UMind: A Unified Multitask Network for Zero-Shot M/EEG Visual Decoding

Chengjian Xu[a,b], Yonghao Song[c], Zelin Liao[d], Haochuan Zhang[d], Qiong Wang[e] and Qingqing Zheng[f,*]

[a]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[b]University of Chinese Academy of Sciences, China
[c]Department of Biomedical Engineering, Tsinghua University, China
[d]School of Automation, Guangdong University of Technology, China
[e]Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, China
[f] Artificial Intelligence Research Institute, Shenzhen University of Advanced Technology, China

ARTICLE INFO

Keywords:
Brain-computer interfaces (BCIs)
Visual stimulus decoding
Electroencephalography(EEG)
Magnetoencephalography (MEG)
Multitask learning
Multimodal alignment

ABSTRACT

Decoding visual information from time-resolved brain recordings, such as EEG and MEG, plays a pivotal role in real-time brain-computer interfaces. However, existing approaches primarily focus on direct brain-image feature alignment and are limited to single-task frameworks or task-specific models. In this paper, we propose a **U**nified **Mult**I**t**ask **N**etwork for zero-shot M/EEG visual **D**ecoding (referred to **UMind**), including visual stimulus retrieval, classification, and reconstruction, where multiple tasks mutually enhance each other. Our method learns robust neural-visual and semantic representations through multimodal alignment with both image and text modalities. The integration of both coarse and fine-grained texts enhances the extraction of these neural representations, enabling more detailed semantic and visual decoding. These representations then serve as dual conditional inputs to a pre-trained diffusion model, guiding visual reconstruction from both visual and semantic perspectives. Extensive evaluations on MEG and EEG datasets demonstrate the effectiveness, robustness, and biological plausibility of our approach in capturing spatiotemporal neural dynamics. Our approach sets a multitask pipeline for brain visual decoding, highlighting the synergy of semantic information in visual feature extraction.

1. Introduction

In the realm of brain-computer interfaces (BCIs), visual information decoding from neural signals is of paramount importance in understanding the intricate processes of visual perception and cognition (Miyawaki et al., 2008). It forms the foundation for advancing our comprehension of how the brain interprets and processes visual stimuli, including tasks such as *visual classification*, *retrieval*, and *reconstruction*. Visual decoding has wide-ranging applications, offering significant potential to provide alternative communication and control mechanisms, thereby enhancing the autonomy and capabilities of individuals with severe motor impairments (Kay et al., 2008). As non-invasive brain imaging technologies, electroencephalography (EEG) and magnetoencephalography (MEG) have shown great potential in the field of visual stimulus decoding (Song et al., 2024; Wang et al., 2025; Borra et al., 2025).

It is noteworthy that learning robust visual representations from M/EEG signals for visual stimulus decoding and reconstruction has become a prominent research focus (Spampinato et al., 2017a; Jiao et al., 2019; Li et al., 2024). Although numerous studies have focused on decoding M/EEG signals, most existing methods are designed for single-task scenarios, such as visual stimulus reconstruction or classification. For instance, Song *et al.* (Song et al., 2024) introduced a simple but efficient contrastive learning framework NICE, which aligns EEG and image representations for decoding images from EEG signals. The NICE demonstrated the feasibility of feature alignment and its biological plausibility. Similarly, Benchetrit *et al.* (Benchetrit et al., 2024) firstly aligns MEG features with CLIP image embeddings for retrieval and reconstruction of real-time visual stimuli. However, these methods remain confined to single-task objectives, lacking a unified framework to mutually reinforce multiple tasks (e.g., retrieval, classification, and reconstruction). Furthermore, they primarily focus on aligning M/EEG signals with visual embeddings, while neglecting the integration of textual semantic information, thereby limiting their ability to utilize rich contextual semantics for accurate cross-modal understanding and inference.

Some studies have attempted to design multitask frameworks for zero-shot visual stimulus decoding. A representative example is the work by Li et al. (Li et al., 2024), which proposes a novel brain decoding framework that enables zero-shot visual stimulus classification, retrieval, and reconstruction using EEG and MEG data. The method achieves results comparable to visual stimulus decoding using fMRI data with a customized M/EEG encoder and a two-stage M/EEG-to-image generation strategy. Although this work presents a multitask model, its underlying implementations still rely on separate M/EEG encoders for each task, thereby preventing cross-task knowledge sharing and mutual reinforcement. Furthermore, while the framework incorporates coarse-grained textual categories (e.g., "cat" or "raspberry") for visual stimulus reconstruction, it fails to perform multi-

*Corresponding author.
zhengqingqing@suat-sz.edu.cn (Qingqing Zheng)

modal alignment between M/EEG signals, images, and text, resulting in generated images or retrieval outcomes with compromised semantic consistency and visual fidelity. Most critically, it fails to utilize fine-grained textual descriptions (e.g., "the aircraft carrier is sailing in the ocean"), which fundamentally constrains their applicability in complex scenarios requiring high-precision representations, such as fine-grained image generation or semantically ambiguous stimulus retrieval.

To address the above challenges, we propose **UMind**, a unified multitask framework for zero-shot visual stimulus retrieval, classification, and reconstruction. This framework learns robust neural visual/semantic representations through cross-modal alignment with visual stimuli and textural semantic information. First, M/EEG signals, images, and text pass through their respective encoders to extract neural features, visual embeddings, and semantic embeddings. For M/EEG-image alignment, it maps neural features to visual embeddings, preserving spatial-temporal dynamics of brain responses. For M/EEG-text alignment, we incorporate a dual-text alignment strategy comprising both *Coarse-grained text* and *Fine-grained text* and align the neural features with semantic embeddings using a text projector. The coarse-grained text contains only category-level labels that provide categorical priors but lacks fine visual details (e.g., color, shape, orientation). In contrast, fine-grained text, generated by pre-trained large multimodal models (LMMs), offers rich perceptual descriptions but may contain ambiguous categorical information. Therefore, coarse-grained and fine-grained text can complement each other, enabling the model to learn more effective neural semantic representations.

Subsequently, to align the M/EEG signals with images and text, we introduce separate image and text projectors following the M/EEG encoder. These projectors extract neural visual and semantic representations, which are then applied to the visual stimulus retrieval and classification tasks. For visual stimulus reconstruction, these neural visual and semantic representations are then mapped to CLIP image embeddings and prompt embeddings, derived from the coarse- and fine-grained text, respectively. Specifically, we employ a diffusion prior (Ramesh et al., 2022) to map the neural visual representations to CLIP image embeddings. Inspired by the learned queries in BLIP2 (Li et al., 2023), we leverage the Q-Former to map the neural semantic representations to the fused prompt embeddings, derived from both types of textural information. Finally, these mapped neural embeddings serve as dual conditions for the pre-trained diffusion model, which guides the image reconstruction process from visual and semantic perspectives.

The main contributions of this paper are as follows:

- We present a unified multitask learning framework that synergistically integrates zero-shot M/EEG-based visual stimulus retrieval, classification, and reconstruction through joint optimization. This multitask framework outperforms conventional single-task approaches, establishing a new paradigm for neural signal processing with mutually reinforcing feature learning.
- We propose a novel multimodal alignment strategy that integrates M/EEG, images, and text to simultaneously extract neural visual and semantic representations. The proposed dual-granularity text integration facilitates multimodal alignment, which not only enables the extraction of neural semantic representations but also enhances the extraction of neural visual representations.
- We separately extract neural visual representations and neural semantic representations and use them as dual conditional inputs to the diffusion model. These representations guide the image generation from both visual and semantic perspectives, ensuring a more comprehensive and accurate reconstruction.
- We conducted extensive quantitative and qualitative validation experiments on two datasets, covering both EEG and MEG modalities. The experimental results demonstrate the superior performance of our method.

The remainder of this paper is structured as follows. We provide a brief overview of the related works in Section 2. Section 3 presents a detailed description of our proposed method. The experiments and results are illustrated in Section 4. A careful discussion is presented in Section 5. Finally, we reach a conclusion in Section 6.

2. Related Works

While early visual decoding predominantly relied on fMRI, recent advances extend to M/EEG paradigms. Therefore, we review related works on fMRI-based and M/EEG-based visual decoding methods

2.1. Visual Decoding from fMRI Data

In earlier studies, researchers employed linear regression to map fMRI data to handcrafted image features or to image features extracted using pre-trained models for visual decoding(Kay et al., 2008; Nishimoto et al., 2011; Wen et al., 2018). With the development of generative adversarial nets (GAN) (Goodfellow et al., 2014), some studies (Ren et al., 2021; Ozcelik et al., 2022; Lin et al., 2022) used the aligned fMRI embeddings as conditional inputs to conditional GAN (Casanova et al., 2021; Karras et al., 2020) for image reconstruction. More recently, diffusion models (Rombach et al., 2022) have emerged as highly effective and powerful tools for high-quality image generation. Numerous studies have leveraged pre-trained diffusion models for fMRI visual stimulus reconstruction, resulting in high-resolution images with exceptional fidelity. Takagi *et al.* (Takagi & Nishimoto, 2023) employed Stable Diffusion for visual stimulus reconstruction with fMRI signals from different visual cortex. Chen *et al.* (Chen et al., 2023) developed a two-stage framework MinD-Vis which first uses mask signal modeling for pre-training to learn fMRI representations and then finetunes a latent diffusion model through

double conditioning for image reconstruction. Ozcelik *et al.* (Ozcelik & VanRullen, 2023) used ridge regression models to map fMRI signals separately to VAE latent variables, CLIP image embeddings, and CLIP text embeddings and then employed a versatile diffusion model (Xu et al., 2023) for image reconstruction. Scotti *et al.* (Scotti et al., 2024a) proposed a framework called MindEye, which consists of two parallel submodules for visual stimulus retrieval and reconstruction. Additionally, the model integrates a high-level semantic pipeline and a low-level perceptual pipeline for image generation. Subsequently, Scotti *et al.* (Scotti et al., 2024b) presented an enhanced model MindEye2 which achieves competitive decoding performance with just one hour of data comparable to using a subject's full dataset through multi-subject pre-training.

2.2. Visual Decoding from M/EEG Data

Due to the high temporal resolution and convenience of EEG signals, researchers have attempted to use EEG signals for real-time visual decoding. Spampinato *et al.* proposed an EEG dataset (Spampinato et al., 2017b) for visual object analysis and conducted a series of works (Palazzo et al., 2017; Kavasidis et al., 2017; Tirupattur et al., 2018; Jiang et al., 2020), including EEG visual stimulus decoding and reconstruction using VAE and GAN on this dataset. However, Li *et al.* (Li et al., 2020) pointed out a flaw in the dataset, namely that all stimuli of a given class are presented together. This results in visual stimulus decoding relying on block-level temporal correlations which are present in all EEG data, rather than the visual information contained in the EEG signals.

Recently, with the release of a new large-scale and rigorous Things-EEG dataset (Gifford et al., 2022), researchers are now able to perform zero-shot visual stimulus decoding and reconstruction on this dataset. Song *et al.* (Song et al., 2024) designed an EEG-image contrastive learning framework, NICE, along with two plug-and-play modules that can capture spatial correlations for EEG visual decoding. NICE++ (Song et al., 2025) further incorporated language guidance to capture the semantic information embedded within EEG signals. Du *et al.* (Du et al., 2023) proposed a multimodal framework BraVL for visual neural representation learning, which utilizes mutual information regularization to align brain signals, images, and text. However, BraVL is primarily designed for fMRI data and is limited to the visual stimulus classification task. Wei *et al.* (Wei et al., 2024) employed contrastive learning and bidirectional cycle consistency to align the feature distributions of EEG and images for visual classification and reconstruction. Based on MEG, Benchetrit*et al.* (Benchetrit et al., 2024) proposed a real-time visual stimulus retrieval and reconstruction framework. Fu *et al.* (Fu et al., 2024) developed a novel EEG image reconstruction framework BrainVis which aligns EEG time-frequency embeddings with the interpolation of both coarse-grained and fine-grained text embeddings. The aligned EEG embeddings serve as conditions of cascaded diffusion models for image reconstruction. Li *et al.* (Li et al., 2024) presented a novel brain decoding framework that can simultaneously perform zero-shot visual stimulus retrieval, classification, and reconstruction tasks based on M/EEG data. Actually, the work trains a separate M/EEG encoder for each of the three tasks, rather than employing a unified multitask framework. In conclusion, these M/EEG-based visual decoding approaches are limited to single or dual-task frameworks and do not leverage text for multimodal alignment to simultaneously extract neural visual representations and neural semantic representations.

3. Methods

3.1. Overall Architecture and Problem Definition

We propose **UMind**, a unified multitask framework for zero-shot visual decoding from M/EEG signals, which simultaneously performs visual stimulus retrieval, classification, and reconstruction. As illustrated in Fig. 1, our architecture comprises three main submodules: a multimodal alignment module, a visual stimulus retrieval and classification module, and a dual conditioned diffusion reconstruction module. In the multimodal alignment module, it disentangles the neurocognitive representations by aligning neural signals $\mathbf{X}_b$ with visual stimuli $\mathbf{X}_v$ and textual information. For textual information, we incorporate dual-textual modality data, including coarse-grained texts ($\mathbf{X}_c$) containing only category information and fine-grained descriptions ($\mathbf{X}_t$) of images generated by a pre-trained image-to-text generation model. This allows our model to extract both neural visual representations and neural semantic representations, which can be used separately for visual stimulus retrieval and classification. In the reconstruction module, the obtained neural visual representations and neural semantic representations are mapped to CLIP image embeddings and prompt embeddings derived from coarse- and fine-grained text, respectively. These mapped embeddings guide the generation of realistic and credible images from both visual and semantic perspectives.

We denote the training set $\mathcal{D}_{train}$ as $(\boldsymbol{X}_b, \boldsymbol{X}_v, \boldsymbol{X}_c, \boldsymbol{X}_t, \boldsymbol{Y}) = \left\{\left(\boldsymbol{x}_b^i, \boldsymbol{x}_v^i, \boldsymbol{x}_c^i, \boldsymbol{x}_t^i, \boldsymbol{y}^i\right)\right\}_{i=1}^{N_{train}}$, where N_{train} is the sample size, $\boldsymbol{x}_b^i \in \mathbb{R}^{C\times T}$ is the i^{th} EEG or MEG trial in training data with C electrode channels and T time points, $\boldsymbol{x}_v^i$ denotes the corresponding i^{th} visual stimulus images, $\boldsymbol{x}_c^i$ denotes the coarse-grained texts containing only the category label, $\boldsymbol{x}_t^i$ denotes the detailed descriptions of the image and $\boldsymbol{y}^i$ denotes the corresponding one-hot category label. Likewise, the test set $\mathcal{D}_{test}$ is defined as $(\boldsymbol{X}_b^{test}, \boldsymbol{X}_v^{test}, \boldsymbol{X}_c^{test}, \boldsymbol{X}_t^{test}, \boldsymbol{Y}^{test})$. The labels of test data are with no overlap with the labels in the training data, namely, $\boldsymbol{Y} \cap \boldsymbol{Y}^{test} = \emptyset$. Therefore, the tasks of visual stimulus retrieval, classification, and reconstruction are zero-shot. The objective is to pre-train a M/EEG encoder using the multimodal training data $\mathcal{D}_{train}$ that includes M/EEG, images, and text, enabling the pre-trained M/EEG encoder to perform zero-shot visual stimulus retrieval, classification, and reconstruction excellently on the test set $\mathcal{D}_{test}$.

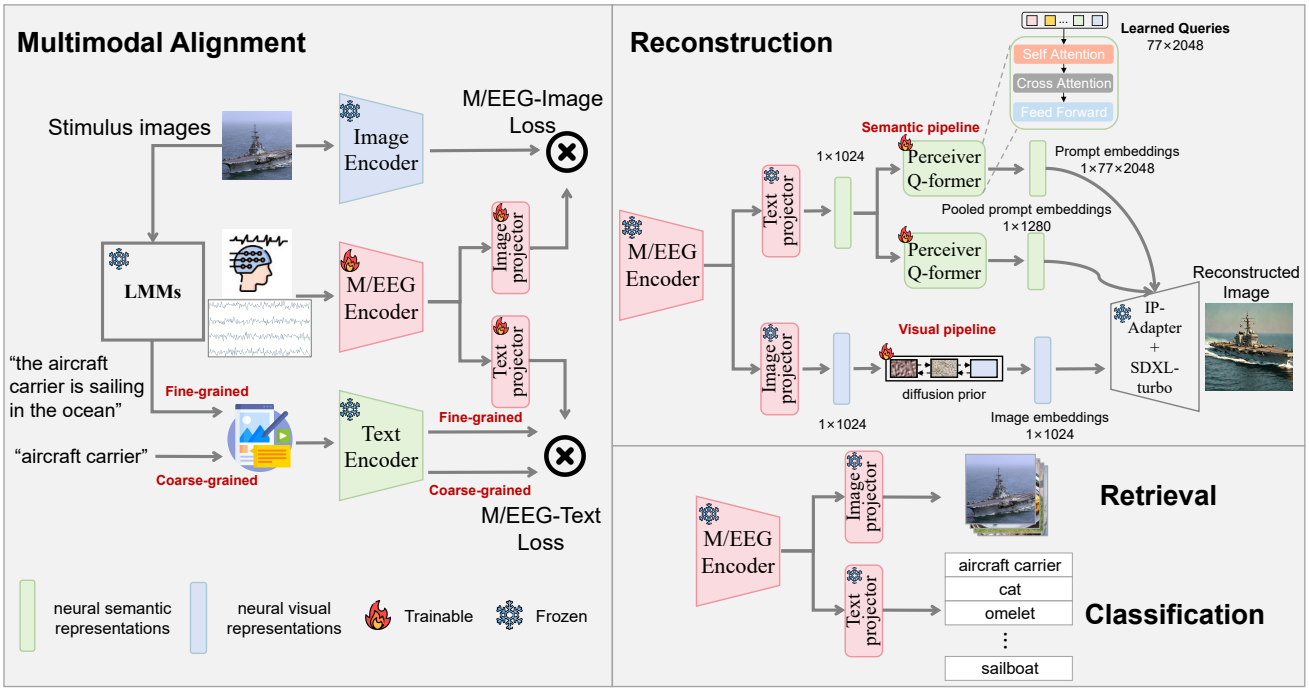

Figure 1: The proposed UMind framework enables zero-shot visual decoding from M/EEG signals, which simultaneously performs visual stimulus retrieval, classification, and reconstruction. It comprises three key components: a multimodal alignment module, a visual stimulus retrieval and classification module, and a dual conditioned diffusion reconstruction module.

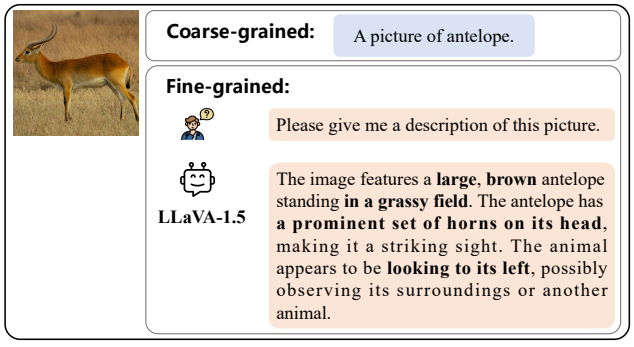

Figure 2: Comparison between fine-grained text generated by LLaVA-1.5 7B and coarse-grained text.

3.2. Multimodal Alignment

To extract both visual and semantic representations from neural signals (M/EEG), we employ a multimodal framework that bridges neural signals with paired images and text. This framework integrates a M/EEG encoder $f_b(\cdot)$ with two parallel projectors: an image projector $p_v(\cdot)$ and a text projector $p_t(\cdot)$. The neural signals are first encoded via $f_b(\cdot)$, then mapped through $p_v(\cdot)$ to derive neural-visual representations $\hat{\boldsymbol{z}}_v^i$ ($\hat{\boldsymbol{z}}_v^i = p_v(f_b(\boldsymbol{x}_b^i))$), which align with the CLIP image embeddings $\boldsymbol{z}_v^i$ from a frozen pre-trained image encoder. Simultaneously, the neural-semantic representations generated by $\hat{\boldsymbol{z}}_s^i$ ($\hat{\boldsymbol{z}}_s^i = p_t(f_b(\boldsymbol{x}_b^i))$) are aligned with the CLIP text embeddings extracted by a frozen pre-trained text encoder.

In aligning M/EEG with textual information, we employ dual-grained supervision: coarse-grained labels and fine-grained captions, as illustrated in Fig. 2. We apply LLaVA-1.5 (Liu et al., 2024) to generate fine-grained captions (enriched with spatial, chromatic, and contextual details). By combining fine-grained and coarse-grained text, we can extract rich semantic information with more detailed descriptions, while also enhancing the extraction of visual representations. We denote the coarse- and fine-grained image embeddings as $\boldsymbol{z}_c^i$ and $\boldsymbol{z}_t^i$, respectively. The neural-visual and neural-semantic representations are aligned with their multimodal counterparts using contrastive learning loss with

$$\mathcal{L}_{CLIP_V} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(s(\hat{\boldsymbol{z}}_v^i, \boldsymbol{z}_v^i)/\tau)}{\sum_{j=1}^{B}\exp(s(\hat{\boldsymbol{z}}_v^i, \boldsymbol{z}_v^j)/\tau)} \tag{1}$$

$$\mathcal{L}_{CLIP_{T1}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(s(\hat{\boldsymbol{z}}_s^i, \boldsymbol{z}_c^i)/\tau)}{\sum_{j=1}^{B}\exp(s(\hat{\boldsymbol{z}}_s^i, \boldsymbol{z}_c^j)/\tau)} \tag{2}$$

$$\mathcal{L}_{CLIP_{T2}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(s(\hat{\boldsymbol{z}}_s^i, \boldsymbol{z}_t^i)/\tau)}{\sum_{j=1}^{B}\exp(s(\hat{\boldsymbol{z}}_s^i, \boldsymbol{z}_t^j)/\tau)} \tag{3}$$

$$\mathcal{L}_{CLIP_T} = \left(\mathcal{L}_{CLIP_{T1}} + \mathcal{L}_{CLIP_{T2}}\right)/2 \tag{4}$$

where $\mathcal{L}_{CLIP_V}$ and $\mathcal{L}_{CLIP_T}$ are the contrastive loss between neural representations with image and text embeddings, respectively. s denotes cosine similarity, τ is the temperature parameter and B is the batch size.

Additionally, we also compute the Mean Squared Error (MSE) loss between the neural representations and the image and text embeddings:

$$\mathcal{L}_{MSE_V} = \frac{1}{Bd}\sum_{i=1}^{B}\|\boldsymbol{z}_v^i - \hat{\boldsymbol{z}}_v^i\|_2^2 \tag{5}$$

$$\mathcal{L}_{MSE_T} = \frac{1}{Bd}\sum_{i=1}^{B}\left(\|\boldsymbol{z}_c^i - \hat{\boldsymbol{z}}_s^i\|_2^2 + \|\boldsymbol{z}_t^i - \hat{\boldsymbol{z}}_s^i\|_2^2\right)/2 \tag{6}$$

where d is the dimension of representations.

During training, both the CLIP image and text encoder are frozen. We optimize the M/EEG encoder, image projector, and text projector by combining the contrastive loss and MSE loss. The overall loss function is:

$$\mathcal{L}_{all} = \alpha(\mathcal{L}_{CLIP_V} + \beta\mathcal{L}_{MSE_V}) + (1-\alpha)(\mathcal{L}_{CLIP_T} + \beta\mathcal{L}_{MSE_T}) \tag{7}$$

where α and β are the hyperparameters that control the balance between the M/EEG-image alignment loss and the M/EEG-text alignment loss, and between the contrastive loss and the MSE loss, respectively.

3.3. Visual Stimulus Retrieval and Classification

In the test phase, after multimodal alignment, we employ cosine similarity to match the extracted neural-visual and semantic representations with image and category templates, facilitating zero-shot visual stimulus retrieval and classification. These templates are constructed using images and coarse-level label texts that do not overlap with the training samples. Each image and its corresponding text are processed separately through pre-trained CLIP image and text encoders to build the templates.

3.4. Dual Guidance for Image Reconstruction

Through the multimodal alignment pre-training, we can simultaneously extract neural-visual representations and neural-semantic representations, which jointly guide visual stimulus reconstruction. Inspired by (Li et al., 2024), we integrate the pre-trained IP-Adapter (Ye et al., 2023) with the SDXL-Turbo (Sauer et al., 2025) to enable dual-modality conditioned image generation. SDXL-Turbo utilizes Adversarial Diffusion Distillation (ADD) to generate high-quality images using only 1–4 sampling steps. The addition of the IP-Adapter enables both text-to-image generation and image prompt capabilities, allowing for guiding image reconstruction from both visual and semantic perspectives using the extracted neural representations.

Visual guidance: While contrastive learning aligns neural-visual representations with CLIP image embeddings, it still lacks explicit spatial reconstruction (Scotti et al., 2024a). Therefore, we train a diffusion prior to map the neural visual representations into the CLIP image space. This prior enables the pre-trained diffusion model to perform image reconstruction from the visual perspective.

Semantic guidance: Direct caption decoding from low-dimensional neural-semantic representations proves unstable due to information loss in alignment (Li et al., 2024). Inspired by BLIP-2 (Li et al., 2023), we project these neural-semantic representations to the CLIP text space using a Q-Former block with a set of learnable query vectors. Specifically, the learnable query vectors $\boldsymbol{z}_q$ first pass through a self-attention layer (Vaswani et al., 2017):

$$\boldsymbol{z}_q' = \text{SelfAttention}(\boldsymbol{z}_q) \tag{8}$$

Next, the learnable query embeddings $\boldsymbol{z}_q'$ attend to neural-semantic features $\hat{\boldsymbol{z}}_s^i$:

$$Q = W_Q \cdot \boldsymbol{z}_q', K = W_K \cdot \hat{\boldsymbol{z}}_s, V = W_V \cdot \hat{\boldsymbol{z}}_s \tag{9}$$

$$\boldsymbol{z}_q'' = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V \tag{10}$$

where W_Q, W_K and W_V are learnable projection matrices.

Finally, the learnable query embeddings $\boldsymbol{z}_q''$ pass through a Feed-Forward Network (FFN) to produce the outputs.

$$\boldsymbol{z}_q^{prompt} = \text{FFN}(\boldsymbol{z}_q'') \tag{11}$$

Additionally, SDXL-Turbo conditions the model on the pooled prompt embedding. Similarly, we utilize the Q-Former to map $\hat{\boldsymbol{z}}_s$ into pooled prompt embeddings, denoted as $\boldsymbol{z}_q^{pool\_prompt}$, which guide image reconstruction from the semantic perspective with the prompt embeddings together. In this way, the proposed dual-guidance approach enhances image reconstruction by leveraging both visual and semantic information.

4. Experiments

4.1. Datasets and Preprocessing

THINGS-EEG The THINGS-EEG dataset (Gifford et al., 2022) comprises EEG recordings from ten participants acquired via 63 electrodes at 1000 Hz during a rapid serial visual presentation (RSVP) task. Each of the 10 participants completed four identical experimental sessions, resulting in a total of 10 datasets. For each dataset, the training set consists of 1654 categories, with 10 images per category, repeated 4 times. The test set includes 200 categories, with one image per category, repeated 80 times. Each dataset contains a total of 82160 image trials.

THINGS-MEG The THINGS-MEG dataset (Hebart et al., 2023) consists of MEG data from 4 subjects with 12 MEG sessions, recorded using 271 channels at a 1200 Hz sampling rate. The image stimuli were presented for 500ms, followed by a variable fixation period of 1000±200ms. The training dataset has 1854 concepts × 12 images × 1 repetition, and the test dataset has 200 concepts × 1 image × 12 repetitions. To create the zero-shot task, we removed 200 test concepts from the training set. The MEG data were segmented into trials, spanning from 0 to 1000 ms after stimulus onset.

Table 1
Retrieval and classification accuracy (%) of different methods on THINGS-EEG dataset

Methods	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
	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
Retrieval accuracy																						
NICE	21.50	53.50	21.00	51.00	27.50	62.50	30.00	63.50	13.00	36.00	22.50	56.00	25.50	**59.50**	38.50	68.50	21.00	57.00	23.00	64.50	24.35	57.20
ATM	20.50	**58.00**	18.00	47.50	25.00	60.00	27.50	58.00	15.50	42.00	27.50	63.50	24.00	53.00	41.00	72.00	21.50	51.00	36.50	69.50	25.70	57.45
MB2C	23.67	56.33	22.67	50.50	26.33	60.17	34.83	67.00	21.33	**53.00**	31.00	62.33	25.00	54.83	39.00	69.33	27.50	59.33	33.17	70.83	28.45	60.37
UMind	**27.00**	56.00	**32.00**	**70.00**	**34.00**	**70.00**	**36.00**	**70.50**	**23.00**	50.50	**32.50**	**68.50**	**28.00**	59.00	**46.50**	**80.00**	**37.50**	**66.50**	**42.00**	**76.00**	**33.85**	**66.70**
Classification accuracy																						
BraVL	6.11	17.89	4.90	14.87	5.58	17.38	4.96	15.11	4.01	13.39	6.01	18.18	6.51	20.35	8.79	23.68	4.34	13.98	7.04	19.71	5.82	17.45
NICE	**9.00**	27.50	**10.50**	24.50	10.00	37.50	11.50	35.50	6.50	20.50	9.00	28.50	9.50	**31.00**	11.50	37.50	7.00	29.00	11.50	31.50	9.60	30.30
ATM	4.50	15.50	1.50	7.00	5.00	17.00	4.50	15.50	2.50	11.00	4.50	12.50	6.50	18.00	8.00	23.00	2.50	10.00	5.50	18.50	4.50	14.80
UMind	7.50	**31.00**	7.50	**32.00**	**15.50**	**37.50**	**17.00**	**38.50**	**8.00**	**22.50**	**10.50**	**37.50**	**11.50**	30.00	**16.00**	**46.00**	**10.00**	**32.50**	**16.50**	**36.50**	**12.00**	**34.40**

Table 2
Retrieval and classification accuracy (%) of different methods on THINGS-MEG dataset

Methods	Subject 1		Subject 2		Subject 3		Subject 4		Average	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
Retrieval accuracy										
NICE	12.00	36.50	23.50	61.00	18.00	50.00	14.00	38.00	16.87	**46.38**
ATM	11.50	32.00	29.00	65.50	24.00	48.50	9.00	30.50	18.38	44.13
UMind	7.50	30.00	38.50	69.50	23.00	49.50	8.50	30.00	**19.38**	44.75
Classification accuracy										
NICE	4.00	16.50	9.00	33.00	10.50	26.50	5.50	17.00	7.25	23.25
ATM	1.00	13.00	10.50	28.50	7.50	21.00	2.50	10.00	5.38	18.13
UMind	5.50	18.50	11.00	34.50	13.00	32.50	6.50	22.50	**9.00**	**27.00**

For preprocessing, the EEG is filtered between 0.1 Hz and 100 Hz. EEG data were sampled from 0 to 1000 ms after stimulus onset, baseline corrected, and downsampled to 250 Hz, followed by multivariate noise normalization (Guggenmos et al., 2018) on the training data. Similarly, MEG signals were filtered between 0.1 Hz and 100 Hz, downsampled from 1200 Hz to 200 Hz. To improve the signal-to-noise ratio, EEG and MEG trials were averaged across repetitions for each image.

4.2. Experiments Settings

Our method was implemented in Python 3.10 using the PyTorch framework, with all experiments conducted on an NVIDIA GeForce RTX 4090 GPU. Following (Song et al., 2024), 740 samples from the training set were randomly selected as a validation set for model optimization. The model was trained using the Adam optimizer with a learning rate of 2×10^{-4} for 100 epochs with a batch size of 256. In the contrastive loss, the temperature parameter τ is a learnable parameter and is initially set to 0.07. The hyperparameters α and β in Eq. (7) were set to 0.5 and 2, respectively.

4.3. Retrieval and Classification Performance

4.3.1. Quantitative Comparison Results

We evaluated our UMind model through zero-shot visual stimulus retrieval and classification tasks on the THINGS-EEG dataset, comparing it with several state-of-the-art approaches, including BraVL (Du et al., 2023), NICE (Song et al., 2024), MB2C (Wei et al., 2024), and ATM (Li et al., 2024). Wilcoxon Signed-Rank Test was employed to evaluate statistical significance. As shown in Table 1, our UMind achieves superior performance for both retrieval and classification tasks. In the retrieval task, our method achieves a

Figure 3: Visualization of top-5 retrieval examples for the retrieval task.

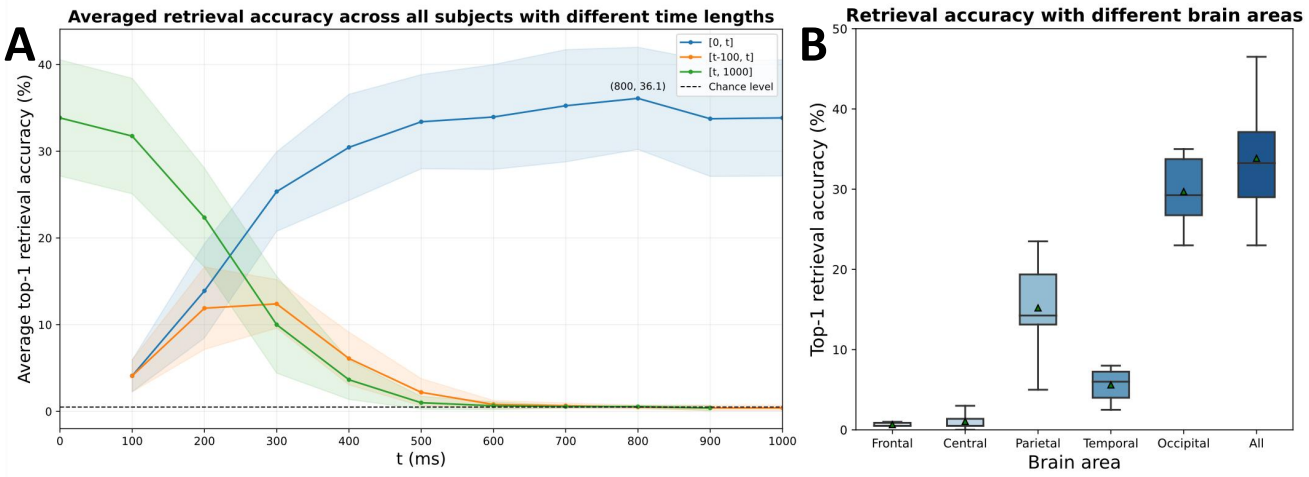

Figure 4: The results of temporal and spatial analysis on THINGS-EEG dataset. (A) The average top-1 retrieval accuracy of all subjects using different EEG time windows: [0, t], [t-100, t], and [t, 1000]. (B) Retrieval performance using electrode channels from different brain areas.

top-1 accuracy of 33.85% and a top-5 accuracy of 66.70%, significantly surpassing the chance levels of 0.5% and 2.5%, respectively. In the classification task, our method achieves a top-1 accuracy of 12.00% and an average top-5 accuracy of 34.40%, outperforming the existing best method by 2.40% ($p < 0.05$) and 4.10% ($p < 0.05$), respectively. These observations highlight the effectiveness of integrating text modality for both retrieval and classification.

To validate generalizability, we tested UMind on the THING-MEG dataset, as shown in Table 2. Our method achieved a top-1 accuracy of 19.38% for the retrieval task and 9.00% for the classification task, respectively. Similarly, our unified multitask framework outperforms individually trained models for each task on the THINGS-MEG dataset, demonstrating the generalizability of our approach.

4.3.2. Quanlitative Comparison Results

We randomly showcase several top-5 retrieval results from the retrieval task, as illustrated in Fig. 3. These results revealed consistent semantic patterns, for example, the retrieval results for “aircraft carrier" are all related to ships, while those for “lettuce" are all associated with vegetable-related items. This indicates that UMind effectively captures semantic details from dual-grained text and neural signals.

4.3.3. Temporal and Spatial Analysis

To investigate the temporal and spatial characteristics of visual stimulus retrieval, we conducted experiments using different EEG time windows and electrode channels from various brain regions. As shown in Fig. 4(A), three types of time windows were employed: expanding time windows [0,t], decreasing time windows [t,1000], and the sliding time windows [$t-100$,t], respectively. It can be observed that the results of the expanding window stabilize at approximately $t = 500ms$. For the sliding window, results exceeding the chance level are primarily observed between $100 - 500ms$. Therefore, the effective information for visual stimulus retrieval is predominantly concentrated within the $0 - 500ms$.

As shown in Fig. 4(B), the frontal and central electrodes contribute minimally to visual stimulus retrieval. In contrast, the occipital, parietal, and temporal electrodes contain varying degrees of effective visual information, ranked from high to low. When using only occipital electrodes, the average top-1 retrieval accuracy is 19.7%, with a decrease of 4.15% ($p > 0.05$) compared to using electrodes from all brain regions. Similar phenomena occur for the classification task.

4.4. Reconstruction Performance

4.4.1. Quantitative Comparison

We further conducted image generation experiments on the THINGS-EEG and THINGS-MEG datasets to investigate its proficiency in learning neural-visual and semantic information. We present the quantitative comparison between our model and recent state-of-the-art methods, including MB2C (Wei et al., 2024), B.D. (Benchetrit et al., 2024), and ATM (Li et al., 2024). We employ various high-level and low-level metrics to evaluate the similarity between the generated images and ground truth images. Low-level metrics include pixel-level correlation (PixCorr), structural similarity index (SSIM), and the second and fifth layers of the AlexNet model (AlexNet(2) and AlexNet(5)), while high-level metrics encompass Inception, CLIP, EfficientNet, and SwAV-ResNet50 (SwAV) to calculate the visual and semantic fidelity. For each category, we generate 10 images and rank them in descending order based on their similarity to the ground truth. Finally, we compute the average metrics across these batches as our final results.

The experimental results are presented in Table 3. It can be observed that our UMind generates images that perform slightly lower on low-level metrics, such as PixCorr and SSIM, compared to other models on the THINGS-EEG dataset. This is because methods like ATM leverage VAE latents to generate low-level images as the initial input for the diffusion model, thereby providing more low-level infor-

Table 3
The quantitative comparison of reconstruction performance between our model and other methods.

Datasets	Methods	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EfficientNet ↓	SwAV ↓
THINGS-EEG	MB2C	**0.188**	0.333	–	–	–	–	–	–
	ATM	0.164	**0.513**	0.583	0.603	0.558	0.577	0.962	0.630
	UMind	0.156	0.390	**0.725**	**0.839**	**0.744**	**0.798**	**0.879**	**0.560**
THINGS-MEG	B.D.	0.076	0.336	**0.736**	**0.826**	0.671	**0.767**	–	**0.584**
	ATM	0.104	0.340	0.613	0.672	0.619	0.603	–	0.651
	UMind	**0.107**	**0.360**	0.700	0.808	**0.679**	0.754	**0.915**	0.601

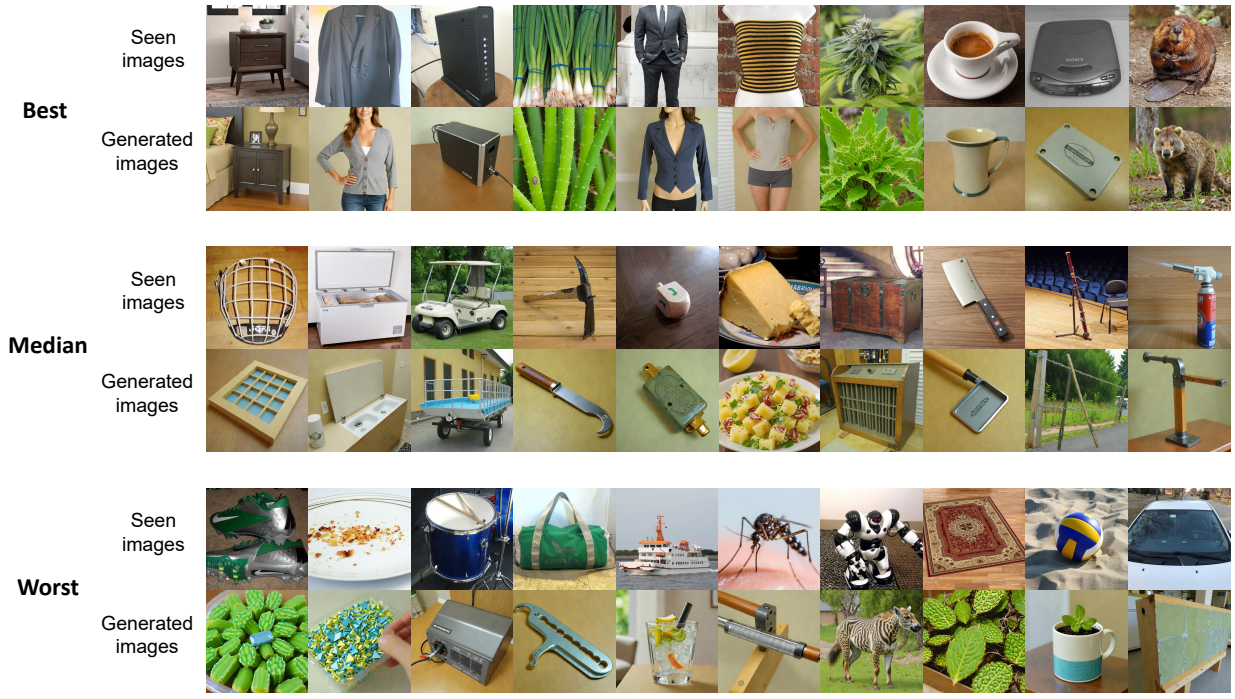

Figure 5: We compare the images reconstructed using UMind with the ground truth images, including those that show the best, median, and worst correspondence to the original stimulus images.

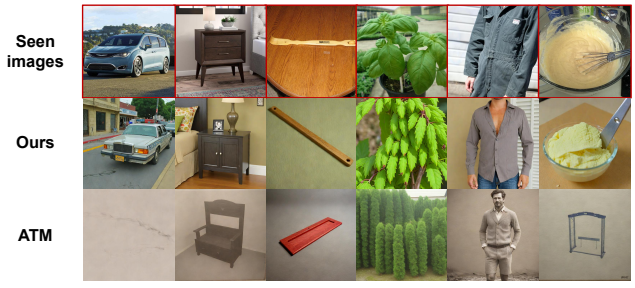

Figure 6: Comparison between the reconstructions using different methods.

mation. However, our method significantly outperforms on high-level metrics. This improvement is attributed to the integration of dual-grained text, along with the guidance from aligned neural-visual representations, enabling the model to learn more detailed visual and semantic information. On the THINGS-MEG dataset, our reconstruction performance surpasses that of ATM and achieves results comparable to B.D. framework.

4.4.2. Qualitative Comparison

We qualitatively compare the images reconstructed by our method with ground truth images. As shown in Fig. 5, the comparison includes the images that best, median, and worst correspond to the original stimulus images. It can be observed that the images generated by our method are visually consistent with the ground truth images, including color, shape, and orientation. Although some reconstructions do not effectively preserve the semantic information of the original images, they still capture detailed features such as similar colors and shapes.

As shown in Fig. 6, we also qualitatively compare the images reconstructed by our method with those by ATM. The images reconstructed using our method achieve greater vi-

Table 4
Retrieval and classification accuracy (%) with different EEG encoder on THINGS-EEG dataset

Methods	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
	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
Retrieval accuracy																						
ShallowNet	14.50	38.50	15.00	35.00	4.00	19.50	14.00	34.50	15.00	42.50	19.50	41.00	9.00	28.50	19.00	52.00	14.50	39.00	24.00	54.50	14.85	38.50
DeepNet	12.00	34.00	16.50	40.00	17.50	45.00	21.00	51.00	11.50	38.00	18.00	46.50	16.00	41.00	24.50	56.50	18.00	44.00	25.00	52.00	18.00	44.80
EEGNet	18.50	46.00	29.50	60.50	27.50	55.00	28.50	63.00	13.00	41.00	32.00	68.50	25.00	59.00	35.00	70.50	26.00	60.00	34.50	69.50	26.95	59.30
Conformer	9.50	30.00	20.50	45.50	20.00	49.50	25.00	55.00	9.50	27.50	29.00	56.00	16.50	43.00	23.50	59.00	15.50	40.00	29.00	57.50	19.80	46.30
TSConv	25.00	54.50	19.00	53.50	28.50	65.50	25.50	64.00	23.50	55.50	22.50	55.50	23.00	59.50	47.50	78.50	25.00	64.50	36.00	71.00	27.55	63.20
ATM	27.00	56.00	32.00	70.00	34.00	70.00	36.00	70.50	23.00	50.50	32.50	68.50	28.00	59.00	46.50	80.00	37.50	66.50	42.00	76.00	**33.85**	**66.70**
Classification accuracy																						
ShallowNet	4.00	14.00	1.00	8.00	1.00	4.50	4.00	13.00	5.50	23.00	1.00	9.00	2.50	10.50	4.50	18.50	5.00	16.00	6.00	19.00	3.45	13.55
DeepNet	6.50	22.50	6.00	23.00	8.50	24.00	9.00	22.50	7.00	18.00	6.50	22.50	6.00	25.00	12.00	27.50	7.00	24.00	7.50	26.00	7.60	23.50
EEGNet	5.00	27.50	8.50	27.00	9.50	30.00	12.50	33.50	8.00	21.50	10.50	31.00	9.50	30.50	16.50	35.50	8.50	25.00	12.00	37.00	10.05	29.85
Conformer	4.50	17.50	8.00	21.50	6.00	28.50	10.50	30.00	5.00	15.00	8.50	25.00	5.50	24.50	9.50	23.50	7.00	22.00	11.00	30.50	7.55	23.80
TSConv	9.50	30.00	7.00	23.00	10.00	31.50	13.50	29.50	6.50	22.50	6.00	23.50	9.50	35.50	15.50	36.50	10.00	35.00	14.00	36.50	10.15	30.35
ATM	7.50	31.00	7.50	32.00	15.50	37.50	17.00	38.50	8.00	22.50	10.50	37.50	11.50	30.00	16.00	46.00	10.00	32.50	16.50	36.50	**12.00**	**34.40**

Table 5
Retrieval and classification accuracy (%) with different pre-trained CLIP on THINGS-EEG dataset

Methods	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
	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
Retrieval accuracy																						
ViT-B/16	29.50	61.50	27.00	55.00	30.50	66.00	39.50	68.50	21.00	50.00	35.50	67.00	30.00	64.00	49.00	82.00	30.00	67.00	33.50	69.50	32.55	65.05
ViT-L/14	28.50	56.00	26.00	58.50	37.00	74.50	42.00	74.00	23.50	54.50	33.00	69.50	27.00	71.50	43.00	74.50	38.00	67.00	40.00	73.00	33.80	67.30
ViT-H/14	27.00	56.00	32.00	70.00	34.00	70.00	36.00	70.50	23.00	50.50	32.50	68.50	28.00	59.00	46.50	80.00	37.50	66.50	42.00	76.00	**33.85**	**66.70**
ViT-G/14	25.50	51.50	33.00	65.50	23.00	52.50	31.50	68.50	20.00	48.50	25.50	58.00	28.00	60.00	48.00	77.50	18.00	46.00	39.50	68.50	29.20	59.65
Classification accuracy																						
ViT-B/16	11.00	34.00	8.50	34.50	13.50	40.00	14.00	40.50	10.50	25.00	10.50	34.00	11.00	35.00	16.50	47.50	15.00	35.50	13.00	37.00	12.35	36.30
ViT-L/14	7.50	24.00	6.50	26.00	10.00	38.50	12.50	40.50	8.00	25.00	8.50	34.50	9.00	35.00	16.00	39.50	11.00	30.50	11.50	40.00	10.05	33.35
ViT-H/14	7.50	31.00	7.50	32.00	15.50	37.50	17.00	38.50	8.00	22.50	10.50	37.50	11.50	30.00	16.00	46.00	10.00	32.50	16.50	36.50	**12.00**	**34.40**
ViT-G/14	8.00	23.50	7.50	29.00	9.50	23.50	13.50	42.00	8.00	28.00	11.50	31.50	11.00	32.00	15.00	37.50	9.00	22.50	11.00	36.50	10.40	30.60

Table 6
Effects of dual grained text on retrieval & classification (%)

coarse-grained	fine-grained	Retrieval		Classification	
		top-1	top-5	top-1	top-5
✗	✗	20.30	44.25	0.55	2.40
✓	✗	30.90	62.45	10.15	31.25
✗	✓	32.35	63.90	7.60	24.85
✓	✓	**33.85**	**66.70**	**12.00**	**34.40**

sual and semantic consistency with the original images. For example, in the second column of Fig. 6, our method accurately reconstructs the nightstand, along with details such as the bed beside it and the lamp on the nightstand. In contrast, the image reconstructed by ATM only resembles a wooden box similar to a nightstand.

4.5. Ablation Study

4.5.1. Effect of different M/EEG encoders

To investigate the impact of different encoders on the experimental results, we compare several representative M/EEG encoders, including EEGNet (Lawhern et al., 2018), DeepConvNet, ShallowNet (Schirrmeister et al., 2017), Conformer (Song et al., 2022), TSConv in NICE (Song et al., 2024), and ATM (Li et al., 2024). The comparison results of different encoders are shown in Table 4. On the THINGS-EEG dataset, the ATM encoder outperforms other methods in both retrieval and classification tasks. ATM achieves a top-1 accuracy of 33.85% in the retrieval task and 12.00% in the classification task, surpassing the best-performing TSConv by 6.30% ($p < 0.01$) and 1.85% ($p < 0.05$), respectively. Therefore, we selected ATM as our M/EEG encoder.

4.5.2. Effect of different image and text encoders

In the multimodal alignment module, we use the pre-trained image encoder and text encoder from Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) to extract image and text features. We systematically evaluate CLIP variants in different scales, including OpenCLIP ViT-B/16, ViT-L/14, ViT-H/14, and ViT-G/14 (Ilharco et al., 2021; Schuhmann et al., 2022). The performance comparison of different pre-trained CLIP models is presented in Table 5. It can be observed that all pre-trained CLIP models demonstrate comparable performance in both retrieval and classification tasks. Among them, OpenCLIP ViT-H/14 was chosen as our frozen image and text encoder because it consistently delivers strong results across both tasks.

Table 7
Effects of coarse-grained text and fine-grained text on reconstruction performance.

coarse-grained	fine-grained	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EfficientNet ↓	SwAV ↓
✗	✓	0.143	0.398	0.669	0.779	0.686	0.756	0.901	0.583
✓	✗	0.148	**0.399**	0.689	0.805	0.709	0.779	0.899	0.584
✓	✓	**0.156**	0.390	**0.725**	**0.839**	**0.744**	**0.798**	**0.879**	**0.560**

Table 8
Effect of loss function for retrieval and classification (%)

Loss	Retrieval		Classification	
	top-1	top-5	top-1	top-5
w/o $\mathcal{L}_{CLIP}$	0.5	2.5	0.5	2.5
w/o $\mathcal{L}_{MSE}$	33.15	64.50	10.75	31.10
overall	**33.85**	**66.70**	**12.00**	**34.40**

Table 9
Single task vs Multitask framework for retrieval & classification (%)

Retrieval	Classification	Retrieval		Classification	
		top-1	top-5	top-1	top-5
✓	✗	31.85	64.15	–	–
✗	✓	–	–	10.50	30.75
✓	✓	**33.85**	**66.70**	**12.00**	**34.40**

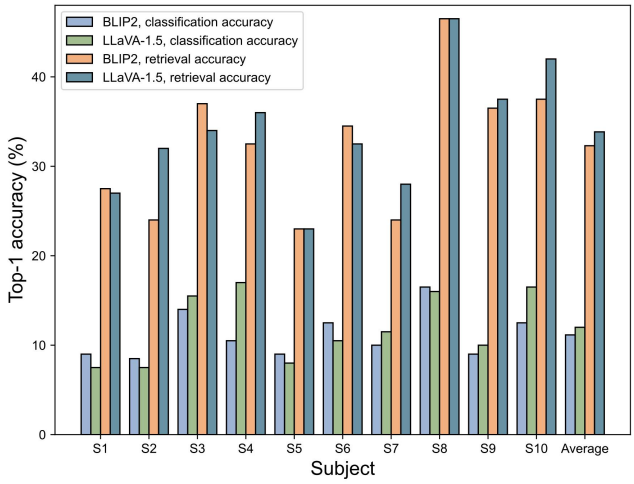

Figure 7: Comparison of results for visual stimulus retrieval and classification using fine-grained text generated by LLaVA-1.5 and BLIP2.

4.5.3. Effect of Coarse-grained and Fine-grained Text

We conducted ablation experiments on the THINGS-EEG dataset to investigate the effects of coarse-grained and fine-grained text. As presented in Table 6, without text modality data, the framework achieves top-1 classification accuracies of only 0.55%, indicating that images provide very limited semantic information for M/EEG visual stimulus classification. After incorporating coarse-grained and fine-grained text, the model's classification performance improves significantly, and retrieval accuracy is also notably enhanced. This demonstrates that introducing the text modality for multimodal alignment not only enables the extraction of detailed semantic information but also enhances the extraction of neural visual representations.

Table 7 also presented the effects of dual-grained text on visual stimulus reconstruction. The results indicate that relying on only one type of text granularity significantly reduces the quality of reconstructed images compared to using both granularity levels together. Specifically, the model achieves better reconstruction performance when both coarse and fine-grained text are employed, as their complementary information allows the model to capture more semantic details. Therefore, our model can learn semantically rich neural representations and reconstruct images that are semantically similar to the ground truth visual stimuli.

4.5.4. Effect of Different Loss Functions

Table 8 shows the effect of different loss functions. As observed, removing the MSE loss resulted in a decrease of 0.7% ($p > 0.05$) and 1.25% ($p < 0.05$) in top-1 accuracy for retrieval and classification, respectively. This indicates that incorporating an appropriately weighted MSE loss for the subsequent reconstruction task can also enhance the model's performance in retrieval and classification. Using only MSE loss is ineffective for multimodal alignment, whereas combining contrastive learning with MSE loss enhances retrieval and classification performance.

4.5.5. Effect of Multitask Framework

To investigate whether our multitask framework can facilitate mutual enhancement among different tasks, we trained separate single-task models using only image and text data for M/EEG visual retrieval and classification, respectively. As shown in Table 9, the results demonstrate that our unified multitask framework outperforms the single-task retrieval and classification frameworks by 2.00% ($p > 0.05$) and 1.50% ($p < 0.05$) in top-1 accuracy, respectively. This suggests that incorporating text for multimodal alignment not only enables the model to handle different tasks but also facilitates mutual enhancement between tasks, leading to more effective learning of neural visual and semantic representations.

4.5.6. *Effect of Different Image-to-Text Generation Models*

To assess the impact of fine-grained text generated by different pre-trained image-to-text models on experimental results, we conducted a comparative analysis using two models: LLaVA-1.5 (Liu et al., 2024) and BLIP2 (Li et al., 2023). The results are shown in Fig. 7. It can be observed that for both retrieval and classification tasks, the average top-1 accuracy of using fine-grained text generated by LLaVA-1.5 surpasses that of BLIP2. This is because the image captions generated by BLIP2 are often short and simplistic, such as "the antelope is brown." In contrast, the text generated by LLaVA-1.5 contains richer details about the image, including information on location, color, and direction, as illustrated in Fig. 2. Consequently, the fine-grained text generated by LLaVA-1.5 better facilitates the model in learning neural semantic representations with more detailed information.

5. Discussion

Current approaches for decoding visual stimuli primarily rely on contrastive learning to align brain signals with images, overlooking the critical role of the textual modality. Additionally, these models are often limited to addressing a single task or require training a separate encoder for each task. To address these limitations, we propose a unified multitask framework UMind capable of simultaneously performing zero-shot visual stimulus retrieval, classification, and reconstruction tasks. By leveraging both coarse-grained and fine-grained text for multimodal alignment, our framework facilitates the learning of neural-visual and neural-semantic representations enriched with finer details. This enhances the extraction of neural representations that are beneficial for visual stimulus retrieval and classification tasks. Furthermore, these extracted neural representations serve as dual conditioning inputs for the pre-trained diffusion model, guiding the generation of realistic and semantically meaningful images from both visual and semantic perspectives.

Nevertheless, there are some limitations in our work. By incorporating coarse-grained and fine-grained text, we have successfully learned more detailed neural semantic representations associated with visual stimuli, which are effective for classification and reconstruction tasks. However, decoding text corresponding to visual stimuli from these semantic representations remains a significant challenge. Additionally, although we have reconstructed images from M/EEG signals, the quality of the reconstructed images still falls short compared to fMRI-based works. Further research is needed to narrow this gap.

6. Conclusion

In this work, we propose **UMind**, a novel multitask multimodal alignment framework that simultaneously addresses zero-shot M/EEG-based visual stimulus retrieval, classification, and reconstruction tasks. By incorporating both coarse-grained and fine-grained text, we can not only extract more detailed neural semantic representations but also enhance the learning of neural visual representations. The extracted neural visual and semantic representations can be used for M/EEG visual stimulus retrieval and classification, respectively. Additionally, they serve as dual conditional inputs to a frozen diffusion model to guide image generation. Extensive quantitative and qualitative experiments demonstrate that our framework achieves state-of-the-art performance across two datasets, highlighting its effectiveness in learning robust visual and semantic representations for visual stimulus decoding.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Benchetrit, Y., Banville, H., & King, J.-R. (2024). Brain decoding: Toward real-time reconstruction of visual perception. URL: https://arxiv.org/abs/2310.19812. arXiv:2310.19812.

Borra, D., Magosso, E., & Ravanelli, M. (2025). A protocol for trustworthy eeg decoding with neural networks. *Neural Networks*, *182*, 106847.

Casanova, A., Careil, M., Verbeek, J., Drozdzal, M., & Romero Soriano, A. (2021). Instance-conditioned gan. In *Advances in Neural Information Processing Systems* (pp. 27517–27529). volume 34.

Chen, Z., Qing, J., Xiang, T., Yue, W. L., & Zhou, J. H. (2023). 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* (pp. 22710–22720).

Du, C., Fu, K., Li, J., & He, H. (2023). Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*, 10760–10777.

Fu, H., Shen, Z., Chin, J. J., & Wang, H. (2024). BrainVis: Exploring the Bridge between Brain and Visual Signals via Image Reconstruction. URL: https://arxiv.org/abs/2312.14871. arXiv:2312.14871.

Gifford, A. T., Dwivedi, K., Roig, G., & Cichy, R. M. (2022). A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, *264*, 119754.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*. volume 27.

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *Neuroimage*, *173*, 434–447.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., & Baker, C. I. (2023). THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, *12*, e82580.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., & Schmidt, L. (2021). Openclip. URL: https://doi.org/10.5281/zenodo.5143773. doi:10.5281/zenodo.5143773.

Jiang, J., Fares, A., & Zhong, S.-H. (2020). A brain-media deep framework towards seeing imaginations inside brains. *IEEE Transactions on Multimedia*, *23*, 1454–1465.

Jiao, Z., You, H., Yang, F., Li, X., Zhang, H., & Shen, D. (2019). Decoding EEG by Visual-guided Deep Neural Networks. In *IJCAI* (pp. 1387–1393). Macao volume 28.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8110–8119).

Kavasidis, I., Palazzo, S., Spampinato, C., Giordano, D., & Shah, M. (2017). *Brain2Image*: Converting Brain Signals into Images. In *Proceedings of the 25th ACM International Conference on Multimedia* (pp. 1809–1817). Mountain View California USA: ACM. doi:10.1145/3123266.3127907.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*, 352–355.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, *15*, 056013.

Li, D., Wei, C., Li, S., Zou, J., & Liu, Q. (2024). Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *Advances in Neural Information Processing Systems* (pp. 102822–102864). volume 37.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning* (pp. 19730–19742). PMLR.

Li, R., Johansen, J. S., Ahmed, H., Ilyevsky, T. V., Wilbur, R. B., Bharadwaj, H. M., & Siskind, J. M. (2020). The perils and pitfalls of block design for EEG classification experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*, 316–333.

Lin, S., Sprague, T., & Singh, A. K. (2022). Mind reader: Reconstructing complex images from brain activities. In *Advances in Neural Information Processing Systems* (pp. 29624–29636). volume 35.

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 26296–26306).

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H. C., Sadato, N., & Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, *60*, 915–929.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, *21*, 1641–1646.

Ozcelik, F., Choksi, B., Mozafari, M., Reddy, L., & VanRullen, R. (2022). Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.

Ozcelik, F., & VanRullen, R. (2023). Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, *13*, 15666.

Palazzo, S., Spampinato, C., Kavasidis, I., Giordano, D., & Shah, M. (2017). Generative adversarial networks conditioned by brain signals. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3410–3418).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748–8763). PMLR.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*, 3. arXiv:2204.06125.

Ren, Z., Li, J., Xue, X., Li, X., Yang, F., Jiao, Z., & Gao, X. (2021). Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, *228*, 117602.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695).

Sauer, A., Lorenz, D., Blattmann, A., & Rombach, R. (2025). Adversarial Diffusion Distillation. In *European Conference on Computer Vision* (pp. 87–103). volume 15144.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, *38*, 5391–5420. doi:10.1002/hbm.23730.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., & Wortsman, M. (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems* (pp. 25278–25294). volume 35.

Scotti, P., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Dempster, A., Verlinde, N., Yundler, E., Weisberg, D., & Norman, K. (2024a). Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Advances in Neural Information Processing Systems*. volume 36.

Scotti, P. S., Tripathy, M., Villanueva, C. K. T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K. A., & Abraham, T. M. (2024b). MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. URL: https://arxiv.org/abs/2403.11207. arXiv:2403.11207.

Song, Y., Liu, B., Li, X., Shi, N., Wang, Y., & Gao, X. (2024). Decoding Natural Images from EEG for Object Recognition. In *International Conference on Learning Representations*.

Song, Y., Wang, Y., He, H., & Gao, X. (2025). Recognizing natural images from EEG with language-guided contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, .

Song, Y., Zheng, Q., Liu, B., & Gao, X. (2022). EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 710–719.

Spampinato, C., Palazzo, S., Kavasidis, I., Giordano, D., Souly, N., & Shah, M. (2017a). Deep learning human mind for automated visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6809–6817).

Spampinato, C., Palazzo, S., Kavasidis, I., Giordano, D., Souly, N., & Shah, M. (2017b). Deep learning human mind for automated visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6809–6817).

Takagi, Y., & Nishimoto, S. (2023). 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* (pp. 14453–14463).

Tirupattur, P., Rawat, Y. S., Spampinato, C., & Shah, M. (2018). ThoughtViz: Visualizing Human Thoughts Using Generative Adversarial Network. In *Proceedings of the 26th ACM International Conference on Multimedia* (pp. 950–958). ACM. doi:10.1145/3240508.3240641.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc. volume 30.

Wang, J., Zhao, S., Luo, Z., Zhou, Y., Li, S., & Pan, G. (2025). Eegmamba: An eeg foundation model with mamba. *Neural Networks*, (p. 107816).

Wei, Y., Cao, L., Li, H., & Dong, Y. (2024). MB2C: Multimodal Bidirectional Cycle Consistency for Learning Robust Visual Neural Representations. In *ACM Multimedia 2024*.

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, *28*, 4136–4160.

Xu, X., Wang, Z., Zhang, G., Wang, K., & Shi, H. (2023). Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7754–7765).

Ye, H., Zhang, J., Liu, S., Han, X., & Yang, W. (2023). IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. doi:10.48550/arXiv.2308.06721. arXiv:2308.06721.