VISUAL-TO-M/EEG BRAIN ENCODING WITH TIME-FREQUENCY DUAL-PATH FINITE SCALAR QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual brain encoding predicts neural responses from visual stimuli, advancing our understanding of visual perception. It is also an indispensable step in visual prostheses, where predicted neural responses guide electrode stimulation patterns. Prior visual brain encoding research has primarily focused on fMRI, whose limited temporal resolution restricts its applicability. While recent work has shifted toward M/EEG-based brain encoding with diffusion models, these approaches treat M/EEG as images—neglecting their frequency-domain structure—and do not explicitly leverage real neural responses in the training objective. We propose a three-stage visual brain encoding framework. In Stage 1, we propose a novel VAE, called *BrainVAE*, with a time-frequency dual-path encoder: the input M/EEG signal is treated as a 2D image (channels $\times$ time) and processed by a Conv2d backbone to extract local spatial-temporal features, while simultaneously transformed by FFT to obtain power spectra. The Conv2d feature tokens and Frequency Prefix Tokens are fused by a Transformer via self-attention, incorporating explicit frequency-domain priors. Finite Scalar Quantization (FSQ) discretizes the latent space without codebook collapse or auxiliary losses. In Stage 2, InfoNCE contrastive learning aligns the quantized neural latent space with I-JEPA image embeddings. In Stage 3, a Transformer-based BrainTokenRegressor predicts FSQ token sequences from I-JEPA image embeddings via cross-attention, with losses computed directly against real neural responses in both token and signal spaces. Experiments on THINGS-EEG2 and THINGS-MEG demonstrate the effectiveness of our approach.

1 INTRODUCTION

Visual brain encoding aims to predict neural responses from the visual stimuli that evoke them, providing crucial insights into how sensory information is represented across neural populations (Naselaris et al., 2011). Beyond its scientific value, brain encoding plays a central role in visual prostheses, where a processing framework converts visual information into electrode stimulation patterns (Soltan et al., 2018). Visual prostheses hold the potential to restore a rudimentary form of vision to millions of people living with incurable blindness (Granley et al., 2022). Thus, brain encoding not only deepens our understanding of human visual perception, but also carries significant practical implications for the development of visual prostheses.

In recent years, neuroscience researchers have primarily focused on visual-to-fMRI brain encoding. Functional magnetic resonance imaging (fMRI) measures neural activity indirectly through blood-oxygen-level-dependent (BOLD) contrast, providing high spatial resolution. However, fMRI suffers from high acquisition cost and inherently limited temporal resolution, restricting its applicability in brain–computer interfaces (Li et al., 2024). In particular, recent studies such as MindSimulator (Bao et al., 2025) and SynBrain (Mai et al., 2026) rely on the Natural Scenes Dataset (NSD) (Allen et al., 2022), which preprocesses fMRI signals by employing single-trial beta weights derived from generalized linear models, along with region-of-interest (ROI) data specific to early and higher (ventral) visual regions. Consequently, the outputs produced by these fMRI-based brain encoding models are not raw neural responses but heavily processed statistical summaries.

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

Recently, Xu *et al.* (Xu et al., 2025) shifted the research focus toward M/EEG-based brain encoding. They observed that existing visual prostheses do not use real neural responses as labels to evaluate the biological similarity of predicted neural responses (Granley et al., 2022; 2023; Wang et al., 2024). To address this, Xu *et al.* (Xu et al., 2025) adopted THINGS-EEG2 (Gifford et al., 2022) and THINGS-MEG (Hebart et al., 2023), two multimodal datasets containing paired images and M/EEG recordings, to implement image-to-M/EEG brain encoding. By using these datasets, the mapping from images to neural responses can be directly learned with ground truth neural responses serving as supervised signals.

However, Xu *et al.* (Xu et al., 2025) employed a diffusion transformer (DiT) (Peebles & Xie, 2023) for M/EEG signal generation, effectively treating M/EEG as two-dimensional images. In practice, M/EEG signals are multi-channel time series that capture not only temporal dynamics but also oscillatory neural activity across different frequency bands, permitting analysis from both temporal and frequency domains (Kida et al., 2016; Keil et al., 2022; Morales & Bowers, 2022; Cohen, 2014). By treating M/EEG as images, this approach fails to exploit the frequency-domain structure of the signals.

Moreover, Xu *et al.* (Xu et al., 2025) rely on diffusion models (Song et al., 2020), whose training objective minimizes the mean squared error (MSE) between the predicted noise and the noise injected into the signal. Although real neural responses participate indirectly in training through the forward noising process, the loss function does not explicitly enforce point-wise correspondence between the model's predictions and the ground truth neural responses. Therefore, the issue raised by Xu *et al.* (Xu et al., 2025) themselves—that prior methods do not use real neural responses as labels to evaluate the biological similarity of predicted neural responses—remains fundamentally unresolved.

To address these limitations, we propose a three-stage visual-to-M/EEG brain encoding framework. Given the limitations of fMRI, we focus on M/EEG signals and validate our method on the EEG dataset THINGS-EEG2 and the MEG dataset THINGS-MEG. In Stage 1, we propose a novel VAE, called *BrainVAE*, with a time-frequency dual-path encoder: the input M/EEG signal is treated as a 2D image (channels $\times$ time) and processed by a Conv2d backbone to extract local spatial-temporal features, while simultaneously transformed by FFT to obtain power spectra. The Conv2d feature tokens and Frequency Prefix Tokens derived from the power spectra are then jointly fed into a Transformer (Vaswani et al., 2017), which captures long-range temporal dependencies beyond the Conv2d receptive field and fuses the two paths via self-attention, incorporating explicit frequency-domain priors without occupying latent dimensions. The encoder output is discretized via Finite Scalar Quantization (FSQ) (Mentzer et al., 2024), which avoids codebook collapse common in traditional vector quantization (VQ) and requires no auxiliary losses. In Stage 2, we fine-tune the BrainVAE with an InfoNCE contrastive loss (Oord et al., 2018) to align the quantized neural latent space with I-JEPA image embeddings, while preserving reconstruction capability. In Stage 3, we train a Transformer-based BrainTokenRegressor that predicts FSQ token sequences from I-JEPA image embeddings via cross-attention, with losses computed in both the token space and the signal space to directly leverage ground truth neural responses.

Our contributions are as follows:

- We propose a time-frequency dual-path encoder in which the input M/EEG signal is simultaneously processed by a Conv2d backbone and FFT; the resulting Conv2d feature tokens and Frequency Prefix Tokens are fused by a Transformer via self-attention, incorporating explicit frequency-domain priors without occupying latent dimensions.
- We adopt Finite Scalar Quantization (FSQ) to construct a discrete neural latent space that avoids codebook collapse, and design a three-stage training pipeline that progressively achieves signal reconstruction, cross-modal alignment, and visual-guided neural encoding.
- We train a Transformer-based BrainTokenRegressor whose losses are computed directly against ground truth neural responses in both token and signal spaces, thereby explicitly incorporating real neural responses as supervised signals throughout training.

2 Related Work

Visual Brain Encoding. Visual brain encoding aims to predict neural responses from visual stimuli (Naselaris et al., 2011). A typical visual brain encoding model first extracts visual stimulus

representations and then maps them to neural responses. Existing visual brain encoding studies have predominantly focused on fMRI. Early approaches employed hand-crafted visual features to predict voxel-wise BOLD responses (Kay et al., 2008). Later studies used deep neural networks to extract hierarchical features from visual stimuli and mapped these features to voxel-wise BOLD responses, revealing an increasing gradient of representational complexity from early to higher-level visual areas (Güçlü & Van Gerven, 2015).

Recent studies have extended visual brain encoding from deterministic prediction to generative fMRI synthesis. MindSimulator combines an fMRI autoencoder with a diffusion-based estimator to synthesize fMRI responses conditioned on visual stimuli (Bao et al., 2025). However, its deterministic autoencoder does not explicitly model the distribution of neural responses. SynBrain addresses this limitation by learning a probabilistic fMRI latent space with a variational model, thereby accounting for the one-to-many relationship between visual stimuli and neural responses (Mai et al., 2026).

Despite these advances, the limited temporal resolution of fMRI motivates a shift to M/EEG, which offers millisecond-level temporal resolution. This work therefore extends visual brain encoding to M/EEG.

Visual Prostheses. Visual prostheses are neural devices intended to provide rudimentary visual perception for patients with incurable blindness caused by retinal degenerative diseases (Busskamp et al., 2010; Cehajic-Kapetanovic et al., 2023; Prévot et al., 2020). A typical visual prosthesis generally follows a three-stage pipeline. First, an image sensor captures the surrounding scene; next, a processing framework converts the acquired visual information into electrode stimulation patterns; finally, an implanted electrode array stimulates ganglion cells to evoke visual perception in the visual cortex (Berry et al., 2017; Sahel et al., 2021; Soltan et al., 2018). The processing framework responsible for encoding visual information is a major determinant of the quality of restored vision and has therefore received increasing research attention (Turner et al., 2019).

Existing processing strategies for visual prostheses can be broadly divided into two categories. The first category employs region-based detection techniques to identify salient areas, structural information, or relevant objects in a visual scene and subsequently converts the extracted features into stimulation patterns using predefined rules or learned transformations (Barnich & Van Droogenbroeck, 2010; Boyle et al., 2008; Guo et al., 2018; 2019). The second category combines neural networks with biophysical phosphene models. These approaches learn stimulation patterns from visual inputs and use phosphene simulation to approximate the visual percepts induced by electrode activation (Küçükoğlu et al., 2022; Granley & Beyeler, 2021; Granley et al., 2022; van der Grinten et al., 2024; van Steveninck et al., 2022). However, most of them are optimized using image-domain reconstruction or simulated perceptual objectives without incorporating real neural responses as direct supervision.

Xu *et al.* (Xu et al., 2025) also recognized this limitation and adopted THINGS-EEG2 and THINGS-MEG, two multimodal datasets containing paired images and neural responses, to introduce measured M/EEG activity into model training. However, their approach is based on a diffusion model whose standard training objective minimizes the MSE between the sampled Gaussian noise and the noise predicted from a corrupted neural response. Although real M/EEG recordings participate indirectly in training through the forward noising process, this objective does not explicitly enforce point-wise correspondence between the predicted neural response and its paired ground truth.

To address this limitation, we replace the diffusion model with a BrainVAE and train a BrainTokenRegressor whose losses are computed directly against ground-truth neural responses, thereby explicitly incorporating real neural responses as supervised signals.

3 METHODOLOGY

3.1 PROBLEM DEFINITION

Let $\mathbf{s} \in \mathbb{R}^{1\times C\times T}$ denote an M/EEG neural response with C channels and T time points, and let $\mathbf{I}$ denote the visual image that elicits it. A pre-trained and frozen I-JEPA (Assran et al., 2023) image encoder Ψ maps the image to an I-JEPA image embedding $\mathbf{v} = \Psi(\mathbf{I}) \in \mathbb{R}^{M\times d_v}$, where M is the number of image tokens and d_v is the embedding dimension. Our goal is to learn a mapping from

visual stimuli to neural responses: given a test image $\mathbf{I}$, generate the corresponding M/EEG signal $\hat{\mathbf{s}}$.

3.2 Stage 1: BrainVAE Pre-training

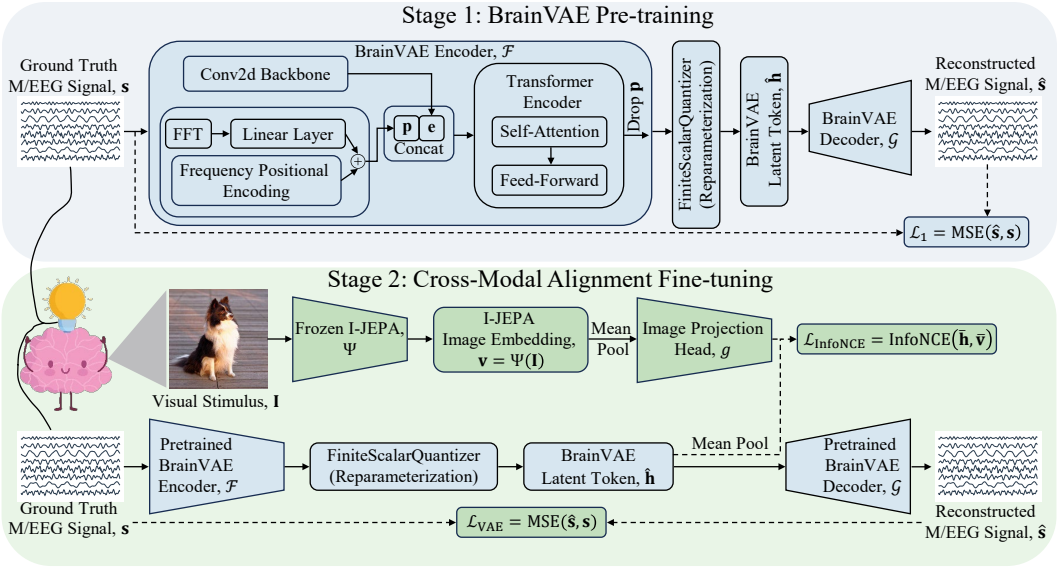

Figure 1: Overview of Stage 1 and Stage 2.

In Stage 1, we train a BrainVAE to reconstruct M/EEG signals through a quantized latent space, learning compact BrainVAE latent tokens without any visual information. An overview of Stage 1 is shown in Figure 1.

Time-Frequency Dual-Path Encoder. The encoder $\mathcal{F}$ processes the M/EEG signal $\mathbf{s}$ through two complementary paths.

The *temporal path* employs a Conv2d backbone consisting of a convolution layer followed by three residual downsampling blocks. The resulting feature map is adaptively pooled to a fixed spatial resolution and flattened into a sequence of N Conv2d feature tokens. Each token captures local spatial-temporal patterns within its receptive field:

$$\mathbf{e} = \text{Flatten}(\text{Pool}(\text{Conv2d}(\mathbf{s}))) \in \mathbb{R}^{N \times D},$$

where $N = 256$ and $D = 256$ for both EEG and MEG modalities.

The *frequency path* extracts spectral features from the raw signal via the Frequency Prefix Extractor $\mathcal{P}$. This module computes the real FFT along the time axis, extracts the power spectrum, removes the DC component, and applies log-normalization for numerical stability:

$$\mathbf{p} = \mathcal{P}(\mathbf{s}) \in \mathbb{R}^{K \times D},$$

where $K = \lfloor T/2 \rfloor$ is the number of frequency bins ($K = 125$ for EEG with $T = 250$; $K = 100$ for MEG with $T = 200$). Each frequency bin's channel-wise power vector is linearly projected to dimension D and augmented with a learnable frequency positional encoding.

The Conv2d feature tokens $\mathbf{e}$ and frequency prefix tokens $\mathbf{p}$ are concatenated along the sequence dimension and passed through a Transformer encoder (Vaswani et al., 2017) consisting of L self-attention layers:

$$[\mathbf{p}'; \mathbf{e}'] = \text{Transformer}([\mathbf{p}; \mathbf{e}]) \in \mathbb{R}^{(K+N) \times D}.$$

Through self-attention, each Conv2d feature token attends to all frequency prefix tokens, effectively absorbing frequency-domain information. After the Transformer, the prefix tokens are discarded

and only the N Conv2d feature tokens are retained as the encoder output:

$$\mathbf{h} = \mathbf{e}' \in \mathbb{R}^{N \times D}.$$

This prefix-based design allows the encoder to incorporate explicit frequency-domain priors—capturing neural oscillatory patterns across frequency bands—without occupying any dimensions in the final latent representation.

Finite Scalar Quantization. The continuous token sequence $\mathbf{h}$ is discretized by the finite scalar quantizer $\mathcal{Q}$ (Mentzer et al., 2024) to produce the *BrainVAE latent tokens*:

$$\hat{\mathbf{h}} = \mathcal{Q}(\mathbf{h}) \in \mathbb{R}^{N \times D}.$$

Unlike traditional VQ, FSQ requires no auxiliary losses and is inherently immune to codebook collapse (Mentzer et al., 2024).

Decoder. The decoder $\mathcal{G}$ mirrors the encoder architecture in reverse. The BrainVAE latent tokens $\hat{\mathbf{h}}$ first pass through a Transformer for global context aggregation, are then reshaped into a 2D feature map, and undergo three residual upsampling blocks to reconstruct the signal:

$$\hat{\mathbf{s}} = \mathcal{G}(\hat{\mathbf{h}}) \in \mathbb{R}^{1 \times C \times T}.$$

Training Objective. The BrainVAE is trained to minimize the reconstruction loss:

$$\mathcal{L}_1 = \text{MSE}(\hat{\mathbf{s}},\, \mathbf{s}). \tag{1}$$

Since FSQ introduces no auxiliary quantization loss, the only training objective is faithful signal reconstruction.

3.3 Stage 2: Cross-Modal Alignment Fine-tuning

In Stage 2, we fine-tune the pre-trained BrainVAE to establish a bridge between the neural latent space and the visual semantic space. The frozen I-JEPA image encoder Ψ extracts I-JEPA image embeddings $\mathbf{v} = \Psi(\mathbf{I}) \in \mathbb{R}^{M \times d_v}$. An overview of Stage 2 is shown in Figure 1.

We derive a global neural embedding by mean-pooling the BrainVAE latent tokens:

$$\bar{\mathbf{h}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{h}}_i \in \mathbb{R}^D.$$

Similarly, the I-JEPA image embeddings are mean-pooled over the M image tokens and projected through a learnable projection head g to produce a projected image embedding:

$$\bar{\mathbf{v}} = g\left(\frac{1}{M} \sum_{j=1}^{M} \mathbf{v}_j \right) \in \mathbb{R}^D.$$

The Stage 2 objective combines the reconstruction loss with an InfoNCE contrastive loss (Oord et al., 2018):

$$\mathcal{L}_2 = \text{MSE}(\hat{\mathbf{s}},\, \mathbf{s}) + \lambda\, \mathcal{L}_{\text{InfoNCE}}(\bar{\mathbf{h}},\, \bar{\mathbf{v}}), \tag{2}$$

where $\lambda = 0.1$ balances reconstruction fidelity with cross-modal alignment.

This stage ensures that the BrainVAE latent space develops semantic structure aligned with I-JEPA image embeddings, while the reconstruction loss prevents degradation of signal fidelity.

3.4 Stage 3: Visual-Guided Neural Encoding

In Stage 3, we train the *BrainTokenRegressor* $\mathcal{R}$, a Transformer-based token regressor that predicts FSQ token sequences directly from I-JEPA image embeddings, enabling end-to-end visual-to-neural conversion. The image encoder Ψ and the BrainVAE encoder $\mathcal{F}$ and quantizer $\mathcal{Q}$ from Stage 2 are frozen, while the decoder $\mathcal{G}$ is unfrozen and jointly fine-tuned with the BrainTokenRegressor. An overview of Stage 3 is shown in Figure 2.

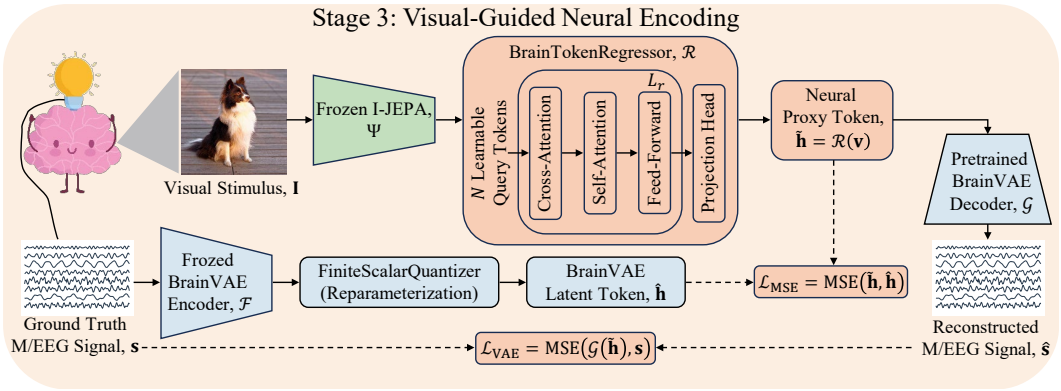

Figure 2: Overview of Stage 3.

BrainTokenRegressor Architecture. The BrainTokenRegressor $\mathcal{R}$ employs N learnable query tokens $\mathbf{q}_0 \in \mathbb{R}^{N \times d_h}$, where d_h is the hidden dimension. The I-JEPA image embeddings $\mathbf{v} = \Psi(\mathbf{I})$ are projected to d_h dimensions and serve as the key-value memory $\mathbf{m}$. In each of the L_r Transformer layers, the query tokens undergo cross-attention to extract visual information, self-attention for inter-token interaction, and a feed-forward network (FFN):

$$\begin{aligned}
\mathbf{q}_l^{\text{CA}} &= \text{LN}(\text{CrossAttn}(\mathbf{q}_{l-1}, \mathbf{m}) + \mathbf{q}_{l-1}), \\
\mathbf{q}_l^{\text{SA}} &= \text{LN}\big(\text{SelfAttn}(\mathbf{q}_l^{\text{CA}}) + \mathbf{q}_l^{\text{CA}}\big), \\
\mathbf{q}_l &= \text{LN}\big(\text{FFN}(\mathbf{q}_l^{\text{SA}}) + \mathbf{q}_l^{\text{SA}}\big),
\end{aligned}$$

where LN denotes layer normalization. The final query tokens are projected to produce the *neural proxy tokens*:

$$\tilde{\mathbf{h}} = \mathcal{R}(\mathbf{v}) = \text{Proj}_{\text{out}}(\mathbf{q}_{L_r}) \in \mathbb{R}^{N \times D}.$$

Training Objective. Given the BrainVAE latent tokens $\hat{\mathbf{h}} = \mathcal{Q}(\mathcal{F}(\mathbf{s}))$ from the frozen BrainVAE encoder, the Stage 3 loss combines a token-space MSE and a signal-space MSE:

$$\mathcal{L}_3 = \text{MSE}(\tilde{\mathbf{h}}, \hat{\mathbf{h}}) + \text{MSE}\Big(\mathcal{G}(\tilde{\mathbf{h}}), \mathbf{s}\Big). \tag{3}$$

The token-space MSE supervises the BrainTokenRegressor to match the BrainVAE latent tokens $\hat{\mathbf{h}}$ encoded from ground truth neural responses, while the signal-space MSE directly compares the generated output $\mathcal{G}(\tilde{\mathbf{h}})$ against the real neural response $\mathbf{s}$, thereby explicitly incorporating real neural responses as supervised signals throughout training.

3.5 Inference

During inference, the three components are composed into a single forward pass. Given a test image $\mathbf{I}$, we first extract I-JEPA image embeddings $\mathbf{v} = \Psi(\mathbf{I})$, then predict the neural proxy tokens $\tilde{\mathbf{h}} = \mathcal{R}(\mathbf{v})$, and finally generate the predicted M/EEG signal via the decoder $\mathcal{G}$:

$$\hat{\mathbf{s}} = \mathcal{G}(\mathcal{R}(\Psi(\mathbf{I}))).$$

4 Experiment

4.1 Main Results

We evaluate our framework on two multimodal datasets that provide paired images and neural responses: the EEG dataset THINGS-EEG2 (Gifford et al., 2022) and the MEG dataset THINGS-MEG (Hebart et al., 2023). We adopt the same preprocessing as (Xu et al., 2025) for both datasets.

Detailed evaluation protocols and implementation settings are provided in Appendix A.1 and Appendix A.2.

We report three complementary metrics between the predicted and ground-truth M/EEG signals: mean squared error (MSE, ↓), Pearson correlation coefficient (PCC, ↑), and cosine similarity (CosSim, ↑). All metrics are computed per trial and then averaged across subjects.

For each subject, we independently train the complete three-stage pipeline (the within-subject setting) and evaluate on the corresponding test set, reporting the mean across subjects as the "Our work" row. To further probe generalization, we additionally report two transfer protocols: the cross-subject setting, where a model trained on one subject is evaluated on another, and the leave-one-subject-out (LOSO) setting. Table 1 reports the quantitative results on THINGS-EEG2 and THINGS-MEG. Under the within-subject setting, our method outperforms all baselines on both datasets across all three metrics, achieving a PCC of 0.666 on THINGS-EEG2 and 0.558 on THINGS-MEG together with the lowest MSE, which demonstrates the effectiveness of directly supervising the model with real neural responses.

Table 1: Quantitative evaluation on THINGS-EEG2 and THINGS-MEG.

(a) THINGS-EEG2

Method	Evaluation Metrics		
	MSE ↓	PCC ↑	CosSim ↑
Güçlü *et al.* (Güçlü & Van Gerven, 2015)	0.211	0.178	0.181
Yamins *et al.* (Yamins et al., 2014)	0.196	0.203	0.209
UNet-based Diffusion Ho et al. (2020)	0.217	0.188	0.185
MindSimulator (Bao et al., 2025)	0.160	0.349	0.345
SynBrain (Mai et al., 2026)	0.156	0.366	0.347
Xu *et al.* (Xu et al., 2025)	0.109	0.425	0.402
Our work	**0.021**	**0.666**	**0.668**
Our work (Cross-subject)	0.367	0.223	0.215
Our work (LOSO Within-subject)	0.343	0.523	0.525
Our work (LOSO Cross-subject)	0.357	0.479	0.480

(b) THINGS-MEG

Method	Evaluation Metrics		
	MSE ↓	PCC ↑	CosSim ↑
Güçlü *et al.* (Güçlü & Van Gerven, 2015)	0.831	0.243	0.254
Yamins *et al.* (Yamins et al., 2014)	0.796	0.269	0.274
UNet-based Diffusion Ho et al. (2020)	0.763	0.122	0.127
MindSimulator (Bao et al., 2025)	0.755	0.281	0.274
SynBrain (Mai et al., 2026)	0.727	0.302	0.326
Xu *et al.* (Xu et al., 2025)	0.663	0.379	0.382
Our work	**0.433**	**0.558**	**0.556**
Our work (Cross-subject)	0.873	0.193	0.137
Our work (LOSO Within-subject)	0.485	0.459	0.457
Our work (LOSO Cross-subject)	0.558	0.337	0.335

Cross-Subject Generalization. The average cross-subject signal correlation is 0.223 on THINGS-EEG2 and 0.193 on THINGS-MEG. Cross-subject performance is markedly lower than the within-subject setting, indicating pronounced inter-individual variability in neural responses (Kanai & Rees, 2011).

Leave-One-Subject-Out Generalization. Motivated by the substantial gap between within-subject and cross-subject performance, we investigate whether increasing the training data by aggregating multiple subjects can help bridge this gap, and therefore adopt a leave-one-subject-out (LOSO) protocol (details in Appendix A.1). Training jointly on multiple subjects substantially improves generalization to an unseen subject: the LOSO Cross-subject PCC reaches 0.479 on THINGS-EEG2 and 0.337 on THINGS-MEG, more than doubling the single-source cross-subject transfer (0.223 and 0.193, respectively). Moreover, the gap between LOSO Within-subject and LOSO Cross-subject is markedly narrower than the original within-subject vs. cross-subject gap. This improvement can be understood through the bottom-up hierarchy of the visual system (DiCarlo & Cox, 2007), in which a visual stimulus is processed sequentially along the ventral stream—by areas V1, V2, and V4 in the occipital cortex and the inferotemporal (IT) cortex in the temporal lobe—for object recognition (Bao et al., 2020; Song et al., 2024). To illustrate this, we plot the topographic maps of Subject 1 and Subject 2 from the THINGS-EEG2 training set as representative examples (Figure 5), computed by averaging all training trials in each 100 ms window. The signal activity underlying object recognition is indeed concentrated in the occipital and temporal regions, yet the precise spatial extent and amplitude vary across individuals. By aggregating data from multiple subjects, the model is exposed to a diverse range of spatial distributions and amplitude patterns, effectively learning to accommodate inter-individual variability, which accounts for the markedly reduced gap between LOSO Within-subject and LOSO Cross-subject performance.

4.2 Ablation Studies

We ablate two key design choices in Table 2: (i) removing Stage 2 cross-modal alignment and (ii) using only one of the two Stage 3 loss terms (Equation 3). Removing Stage 2 lowers PCC to

0.606/0.515 on EEG/MEG, confirming that contrastive alignment provides an effective initialization for Stage 3. For the Stage 3 loss, using only the token-space MSE causes the most severe drop (PCC 0.381/0.360), demonstrating that direct supervision against real neural responses is indispensable, while using only the signal-space MSE also degrades performance (0.628/0.526), confirming that token-level supervision provides complementary guidance.

Table 2: Ablation studies on THINGS-EEG2 and THINGS-MEG.

(a) THINGS-EEG2

Setting	Evaluation Metrics		
	MSE ↓	PCC ↑	CosSim ↑
w/o Stage 2	0.025	0.606	0.608
Token-space MSE only	0.051	0.381	0.368
Signal-space MSE only	0.025	0.628	0.630
Full model (Our work)	**0.021**	**0.666**	**0.668**

(b) THINGS-MEG

Setting	Evaluation Metrics		
	MSE ↓	PCC ↑	CosSim ↑
w/o Stage 2	0.486	0.515	0.513
Token-space MSE only	0.565	0.360	0.343
Signal-space MSE only	0.484	0.526	0.524
Full model (Our work)	**0.433**	**0.558**	**0.556**

4.3 Frequency-Band Analysis

Since the BrainVAE encoder incorporates frequency-domain priors through the frequency prefix tokens, we examine reconstruction quality across neural oscillatory bands. We decompose both the predicted and ground-truth signals into five canonical bands—delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–100 Hz)—using a zero-phase fourth-order Butterworth band-pass filter, and compute the metrics per band, with the full-band (unfiltered) signal as reference. As shown in Table 3, on THINGS-EEG2 the theta band is reconstructed most faithfully (PCC 0.899), followed by alpha (0.785), while gamma is the hardest (0.032); on THINGS-MEG the delta band attains the highest PCC (0.502), and gamma is again the most difficult (0.015). The high-frequency gamma band remains challenging, and following Song et al. (2024), two factors likely contribute to this difficulty: (i) pure neuronal gamma oscillations are hard to recover, as the high-frequency range is highly susceptible to artifacts such as miniature saccades and muscle activity (Fries et al., 2008; Yuval-Greenberg et al., 2008; Hipp & Siegel, 2013); and (ii) the gamma-band amplitude is intrinsically low and readily modulated by cognitive processes such as attention and memory (Herrmann & Demiralp, 2005).

Table 3: Frequency-band analysis on THINGS-EEG2 and THINGS-MEG. The best value in each metric column is highlighted in bold, determined from the original six-decimal results.

(a) THINGS-EEG2

Band	Freq. Range	MSE ↓	PCC ↑	CosSim ↑
delta	0.5–4 Hz	0.004	0.551	0.551
theta	4–8 Hz	0.001	**0.899**	**0.899**
alpha	8–13 Hz	**0.001**	0.785	0.785
beta	13–30 Hz	0.001	0.611	0.611
gamma	30–100 Hz	0.002	0.032	0.032
Full-band (Our work)	–	0.021	0.666	0.668

(b) THINGS-MEG

Band	Freq. Range	MSE ↓	PCC ↑	CosSim ↑
delta	0.5–4 Hz	0.096	**0.502**	**0.502**
theta	4–8 Hz	0.029	0.441	0.441
alpha	8–13 Hz	0.040	0.303	0.303
beta	13–30 Hz	0.033	0.216	0.216
gamma	30–100 Hz	**0.017**	0.015	0.015
Full-band (Our work)	–	0.433	0.558	0.556

4.4 Temporal-Window Analysis

As M/EEG signals are inherently time-resolved, we evaluate reconstruction quality at different post-stimulus stages by partitioning the 0–1000 ms interval into three windows—0–300 ms, 300–500 ms, and 500–1000 ms—in addition to the full trial. As reported in Table 4, on both datasets the early and middle windows outperform the late window: on THINGS-EEG2 the PCC decreases from 0.656 (0–300 ms) and 0.654 (300–500 ms) to 0.605 (500–1000 ms), and on THINGS-MEG from 0.571 (0–300 ms) and 0.585 (300–500 ms) to 0.495 (500–1000 ms). According to (Gifford et al., 2022; Hebart et al., 2023), the visually evoked signal reaches an early peak ($\sim$ 100 ms) and, while still above chance until the end of the epoch (800–1000 ms), gradually decays over time. This is mirrored by our results, where the early (0–300 ms) and middle (300–500 ms) windows both perform well while the late window (500–1000 ms) drops markedly.

Table 4: Temporal-window analysis on THINGS-EEG2 and THINGS-MEG.

(a) THINGS-EEG2

Time Window	MSE ↓	PCC ↑	CosSim ↑
0–300 ms	**0.016**	**0.656**	**0.655**
300–500 ms	0.019	0.654	0.653
500–1000 ms	0.025	0.605	0.610
Full (Our work)	0.021	0.666	0.668

(b) THINGS-MEG

Time Window	MSE ↓	PCC ↑	CosSim ↑
0–300 ms	**0.289**	0.571	0.569
300–500 ms	0.397	**0.585**	**0.581**
500–1000 ms	0.517	0.495	0.492
Full (Our work)	0.433	0.558	0.556

4.5 Brain-Region Analysis

To examine whether the learned representations are spatially consistent with visual neuroscience priors, we group channels into five regions—frontal, central, temporal, parietal, and occipital—and compute the metrics per region. As shown in Figure 3, the occipital region is reconstructed best on both datasets (PCC 0.758 for EEG and 0.435 for MEG), followed by the parietal region, whereas the frontal and central regions are the least accurate (detailed numerical results are provided in Table 5 in Appendix A.4). This spatial pattern aligns with the bottom-up hierarchy of the visual system (DiCarlo & Cox, 2007) discussed in Section 4.1, where the occipital and temporal regions are the primary cortical areas for object recognition. The strong reconstruction in these regions suggests that our model learns a biologically plausible spatial structure.

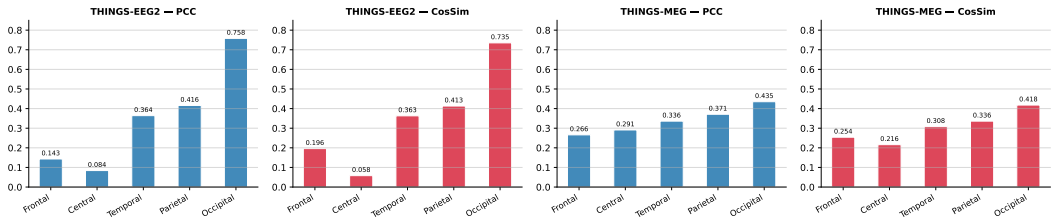

Figure 3: Brain-region analysis on THINGS-EEG2 and THINGS-MEG.

4.6 Image Encoder Comparison

We compare four image encoders—I-JEPA (Assran et al., 2023), CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2023; Darcet et al., 2024), and SigLIP2 (Tschannen et al., 2025)—under the within-subject setting, keeping all other components fixed. As shown in Figure 4, all four encoders achieve closely matched performance on both datasets, demonstrating that our framework is largely image-encoder-agnostic (detailed numerical results are provided in Table 6 in Appendix A.5).

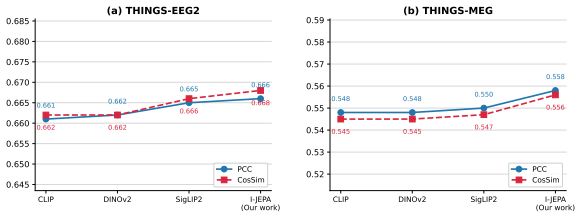

Figure 4: Image encoder comparison on THINGS-EEG2 and THINGS-MEG.

5 Conclusion

We presented a three-stage framework for visual-to-M/EEG brain encoding: a BrainVAE quantizes M/EEG signals into a latent space, contrastive learning aligns it with I-JEPA image embeddings, and a BrainTokenRegressor predicts neural response tokens under direct supervision from real neural responses. Extensive experiments on THINGS-EEG2 and THINGS-MEG demonstrate the effectiveness of the proposed method.

AI USE STATEMENT

In this work, we used generative AI tools for tasks in the recommended-disclosure category. Specifically, we used them to: (i) edit and polish the manuscript for readability and conciseness; (ii) generate and refine the plotting scripts for the figures reported in this paper; (iii) assist in drafting and formatting LaTeX content, including tables, pseudocode; and (iv) suggest section structure. We have not used generative AI tools for tasks in the required-disclosure category—including proposing or refining research hypotheses, designing the methodology or experiments, implementing the core method, generating synthetic data, or interpreting experimental results—all of which were carried out by the authors. We have reviewed all AI-assisted work: all AI-edited text was checked for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Guangyin Bao, Qi Zhang, Zixuan Gong, Zhuojia Wu, and Duoqian Miao. Mindsimulator: Exploring brain concept localization via synthetic fmri. *arXiv preprint arXiv:2503.02351*, 2025.

Pinglei Bao, Liang She, Mason McGill, and Doris Y Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, 2020.

Olivier Barnich and Marc Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image processing*, 20(6):1709–1724, 2010.

Michael H Berry, Amy Holt, Joshua Levitz, Johannes Broichhagen, Benjamin M Gaub, Meike Visel, Cherise Stanley, Krishan Aghi, Yang Joon Kim, Kevin Cao, et al. Restoration of patterned vision with an engineered photoactivatable g protein-coupled receptor. *Nature communications*, 8(1): 1862, 2017.

Justin R Boyle, Anthony J Maeder, and Wageeh W Boles. Region-of-interest processing for electronic visual prostheses. *Journal of Electronic Imaging*, 17(1):013002–013002, 2008.

Volker Busskamp, Jens Duebel, David Balya, Mathias Fradot, Tim James Viney, Sandra Siegert, Anna C Groner, Erik Cabuy, Valérie Forster, Mathias Seeliger, et al. Genetic reactivation of cone photoreceptors restores visual responses in retinitis pigmentosa. *science*, 329(5990):413–417, 2010.

Jasmina Cehajic-Kapetanovic, Mandeep S Singh, Eberhart Zrenner, and Robert E MacLaren. Bioengineering strategies for restoring vision. *Nature biomedical engineering*, 7(4):387–404, 2023.

Mike X Cohen. *Analyzing neural time series data: theory and practice*. The MIT press, 2014.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

Pascal Fries, René Scheeringa, and Robert Oostenveld. Finding gamma. *Neuron*, 58(3):303–305, 2008.

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.

Jacob Granley and Michael Beyeler. A computational model of phosphene appearance for epiretinal prostheses. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 4477–4481. IEEE, 2021.

Jacob Granley, Lucas Relic, and Michael Beyeler. Hybrid neural autoencoders for stimulus encoding in visual and other sensory neuroprostheses. *Advances in Neural Information Processing Systems*, 35:22671–22685, 2022.

Jacob Granley, Tristan Fauvel, Matthew Chalk, and Michael Beyeler. Human-in-the-loop optimization for deep stimulus encoding in visual prostheses. *Advances in neural information processing systems*, 36:79376–79398, 2023.

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

Fei Guo, Yuan Yang, and Yong Gao. Optimization of visual information presentation for visual prosthesis. *International journal of biomedical imaging*, 2018(1):3198342, 2018.

Fei Guo, Yuan Yang, Yang Xiao, Yong Gao, and Ningmei Yu. Recognition of moving object in high dynamic scene for visual prosthesis. *IEICE transactions on information and systems*, 102 (7):1321–1331, 2019.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

CS Herrmann and Tamer Demiralp. Human eeg gamma oscillations in neuropsychiatric disorders. *Clinical neurophysiology*, 116(12):2719–2733, 2005.

Joerg F Hipp and Markus Siegel. Dissociating neuronal gamma-band activity from cranial and ocular muscle activity in eeg. *Frontiers in human neuroscience*, 7:338, 2013.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ryota Kanai and Geraint Rees. The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience*, 12(4):231–242, 2011.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

Andreas Keil, Edward M Bernat, Michael X Cohen, Mingzhou Ding, Monica Fabiani, Gabriele Gratton, Emily S Kappenman, Eric Maris, Kyle E Mathewson, Richard T Ward, et al. Recommendations and publication guidelines for studies using frequency domain and time-frequency domain analyses of neural time series. *Psychophysiology*, 59(5):e14052, 2022.

Tetsuo Kida, Emi Tanaka, and Ryusuke Kakigi. Multi-dimensional dynamics of human electromagnetic brain activity. *Frontiers in human neuroscience*, 9:713, 2016.

Burcu Küçükoğlu, Bodo Rueckauer, Nasir Ahmad, Jaap de Ruyter van Steveninck, Umut Güçlü, and Marcel van Gerven. Optimization of neuroprosthetic vision via end-to-end deep reinforcement learning. *International Journal of Neural Systems*, 32(11):2250052, 2022.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024.

Weijian Mai, Jiamin Wu, Yu Zhu, Zhouheng Yao, Dongzhan Zhou, Andrew Luo, Qihao Zheng, Wanli Ouyang, and Chunfeng Song. Synbrain: Enhancing visual-to-fmri synthesis via probabilistic representation learning. *Advances in Neural Information Processing Systems*, 38:29253–29284, 2026.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *International Conference on Learning Representations*, volume 2024, pp. 51772–51783, 2024.

Santiago Morales and Maureen E Bowers. Time-frequency analysis methods and their application in developmental eeg data. *Developmental cognitive neuroscience*, 54:101067, 2022.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

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

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Paul-Henri Prévot, Kevin Gehere, Fabrice Arcizet, Himanshu Akolkar, Mina A Khoei, Kévin Blaize, Omar Oubari, Pierre Daye, Marion Lanoë, Manon Valet, et al. Behavioural responses to a photovoltaic subretinal prosthesis implanted in non-human primates. *Nature biomedical engineering*, 4(2):172–180, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

José-Alain Sahel, Elise Boulanger-Scemama, Chloé Pagot, Angelo Arleo, Francesco Galluppi, Joseph N Martel, Simona Degli Esposti, Alexandre Delaux, Jean-Baptiste de Saint Aubert, Caroline de Montleau, et al. Partial recovery of visual function in a blind patient after optogenetic therapy. *Nature medicine*, 27(7):1223–1229, 2021.

Ahmed Soltan, John Martin Barrett, Pleun Maaskant, Niall Armstrong, Walid Al-Atabany, Lionel Chaudet, Mark Neil, Evelyne Sernagor, and Patrick Degenaar. A head mounted device stimulator for optogenetic retinal prosthesis. *Journal of neural engineering*, 15(6):065002, 2018.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. In *International conference on learning representations*, volume 2024, pp. 47648–47665, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Maxwell H Turner, Luis Gonzalo Sanchez Giraldo, Odelia Schwartz, and Fred Rieke. Stimulus-and goal-oriented frameworks for understanding natural vision. *Nature neuroscience*, 22(1):15–24, 2019.

Maureen van der Grinten, Jaap de Ruyter van Steveninck, Antonio Lozano, Laura Pijnacker, Bodo Rueckauer, Pieter Roelfsema, Marcel van Gerven, Richard van Wezel, Umut Güçlü, and Yağmur Güçlütürk. Towards biologically plausible phosphene simulation for the differentiable optimization of visual cortical prostheses. *Elife*, 13:e85812, 2024.

Jaap de Ruyter van Steveninck, Umut Güçlü, Richard van Wezel, and Marcel van Gerven. End-to-end optimization of prosthetic vision. *Journal of Vision*, 22(2):20–20, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chuanqing Wang, Di Wu, Chaoming Fang, Jie Yang, and Mohamad Sawan. Exploring effective stimulus encoding via vision system modeling for visual prostheses. In *International Conference on Learning Representations*, volume 2024, pp. 15384–15396, 2024.

Ganxi Xu, Zhao-Rong Lai, Yuting Tang, Yonghao Song, Guoxu Zhou, Jian Zhu, Jinyi Long, et al. Image-to-brain signal generation for visual prosthesis with clip guided multimodal diffusion models. *arXiv preprint arXiv:2509.00787*, 2025.

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

Shlomit Yuval-Greenberg, Orr Tomer, Alon S Keren, Israel Nelken, and Leon Y Deouell. Transient induced gamma-band response in eeg as a manifestation of miniature saccades. *Neuron*, 58(3): 429–441, 2008.

A Appendix

A.1 Evaluation Protocols

THINGS-EEG2 contains data from 10 subjects and THINGS-MEG from 4 subjects, each split into training and test sets. For the "Our work" (within-subject) setting, we train a separate model for each subject using its own training data and evaluate on its own test data, reporting the mean across all subjects. For the "Our work (Cross-subject)" setting, each per-subject model is evaluated on the test data of all other subjects.

We additionally adopt a leave-one-subject-out (LOSO) protocol to mitigate cross-subject performance degradation. Taking THINGS-EEG2 as an example: in each fold we hold out one subject and combine the training data of the remaining 9 subjects into a unified training set, with their test data forming a unified test set. A single model is trained on this combined training set and then evaluated in two ways: (1) on the combined test set of the 9 training subjects, yielding the "LOSO Within-subject" result, and (2) on the test data of the held-out subject, yielding the "LOSO Cross-subject" result. Averaging across all 10 folds produces the final LOSO results. The same protocol is applied to THINGS-MEG with its 4 subjects.

A.2 Implementation Details

- **Image encoder.** We use a frozen I-JEPA ViT-H/14 (Assran et al., 2023), which outputs an image embedding $\mathbf{v} \in \mathbb{R}^{256\times1280}$.
- **BrainVAE encoder.** The M/EEG signal is treated as a single-channel 2D image (channels $\times$ time; 63×250 for EEG and 271×200 for MEG). The spatial-temporal path uses a Conv2d backbone with an initial 3×3 convolution ($1 \to 32$) followed by three stride-2 downsampling residual blocks ($32 \to 64 \to 128 \to 256$), where each block applies 3×3 convolutions with BatchNorm and ELU. The resulting feature map is pooled to a fixed grid via adaptive average pooling (8×32 for EEG and 16×16 for MEG), flattened and linearly projected into $N = 256$ tokens of dimension $D = 256$, with a learnable spatial-temporal positional embedding added.
- **Frequency prefix.** The frequency path applies an FFT along the time axis to the raw signal, takes the power spectrum, discards the DC component, and applies $\log(1 + \cdot)$ normalization. The per-frequency channel-power vectors are linearly projected to dimension D and augmented with a learnable frequency positional embedding, yielding K Frequency Prefix Tokens ($K = 125$ for EEG and $K = 100$ for MEG). These prefix tokens are concatenated with the 256 Conv2d feature tokens and fused by a 4-layer Pre-LN Transformer (8 heads, MLP ratio 4, dropout 0.1) via self-attention; the prefix tokens are then discarded, retaining the 256 fused neural tokens.

- **FSQ quantizer.** The quantizer linearly projects each token to a quantization dimension of $d_{\text{fsq}} = 4$, bounds it with $\tanh$, and rounds to the nearest grid point using a straight-through estimator under the level configuration $[8, 5, 5, 5]$ (an implicit codebook of size 1000), before projecting back to dimension D. FSQ requires no auxiliary codebook loss and avoids codebook collapse.
- **BrainVAE decoder.** The decoder mirrors the encoder: a 4-layer Transformer processes the quantized tokens, which are reshaped into a 2D feature map and progressively upsampled by three residual blocks (bilinear interpolation followed by 3×3 convolution), then cropped to the original signal size.
- **Optimization.** We adopt the AdamW optimizer with a learning rate of 1×10^{-4}, a weight decay of 1×10^{-5}, and cosine annealing (down to 1×10^{-6}) for 100 epochs at each stage. The batch size is 64 for EEG and 16 for MEG.
- **Stage 1 training.** The BrainVAE is trained with a signal-space reconstruction MSE loss.
- **Stage 2 training.** We fine-tune the BrainVAE with the reconstruction loss plus an InfoNCE contrastive loss (temperature 0.07, weight $\lambda = 0.1$) that aligns the global neural embedding $\bar{\mathbf{h}}$ with the projected image embedding $\bar{\mathbf{v}}$.
- **Stage 3 training.** We use a joint token-space MSE and signal-space MSE objective, unfreezing the decoder and jointly fine-tuning it at $1/10$ of the main learning rate.
- **Hardware.** All experiments are conducted on 4 NVIDIA V100S GPUs using PyTorch.

A.3 Training Algorithm

Algorithm 1 summarizes the three-stage training procedure, following the notation of Section 3.2–3.4: $\mathcal{F}$ is the BrainVAE encoder, $\mathcal{Q}$ the finite scalar quantizer, $\mathcal{G}$ the decoder, Ψ the frozen I-JEPA image encoder, g the projection head, and $\mathcal{R}$ the BrainTokenRegressor.

Algorithm 1 Three-Stage Training of the Visual-to-M/EEG Brain Encoding Framework

Stage 1: BrainVAE Pre-training
  Input: neural responses $\{\mathbf{s}\}$; modules $\mathcal{F}, \mathcal{Q}, \mathcal{G}$
for each mini-batch $\mathbf{s}$ **do**
  $\hat{\mathbf{h}} \leftarrow \mathcal{Q}(\mathcal{F}(\mathbf{s}))$
  $\hat{\mathbf{s}} \leftarrow \mathcal{G}(\hat{\mathbf{h}})$
  $\mathcal{L}_1 \leftarrow \text{MSE}(\hat{\mathbf{s}}, \mathbf{s})$
  update $\mathcal{F}, \mathcal{Q}, \mathcal{G}$ by minimizing $\mathcal{L}_1$
end for

Stage 2: Cross-Modal Alignment Fine-tuning
  Input: paired data $\{(\mathbf{s}, \mathbf{I})\}$; pre-trained $\mathcal{F}, \mathcal{Q}, \mathcal{G}$; frozen Ψ; projection head g
for each mini-batch $(\mathbf{s}, \mathbf{I})$ **do**
  $\hat{\mathbf{h}} \leftarrow \mathcal{Q}(\mathcal{F}(\mathbf{s}))$; $\hat{\mathbf{s}} \leftarrow \mathcal{G}(\hat{\mathbf{h}})$
  $\bar{\mathbf{h}} \leftarrow \frac{1}{N}\sum_i \hat{\mathbf{h}}_i$
  $\bar{\mathbf{v}} \leftarrow g\Big(\frac{1}{M}\sum_j \Psi(\mathbf{I})_j\Big)$
  $\mathcal{L}_2 \leftarrow \text{MSE}(\hat{\mathbf{s}}, \mathbf{s}) + \lambda\, \mathcal{L}_{\text{InfoNCE}}(\bar{\mathbf{h}}, \bar{\mathbf{v}})$
  update $\mathcal{F}, \mathcal{Q}, \mathcal{G}, g$ by minimizing $\mathcal{L}_2$
end for

Stage 3: Visual-Guided Neural Encoding
  Input: paired data $\{(\mathbf{s}, \mathbf{I})\}$; frozen $\mathcal{F}, \mathcal{Q}, \Psi$; BrainTokenRegressor $\mathcal{R}$; decoder $\mathcal{G}$ (unfrozen)
for each mini-batch $(\mathbf{s}, \mathbf{I})$ **do**
  $\hat{\mathbf{h}} \leftarrow \mathcal{Q}(\mathcal{F}(\mathbf{s}))$
  $\tilde{\mathbf{h}} \leftarrow \mathcal{R}(\Psi(\mathbf{I}))$
  $\mathcal{L}_3 \leftarrow \text{MSE}(\tilde{\mathbf{h}}, \hat{\mathbf{h}}) + \text{MSE}(\mathcal{G}(\tilde{\mathbf{h}}), \mathbf{s})$
  update $\mathcal{R}, \mathcal{G}$ by minimizing $\mathcal{L}_3$
end for

A.4 Brain-Region Analysis Details

Table 5 provides the detailed numerical results corresponding to the brain-region analysis in Section 4.5.

Table 5: Brain-region analysis on THINGS-EEG2 and THINGS-MEG.

(a) THINGS-EEG2

Brain Region	MSE ↓	PCC ↑	CosSim ↑
Frontal	0.081	0.143	0.196
Central	**0.030**	0.084	0.058
Temporal	0.041	0.364	0.363
Parietal	0.058	0.416	0.413
Occipital	0.075	**0.758**	**0.735**
All (Our work)	0.021	0.666	0.668

(b) THINGS-MEG

Brain Region	MSE ↓	PCC ↑	CosSim ↑
Frontal	0.534	0.266	0.254
Central	**0.459**	0.291	0.216
Temporal	0.576	0.336	0.308
Parietal	0.487	0.371	0.336
Occipital	0.598	**0.435**	**0.418**
All (Our work)	0.433	0.558	0.556

For the brain-region analysis in Section 4.5, we partition the recording channels into five regions—frontal, central, temporal, parietal, and occipital.

EEG Configuration. For THINGS-EEG2 with 63 electrodes, the regions are defined based on the standard 10–10 electrode naming system:

- **Frontal** (15 channels): Fp1, Fp2, AF7, AF3, AFz, AF4, AF8, F7, F5, F3, F1, F2, F4, F6, F8
- **Central** (21 channels): FC5, FC3, FC1, FCz, FC2, FC4, FC6, C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPz, CP2, CP4, CP6
- **Temporal** (14 channels): FT9, FT7, FT8, FT10, T7, T8, TP9, TP7, TP8, TP10, P7, P8, PO7, PO8
- **Parietal** (10 channels): P5, P3, P1, Pz, P2, P4, P6, PO3, POz, PO4
- **Occipital** (3 channels): O1, Oz, O2

MEG Configuration. For THINGS-MEG with 271 magnetometers (after excluding one faulty channel, MRO11), channels are assigned to the five regions according to their physical sensor locations. The assignment is based on spatial position rather than the label prefix. For brevity, we omit the common acquisition suffix "-1609" from all channel names below.

- **Frontal** (95 channels): MLC17, MLC21, MLC22, MLC53, MLC54, MLC55, MLC61, MLC62, MLC63, MLF11, MLF12, MLF13, MLF14, MLF21, MLF22, MLF23, MLF24, MLF34, MLF52, MLF64, MLF65, MLF66, MLF67, MLO11, MLO12, MLO13, MLO14, MLO21, MLO22, MLO23, MLO24, MLO31, MLO32, MLO33, MLO34, MLO41, MLO42, MLO43, MLO44, MLO51, MLO52, MLO53, MLP11, MLP21, MLP22, MLP23, MLP31, MLT52, MLT53, MRC11, MRC12, MRC21, MRC22, MRC41, MRC42, MRC61, MRF21, MRF22, MRF23, MRF24, MRF33, MRF52, MRF64, MRF65, MRF66, MRF67, MRO12, MRO14, MRO21, MRO22, MRO23, MRO24, MRO31, MRO32, MRO33, MRO34, MRO41, MRO42, MRO43, MRO44, MRO51, MRO52, MRO53, MRP11, MRP12, MRP22, MRP23, MRP31, MRP32, MRT52, MRT53, MZC01, MZC02, MZO01, MZO02
- **Central** (16 channels): MLC31, MLC41, MLF35, MLF46, MLF51, MLF53, MLF54, MLF61, MLF62, MRF34, MRF46, MRF51, MRF53, MRF54, MRF61, MRF62
- **Temporal** (65 channels): MLC23, MLC51, MLF31, MLF32, MLF33, MLF41, MLF42, MLF43, MLF44, MLF45, MLP12, MLP32, MLP33, MLP34, MLT25, MLT33, MLT37, MLT41, MLT47, MLT51, MLT54, MLT55, MLT56, MRC13, MRC14, MRC15, MRC16, MRC23, MRC24, MRC25, MRC31, MRC51, MRC52, MRC53, MRC62, MRC63, MRF11, MRF12, MRF25, MRF31, MRF32, MRF35, MRF41, MRF42, MRF44, MRF45, MRP21, MRP33, MRP34, MRP35, MRT25, MRT33, MRT37, MRT41, MRT47, MRT51, MRT54, MRT55, MRT56, MZC03, MZC04, MZF01, MZF02, MZO03, MZP01

- **Parietal** (21 channels): MLC11, MLC12, MLC13, MLC14, MLC15, MLC16, MLC24, MLC25, MLC32, MLC42, MLC52, MLF55, MLF56, MLF63, MLT21, MLT22, MRF55, MRF56, MRF63, MRT21, MRT22
- **Occipital** (74 channels): MLP35, MLP41, MLP42, MLP43, MLP44, MLP45, MLP51, MLP52, MLP53, MLP54, MLP55, MLP56, MLP57, MLT11, MLT12, MLT13, MLT14, MLT15, MLT16, MLT23, MLT24, MLT26, MLT27, MLT31, MLT32, MLT34, MLT35, MLT36, MLT42, MLT43, MLT44, MLT45, MLT46, MLT57, MRC17, MRC32, MRC54, MRC55, MRF13, MRF14, MRP41, MRP42, MRP43, MRP44, MRP45, MRP51, MRP52, MRP53, MRP54, MRP55, MRP56, MRP57, MRT11, MRT12, MRT13, MRT14, MRT15, MRT16, MRT23, MRT24, MRT26, MRT27, MRT31, MRT32, MRT34, MRT35, MRT36, MRT42, MRT43, MRT44, MRT45, MRT46, MRT57, MZF03

A.5 Image Encoder Comparison Details

Table 6 provides the detailed numerical results corresponding to the image encoder comparison in Section 4.6.

Table 6: Image encoder comparison on THINGS-EEG2 and THINGS-MEG.

(a) THINGS-EEG2

Image Encoder	MSE ↓	PCC ↑	CosSim ↑
CLIP	0.021	0.661	0.662
DINOv2	0.021	0.662	0.662
SigLIP2	0.021	0.665	0.666
I-JEPA (Our work)	**0.021**	**0.666**	**0.668**

(b) THINGS-MEG

Image Encoder	MSE ↓	PCC ↑	CosSim ↑
CLIP	0.435	0.548	0.545
DINOv2	**0.422**	0.548	0.545
SigLIP2	0.444	0.550	0.547
I-JEPA (Our work)	0.433	**0.558**	**0.556**

A.6 Per-Subject Topographic Maps

As discussed in Section 4.1, we plot the topographic maps of Subject 1 and Subject 2 from the THINGS-EEG2 training set as representative examples. Each topography is obtained by averaging all training trials within each 100 ms window.

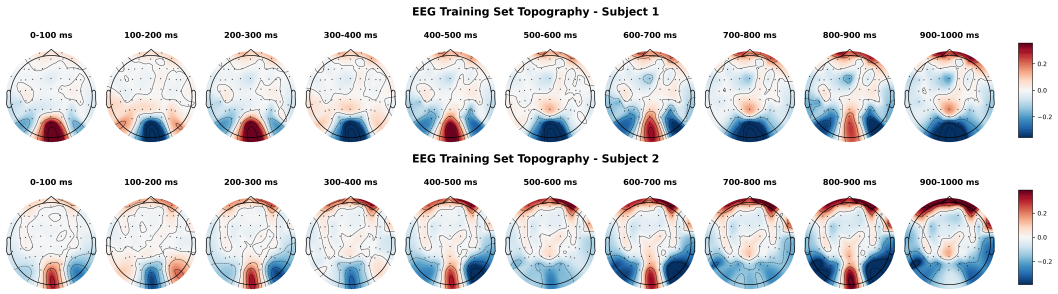

Figure 5: Per-subject topographic maps of Subject 1 and Subject 2 from THINGS-EEG2.