	58.0	18.0	50.7	26.8	57.8	20.3	49.9
TSConv	24.5	52.0	20.0	50.5	26.0	57.7	27.5	61.0	20.5	48.8	25.3	58.5	24.5	51.0	31.0	63.0	22.7	55.0	29.7	64.5	25.2	56.2
w/ UBP	31.5	63.5	21.2	54.7	32.8	64.3	32.2	67.0	23.7	55.8	33.7	69.5	32.8	64.0	37.0	77.2	29.3	62.3	39.0	73.3	31.3	65.1
EEGProject	20.3	48.5	23.5	52.3	30.0	59.7	28.5	62.7	23.3	48.3	29.0	60.2	28.2	54.8	30.7	64.0	29.0	57.8	30.5	65.5	27.3	57.4
w/ UBP	29.0	62.0	38.5	67.5	42.2	77.2	41.0	74.8	31.0	59.5	44.3	74.8	37.3	71.0	48.5	81.0	42.0	69.7	50.2	81.2	40.4	71.9

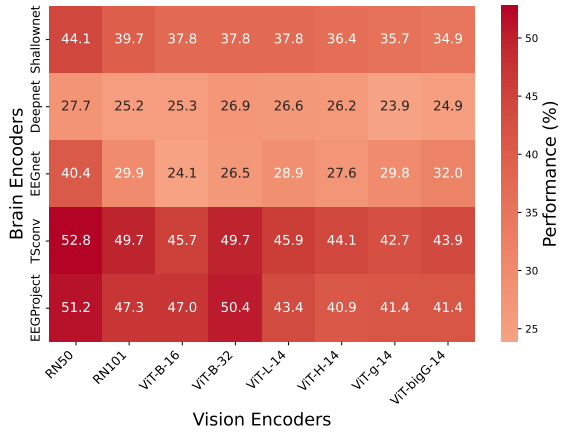
Table 12. Top-1 and Top-5 Accuracy (%) on THINGS-EEG with CLIP **ViT-bigG-14** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Avg	
Backbone	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
ShallowNet	7.8	26.5	6.3	19.0	9.0	33.8	9.8	33.0	6.0	20.0	10.2	29.7	7.8	32.3	12.3	39.0	9.2	32.3	13.0	40.3	9.1	30.6
w/ UBP	19.5	52.8	19.0	57.5	29.7	60.7	34.3	68.3	17.5	46.2	34.0	67.0	28.2	59.7	33.2	71.0	27.2	59.0	31.2	72.0	27.4	61.4
DeepNet	7.3	21.5	8.0	27.8	9.0	31.5	10.5	32.5	7.5	21.7	10.5	29.3	10.8	29.7	10.0	29.7	9.8	32.0	11.0	34.7	9.4	29.1
w/ UBP	7.3	21.5	8.0	27.8	9.0	31.5	10.5	32.5	7.5	21.7	10.5	29.3	10.8	29.7	10.0	29.7	9.8	32.0	11.0	34.7	9.4	29.1
EEGNet	13.0	36.2	13.0	43.0	20.0	47.2	23.2	51.5	17.0	36.7	18.8	50.7	16.8	42.5	25.5	58.0	19.3	50.5	23.8	54.5	19.0	47.1
w/ UBP	16.3	40.7	20.5	53.7	24.0	55.8	28.0	58.8	18.5	45.0	28.0	59.7	23.2	51.2	28.7	62.5	24.5	55.5	25.5	62.7	23.7	54.6
TSConv	22.0	48.0	18.2	57.2	26.7	58.5	31.2	61.0	19.8	46.0	32.5	63.5	24.7	53.7	38.5	68.0	24.8	58.0	35.8	69.5	27.4	58.3
w/ UBP	30.0	62.0	28.5	64.5	32.0	66.2	36.8	70.7	23.7	54.7	43.3	74.3	32.2	67.5	42.0	75.2	30.5	65.5	37.3	76.0	33.6	67.7
EEGProject	22.0	52.5	29.7	58.5	31.5	63.7	29.7	63.7	25.2	53.5	36.2	67.3	32.5	63.0	37.5	68.2	26.3	61.3	39.2	72.0	31.0	62.4
w/ UBP	33.8	63.7	41.8	78.2	43.8	78.2	39.7	73.8	34.0	64.0	55.8	82.7	38.7	74.8	49.0	80.3	43.5	74.3	53.2	85.5	43.3	75.6

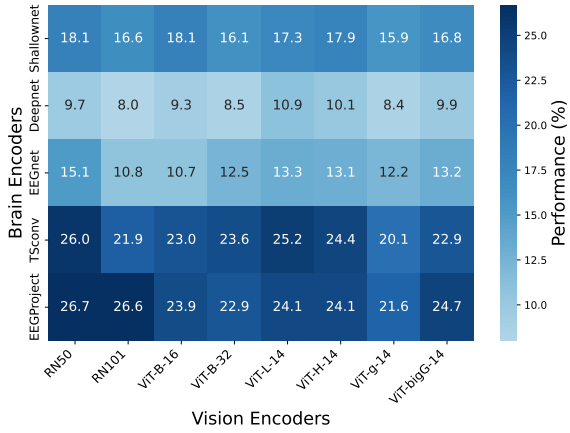
B.3. THINGS-MEG Results



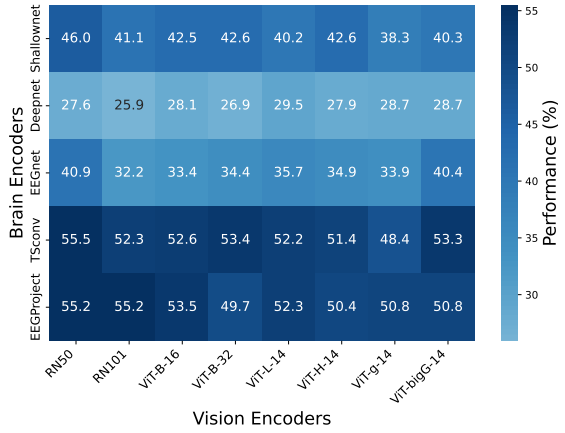
(a) **Top-1** accuracy (%) of **Vanilla** on the THINGS-MEG dataset.



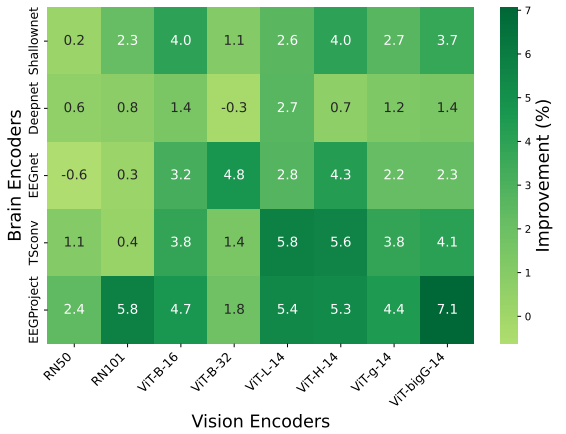
(b) **Top-5** accuracy (%) of **Vanilla** on the THINGS-MEG dataset.



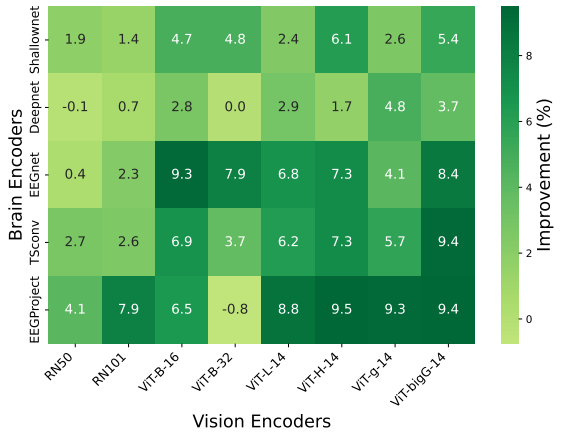
(c) **Top-1** accuracy (%) of **UBP** on the THINGS-MEG dataset.



(d) **Top-5** accuracy (%) of **UBP** on the THINGS-MEG dataset.



(e) **Top-1** accuracy **improvement** (%) on the THINGS-MEG dataset.



(f) **Top-5** accuracy **improvement** (%) on the THINGS-MEG dataset.

Figure 11. Results on the THINGS-EEG dataset.

Table 13. Top-1 and Top-5 Accuracy (%) on THINGS-MEG with CLIP **RN50** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Avg	
Backbone	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
ShallowNet	9.8	32.0	27.3	58.5	25.2	56.2	9.3	29.7	17.9	44.1
w/ UBP	10.7	36.0	27.0	65.3	22.2	51.8	12.5	31.0	18.1	46.0
DeepNet	4.3	17.0	13.3	40.0	12.8	34.7	6.2	19.2	9.1	27.7
w/ UBP	4.5	14.8	16.5	40.5	11.3	36.0	6.7	19.2	9.7	27.6
EEGNet	10.5	26.8	25.2	59.3	19.3	47.5	8.0	28.2	15.7	40.4
w/ UBP	7.5	24.5	29.7	66.0	14.5	46.8	8.7	26.3	15.1	40.9
TSconv	14.8	40.5	39.2	73.5	30.2	60.8	15.5	36.2	24.9	52.8
w/ UBP	18.3	45.0	40.7	76.5	28.0	59.0	17.0	41.5	26.0	55.5
EEGProject	13.2	34.5	43.3	74.5	27.0	57.0	13.5	38.8	24.2	51.2
w/ UBP	15.0	38.0	46.0	80.5	27.3	59.0	18.5	43.5	26.7	55.2

Table 14. Top-1 and Top-5 Accuracy (%) on THINGS-MEG with CLIP **RN101** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Avg	
Backbone	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
ShallowNet	8.5	27.8	20.0	53.0	21.0	49.7	7.7	28.2	14.3	39.7
w/ UBP	8.7	30.0	25.0	55.3	21.5	46.3	11.0	33.0	16.6	41.1
DeepNet	4.8	15.0	11.5	36.2	8.3	30.5	4.3	19.3	7.2	25.2
w/ UBP	3.5	17.0	12.3	38.5	9.5	27.5	6.8	20.7	8.0	25.9
EEGNet	6.5	18.0	17.2	47.2	14.2	36.5	4.0	17.7	10.5	29.9
w/ UBP	6.0	20.7	19.3	49.8	11.8	36.5	6.0	21.8	10.8	32.2
TSconv	15.0	37.7	32.0	67.2	24.7	55.5	14.3	38.5	21.5	49.7
w/ UBP	13.5	40.5	31.5	68.8	25.5	58.3	17.0	41.7	21.9	52.3
EEGProject	10.7	33.2	37.3	69.5	26.2	55.5	9.0	31.0	20.8	47.3
w/ UBP	15.3	37.7	44.8	76.5	31.0	63.5	15.5	43.3	26.6	55.2

Table 15. Top-1 and Top-5 Accuracy (%) on THINGS-MEG with CLIP **ViT-B-16** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Avg	
Backbone	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
ShallowNet	8.2	28.5	21.5	49.8	20.5	45.0	6.2	27.8	14.1	37.8
w/ UBP	9.8	28.8	25.0	59.7	25.0	52.8	12.8	28.7	18.1	42.5
DeepNet	4.0	15.3	14.8	35.3	9.0	32.2	3.5	18.3	7.8	25.3
w/ UBP	6.8	19.0	14.8	35.3	8.0	35.8	7.5	22.3	9.3	28.1
EEGNet	3.0	12.8	14.5	40.3	9.5	29.5	2.7	13.8	7.4	24.1
w/ UBP	5.3	20.8	18.5	54.5	13.7	38.5	5.3	19.7	10.7	33.4
TSconv	10.8	33.3	29.3	62.3	26.0	54.8	10.7	32.5	19.2	45.7
w/ UBP	13.5	40.3	35.3	70.3	29.0	62.3	14.3	37.5	23.0	52.6
EEGProject	14.2	35.8	27.8	65.3	23.7	52.0	11.0	35.0	19.2	47.0
w/ UBP	16.7	39.2	38.2	72.5	23.7	62.3	16.8	40.0	23.9	53.5

Table 16. Top-1 and Top-5 Accuracy (%) on THINGS-MEG with CLIP **ViT-B-32** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Avg	
Backbone	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
ShallowNet	8.2	24.2	24.0	54.8	18.3	46.0	9.5	26.2	15.0	37.8
w/ UBP	7.5	30.0	23.3	54.8	21.8	55.8	11.7	29.7	16.1	42.6
DeepNet	5.0	18.5	12.5	38.5	12.2	32.2	5.3	18.2	8.7	26.9
w/ UBP	4.7	18.3	13.3	37.5	10.7	32.8	5.2	19.0	8.5	26.9
EEGNet	4.5	15.7	15.7	45.5	8.5	31.2	2.2	13.5	7.7	26.5
w/ UBP	8.0	22.5	24.5	58.0	13.3	41.0	4.2	16.3	12.5	34.4
TSconv	16.3	38.5	34.3	68.5	24.2	54.0	14.0	37.7	22.2	49.7
w/ UBP	14.8	42.2	37.3	69.2	27.8	62.0	14.8	40.2	23.6	53.4
EEGProject	10.0	35.8	34.7	70.3	25.5	54.5	14.0	41.3	21.1	50.4
w/ UBP	15.0	38.0	34.0	66.5	29.7	63.3	12.7	31.0	22.9	49.7

Table 17. Top-1 and Top-5 Accuracy (%) on THINGS-MEG with CLIP **ViT-L-14** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Avg	
Backbone	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
ShallowNet	6.5	26.0	23.0	51.7	19.0	44.0	10.2	29.5	14.7	37.8
w/ UBP	8.7	27.5	30.8	57.7	17.5	48.0	12.0	27.5	17.3	40.2
DeepNet	4.5	18.0	12.3	37.3	11.3	34.3	5.0	16.8	8.3	26.6
w/ UBP	6.2	18.8	18.5	44.0	12.8	34.5	6.2	20.7	10.9	29.5
EEGNet	6.2	18.5	18.2	45.0	11.7	33.8	6.0	18.5	10.6	28.9
w/ UBP	5.5	24.7	27.5	55.5	14.7	42.2	5.5	20.5	13.3	35.7
TSconv	15.0	34.8	28.8	64.0	21.5	50.0	12.5	35.0	19.4	45.9
w/ UBP	15.5	41.2	39.5	69.0	28.2	57.2	17.8	41.2	25.2	52.2
EEGProject	12.2	31.2	32.5	64.0	16.8	47.0	13.0	31.5	18.6	43.4
w/ UBP	14.8	37.5	39.0	72.5	24.5	56.2	18.0	42.8	24.1	52.3

Table 18. Top-1 and Top-5 Accuracy (%) on THINGS-MEG with CLIP **ViT-H-14** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Avg	
Backbone	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
ShallowNet	7.5	26.0	24.5	50.5	14.5	43.8	9.0	25.5	13.9	36.4
w/ UBP	11.0	34.5	28.7	55.8	20.0	53.7	11.7	26.2	17.9	42.6
DeepNet	5.3	18.3	16.0	39.2	11.0	30.0	5.0	17.2	9.3	26.2
w/ UBP	4.3	19.5	18.0	41.8	12.7	33.2	5.2	17.0	10.1	27.9
EEGNet	5.7	18.3	16.7	45.0	8.2	31.2	4.5	15.7	8.8	27.6
w/ UBP	7.5	25.0	23.2	53.7	16.0	41.5	5.5	19.2	13.1	34.9
TSconv	12.0	31.5	30.0	58.7	20.3	51.5	13.0	34.8	18.8	44.1
w/ UBP	14.5	39.5	39.0	70.0	30.5	58.0	13.7	38.0	24.4	51.4
EEGProject	11.0	28.0	30.7	62.3	22.7	44.8	10.8	28.7	18.8	40.9
w/ UBP	15.5	37.3	41.0	74.8	24.8	55.3	15.3	34.5	24.1	50.4

Table 19. Top-1 and Top-5 Accuracy (%) on THINGS-MEG with CLIP **ViT-g-14** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Avg	
Backbone	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
ShallowNet	9.5	25.7	17.5	47.0	17.3	42.7	8.5	27.3	13.2	35.7
w/ UBP	10.5	28.2	25.5	51.0	19.2	46.5	8.3	27.5	15.9	38.3
DeepNet	5.0	16.0	11.5	33.7	9.8	29.7	2.5	16.0	7.2	23.9
w/ UBP	4.3	17.2	14.2	42.5	9.0	33.8	6.2	21.2	8.4	28.7
EEGNet	6.5	20.3	17.0	46.3	12.0	32.5	4.5	20.3	10.0	29.8
w/ UBP	6.2	24.0	23.3	53.8	13.7	37.5	5.5	20.2	12.2	33.9
TSconv	10.8	33.8	26.0	59.3	18.2	48.5	10.0	29.3	16.2	42.7
w/ UBP	11.5	34.7	32.0	65.0	24.2	56.5	12.5	37.3	20.1	48.4
EEGProject	11.0	28.7	27.5	59.5	19.5	44.3	10.7	33.2	17.2	41.4
w/ UBP	13.5	40.0	35.8	69.8	23.8	55.0	13.5	38.3	21.6	50.8

Table 20. Top-1 and Top-5 Accuracy (%) on THINGS-MEG with CLIP **ViT-bigG-14** With/Without UBP.

	Subject 1		Subject 2		Subject 3		Subject 4		Avg	
Backbone	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
ShallowNet	9.0	25.8	18.5	47.5	16.3	42.2	8.5	24.0	13.1	34.9
w/ UBP	6.5	25.5	25.5	55.8	25.5	51.2	9.8	28.7	16.8	40.3
DeepNet	4.2	18.2	13.0	36.0	11.3	29.0	5.5	16.5	8.5	24.9
w/ UBP	4.7	16.3	16.0	43.0	12.0	35.3	7.0	20.2	9.9	28.7
EEGNet	5.7	23.0	19.0	49.2	13.3	38.2	5.7	17.5	10.9	32.0
w/ UBP	8.2	28.5	23.7	61.0	16.3	50.2	4.5	22.0	13.2	40.4
TSconv	11.2	34.5	28.5	59.5	23.7	49.8	11.7	32.0	18.8	43.9
w/ UBP	13.2	40.5	36.5	68.5	28.7	65.7	13.0	38.5	22.9	53.3
EEGProject	9.0	30.0	32.8	59.5	20.3	46.5	8.7	29.5	17.7	41.4
w/ UBP	13.5	33.3	40.3	70.0	26.7	63.2	18.5	36.8	24.7	50.8

Table 21. Details of different EEG encoders with Emb dimension of 1024

Brain Encoder	**Params**
ShallowNet	2.56 M
DeepNet	2.76 M
EEGNet	2.34 M
TSConv	2.56 M
EEGProject	5.40 M

Table 22. Details of different Vision encoders

Vision Encoder	**Params**	**Emb dim**
RN50	38.32 M	1024
RN101	56.26 M	512
ViT-B-16	86.19 M	512
ViT-B-32	87.85 M	512
ViT-L-14	303.97 M	768
ViT-H-14	632.08 M	1024
ViT-g-14	1012.65 M	1024
ViT-bigG-14	1844.91 M	1280