D^2-FOSA: Dual-Diffusion Guided EEG-to-Image Reconstruction with Frequency-Oriented Semantic Alignment

Supplementary Material

A. Overview

This supplementary material provides a comprehensive overview of the datasets and preprocessing pipelines (Section B), methodological details (Section C.1), evaluation metrics (Section C.2 C.3), algorithmic implementations (Section C.4), and experimental setups and additional results (Section D) used in this study. Specifically, we detail the processing and evaluation protocols for the THINGS-EEG, THINGS-MEG, EEGCVPR40, and EEGImageNet datasets, ensuring reproducibility and consistency with prior works. Furthermore, we present an in-depth explanation of our proposed D^2-FOSA framework, including training and inference procedures, sensitivity analyses, and computational resource comparisons. Additional experimental results, such as retrieval and generation performance, are included to further validate the robustness and effectiveness of our approach. For the convenience of the research community, we have also included the code in the supplementary package, enabling easy reproduction and extension of our work.

B. Dataset

THINGS-EEG [4] is a large-scale benchmark for decoding visual perception from brain activity. The dataset comprises EEG recordings from 10 participants exposed to 1,854 real-world object categories under a rapid serial visual presentation (RSVP) paradigm [11]. EEG signals were recorded from 64 channels (63 active + 1 reference) at a sampling rate of 1,000 Hz. Each participant completed 82,160 trials in total. The training set contains 1,654 categories, each associated with 10 unique images presented 4 times, while the test set includes 200 categories with 1 image repeated 80 times. For each image, we extract EEG epochs from 0 to 1,000 ms post-stimulus, and average repeated trials to enhance signal-to-noise ratio. The EEG signals are then downsampled to 250 Hz, resulting in a final representation of 63×250 (channels × time points) per trial. This dataset offers high temporal resolution and fine-grained object diversity, making it a suitable benchmark for evaluating EEG-to-image generation and cross-modal representation learning. Table 2 presents the results of re-classifying test set images into their coarse categories, which serves as a proxy to evaluate cross-modal alignment. During EEG-to-image retrieval, if the retrieved image is classified into the same coarse category as the EEG label, the retrieval is considered successful. This evaluation emphasizes the semantic consistency between EEG inputs and the retrieved images, highlighting the effectiveness of the learned alignment.

THINGS-MEG [5] is a large-scale benchmark for high-resolution visual neuroscience. MEG responses were collected from four healthy participants using a 275-channel CTF system (272 usable sensors) while they viewed 22,248 unique natural images spanning 1,854 THINGS object concepts. Each image appeared once in the training set, whereas 200 held-out concepts were presented 12 times to enable noise-ceiling estimation, with repeated trials averaged to improve SNR following [13]. Participants performed passive viewing with an oddball detection task, and eye position was continuously monitored (EyeLink 1000 Plus, 1,200 Hz). Preprocessing followed standard pipelines, including 0.1–40 Hz filtering, epoching from -100 to 1,300 ms, baseline correction, bad-sensor removal, and downsampling to 200 Hz. The dataset adheres to the BIDS-MEG standard and provides semantic labels, concept IDs, structural MRI, and repeated test trials, offering a comprehensive resource for studying cross-modal alignment and neural representation learning.

EEGCVPR40 [14] is a visual EEG benchmark derived from a curated 40-class subset of ImageNet [2]. EEG was recorded from six healthy participants at 1,000 Hz using a 128-channel actiCAP system while they passively viewed 2,000 natural images (0.5 s per trial), with randomized presentation order and an oddball detection task to maintain fixation. Preprocessing follows the protocol of: 49–51 Hz notch filtering, 14–71 Hz band-pass filtering, baseline correction, ICA-based artifact removal, and extraction of the 20–460 ms post-stimulus window. After artifact rejection, the dataset provides 11,940 clean EEG–image pairs, each represented as a 128×440 spatiotemporal matrix. Data from all subjects are concatenated and split into train/validation/test sets following prior work, supporting standardized evaluation for EEG-based visual decoding and cross-modal learning.

EEGImageNet [19] is a recent large-scale benchmark for EEG-based visual decoding and reconstruction, comprising 63,850 EEG–image pairs collected from 16 healthy participants at the Tsinghua University DCST Lab. EEG signals were recorded using a 64-channel Quik-Cap system (62 usable channels, 1,000 Hz, 16-bit) while subjects viewed 4,000 high-quality ImageNet images spanning 80 categories, organized into 40 coarse-grained and 40 fine-grained classes. Each image was presented for 500 ms with

Table 1. Summary of the four benchmark datasets used in this study.

Dataset	Participants	Total Images	Categories	Stimulus Duration	Sampling Rate	Channels/ Sensors	Trials/ Repeats
THINGS-EEG [4]	10	18,540	1,854	1,000 ms	1,000 Hz	64 (63+1)	4/80
THINGS-MEG [5]	4	22,248	1,854	1,300 ms	200 Hz	275 (272)	1/12
EEGCVPR40 [14]	6	2,000	40	500 ms	1,000 Hz	128	1
EEGImageNet [19]	16	4,000	80	500 ms	1,000 Hz	64 (62)	1

a 500 ms inter-stimulus interval, following a 30/20 train–test split per category. Standard preprocessing includes average mastoid re-referencing, 0.5–80 Hz band-pass filtering, 50 Hz notch filtering, and removal of ocular and motion artifacts, yielding clean spatiotemporal EEG segments aligned with stimulus onset. The dataset is provided in PyTorch format with WordNet category labels and image indices, offering a robust foundation for training and evaluating models in visual decoding, reconstruction, and multimodal EEG representation learning.

Data Preprocessing. As summarized in Table 1, we utilize four benchmark datasets: THINGS-EEG, THINGS-MEG, EEGCVPR40, and EEGImageNet. Each dataset undergoes a standardized preprocessing pipeline to ensure high-quality neural signals for robust EEG-to-image retrieval, following established methodologies [8, 13, 17, 19]. The preprocessing includes four key steps: (1) re-referencing to reduce spatial noise, (2) band-pass and notch filtering to isolate relevant frequency bands and remove powerline interference, (3) independent component analysis (ICA) to eliminate artifacts (e.g., eye blinks, muscle movements), and (4) extraction of stimulus-aligned time windows to capture neural responses to visual stimuli. These steps ensure clean, spatiotemporally aligned signals optimized for downstream tasks such as neural decoding, cross-modal alignment, and EEG-to-image retrieval. This preprocessing framework enhances the reliability and reproducibility of our experiments, providing a solid foundation for advancing EEG-based visual decoding research.

C. More Implementation Details

C.1. FOMamba | Detailed derivation

Classical State Space Models (SSMs), such as S4 and Mamba, describe continuous-time latent dynamics as

$$\dot{h}(t) = Ah(t) + Bx(t), \tag{1}$$

where $h(t) \in \mathbb{R}^{D_{\text{inner}}}$ denotes the latent state, $x(t) \in \mathbb{R}^{D_{\text{in}}}$ is the input signal, $A \in \mathbb{R}^{D_{\text{inner}} \times D_{\text{inner}}}$ governs the system dynamics, and $B \in \mathbb{R}^{D_{\text{inner}} \times D_{\text{in}}}$ modulates the input influence. Standard SSMs assume A is diagonal or low-rank, implying independent latent channels. However, this assumption is incompatible with oscillatory neural signals (e.g., EEG), where frequency bands (α, β, γ) exhibit structured modal coupling.

To address this limitation, we introduce a **2×2 block-diagonal modal parameterization** that explicitly encodes oscillatory frequency and damping dynamics within each latent pair. We decompose A into $D_{\text{inner}}/2$ complex-conjugate blocks:

$$A = \text{blkdiag}(A_1, A_2, \ldots, A_{D_{\text{inner}}/2}), \tag{2}$$

where each block is defined as

$$A_k = \begin{bmatrix} -\rho_k & -\omega_k \\ \omega_k & -\rho_k \end{bmatrix}, \tag{3}$$

with $\rho_k > 0$ controlling modal damping and $\omega_k > 0$ denoting the angular frequency of the k-th oscillatory mode. This structure yields eigenvalues $\lambda_k = -\rho_k \pm j\omega_k$, producing a decaying oscillatory system.

For the k-th modal pair, the dynamics become

$$\frac{d}{dt}\begin{bmatrix} h_{2k-1} \\ h_{2k} \end{bmatrix} = \begin{bmatrix} -\rho_k & -\omega_k \\ \omega_k & -\rho_k \end{bmatrix}\begin{bmatrix} h_{2k-1} \\ h_{2k} \end{bmatrix} + B_k x(t), \tag{4}$$

where $B_k \in \mathbb{R}^{2 \times D_{\text{in}}}$ denotes the corresponding input projection.

Following the implicit SSM formulation in Mamba, we use a learnable step size Δt and apply the bilinear (Tustin) discretization:

$$A_d = (I - \tfrac{\Delta t}{2}A)^{-1}(I + \tfrac{\Delta t}{2}A), \tag{5}$$

$$B_d = (I - \tfrac{\Delta t}{2}A)^{-1}\Delta t B, \tag{6}$$

where A_d and B_d denote the discretized transition and input matrices, respectively. For each modal block, we obtain the closed-form discrete update:

$$A_{d,k} = e^{-\rho_k \Delta t}\begin{bmatrix} \cos(\omega_k \Delta t) & -\sin(\omega_k \Delta t) \\ \sin(\omega_k \Delta t) & \cos(\omega_k \Delta t) \end{bmatrix}, \tag{7}$$

$$B_{d,k} = \Delta t\, e^{-\rho_k \Delta t} B_k, \tag{8}$$

Table 2. Complete list of the 200 stimulus categories used for test-time reclassification in the THINGS-EEG and THINGS-MEG datasets.

Items 1–50		Items 51–100		Items 101–150		Items 151–200	
Name	**Category**	**Name**	**Category**	**Name**	**Category**	**Name**	**Category**
Aircraft Carrier	Transportation	Cordon Bleu	Food	Jelly Bean	Food	Robot	Household
Antelope	Animals	Coverall	Clothes	Jukebox	Household	Rooster	Animals
Backscratcher	Household	Crab	Animals	Kettle	Household	Rug	Household
Balance Beam	Household	Creme Brulee	Food	Kneepad	Clothes	Sailboat	Transportation
Banana	Food	Crepe	Food	Ladle	Tools	Sandal	Clothes
Baseball Bat	Household	Crib	Household	Lamb	Animals	Sandpaper	Tools
Basil	Food	Croissant	Food	Lampshade	Household	Sausage	Food
Basketball	Household	Crow	Animals	Laundry Basket	Household	Scallion	Food
Bassoon	Household	Cruise Ship	Transportation	Lettuce	Food	Scallop	Animals
Baton4	Household	Crumb	Food	Lightning Bug	Animals	Scooter	Transportation
Batter	Food	Cupcake	Food	Manatee	Animals	Seagull	Animals
Beaver	Animals	Dagger	Tools	Marijuana	Nature	Seaweed	Nature
Bench	Household	Dalmatian	Animals	Meatloaf	Food	Seed	Nature
Bike	Transportation	Dessert	Food	Metal Detector	Tools	Skateboard	Transportation
Birthday Cake	Food	Dragonfly	Animals	Minivan	Transportation	Sled	Transportation
Blowtorch	Tools	Dreidel	Household	Modem	Household	Sleeping Bag	Household
Boat	Transportation	Drum	Household	Mosquito	Animals	Slide	Household
Bok Choy	Food	Duffel Bag	Clothes	Muff	Clothes	Slingshot	Tools
Bonnet	Clothes	Eagle	Animals	Music Box	Household	Snowshoe	Clothes
Bottle Opener	Tools	Eel	Animals	Mussel	Animals	Spatula	Tools
Brace	Clothes	Egg	Food	Nightstand	Household	Spoon	Household
Bread	Food	Elephant	Animals	Okra	Food	Station Wagon	Transportation
Breadbox	Household	Espresso	Food	Omelet	Food	Stethoscope	Tools
Bug	Animals	Face Mask	Clothes	Onion	Food	Strawberry	Food
Buggy	Transportation	Ferry	Transportation	Orange	Food	Submarine	Transportation
Bullet	Household	Flamingo	Animals	Orchid	Nature	Suit	Clothes
Bun	Food	Folder	Household	Ostrich	Animals	T-shirt	Clothes
Bush	Nature	Fork	Household	Pajamas	Clothes	Table	Household
Calamari	Food	Freezer	Household	Panther	Animals	Taillight	Transportation
Candlestick	Household	French Horn	Household	Paperweight	Household	Tape Recorder	Household
Cart	Transportation	Fruit	Food	Pear	Food	Television	Household
Cashew	Food	Garlic	Food	Pepper1	Food	Tiara	Clothes
Cat	Animals	Glove	Clothes	Pheasant	Animals	Tick	Animals
Caterpillar	Animals	Golf Cart	Transportation	Pickax	Tools	Tomato Sauce	Food
Cd Player	Household	Gondola	Transportation	Pie	Food	Tongs	Tools
Chain	Household	Goose	Animals	Pigeon	Animals	Tool	Tools
Chaps	Clothes	Gopher	Animals	Piglet	Animals	Top Hat	Clothes
Cheese	Food	Gorilla	Animals	Pocket	Clothes	Treadmill	Household
Cheetah	Animals	Grasshopper	Animals	Pocketknife	Tools	Tube Top	Clothes
Chest2	Household	Grenade	Household	Popcorn	Food	Turkey	Animals
Chime	Household	Hamburger	Food	Popsicle	Food	Unicycle	Transportation
Chopsticks	Household	Hammer	Tools	Possum	Animals	Vise	Tools
Cleat	Clothes	Handbrake	Transportation	Pretzel	Food	Volleyball	Household
Cleaver	Tools	Headscarf	Clothes	Pug	Animals	Wallpaper	Household
Coat	Clothes	Highchair	Household	Punch2	Food	Walnut	Food
Cobra	Animals	Hoodie	Clothes	Purse	Clothes	Wheat	Nature
Coconut	Food	Hummingbird	Animals	Radish	Food	Wheelchair	Transportation
Coffee Bean	Food	Ice Cube	Food	Raspberry	Food	Windshield	Transportation
Coffeemaker	Household	Ice Pack	Household	Recorder	Household	Wine	Food
Cookie	Food	Jeep	Transportation	Rhinoceros	Animals	Wok	Household

which separately preserves exponential damping and rotational oscillation—two key properties of EEG temporal dynamics.

To incorporate neural spectral priors, we introduce two frequency-aware learnable modifications:

- **Static Log-Frequency Bias** $F_{\log,k}$ adjusts each mode's frequency preference:

$$\tilde{\omega}_k = \mathrm{softplus}(\omega_k + F_{\log,k}), \tag{9}$$

  where $\tilde{\omega}_k$ is the refined angular frequency.
- **Stochastic Damping Regularization** improves robust-

ness under spectral noise:

$$\rho_k = \rho_{0,k} + \epsilon_\rho, \qquad \epsilon_\rho \sim \mathcal{N}(0, \sigma_\rho^2), \tag{10}$$

where $\rho_{0,k}$ is the base damping parameter.

The augmented modal block therefore becomes

$$A_k(F_{\log}, \rho_k, \omega_k) = \begin{bmatrix} -\rho_k & -\tilde{\omega}_k \\ \tilde{\omega}_k & -\rho_k \end{bmatrix}, \tag{11}$$

which captures adaptive oscillatory behavior.

Let (A_d, B_d, C, D) denote the discretized SSM. Under Mamba's selective scan, the hidden update is

$$h_{t+1} = A_d \odot h_t + B_d \odot x_t, \qquad y_t = C \odot h_t + D \odot x_t, \tag{12}$$

where $\odot$ denotes block-wise fusion based on modal gating. Equivalently, the recurrence can be expressed as

$$H_{t+1} = \Phi_\rho(\Delta t) \odot R_\omega(\Delta t) \odot H_t + \Psi_\rho(\Delta t) \odot X_t, \tag{13}$$

with

$$\Phi_\rho(\Delta t) = e^{-\rho \Delta t}, \tag{14}$$

$$R_\omega(\Delta t) = \begin{bmatrix} \cos(\omega \Delta t) & -\sin(\omega \Delta t) \\ \sin(\omega \Delta t) & \cos(\omega \Delta t) \end{bmatrix}, \tag{15}$$

representing damping and rotation operators, respectively.

To improve stability, we use bounded differentiable reparameterizations:

$$\rho_k = \mathrm{softplus}(r_k + \epsilon_\rho) + \rho_{\min}, \tag{16}$$

$$\omega_k = \pi \cdot \mathrm{sigmoid}(o_k + F_{\log,k}), \tag{17}$$

where (r_k, o_k) are unconstrained learnable parameters. Their Jacobians are smooth:

$$\frac{\partial \rho_k}{\partial r_k} = \sigma(r_k), \qquad \frac{\partial \omega_k}{\partial o_k} = \pi \sigma(o_k)(1 - \sigma(o_k)). \tag{18}$$

The complete frequency-aware continuous-time SSM is given by

$$\dot{h}_t = \mathcal{A}(F_{\log}, \rho, \omega) h_t + B x_t, \tag{19}$$

$$h_{t+1} = e^{\mathcal{A}(F_{\log}, \rho, \omega)\Delta t} h_t + \int_0^{\Delta t} e^{\mathcal{A}(F_{\log}, \rho, \omega) s} B x_t \, ds, \tag{20}$$

$$y_t = C h_t + D x_t, \tag{21}$$

where the augmented dynamics matrix is

$$\mathcal{A}(F_{\log}, \rho, \omega) = \mathrm{blkdiag}\left(\begin{bmatrix} -\rho_k & -\tilde{\omega}_k \\ \tilde{\omega}_k & -\rho_k \end{bmatrix} \right)_{k=1}^{D_{\text{inner}}/2}. \tag{22}$$

This formulation unifies continuous-time modal dynamics with discrete selective scanning, yielding a stable, interpretable, and frequency-sensitive Mamba variant optimized for modeling oscillatory neural time series such as EEG.

C.2. Evaluation Metric Implementation

Classification Accuracy. To evaluate the classification performance of our model, we adopt a contrastive learning framework that integrates EEG signals and visual stimuli (images) for robust representation learning. Specifically, we utilize the pre-trained FSTDE model for zero-shot EEG classification across four diverse datasets: THINGS-EEG, THINGS-MEG, EEGCVPR40, and EEGImageNet. The model's performance is quantified using Top-K accuracy metrics, focusing on Top-1 and Top-5 predictions.For the THINGS-EEG and THINGS-MEG datasets, we conduct evaluations under two challenging scenarios: within-subject and leave-one-subject-out. Additionally, we assess N-Way classification accuracy (N=50, 200) by tasking the model to identify the correct class among N-1 unrelated "distractor" samples from the test set, providing a comprehensive analysis of the model's robustness in large-scale classification tasks.For the EEGCVPR40 and EEGImageNet datasets, due to the limited number of classes, we report only Top-1 and Top-5 accuracy. These evaluations highlight the model's ability to generalize across datasets with varying class distributions and experimental conditions, demonstrating its versatility and robustness in EEG-to-class mapping tasks.

Retrieval Accuracy. The retrieval task is designed to identify the Top-K images that most closely align with a given EEG signal, leveraging a shared latent space for cross-modal representation learning. Specifically, we extract image embeddings by processing candidate images through a pre-trained CLIP image encoder, which captures high-level semantic features. For EEG signals, we utilize the representations generated by the pre-trained FSTDE model, which encodes neural activity into a rich latent representation. These EEG and image embeddings are projected into the same latent space, where similarity metrics are used to measure their alignment. By bridging the gap between neural activity and visual semantics, our framework enables accurate and efficient retrieval, demonstrating its robustness in cross-modal understanding.

Generation Accuracy. The generation task represents the most challenging evaluation scenario, aiming to reconstruct meaningful visual content directly from EEG signals. Leveraging the EEG representations derived from the pre-trained FSTDE model, our framework generates images that align with the neural activity patterns captured in the EEG data. This task is evaluated across four diverse datasets: THINGS-EEG, THINGS-MEG, EEGCVPR40, and EEGImageNet, ensuring a comprehensive analysis of the model's generalization capabilities. To assess the quality of the generated images, we employ a suite of quantitative metrics that evaluate both perceptual quality and semantic fidelity. These metrics provide a holistic evaluation

Table 3. Top-1(T1) and Top-5(T5) accuracy (%) for 50-way zero-shot retrieval on THINGS-EEG.

Method	Year	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Avg	
		T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5	T1	T5
Intra-subject: train and test on one subject																							
BraVL [3]	2023	14.8	41.5	12.9	39.2	15.0	40.9	12.4	36.5	10.5	33.8	15.1	41.2	15.1	42.4	20.3	49.8	10.6	34.1	16.8	43.6	14.3	40.3
ATM [8]	2024	44.0	80.0	44.0	82.5	53.5	88.0	52.0	86.5	39.5	74.0	47.5	87.0	47.0	83.5	67.0	90.5	46.0	82.0	52.5	86.5	49.3	84.1
MB2C [17]	2024	41.3	83.3	38.7	82.7	48.7	84.7	56.0	84.7	39.3	70.0	54.7	86.7	45.3	80.7	68.7	89.3	53.3	89.3	58.7	90.7	50.5	84.2
Ours	2025	**48.3**	**86.5**	**44.8**	**84.1**	**55.6**	**89.0**	**57.2**	**89.4**	**40.3**	**75.5**	**55.7**	**87.4**	**51.3**	**88.6**	**69.1**	**92.3**	**53.2**	**89.5**	**55.1**	**88.6**	**53.1**	**87.1**
Inter-subject: leave one subject out for test																							
BraVL [3]	2023	6.4	23.0	5.0	20.7	3.9	17.8	5.6	18.6	4.7	19.4	5.7	23.1	6.3	24.1	6.0	23.9	4.6	18.7	5.9	22.8	5.4	21.2
ATM [8]	2024	25.0	58.5	22.5	51.5	18.0	42.0	21.0	56.0	16.5	44.5	26.0	56.0	22.0	61.0	20.0	52.5	16.5	45.0	28.5	64.0	21.6	53.1
MB2C [17]	2024	25.3	68.0	34.7	74.0	18.0	59.3	29.3	63.3	22.0	59.3	20.0	54.7	22.7	59.3	26.7	48.7	23.3	62.7	31.3	77.3	25.3	62.7
Ours	2025	**26.1**	**68.9**	**37.3**	**74.4**	**21.1**	**61.3**	**29.7**	**63.8**	**24.7**	**62.5**	**25.3**	**60.7**	**23.3**	**60.6**	**27.5**	**54.6**	**23.7**	**63.9**	**33.3**	**79.8**	**27.2**	**65.1**

of the alignment between the generated images and the original visual stimuli, offering insights into the model's ability to bridge the gap between neural signals and visual semantics. By reconstructing high-quality and semantically meaningful images, our framework demonstrates its potential in advancing neural decoding and cross-modal generation tasks.

C.3. Image Generation Evaluation Metrics

Inception Score (IS) [12] evaluates the quality and diversity of generated images. A higher score indicates better performance. The IS is defined as:

$$\text{IS} = \exp\left(\mathbb{E}_{\mathbf{x}\sim p_g}\left[D_{KL}\left(p(y|\mathbf{x})||p(y)\right)\right]\right), \tag{23}$$

where $\mathbf{x}$ represents a generated image, p_g is the distribution of the generator, $p(y|\mathbf{x})$ is the conditional class distribution predicted by a pre-trained Inception model, $p(y)$ is the marginal class distribution, and D_{KL} denotes the Kullback-Leibler divergence.

Fréchet Inception Distance (FID) [6] measures the similarity between the feature distributions of real and generated images. A lower FID indicates better alignment. The FID is computed as:

$$\text{FID} = ||\boldsymbol{\mu}_r - \boldsymbol{\mu}_g||^2 + \text{Tr}\left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g)^{1/2}\right), \tag{24}$$

where $\boldsymbol{\mu}_r$ and $\boldsymbol{\mu}_g$ are the mean feature vectors of real and generated images, respectively, and $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_g$ are their corresponding covariance matrices. The trace operator, Tr, computes the sum of the diagonal elements of a matrix.

Kernel Inception Distance (KID) [1] quantifies the distance between the distributions of real and generated images using the squared Maximum Mean Discrepancy (MMD). A lower KID value indicates better performance. The KID is defined as:

$$\begin{aligned}\text{KID} &= \text{MMD}^2 \\ &= \mathbb{E}_{\substack{\mathbf{x}_r,\mathbf{x}_r'\sim p_r \\ \mathbf{x}_g,\mathbf{x}_g'\sim p_g}}\left[k(\mathbf{x}_r,\mathbf{x}_r') - 2k(\mathbf{x}_r,\mathbf{x}_g) + k(\mathbf{x}_g,\mathbf{x}_g')\right],\end{aligned} \tag{25}$$

where p_r and p_g are the distributions of real and generated images, respectively, and k is a polynomial kernel function.

Structural Similarity Index Measure (SSIM) [16] evaluates the perceptual similarity between two images by considering luminance, contrast, and structural information. A higher SSIM indicates better similarity. The SSIM is calculated as:

$$\text{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{26}$$

where μ_x and μ_y are the mean pixel values of images $\mathbf{x}$ and $\mathbf{y}$, σ_x^2 and σ_y^2 are their variances, σ_{xy} is the covariance between the two images, and C_1 and C_2 are small constants to stabilize the division.

Pearson Correlation Coefficient (PCC) measures the linear correlation between the pixel values of two images. A higher PCC indicates stronger correlation. The PCC is defined as:

$$\text{PCC}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{27}$$

where x_i and y_i are the pixel values of images $\mathbf{x}$ and $\mathbf{y}$, respectively, and $\bar{x}$ and $\bar{y}$ are their mean pixel values.

C.4. Algorithmic Overview

The proposed D^2-FOSA introduces a cross-modal learning framework that bridges EEG signals and visual images through a dual-path diffusion mechanism and semantic alignment in a shared latent space. The detailed procedures of FSTDE and D^2-FOSA are provided in Algorithm 1 and

Algorithm 1 Frequency-Spatio-Temporal Dynamics Encoder (FSTDE)

Input: EEG signal $X \in \mathbb{R}^{C \times T}$
Output: EEG embedding $\mathbf{X}_e \in \mathbb{R}^d$

// FOMamba
Construct block-diagonal matrix $A = \text{blkdiag}(A_1, \ldots, A_{D_{\text{inner}}/2})$, where:

$$A_k = \begin{bmatrix} -\rho_k & -\tilde{\omega}_k \\ \tilde{\omega}_k & -\rho_k \end{bmatrix}, \quad \tilde{\omega}_k = \text{softplus}(\omega_k + F_{\log,k})$$

Discretize A using bilinear transformation:

$$A_d = (I - \tfrac{\Delta t}{2} A)^{-1}(I + \tfrac{\Delta t}{2} A),$$
$$B_d = (I - \tfrac{\Delta t}{2} A)^{-1} \Delta t B$$

for $t = 1$ to T **do**
  Update hidden state: $h_t = A_d h_{t-1} + B_d x_t$
  Compute output: $y_t = C h_t + D x_t$
end for
Extract temporal features: $\mathbf{H}_t = \{y_1, \ldots, y_T\}$
// Neural Graph Structure Extraction
Construct graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ based on EEG topology
Propagate features using GCN:

$$\mathbf{H}_s = \sigma\left(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{H}_t\mathbf{W}\right)$$

// Spatio-Temporal Feature Fusion
Apply depthwise and separable convolutions to $\mathbf{H}_s$
Flatten and project features via MLP:

$$\mathbf{X}_e = \text{MLP}(\text{Flatten}(\mathbf{H}_s))$$

return $\mathbf{X}_e$

Algorithm 2 Training and Inference Procedures of D^2-FOSA

Input: EEG signal X, paired image I
Output (in inference): Generated image $\hat{I}$

// Training Phase
Extract EEG latent $\mathbf{X}_e = f_{\text{FSTDE}}(X)$
Extract image latent $\mathbf{X}_i = f_{\text{CLIP}}(I)$
Compute contrastive loss: $\mathcal{L}_{\text{InfoNCE}}(\mathbf{X}_e, \mathbf{X}_i)$
{**I2E-DLG**: Align image latent to EEG latent}
Add noise: $\tilde{\mathbf{X}}_i = q(\mathbf{X}_i|t)$
for $t = T$ to 1 **do**
  Predict noise: $\hat{\epsilon}_\theta^{\text{I2E}}(\tilde{\mathbf{X}}_i, t, \mathbf{X}_e)$
  Update latent: $\tilde{\mathbf{X}}_i \leftarrow \text{ReverseStep}(\tilde{\mathbf{X}}_i, \hat{\epsilon}_\theta^{\text{I2E}}, t)$
end for
Reconstructed EEG latent: $\hat{\mathbf{X}}_e = \tilde{\mathbf{X}}_i$
Compute latent loss: $\mathcal{L}_{\text{latent}}^{i \to e} = \|\hat{\mathbf{X}}_e - \mathbf{X}_e\|_2^2$
{**E2I-DLG**: Align EEG latent to image latent}
Add noise: $\tilde{\mathbf{X}}_e = q(\mathbf{X}_e|t)$
for $t = T$ to 1 **do**
  Predict noise: $\hat{\epsilon}_\theta^{\text{E2I}}(\tilde{\mathbf{X}}_e, t, \mathbf{X}_i)$
  Update latent: $\tilde{\mathbf{X}}_e \leftarrow \text{ReverseStep}(\tilde{\mathbf{X}}_e, \hat{\epsilon}_\theta^{\text{E2I}}, t)$
end for
Reconstructed image latent: $\hat{\mathbf{X}}_i = \tilde{\mathbf{X}}_e$
Compute latent loss: $\mathcal{L}_{\text{latent}}^{e \to i} = \|\hat{\mathbf{X}}_i - \mathbf{X}_i\|_2^2$
Compute total loss:
  $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{InfoNCE}} + \lambda_1 \mathcal{L}_{\text{latent}}^{i \to e} + \lambda_2 \mathcal{L}_{\text{latent}}^{e \to i}$
Update all parameters using AdamW [9]
// Inference Phase
Given test EEG X, extract EEG latent: $\mathbf{X}_e = f_{\text{FSTDE}}(X)$
Initialize noise: $\mathbf{X} \sim \mathcal{N}(0, \mathbf{I})$
for $t = T$ to 1 **do**
  Predict noise: $\hat{\epsilon}_\theta^{\text{E2I}}(\mathbf{X}, t, \mathbf{X}_e)$
  Update latent: $\mathbf{X} \leftarrow \text{ReverseStep}(\mathbf{X}, \hat{\epsilon}_\theta^{\text{E2I}}, t)$
end for
Generate image: $\hat{I} = \text{SDXL}(\mathbf{X})$ [10]
return Generated image $\hat{I}$

Algorithm 2, respectively. The procedure involves the following steps:

- **Step 1: EEG Feature Encoding.** The input EEG signal $X \in \mathbb{R}^{C \times T}$ is first processed by the FOMamba encoder, which extracts fine-grained temporal features $\mathbf{H}_t$. These temporal features are then passed to the Neural Graph Structure Extractor, which models the spatial relationships of EEG electrodes based on their natural graph structure (e.g., the 10-20 system), producing spatially-aware features $\mathbf{H}_s$. Finally, the Spatio-Temporal Feature Extractor combines these spatial and temporal features to learn dynamic spatio-temporal patterns. The resulting high-level EEG embedding $\mathbf{X}_e \in \mathbb{R}^d$ is projected into the latent space for cross-modal alignment.
- **Step 2: Semantic Alignment.** The EEG embedding $\mathbf{X}_e$ and the image embedding $\mathbf{X}_i$ are projected into a shared latent space and aligned using a contrastive learning framework inspired by CLIP. Specifically, we employ an InfoNCE-based contrastive loss [15] to maximize the similarity between positive EEG-image pairs while minimizing it for negative pairs. This alignment facilitates robust cross-modal understanding, enabling the model to bridge the semantic gap between neural signals and visual representations.
- **Step 3: Dual-Path Diffusion Generation.** To achieve robust cross-modal generation, we employ two conditional latent diffusion models [7] that operate bidirectionally in

the shared latent space:

– **E2I-DLG:** This module generates an image latent representation $\hat{\mathbf{X}}_i$ from the EEG latent embedding $\mathbf{X}_e$, conditioned on Gaussian noise. The generated latent is then decoded into the final image $\hat{I}$ using a pretrained SDXL decoder [10].

– **I2E-DLG:** This module reconstructs the EEG latent embedding $\hat{\mathbf{X}}_e$ from the image latent representation $\mathbf{X}_i$, enforcing bidirectional consistency and ensuring alignment in the shared latent space.

This dual-path framework not only enhances the fidelity of generated outputs but also ensures semantic consistency between EEG and image modalities.

- **Step 4: Training Objective.** The overall training objective is designed to ensure robust cross-modal alignment and accurate reconstruction. It integrates the following components:
  – **Contrastive Loss:** $\mathcal{L}_{\text{InfoNCE}}(\mathbf{X}_e, \mathbf{X}_i)$, which maximizes the similarity between aligned EEG and image embeddings while minimizing it for unaligned pairs.
  – **EEG Reconstruction Loss:** $\mathcal{L}_{\text{latent}}^{i \to e} = \|\hat{\mathbf{X}}_e - \mathbf{X}_e\|_2^2$, which ensures the accurate reconstruction of EEG embeddings from image embeddings.
  – **Image Reconstruction Loss:** $\mathcal{L}_{\text{latent}}^{e \to i} = \|\hat{\mathbf{X}}_i - \mathbf{X}_i\|_2^2$, which ensures the accurate reconstruction of image embeddings from EEG embeddings.

  The model parameters are optimized using the AdamW optimizer [9], ensuring stable and efficient convergence.
- **Step 5: Inference Phase.** During inference, the framework focuses on efficient EEG-to-image generation by utilizing only the EEG encoder and the E2I diffusion generator. Specifically, a test EEG input X is encoded into its latent representation $\mathbf{X}_e = f_{\text{FSTDE}}(X)$ through the FSTDE encoder. This latent embedding $\mathbf{X}_e$ serves as the conditioning signal for the E2I-DLG, which generates the image latent $\hat{\mathbf{X}}_i$. Finally, the generated latent representation $\hat{\mathbf{X}}_i$ is decoded into a high-quality, realistic image $\hat{I}$ using a pretrained SDXL image decoder [10]. This streamlined process ensures robust and accurate EEG-to-image reconstruction.

D. Additional results

D.1. Effectiveness of FOMamba

THINGS-EEG. Figure 1 presents a detailed comparative topographical analysis to illustrate how our proposed FOMamba model extracts both temporal and frequency-domain information from raw EEG signals, in contrast to the standard Mamba model. This analysis is crucial for understanding the model's ability to enhance neural representations for downstream EEG-to-image reconstruction tasks. The figure is organized into two columns: the first column (a, c, e) displays the mean scalp activity for each signal type,

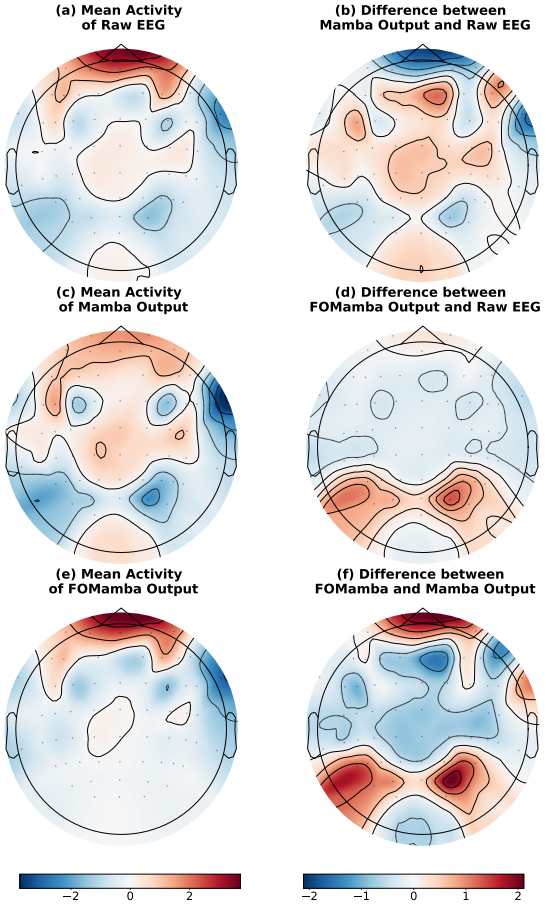

Figure 1. Topographical analysis of EEG signal processing on the THINGS-EEG dataset. This figure compares the mean scalp activity and difference maps across three signal types: Raw EEG, Mamba output, and our proposed FOMamba output, evaluated on the THINGS-EEG dataset. The first column **(a, c, e)** illustrates the mean scalp activity for each signal type, highlighting that FOMamba preserves more fine-grained spatial patterns, particularly in the frontal and occipital regions. The second column **(b, d, f)** presents the difference maps, showing that FOMamba significantly amplifies neural activity in vision-critical areas, such as the occipital and parietal regions. This targeted enhancement demonstrates FOMamba's superior ability to extract neural features relevant for EEG-to-image reconstruction tasks on the THINGS-EEG dataset.

while the second column (b, d, f) illustrates the difference maps. Figure 1 (a) displays the baseline mean activity of Raw EEG, revealing a broad and diffuse activation pattern across the scalp. This serves as the reference for evaluating subsequent processing steps. Figure 1 (c) presents the mean activity of Mamba Output, which demonstrates a slightly refined representation but lacks significant spatial focus. Figure 1 (e) shows the mean activity of *FOMamba Output*,

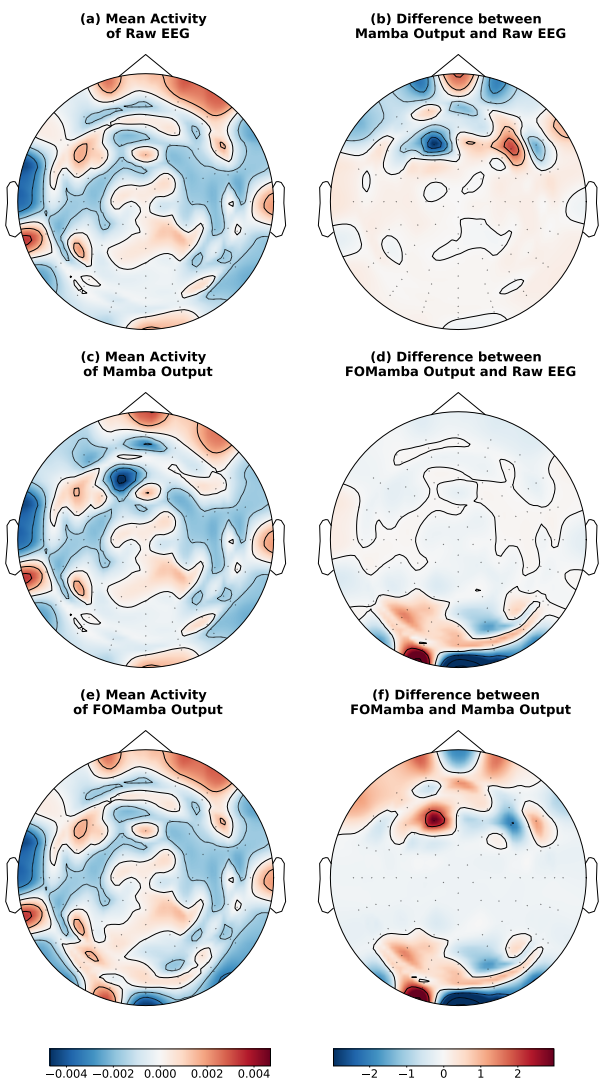

Figure 2. Topographical analysis of EEG signal processing on the EEGCVPR40 dataset. This figure compares the mean scalp activity and difference maps across three signal types: Raw EEG, Mamba output, and our proposed FOMamba output, evaluated on the EEGCVPR40 dataset. The first column **(a, c, e)** illustrates the mean scalp activity for each signal type, highlighting that FOMamba preserves more fine-grained spatial patterns, particularly in the frontal and occipital regions. The second column **(b, d, f)** presents the difference maps, showing that FOMamba significantly amplifies neural activity in vision-critical areas, such as the occipital and parietal regions. This targeted enhancement demonstrates FOMamba's superior ability to extract neural features relevant for EEG-to-image reconstruction tasks on the EEGCVPR40 dataset.

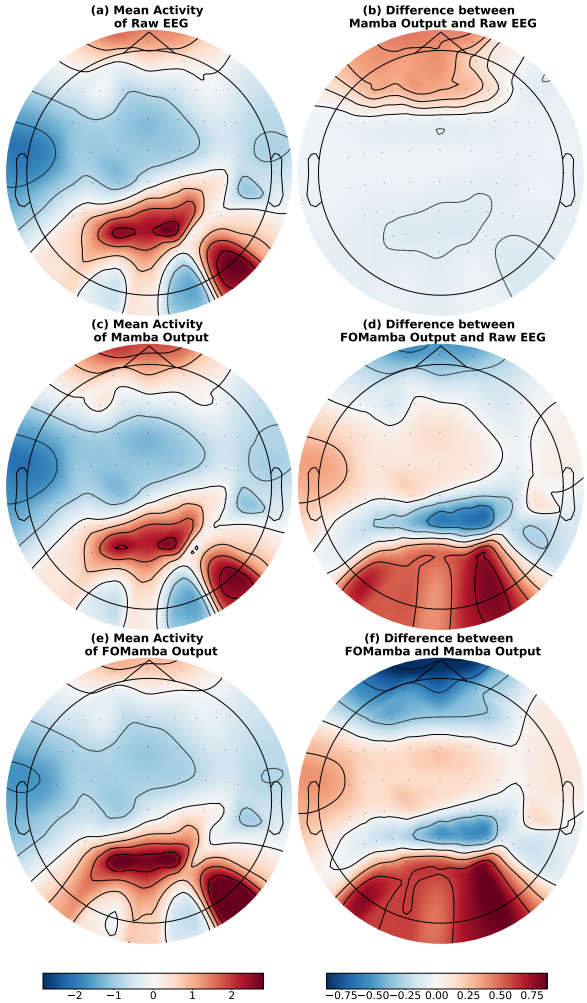

Figure 3. Topographical analysis of EEG signal processing on the EEGImageNet dataset. This figure compares the mean scalp activity and difference maps across three signal types: Raw EEG, Mamba output, and our proposed FOMamba output, evaluated on the EEGImageNet dataset. The first column **(a, c, e)** illustrates the mean scalp activity for each signal type, highlighting that FOMamba preserves more fine-grained spatial patterns, particularly in the frontal and occipital regions. The second column **(b, d, f)** presents the difference maps, showing that FOMamba significantly amplifies neural activity in vision-critical areas, such as the occipital and parietal regions. This targeted enhancement demonstrates FOMamba's superior ability to extract neural features relevant for EEG-to-image reconstruction tasks on the EEGImageNet dataset.

which produces a cleaner and more spatially focused activation map compared to both Raw EEG and Mamba Output. Figure 1 (b) shows the difference between Mamba Output and Raw EEG, indicating that Mamba largely preserves the original spatial distribution while introducing a general smoothing effect. Figure 1 (d) visualizes the differ-

ence between FOMamba Output and Raw EEG, highlighting a pronounced enhancement in the occipital and parietal regions (strong red clusters). These regions are critical for visual perception and mental imagery, indicating that FOMamba selectively amplifies neural signals relevant to visual content. Figure 1 (f) compares FOMamba Output to Mamba Output, further isolating the contribution of FOMamba. The strong positive differences in the occipital lobe confirm that FOMamba's architectural innovations are responsible for targeted enhancement of vision-related neural activity. Overall, FOMamba's frequency-oriented design actively restructures the spatial representation of EEG signals, selectively boosting activation in vision-critical cortical regions (e.g., the occipital lobe). This results in neural embeddings that are not only more discriminative but also more effective for high-fidelity EEG-to-image synthesis.

Figure 6 illustrates the temporal evolution of EEG topographic maps across ten time windows (0–1000 ms), comparing the raw EEG with the outputs of Mamba and our proposed FOMamba model for the EEG-to-image reconstruction task. The figure is organized into six rows, with each row corresponding to a specific signal type or difference map. Figure 6 (a) displays the ground-truth spatiotemporal neural activity patterns of the raw EEG, revealing broad and diffuse activation across the scalp. These patterns serve as the baseline for evaluating the reconstruction quality of subsequent models. Figure 6 (b) presents the Mamba output, which recovers the coarse spatial structure of the EEG signals but suffers from noticeable blurring and loss of high-frequency details, particularly in the occipital and frontal regions. Figure 6 (c) demonstrates the FOMamba output, which produces sharper and more temporally consistent activation patterns. FOMamba effectively preserves fine-grained spatial features, especially in vision-critical regions such as the occipital and frontal cortices, highlighting its superior ability to capture neural dynamics relevant to visual processing. Figure 6 (d) visualizes the difference map between Mamba output and raw EEG, showing large residuals across multiple regions. These residuals indicate that Mamba fails to restore fine-grained neural signals, particularly in areas associated with transient visual-related activity. Figure 6 (e) shows the difference map between FOMamba output and raw EEG, demonstrating that FOMamba substantially reduces residuals and achieves a closer match to the true neural responses. This improvement highlights FOMamba's ability to capture transient visual-related activity with higher fidelity. Figure 6 (f) compares FOMamba output directly to Mamba output, revealing positive differences concentrated in the occipital cortex. These differences confirm that FOMamba's frequency-optimized design enables it to better preserve vision-critical spatiotemporal structures, resulting in superior neural representations for subsequent image reconstruction. Overall, the figure highlights FOMamba's significant advancements in extracting spatiotemporal EEG features as well as frequency-domain characteristics. By integrating temporal, spatial, and frequency information, FOMamba captures neural dynamics more comprehensively, enabling the generation of more accurate and generative neural embeddings across all ten time windows for high-fidelity EEG-to-image synthesis.

EEGCVPR40. Figure 2 presents the topographical analysis of EEG signal processing on the EEGCVPR40 dataset, comparing the mean scalp activity and difference maps across three signal types: Raw EEG, Mamba output, and FOMamba output. The results highlight the superior ability of FOMamba to extract and enhance frequency-specific neural activity relevant to visual processing. Figure 2 (a) shows the mean scalp activity of Raw EEG, which serves as the baseline for comparison. The activation patterns are broad and diffuse, reflecting the unprocessed nature of the signals and the presence of noise across all frequency bands. Figure 2 (b) illustrates the difference map between Mamba output and Raw EEG. While Mamba preserves the overall spatial structure of the signals, it fails to selectively enhance vision-critical frequency bands, resulting in smoothing effects in the occipital and parietal regions. Figure 2 (c) depicts the mean scalp activity of Mamba output. Although the spatial patterns are more refined compared to Raw EEG, the lack of targeted enhancement in frequency bands associated with visual processing limits its effectiveness. Figure 2 (d) presents the difference map between FOMamba output and Raw EEG. FOMamba demonstrates a significant ability to amplify neural activity in the occipital and parietal regions, which are strongly associated with visual perception. This enhancement is achieved by explicitly modeling oscillatory dynamics in vision-relevant frequency bands (e.g., alpha, beta, and gamma). Figure 2 (e) visualizes the mean scalp activity of FOMamba output. The results highlight FOMamba's capability to preserve fine-grained spatial features while selectively enhancing neural signals in the frontal and occipital regions, which are critical for visual decoding. Figure 2 (f) illustrates the difference map between FOMamba output and Mamba output. The map emphasizes FOMamba's superior ability to extract and amplify vision-critical neural features, particularly in the occipital cortex, by leveraging its frequency-aware design. Overall, the results demonstrate that FOMamba achieves significant advancements in capturing spatiotemporal EEG patterns and selectively enhancing visual frequency information, making it highly effective for EEG-to-image reconstruction tasks on the EEGCVPR40 dataset.

Figure 7 presents the time-resolved topographical analysis of EEG signal processing on the EEGCVPR40 dataset, comparing the mean scalp activity and difference maps across three signal types: Raw EEG, Mamba output, and

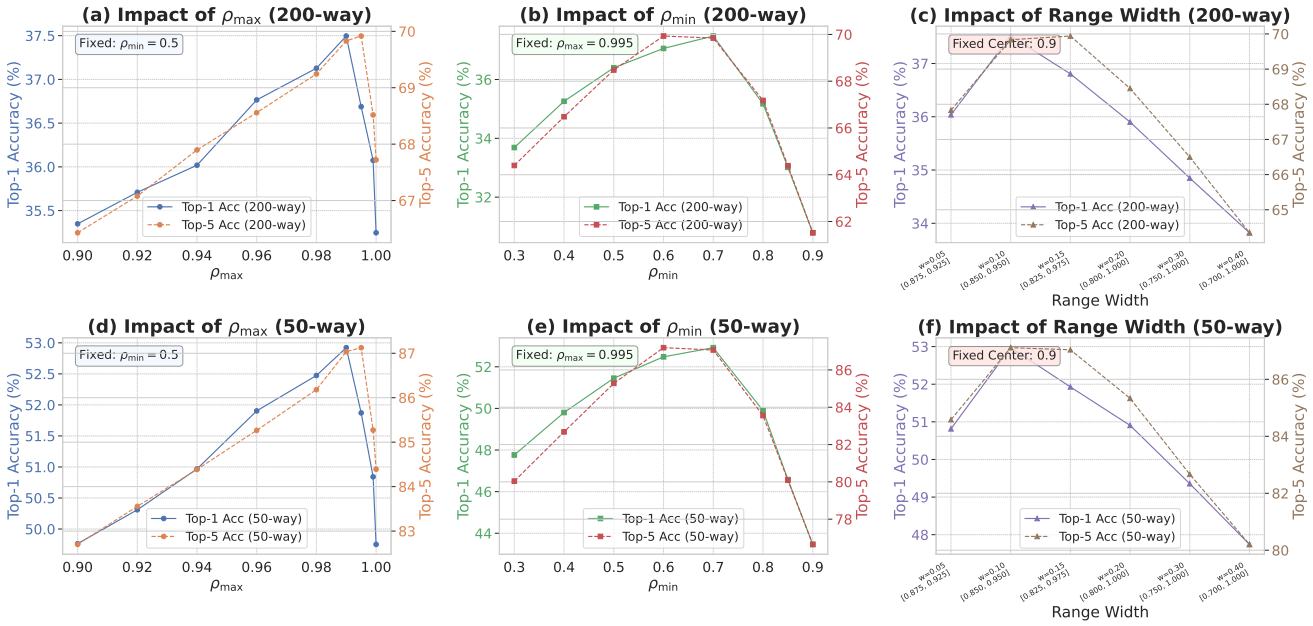

Figure 4. Impact of key hyperparameters (ρ_{max}, ρ_{min}, and range width) on Top-1 and Top-5 accuracy for EEG-to-image retrieval tasks under 200-way and 50-way settings: (a) and (d): Increasing ρ_{max} improves accuracy up to a threshold (around 0.98), after which performance drops due to overfitting. (b) and (e): Accuracy peaks when ρ_{min} is set to approximately 0.7, indicating an optimal balance between signal suppression and retention. (c) and (f): Narrower ranges centered around 0.9 yield better accuracy, suggesting that focusing on specific EEG signal ranges enhances retrieval performance.

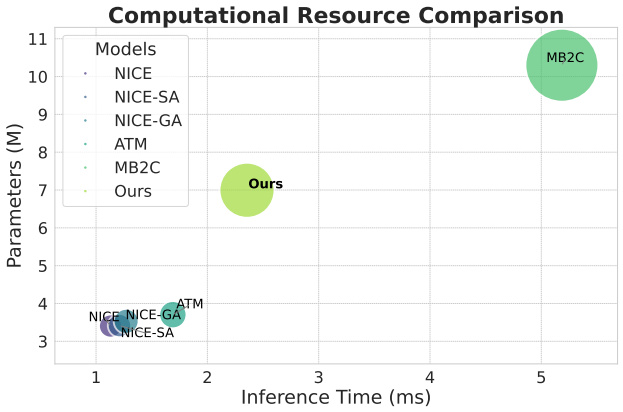

Figure 5. Computational resource comparison for EEG-to-image reconstruction models. The x-axis shows inference time (ms), and the y-axis shows parameters (M). NICE variants are efficient but limited, ATM moderately increases resources for better performance, MB2C is the most resource-intensive, while our model achieves a balance with 7M parameters and 3 ms inference time.

FOMamba output. The analysis is performed over five distinct time windows (0-100 ms, 100-200 ms, 200-300 ms, 300-400 ms, and 400-500 ms), highlighting the temporal evolution of neural activity. The results demonstrate the superior ability of FOMamba to extract and enhance frequency-specific neural activity relevant to visual processing. Figure 7 (a) illustrates the mean scalp activity of Raw EEG across the five time windows. The activation patterns are broad and diffuse, reflecting the unprocessed nature of the signals and the presence of noise across all frequency bands. Temporal dynamics are visible but lack spatial specificity. Figure 7 (b) shows the mean scalp activity of Mamba output. While Mamba refines the spatial patterns compared to Raw EEG, it fails to selectively enhance vision-critical frequency bands, resulting in less pronounced activity in the occipital and parietal regions across all time windows. Figure 7 (c) visualizes the mean scalp activity of FOMamba output. FOMamba demonstrates its ability to preserve fine-grained spatial features while selectively enhancing neural signals in the frontal and occipital regions. This enhancement is particularly evident in the 100-300 ms time window, which is strongly associated with visual perception. The improvement is achieved by explicitly modeling oscillatory dynamics in vision-relevant frequency bands (e.g., alpha, beta, and gamma). Figure 7 (d) presents the difference maps between Mamba output and Raw EEG across the time windows. The maps highlight that Mamba introduces smoothing effects, particularly in the occipital and parietal regions, but fails to amplify neural activity in frequency bands critical for visual processing. Figure 7 (e) shows the difference maps between FOMamba output and

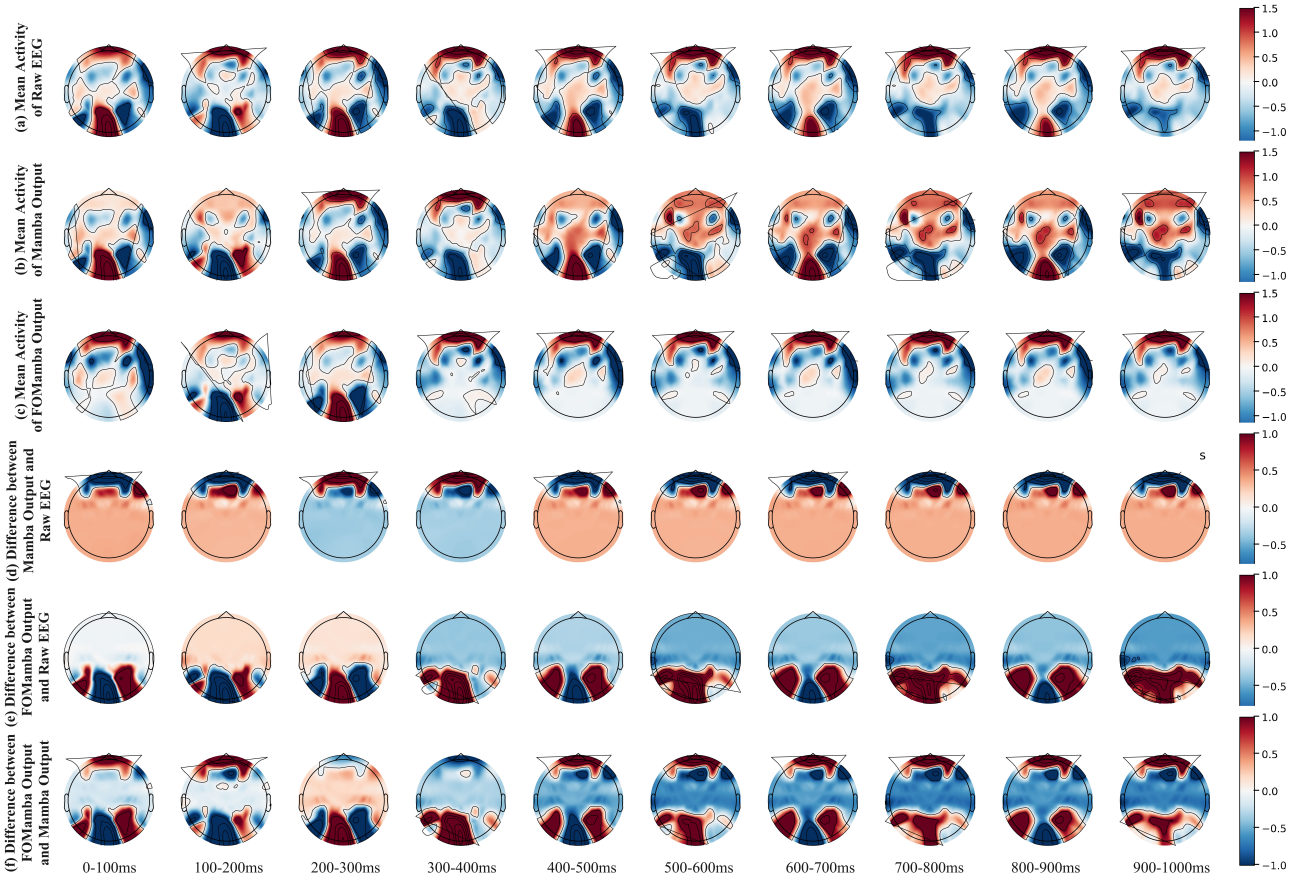

Figure 6. Temporal evolution of EEG topographic maps for raw EEG, Mamba output, and FOMamba output across nine time windows (0-1000 ms) on the THINGS-EEG dataset. Rows (a)-(c) show the mean activity of raw EEG, Mamba, and FOMamba outputs, respectively, evaluated on the THINGS-EEG dataset. Rows (d)-(f) illustrate the difference maps between model outputs and raw EEG, as well as between FOMamba and Mamba outputs. FOMamba output (row c) preserves more fine-grained spatial and temporal EEG patterns, especially in the frontal and occipital regions. The difference maps (row f) further highlight that FOMamba captures neural dynamics closer to the raw EEG, demonstrating its superior ability for EEG-to-image reconstruction tasks on the THINGS-EEG dataset.

Raw EEG. FOMamba significantly amplifies neural activity in the occipital and parietal regions, particularly during the 100-300 ms time window, demonstrating its ability to enhance signals relevant to visual processing by leveraging its frequency-aware design. Figure 7 (f) illustrates the difference maps between FOMamba output and Mamba output. The maps emphasize FOMamba's superior ability to extract and amplify vision-critical neural features, particularly in the occipital cortex, compared to Mamba. This advantage is consistent across all time windows, with the most significant improvements observed in the 100-300 ms range. Overall, the results demonstrate that FOMamba achieves significant advancements in capturing spatiotemporal EEG patterns and selectively enhancing visual frequency information, making it highly effective for EEG-to-image reconstruction tasks on the EEGCVPR40 dataset.

EEGImageNet. Figure 3 presents the topographical analysis of EEG signal processing on the EEGImageNet dataset, comparing the mean scalp activity and difference maps across three signal types: Raw EEG, Mamba output, and FOMamba output. The visualization highlights the superior ability of FOMamba to extract and enhance frequency-specific neural activity relevant to visual processing. Figure 3 (a) illustrates the mean scalp activity of Raw EEG, which serves as the baseline for comparison. The activation patterns are broad and diffuse, reflecting the unprocessed nature of the signals and the presence of noise across all frequency bands. Figure 3 (b) presents the difference map between Mamba output and Raw EEG. While Mamba preserves the overall spatial structure of the signals, it fails to selectively enhance vision-critical frequency bands, resulting in smoothing effects in the occipital and parietal regions. Figure 3 (c) depicts the mean scalp activity of Mamba output. Although the spatial patterns are more refined com-

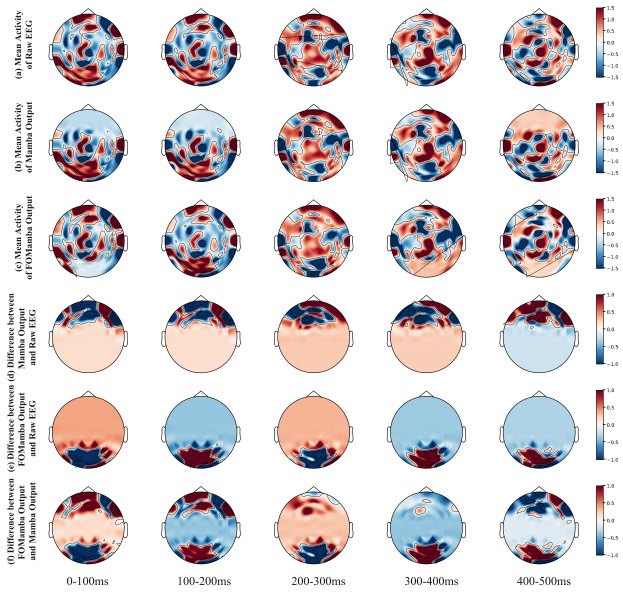

Figure 7. Temporal evolution of EEG topographic maps for raw EEG, Mamba output, and FOMamba output across nine time windows (0-1000 ms) on the EEGCVPR40 dataset. Rows (a)-(c) show the mean activity of raw EEG, Mamba, and FOMamba outputs, respectively, evaluated on the EEGCVPR40 dataset. Rows (d)-(f) illustrate the difference maps between model outputs and raw EEG, as well as between FOMamba and Mamba outputs. FOMamba output (row c) preserves more fine-grained spatial and temporal EEG patterns, especially in the frontal and occipital regions. The difference maps (row f) further highlight that FOMamba captures neural dynamics closer to the raw EEG, demonstrating its superior ability for EEG-to-image reconstruction tasks on the EEGCVPR40 dataset.

pared to Raw EEG, the lack of targeted enhancement in frequency bands associated with visual processing limits its effectiveness. Figure 3 (d) shows the difference map between FOMamba output and Raw EEG. FOMamba demonstrates a significant ability to amplify neural activity in the occipital and parietal regions, which are strongly associated with visual perception. This enhancement is achieved by explicitly modeling oscillatory dynamics in vision-relevant frequency bands (e.g., alpha, beta, and gamma). Figure 3 (e) visualizes the mean scalp activity of FOMamba output. The results highlight FOMamba's capability to preserve fine-grained spatial features while selectively enhancing neural signals in the frontal and occipital regions, which are critical for visual decoding. Figure 3 (f) illustrates the difference map between FOMamba output and Mamba output. The map emphasizes FOMamba's superior ability to extract and amplify vision-critical neural features, particularly in the occipital cortex, by leveraging its frequency-aware design. Overall, the results demonstrate that FOMamba achieves significant advancements in capturing spatiotemporal EEG patterns and selectively enhancing visual frequency information, making it highly effective for EEG-to-image reconstruction tasks on the EEGImageNet dataset.

Figure 8 presents the time-resolved topographical analysis of EEG signal processing on the EEGImageNet dataset, comparing the mean scalp activity and difference maps across three signal types: Raw EEG, Mamba output, and FOMamba output. The analysis spans ten distinct time windows (0-100 ms, 100-200 ms, ..., 900-1000 ms), capturing the temporal evolution of neural activity. The results highlight the superior ability of FOMamba to extract and enhance frequency-specific neural activity relevant to visual processing. Figure 8 (a) illustrates the mean scalp activity of Raw EEG across the ten time windows. The activation patterns are broad and diffuse, reflecting the unprocessed nature of the signals and the presence of noise across all frequency bands. Temporal dynamics are visible but lack spatial specificity, particularly in vision-critical regions such as the occipital cortex. Figure 8 (b) shows the mean scalp activity of Mamba output. While Mamba refines the spatial patterns compared to Raw EEG, it fails to selectively enhance vision-critical frequency bands, resulting in less pronounced activity in the occipital and parietal regions across all time windows. Figure 8 (c) visualizes the mean scalp activity of FOMamba output. FOMamba demonstrates its ability to preserve fine-grained spatial features while selectively enhancing neural signals in the frontal and occipital regions. This enhancement is particularly evident in the 100-300 ms time window, which is strongly associated with visual perception. The improvement is achieved by explicitly modeling oscillatory dynamics in vision-relevant frequency bands (e.g., alpha, beta, and gamma). Figure 8 (d) presents the difference maps between Mamba output and Raw EEG across the ten time windows. The maps highlight that Mamba introduces smoothing effects, particularly in the occipital and parietal regions, but fails to amplify neural activity in frequency bands critical for visual processing. Figure 8 (e) shows the difference maps between FOMamba output and Raw EEG. FOMamba significantly amplifies neural activity in the occipital and parietal regions, particularly during the 100-300 ms time window, demonstrating its ability to enhance signals relevant to visual processing by leveraging its frequency-aware design. Figure 8 (f) illustrates the difference maps between FOMamba output and Mamba output. The maps emphasize FOMamba's superior ability to extract and amplify vision-critical neural features, particularly in the occipital cortex, compared to Mamba. This advantage is consistent across all time windows, with the most significant improvements observed in the 100-300 ms range. Overall, the results demonstrate that FOMamba achieves significant advancements in capturing spatiotemporal EEG patterns and selectively enhancing visual frequency information, making it highly effective for

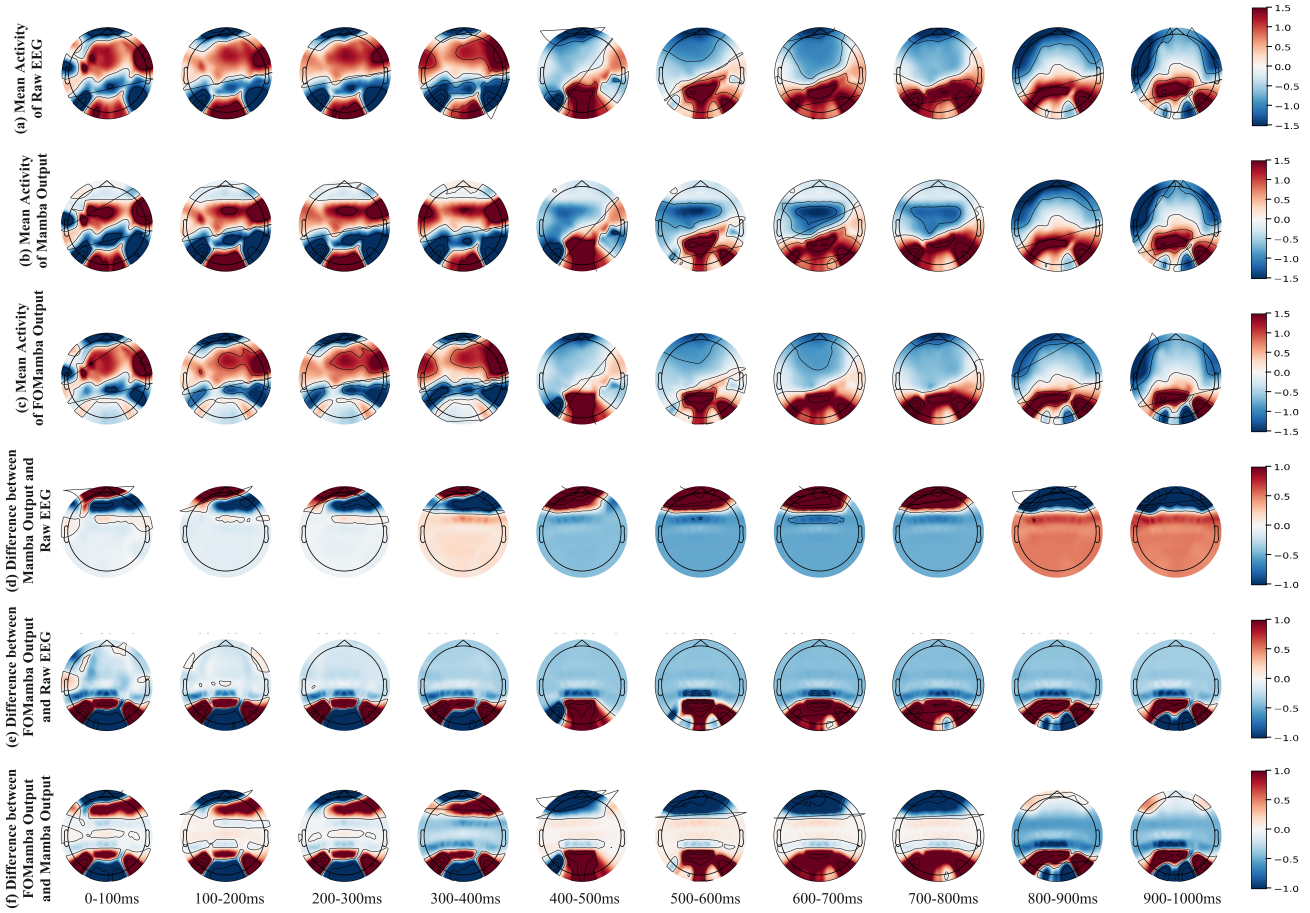

Figure 8. Temporal evolution of EEG topographic maps for raw EEG, Mamba output, and FOMamba output across nine time windows (0-1000 ms) on the EEGImageNet dataset. Rows (a)-(c) show the mean activity of raw EEG, Mamba, and FOMamba outputs, respectively, evaluated on the EEGImageNet dataset. Rows (d)-(f) illustrate the difference maps between model outputs and raw EEG, as well as between FOMamba and Mamba outputs. FOMamba output (row c) preserves more fine-grained spatial and temporal EEG patterns, especially in the frontal and occipital regions. The difference maps (row f) further highlight that FOMamba captures neural dynamics closer to the raw EEG, demonstrating its superior ability for EEG-to-image reconstruction tasks on the EEGImageNet dataset.

EEG-to-image reconstruction tasks on the EEGImageNet dataset.

D.2. Retrieval Performance

Table 3 reports the Top-1 (T1) and Top-5 (T5) accuracy for 50-way zero-shot retrieval across ten subjects on the THINGS-EEG dataset. Both intra-subject (train and test on the same subject) and inter-subject (leave-one-subject-out testing) settings are considered. In the intra-subject setting, FOMamba achieves the highest average Top-1 (53.1%) and Top-5 (87.1%) accuracy, outperforming prior methods such as BraVL, ATM, and MB2C. Notably, FOMamba demonstrates consistent improvements across all subjects, with particularly strong performance on Subject 8 (T1: 69.1%, T5: 92.3%), highlighting its ability to capture subject-specific neural dynamics. In the inter-subject setting, FOMamba also achieves the best average Top-1 (27.2%) and Top-5 (65.1%) accuracy, surpassing the previ-

Table 4. Top-1(T1) and Top-5(T5) accuracy (%) on THINGS-MEG. Results for 200-way and 50-way retrieval are marked with † and ‡, respectively.

Method	Year	Subject 1		Subject 2		Subject 3		Subject 4		Avg	
		T1	T5	T1	T5	T1	T5	T1	T5	T1	T5
Intra-subject: train and test on one subject											
NICE [13]†	2024	9.6	27.8	18.5	47.8	14.2	41.6	9.0	26.6	12.8	36.0
NICE-SA [13]†	2024	9.8	27.8	18.6	46.4	10.5	38.4	11.7	27.2	12.7	35.0
NICE-GA [13]†	2024	8.7	30.5	21.8	56.6	16.5	49.7	10.3	32.3	14.3	42.3
UBP [18]†	2025	15.0	**38.0**	46.0	80.5	27.3	**59.0**	18.5	43.5	26.7	55.2
Ours†	2025	**16.2**	**37.5**	**47.8**	**80.9**	**27.5**	**58.9**	**18.8**	**45.7**	**27.5**	**55.7**
Ours‡	2025	**31.5**	**71.3**	**67.9**	**88.4**	**48.2**	**87.6**	**33.7**	**87.0**	**45.3**	**83.5**
Inter-subject: leave one subject out for test											
UBP [18]†	2025	2.0	5.7	1.5	**17.2**	2.7	10.5	2.5	8.0	2.2	10.4
Ours†	2025	**3.2**	**6.3**	**2.1**	14.8	**3.6**	**13.7**	**4.2**	**8.6**	**3.2**	**10.8**
Ours‡	2025	**7.4**	**13.6**	**4.3**	**23.7**	**6.2**	**21.8**	**7.9**	**17.3**	**6.4**	**19.1**

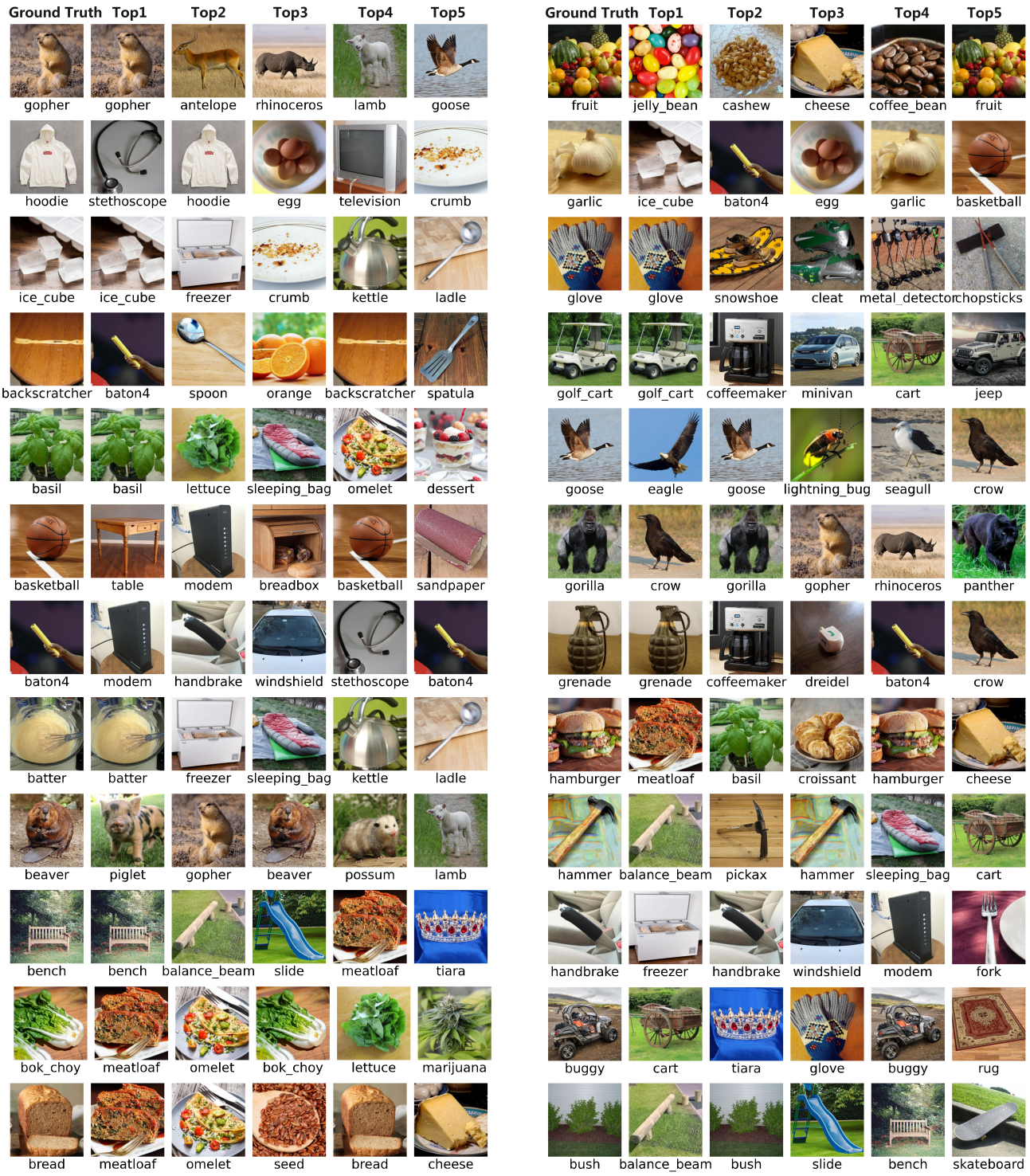

Figure 9. Additional retrieval results.

ous best method (MB2C) by a significant margin. The improvements in cross-subject generalization underscore FO-Mamba's robustness and its superior representation learning capabilities for EEG-to-image reconstruction.

Table 4 presents the results on the THINGS-MEG dataset for both 200-way and 50-way retrieval tasks. In the 200-

way retrieval task, FOMamba achieves the highest Top-1 and Top-5 accuracy in both intra- and inter-subject scenarios, outperforming NICE, NICE-SA, NICE-GA, and UBP baselines. For example, in the intra-subject setting, FOMamba achieves an average Top-1 accuracy of 27.5% and Top-5 accuracy of 55.7%, demonstrating its ability to handle the increased complexity of the 200-way retrieval task. In the 50-way retrieval task, FOMamba further extends its lead, achieving an average Top-1 accuracy of 45.3% and Top-5 accuracy of 83.5% in the intra-subject setting. In the inter-subject setting, FOMamba achieves an average Top-1 accuracy of 6.4% and Top-5 accuracy of 19.1%, significantly outperforming UBP and other baselines. These results highlight FOMamba's effectiveness in addressing neural variability across subjects and its robustness in challenging retrieval scenarios.

Overall, the results in Table 3 and Table 4 demonstrate that FOMamba achieves state-of-the-art performance in EEG/MEG-based image retrieval tasks, excelling in both accuracy and generalization. These findings validate the model's ability to learn robust and transferable neural representations for cross-modal decoding. In addition, Figure 9 presents additional examples of EEG-to-image retrieval results. These cases further demonstrate the model's cross-modal retrieval capability across various categories. Even in challenging scenarios with large intra-class variations or visual distractions, the model consistently retrieves semantically aligned images, highlighting the robustness and generalization of the learned representations.

D.3. Latent Diffusion Generation

Figure 10, 11, and 12 showcase representative EEG-to-image generation results produced by our D^2-FOSA model, highlighting its ability to decode neural signals into visually coherent representations. Figure 10 presents high-quality reconstructions that closely align with the semantic content of the original visual stimuli. These generated images exhibit clear object shapes, textures, and spatial arrangements, demonstrating the model's capability to capture fine-grained neural representations and translate them into accurate visual outputs. Figure 11 illustrates medium-quality reconstructions where the global structure of the original stimuli is reasonably preserved, but fine-grained details such as textures and contours are less precise or slightly distorted. These examples reflect the model's ability to maintain semantic coherence despite challenges in decoding subtle neural patterns. Figure 12 displays low-quality reconstructions where the generated images deviate significantly from the target semantics. These examples exhibit vague contours, weak structural coherence, and reduced fidelity to the original stimuli, highlighting the inherent variability in EEG signal quality and the challenges of cross-modal generation from brain signals. Overall, these

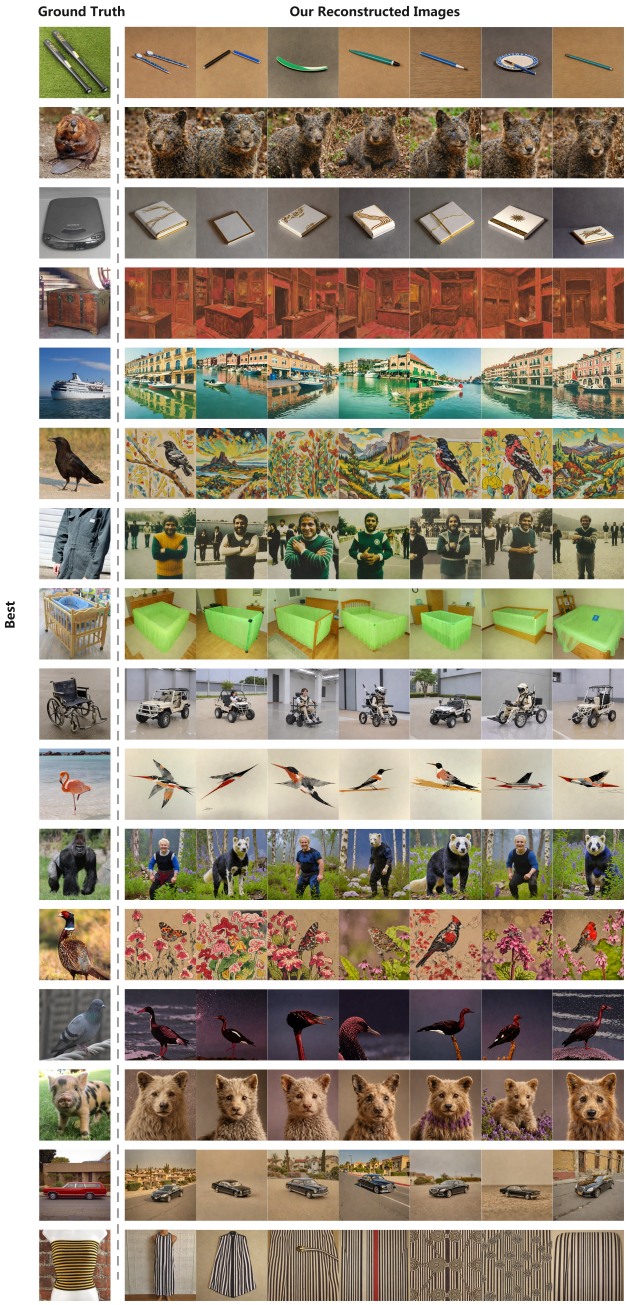

Figure 10. Additional image reconstruction results with the best alignment to original images.

results demonstrate the robustness of our D^2-FOSA model in handling diverse neural signal qualities and its potential for advancing EEG-based visual decoding. The variability across different reconstruction qualities underscores the complexity of translating brain signals into meaningful visual representations, paving the way for future improvements in cross-modal generation frameworks.

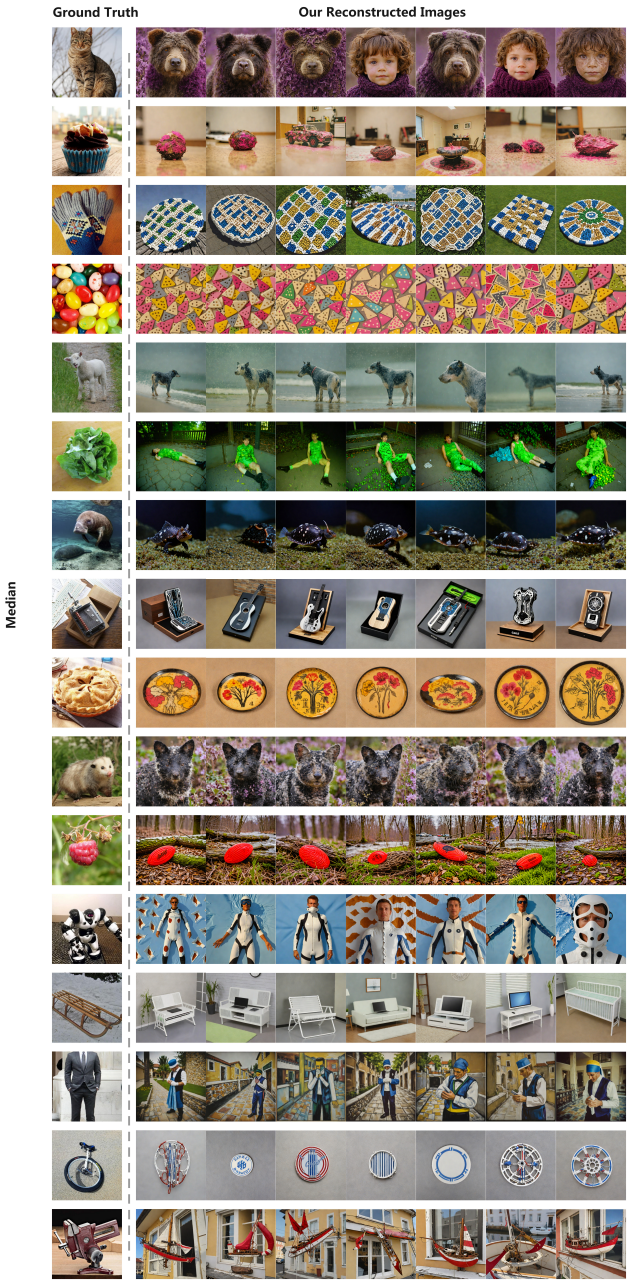

Figure 11. Additional image reconstruction results with the median alignment to original images.

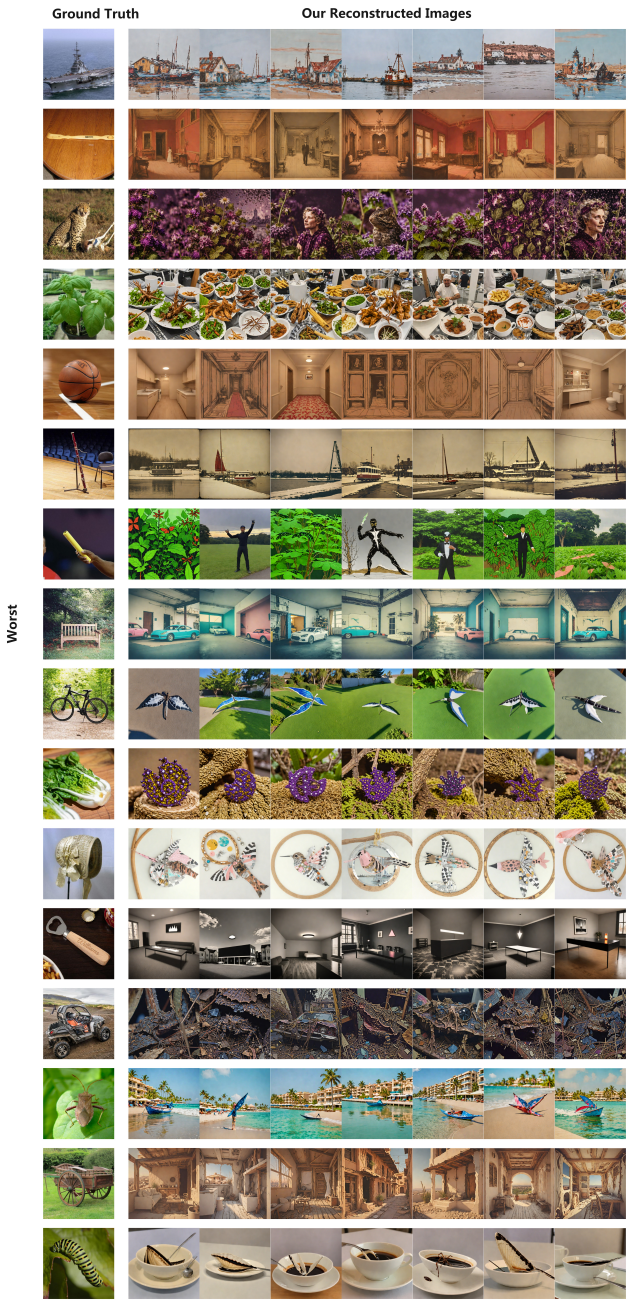

Figure 12. Additional image reconstruction results with the worst alignment to original images.

D.4. Sensitivity Analysis

Figure 4 provides a comprehensive sensitivity analysis of the key hyperparameters in the FOMamba frequency module, including $\rho_{\max}$, $\rho_{\min}$, and the range width, and their impact on Top-1 and Top-5 retrieval accuracy under both 200-way and 50-way settings. Figure 4 (a) and (d) demonstrate the effect of $\rho_{\max}$, showing that increasing $\rho_{\max}$ enhances retrieval accuracy up to approximately 0.98. Beyond this threshold, performance declines due to over-regularization, which suppresses high-frequency dynamics critical for capturing fine-grained neural patterns. This highlights the importance of balancing high-frequency preservation and regularization. Figure 4 (b) and (e) analyze the impact of $\rho_{\min}$, revealing that optimal accuracy is achieved near 0.7. This suggests that $\rho_{\min}$ effectively balances noise suppression and the retention of informative temporal structures,

ensuring robust neural decoding across diverse signal qualities. Figure 4 (c) and (f) explore the influence of range width, indicating that narrower ranges centered around 0.9 consistently yield superior retrieval accuracy. This result underscores the importance of constraining the damping parameters within a focused band to enhance the model's ability to capture meaningful oscillatory patterns and neural dynamics. Overall, these findings emphasize the critical role of precise hyperparameter tuning in the FOMamba frequency module for maximizing EEG-to-image retrieval performance. By optimizing $\rho_{\max}$, $\rho_{\min}$, and range width, the model achieves a fine balance between noise suppression, temporal structure preservation, and high-frequency dynamics, enabling robust cross-modal decoding.

D.5. Computation Resource Analysis

Figure 5 presents a comprehensive comparison of computational efficiency and modeling capacity across state-of-the-art EEG-to-image reconstruction models. The x-axis represents inference latency (ms), while the y-axis denotes the number of parameters (in millions). The bubble size reflects computational complexity, highlighting the trade-offs between efficiency and performance. NICE, NICE-SA, and NICE-GA occupy the low-resource region, characterized by minimal parameter counts and fast inference times. However, their limited modeling capacity constrains their ability to handle complex reconstruction tasks, resulting in suboptimal performance. ATM moderately increases computational cost, leveraging architectural improvements to achieve better reconstruction accuracy while maintaining reasonable efficiency. MB2C achieves the highest reconstruction quality among existing methods but at the expense of significant computational overhead, with over 10 million parameters and the slowest inference time, making it less suitable for real-time applications. In contrast, our proposed model strikes an optimal balance between efficiency and effectiveness. With approximately 7 million parameters and an inference latency of around 3 ms, it delivers superior reconstruction quality and retrieval accuracy, outperforming prior methods in both computational efficiency and modeling capacity. This balance positions our model as the most practical and scalable solution for real-time EEG-to-image decoding, paving the way for advancements in cross-modal neural representation learning.

References

[1] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 5

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1

[3] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 5

[4] Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 2022. 1, 2

[5] 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*, 2023. 1, 2

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 5

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 6

[8] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. In *Advances in Neural Information Processing Systems*, 2024. 2, 5

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6, 7

[10] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 6, 7

[11] Mary C Potter and Ellen I Levy. Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 1976. 1

[12] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016. 5

[13] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. In *International Conference on Learning Representations*, 2024. 1, 2, 13

[14] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2

[15] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[16] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 5

[17] Yayun Wei, Lei Cao, Hao Li, and Yilin Dong. Mb2c: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In *Proceedings of the ACM International Conference on Multimedia*, 2024. 2, 5

[18] Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 13

[19] Shuqi Zhu, Ziyi Ye, Qingyao Ai, and Yiqun Liu. Eegimagenet: An electroencephalogram dataset and benchmarks with image visual stimuli of multi-granularity labels. *arXiv preprint arXiv:2406.07151*, 2024. 1, 2