A CLOSED-LOOP EEG-BASED VISUAL STIMULATION FRAMEWORK FROM CONTROLLABLE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in artificial neural networks (ANNs) have significantly improved our ability to predict neural activities in the ventral visual stream of the human brain in response to visual stimuli. However, designing visual stimuli to elicit specific neural responses remains a considerable challenge due to high experimental costs, the high dimensionality of stimuli, and an incomplete understanding of neuronal selectivity. To address these challenges, we propose a novel **electroencephalography (EEG)-based closed-loop framework for visual stimulus optimization**. This framework integrates an EEG encoder, treated as a non-differentiable black-box model, to predict neural activity evoked by visual stimuli. By leveraging this encoder, we directly analyze the relationship between the visual stimuli and the desired neural responses. Through the combination of EEG feature extractors and a generation/retrieval module, the framework is theoretically capable of exploring an infinite space of natural image stimuli to identify the one that maximally activates neural activity aligned with a targeted brain state. Our experimental validation demonstrates that, regardless of the precision of ANN-predicted brain coding, the proposed framework effectively identifies the theoretically optimal natural image stimulus within a fixed number of iterations. Furthermore, the approach exhibits strong generalization across various target neural activity patterns, highlighting its robustness and potential for broader applications in brain-inspired stimulus optimization. Our code is available at `https://anonymous.4open.science/status/closed-loop-F2E9`.

1 INTRODUCTION

Previous research has demonstrated that human visual perception exhibits a certain degree of selectivity. (Epstein & Kanwisher, 1998; Qiu et al., 2023) indicate that higher-level visual cortex regions preferentially process complex semantic categories. However, attempts to map this selectivity based on responses to a fixed set of stimuli face inherent limitations, as they only reveal selectivity for the specific attributes represented within the sampled stimuli (Luo et al., 2024a). Furthermore, using manually selected synthetic stimuli can introduce biases and may not fully capture the rich complexity and variability of natural scenes. To overcome these challenges, an alternative approach is to use counterfactual reasoning by generating visual stimuli tailored to evoke specific, desired neural features, we can greatly reduce bias introduced by human factors.

The idea of eliciting and regulating specific brain activity holds significant potential for various applications such as clinical disease treatment and therapeutic interventions. As a result, an increasing number of researchers are working to develop frameworks aimed at achieving precise control of brain activity. For example, (Ponce et al., 2019; Walker et al., 2019) aim to regulate activity at the neuronal level, though these methods often lack generalization and fail to encompass the full range of visual features due to biases in training datasets. More recently, (Luo et al., 2024b) introduced VEP Booster, an AI framework designed to generate reliable and stable EEG biomarkers under visual stimulation protocols. Although stroboscopic visual stimulation methods like this advance neural targeting, they may still lack alignment with the prior knowledge embedded in the human visual perception system.

A series of studies on extensive natural image datasets (Hebart et al., 2019) and pre-trained image generation models (Rombach et al., 2022) allow us to further leverage state-of-the-art diffusion mod-

els to identify fine-grained brain functional specializations in an objective and data-driven manner. In this work, we established a novel closed-loop framework as illustrated in Figure 1 containing the black box modeling, feature extraction, and visual stimulus retrieval or generation. Our contributions are summarized as follows:

- We present a cutting-edge closed-loop visual neurofeedback framework that synthesizes high-level images to achieve the control objective on brain activity signatures. Our framework can direct mapping between synthetic visual stimuli and specific brain signatures in visual processing regions.
- By replacing traditional human EEG experiments with brain activity predicted by the black-box model (as the surrogate brain), we minimize dataset bias and enhance the model's ability to generalize to novel stimuli, providing insights for experiments on human subjects.
- We leverage state-of-the-art diffusion models to identify fine-grained brain functional specializations, and incorporate natural image priors to improve generalization, which can be flexibly designed according to the specific control goal, such as image retrieval to approximate the neural activity generated by reference image.

2 Related Work

Mapping Selectivity and Invariance from EEG. Modern neuroscience posits that specific regions of the brain exhibit distinct sensitivities or preferences for particular types of stimuli (Tesileanu et al., 2022). Selectivity refers to the phenomenon where neurons or neural networks in these regions display a marked preference for specific visual inputs, responding more strongly or consistently to them. For example, (Luo et al., 2024a) refers to the phenomenon where neurons or neural networks in these regions display a marked preference for specific visual inputs, responding more strongly or consistently to them. On the other hand, invariance refers to the brain's ability to maintain consistent neural responses to different stimuli that effectively convey the same information. In other words, multiple distinct stimuli can elicit similar brain activities (Baroni et al., 2023). In order to investigate the intrinsic invariance shared between artificial neural networks and the brain, (Feather et al., 2023) proposed a method to generate model equivalent stimuli (also known as model metamers). These stimuli produce the same neuronal activation as a reference stimulus, enabling the exploration of the internal states of AI models and their alignment with neural processes.

Closed-loop Control of Brain Activity. Closed-loop control of brain activity is a sophisticated approach for regulating brain function by leveraging real-time monitoring and feedback mechanisms. This method holds great promise in neuroscience and neural engineering, particularly for developing advanced treatments for neurological disorders such as epilepsy, Parkinson's disease, and depression. Traditionally, studies in this area have employed cutting-edge algorithms to enhance the efficiency and precision of signal processing and decision-making, thereby advancing the intelligence of closed-loop systems. (Bashivan et al., 2019) applied gradient descent to optimize the characteristics of target neuron excitation or inhibition and used the resulting gradients to update the ANN-based stimulus image generator, effectively regulating the activity of specific target neurons. (Walker et al., 2019) proposed an innovative experimental paradigm called "inception loops", which combines *in vivo* recordings with in silico modeling to synthesize optimal visual stimuli that can stimulate specific neuronal responses. (Luo et al., 2024b) employed a closed-loop strategy wherein a trained generative model to continuously refine the VEP image of the biomarker. This iterative process produced higher-quality EEG data, demonstrating the utility of closed-loop methods in improving biomarker-driven optimization framework.

Brain-conditioned Image generation. Gradient-based brain condition generation is becoming a pivotal technique in optimizing visual stimulus design, particularly for neurofeedback and brain-computer interface (BCI) applications (Luo et al., 2024b;a). This method relies on iteratively refining stimuli by backpropagating the gradients of neural activity representations to steer brain states toward desired conditions or achieve specific cognitive effects. Such an approach enables precise, adaptive stimulus optimization in response to real-time neural feedback, forming the basis for personalized brain modulation.

ICLR_2025/figures/fig1.pdf

Figure 1: **Conceptualization.** The closed-loop visual stimulation framework includes three core components. (1) The *black box model* is used as a surrogate brain to generate neural responses to visual stimulation, and can be replaced by EEG data recorded from human participants in real closed-loop experiments. (2) The *feature extractor* extracts the brain features associated with the target neural activity, which can be designed flexibly according to specific control goals. (3) The *controllable image generator* to synthesize some candidate images. Through closed-loop iteration, the system continuously refines the visual stimulation to achieve the desired brain response.

Recent advances have expanded the scope of gradient-based techniques by integrating more sophisticated neural models and leveraging high-dimensional neural representations captured by EEG, fMRI (Gu et al., 2023), and other brain imaging modalities. These advances have significantly enhanced the precision of stimulus generation, accounting for individual variability in neural responses. Moreover, by incorporating deep learning models, such as guided diffusion models (Ye et al., 2023), researchers can now generate highly detailed and context-specific stimuli tailored to align closely with target neural states, further advancing the field of brain condition generation.

3 METHOD

We aim to find the optimal stimulus image through either image retrieval or editing within a searchable space to produce specific neural activity in the brain. This closed-loop system is adaptable to various control objectives, enabling it to perform a wide range of tasks.Configuring the feature extractor for semantic representations enables image retrieval tasks, allowing the system to progressively locate the optimal stimulus image from a dataset. Alternatively, aligning the feature extractor with Power Spectral Density (PSD) features facilitates iterative image generation, producing stimuli

tailored to evoke the desired neural activity. Specifically, if the roulette wheel algorithm repeatedly selects images with specific colors or textures, the system recognizes how relevant of these features to the target class and assigns them greater weight in subsequent iterations. Through this closed-loop process, the system refines the visual stimulus to evoke the desired EEG responses. We illustrate our overall framework in Figure 1.

3.1 Closed-loop Framework

Let the EEG signals be denoted as X, where T represents the length of the time window of the data, C is the number of EEG channels, and Ω denotes a database of N images, labeled $1, 2, ..., N$ for simplicity. Concurrently, we use the encoding model g to predict brain activity signal $X = g(U) \in \mathbb{R}^{N\times C\times T}$. Our objective is to derive brain activity embeddings $Y = f(g(U)) \in \mathbb{R}^{N\times F}$ from the images $I \in \mathbb{R}^{N\times 3\times H\times W}$, where f is the feature mapping function from X to Y, U is the set of stimulus images set, and F represents the dimension of embedding. Our iteration process can be approximated as a value-based iterative Markov Decision Process (MDP). The state is represented as the probability distribution of each image $P(u)$ in the image database belonging to target category u_{target}. The state updated after each iteration corresponds to a state transition in the MDP. In each iteration, the framework determines which image to select, represented as an action in the MDP. In our model, let $j \in [\![1, N]\!]$, the reward is defined as the similarity score between the selected or generated image u_i from database and the features of the target category u_{target}:

$$sim\langle u_j, u_{\text{target}}\rangle = \frac{f\left(g(u_j)\right) \cdot f\left(g(u_{\text{target}})\right)}{\|f\left(g(u_j)\right)\| \|f\left(g(u_{\text{target}})\right)\|} \tag{1}$$

Let u_i be any image in the search space, which is the target of model evaluation. During the iteration of the t to $t+1$ step, we update $S_{t+1}(u_i)$ based on u_i. The weight coefficient α controls the cumulative probability increment. Let u_+ be the image that the system considers to be closest to the target category by computing EEG feature similarity. For the history subset H of selected images k, the posterior probability that u_i is the most similar to the target image is updated as follows:

$$S_{t+1}(u_i) = \alpha \cdot S_t(u_i) + (1-\alpha) \cdot \frac{\exp\left(s(u_+, u_i)\right)}{\sum_{k=1}^{H} \exp\left(s(u_+^k, u_i)\right)} \cdot S_t(u_i) \tag{2}$$

where s is the cosine similarity of CLIP (Radford et al., 2021) embedding. The update probability $P_{t+1}(u_i)$ for u_i is computed by normalizing the exponentiated value of the updated score $S_{t+1}(u_i)$ over the sum of exponentiated scores for all u_j in the dataset, ensuring that the probabilities across all u_i sum to 1:

$$P_{t+1}(u_i) = \frac{\exp\left(S_{t+1}(u_i)\right)}{\sum_{j=1}^{N} \exp\left(S_{t+1}(u_j)\right)} \tag{3}$$

In step t iteration, our framework operates as follows. First, we initialize a set of random images $U_0 = \{u_1, u_2, \ldots, u_j\}$. Using the pretrained encoding model g to synthesize EEG signals X_i from these stimuli. Second, for any given representation function Y_i, we calculate the neural activity representation $Y_i = f(g(U_i)) \in \mathbb{R}^{N\times F}$ from the predicted signal x_i, to estimate the difference based on the target neural representation Y_{target}. Third, the similarity score $sim\langle u_j, u_{\text{target}}\rangle$ between each neural representation derived from each current stimulus u_j and the target representation is computed. Subsequently, stimulus images exhibiting higher similarity scores are more likely to be selected. Based on $sim\langle u_j, u_{\text{target}}\rangle$, stimulation is probabilistically sampled, favoring images that are closer to the target representation. Finally, the sampled images are used to retrieve similar images for the step $t+1$ or input into the diffusion model to generate new stimulus samples.

3.2 Black-box Encoding Model

An image-computable brain encoder is a learned function g_θ that maps an image $I_i \in \mathbb{R}^{3\times H\times W}$ to a synthetic EEG X_i. Instead of collecting real EEG, we employ encoding models to generate neural responding to visual stimuli, which can later be replaced with EEG recordings obtained from human participants in real experiments. We assume that our framework remains effective regardless of the specific structure of the encoding model, allowing us to focus on the advancements of the framework itself rather than the details of the encoding model's architecture. To ensure robustness,

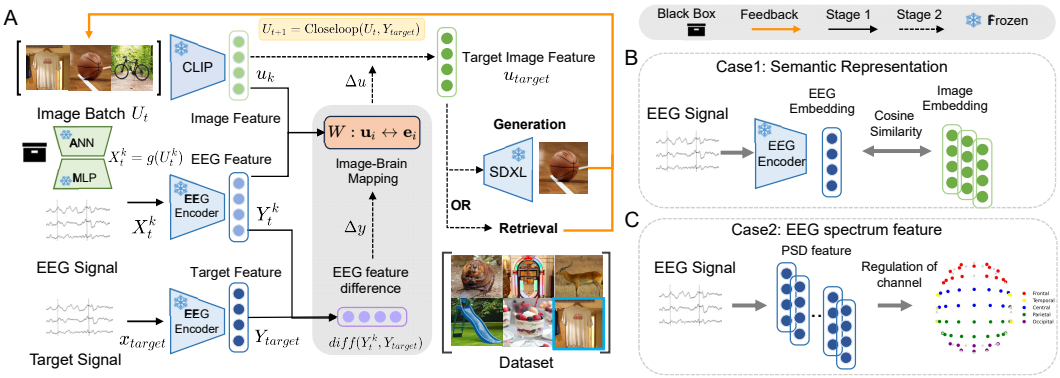

Figure 2: **Closed-loop EEG-based visual stimulation framework for controllable generation.** (A) Our framework relies on a closed-loop iterative algorithm to approximate neural features. The image with a higher similarity score is selected through heuristic algorithm and passed back to the image generator to generate optimized stimuli with a natural image. (B) A case of semantic feature from pre-trained EEG encoder, which is aligned with images. (C) The other case of channel-wise energy, using PSD feature.

we use two different CNN models, AlexNet (Krizhevsky, 2014) and CORnet-S (Kubilius et al., 2019), as feature extractors and train regressors to predict the ground truth of EEG $\hat{X}$.

In the encoding model, we modify the CNN's 1000-neuron output layer to a $C \times T$-neuron layer, where each neuron corresponds to one of the flattened EEG data ponits $C \times T$. Each subject is assigned unique model parameters, achieved by randomly initializing independent instances of the model for each participant and across all EEG time points T. Given the input training images I and the corresponding target EEG data $\hat{X}$, the model updates its weights by minimizing the mean squared error (MSE) between predicted EEG X and the $\hat{X}$. This setup ensures a personalized and accurate prediction of synthetic neural activity.

3.3 Interactive Search

In order to find the optimal stimulus image that causes the target neural activity, we search for images that produce similar neural activity based on the target neural feature. However, the entire target query image is assumed to be unknown. In this experiment, we set the transition probability as the global cumulative probability and then sample new image stimuli in each step using a roulette wheel method. To solve the problem of how to start retrieval when there is no clear query image, we use the mathematical framework of (Ferecatu & Geman, 2007), based on mind matching, starting from a random sample of images, and iteratively let the user select the image that is closest to the category in his mind. In our question, Our specific algorithm process is shown in Algorithm 1.

In our framework, the *Closed-loop Retrieval Iteration Algorithm* can be interpreted as a series of state transitions aimed at maximizing the similarity between current neural feature and a target neural response. Initially, our framework randomly selects a set of images U_0, without knowing the specific image features of the target class. We use a roulette wheel algorithm to select from current images according to $sim\langle u_j, u_{\text{target}}\rangle$. The system updates the probability $p_t(u_j)$ of each image in the database belonging to the target class based on the response model's prediction $Y = f(g(U)) \in \mathbb{R}^{N \times F}$. Once one image is selected, the system increases the probability that the image belongs to the target class. The system calculates the distance between the brain activity feature vector of the target image and the brain activity feature vector predicted by the image selected by the roulette wheel algorithm (i.e., the image that is considered to be closer to the target class).

The algorithm begins by initializing equal selection probabilities for each image in the candidate set, denoted as $p_0(u) = \frac{1}{N}$, where N is the total number of images in the retrieval set. This initialization phase serves as an exploratory step, with equal probabilities reflecting the absence of prior information. In each iteration (representing a state in the MDP framework), a subset of

Algorithm 1 Closed-loop Retrieval Iteration Algorithm

1: **Initialize:** Set initial set $U_0 = \{u_1, u_2, \ldots, u_k\}$, where $U_0 \subseteq \Omega$.
2: **repeat**
3: **Action Selection:** $U_t = \{u_1, u_2, \ldots, u_k\}$ from Ω based on $p_t(u)$.
4: **Reward Calculation:**

$$sim_{\max} = \max\ sim\langle u_k, u_{\text{target}}\rangle$$

5: **if** $sim_{\max} < threshold_1$:
6: Go to Step 3.
7: **else**:
8: **Optimal Action Reference:**

$$\{u_{\text{top1}}, u_{\text{top2}}\} = \arg\max_{\substack{u_k \in U_t \\ \text{top 2}}} \frac{\exp\left(sim\langle u_k, u_{\text{target}}\rangle\right)}{\sum_{u_h \in H} \exp\left(sim\langle u_h, u_{\text{target}}\rangle\right) + \sum_{u_k \in U_t} \exp\left(sim\langle u_k, u_{\text{target}}\rangle\right)}$$

9: **if** $sim\langle u_{target}, u_{top1}\rangle$ or $sim\langle u_{target}, u_{top2}\rangle > threshold_2$:
10: **CLIP-based Retrieval:** Using u_{top1} and u_{top2}, retrieve the top-k images $\{u'_1, u'_2, \ldots, u'_k\}$ from Ω that have the highest similarity s:

$$u'_k = \arg\max_{u \in U} \{s(u, u_{\text{top1}}), s(u, u_{\text{top2}})\}.$$

11: **Update Action Set:** Update the subset U_{t+1}:

$$U_{t+1} = \{u'_1, u'_2, \ldots, u'_k\}.$$

12: **Recurse on U_{i+1}:** Repeat the process for the new action set U_{t+1}, treating it as the current action set U_t for the next iteration.
13: **until** $s_{\max} \geq threshold_{primary}$
14: **Return:** Return the best action set U_t as the final set of retrieved images.

images $U_t = \{u_1, u_2, \ldots, u_j\}$ is selected from the candidate images set U based on the current selection probabilities $p_t(u)$.

For each image u_j in the subset U_t the algorithm computes a similarity score $sim\langle u_j, u_{\text{target}}\rangle$ by comparing the image's representation with the target. This similarity score acts as an immediate reward signal within the MDP framework. The maximum similarity score among the subset is identified as a measure of the effectiveness of the current action. If $sim_{\max}$ does not meet a predefined $threshold_1$, the reward is considered insufficient, and the algorithm returns to the image selection step, effectively trying a new action within the same state. If $sim_{\max}$ meets or exceeds the threshold, the algorithm proceeds to identify the two images u_{top1} and u_{top2} with the highest similarity scores. These two images act as reference points for updating the probabilities of other images in the subsequent state.

As for each image u_j in U that surpasses $threshold_2$ with either u_{top1} or u_{top2}, its selection probability $P_{t+1}(u_j)$ is updated by multiplying with a constant factor, representing a policy improvement step that prioritizes images likely to yield higher rewards. After updating, a Softmax function is applied to normalize the probabilities, focusing selection weight on images more similar to the target. This normalization step reflects the transition to a new state with an updated policy. The iteration continues, with the algorithm transitioning through states by selecting new subsets based on the refined probabilities, until $sim_{\max}$ reaches $threshold_{primary}$. At this point, the loop stops, as the algorithm has effectively found an optimal subset of images that maximizes the similarity reward with respect to the target.

3.4 Heuristic Solution

Retrieving the optimal image stimulus only in the image feature space limits the potential to get closer to the target brain activity. To design an optimal stimulus to the greatest extent, we use StableDiffusion XL-turbo for gradient-guided optimal stimulus generation. The pretrained guided diffusion model $G(U_t)$ generates new visual stimuli via image-to-image. Based on MDP, we use

Algorithm 2 Closed-loop Generative Iteration Algorithm

1: **Initialize:** Set initial set $U_0 = \{u_1, u_2, \ldots, u_k\}$, where $U_0 \subseteq \Omega$.
2: **repeat**
3: **Selection:** $U_t = \{u_1, u_2, \ldots, u_k\}$ from Ω based on $p_t(u)$.
4: **Sampling:** Based on the calculated similarity scores, sample from U_t using:

$$P(u_k) = \frac{\exp\left(sim\langle u_k, u_{\text{target}}\rangle\right)}{\sum_{u_{k'} \in U_t} \exp\left(sim\langle u_{k'}, u_{\text{target}}\rangle\right)}$$

where $P(u_k)$ is the sampling probability for each $u_k \in U_t$.
5: **Crossover:** Draw two distinct samples u_a, u_b from U_t based on $P(u_k)$, and output new samples by combining the partial embedding of u_a and u_b:

$$F(u_{\text{tmp}}^{(1)}) \leftarrow \alpha \cdot F(u_a) + (1-\alpha) \cdot F(u_b)$$

$$F(u_{\text{tmp}}^{(2)}) \leftarrow \alpha \cdot F(u_b) + (1-\alpha) \cdot F(u_a)$$

where α is a crossover control factor.
6: **Mutation:** Based on $P(u_k)$, apply mutation to the drawn images u_{c} from U_t, and another image u_{d} is drawn from the remaining U_t (i.e., $U_t \setminus \{u_{\text{c}}\}$):

$$F(u_{\text{tmp}}^{(3)}) \leftarrow \beta \cdot F(u_c) + (1-\beta) \cdot F(u_d)$$

where β is a mutation control factor.
7: **Generation:** Generate a new set of images $U_{\text{gen}} = \{u_{\text{gen}}^{(1)}, u_{\text{gen}}^{(2)}, u_{\text{gen}}^{(3)}\}$ according to the outputs of crossover and mutation phase.
8: **Selection:** Combine U_{gen} with U_t and randomly selected samples $U_{\text{random}} = \{u_{\text{ran}}^{(1)}, u_{\text{ran}^{(2)}}, \ldots, u_{\text{ran}}^{(n)}\}$, where $U_0 \subseteq \Omega$.
9: **Update Action Set:** Update the subset U_{t+1}:

$$U_{t+1} \leftarrow \{U_t, U_{\text{gen}}, U_{\text{random}}\}$$

10: Replace the old population with the new set of images U_{i+1}.
11: **until** similarity score converges or reach the maximum number of cycles.

a genetic algorithm to assist the generator in generating image stimuli in the direction of the target neural activity while ensuring global optimality. Our specific algorithm process is shown in Algorithm 2. Unlike Algorithm 1, after sampling the stimulus image in each step of roulette, we partially cross the image features, and randomly sample new image samples from the image space. Mutation is performed based on the obtained image features and the original image features retained by the population to ensure that normal semantic images that are understandable to humans can still be generated after mutation.

4 EXPERIMENTS

4.1 SETUP

Datasets We conducted our experiments using the training set of the THINGS-EEG2 dataset (Gifford et al., 2022; Grootswagers et al., 2022), which consists of a large EEG corpus from 10 human subjects performing a visual task. The experiments used the Rapid Serial Visual Presentation (RSVP) paradigm for orthogonal target detection tasks to ensure participants' attention to the visual stimuli. All 10 participants underwent 4 equivalent experiments, resulting in 10 datasets with 16,540 unique training image conditions, each repeated 4 times, and 200 unique testing image conditions, each repeated 80 times. In total, this yielded (16,540 training image conditions $\times$ 4 repetitions) + (200 testing image conditions $\times$ 80 repetitions) = 82,160 image trials. The original data were recorded using a 64-channel EEG system with a 1000 Hz sampling rate. After signal denoising, the data were downsampled to 100 Hz, focusing on 17 channels over the occipital and parietal regions. For preprocessing, we segmented the EEG data into trials from 0 to 1000 ms post-stimulus

onset, with baseline correction applied using the mean of the 200 ms pre-stimulus period. All electrodes were retained, and the data were downsampled to 250 Hz for analysis. Multivariate noise normalization was applied to the training data (Guggenmos et al., 2018).

Encoding Model In the training phase, we used a batch size of 64 images and the Adam optimizer with a learning rate of 10^{-5}, a weight decay term of 0, and default values for other hyperparameters. Training was conducted over 50 epochs, with EEG responses for test image conditions synthesized using the model weights from the epoch that yielded the lowest validation loss. For each participant, the models generated EEG signals with a shape of 17 EEG channels $\times$ 250 EEG time points as the output corresponding to the input images.

4.2 Regulation of Brain Semantic Representation

In order to verify the effectiveness of our EEG-based closed-loop visual stimulation framework for achieving the target neural activity representation, we first conducted a retrieval task in the image space. We regarded the encoding model g as a black box model to ensure that the gradient is not used to update the parameters of the encoding model, so as to better focus on the closed-loop regulation framework itself. We performed the retrieval task in the test set of THINGS-EEG2 dataset with 200 $\times$ 12 = 2400 images. We use the EEG encoder in (Li et al., 2024) to obtain EEG semantic representations aligned with 1×1024 CLIP image features. Before the retrieval begins, random initialization ensures 10 initial points are scattered as much as possible in the image feature space. During the search process, each initial image sample calculates the cosine similarity with the global image features, and uses the cumulative probability to have more reasonable opportunities to select image samples that can produce new and closer to the target EEG neural representation. In the image feature space, through the initial initial image sample point, it continuously expands to form a small area and iterates, and finally approaches the theoretically optimal stimulus image sample. The condition for the iteration to terminate is similarity $s(u_+, u_i) > 0.97$.

Based on semantic representation, our retrieval results are shown in Figure 3. In Figure 3(A), we plotted the similarity scores of stimuli and random stimuli at different time steps of the iteration process. Figure 3(B) shows the average similarity and mean square error with the expected EEG features at different iteration time steps for subject 8. Figure 3(C) illustrates the convergence patterns from initial to final positions for selected iterations (e.g., iterations 1 and 10) over multiple cycles. In each iteration, ten images are viewed, with points representing the closest match to the target stimulus at each step. Notably, these points show a gradual approach toward the target stimulus, marked by a red pentagram, across successive iterations. For a given target neural activity representation, our framework iteratively predicts intermediate EEG results and retrieves stimulus images at each iteration. Notably, only the neural activity representation evoked by the reference image is known throughout this process. Through successive iterations in 3(D), the framework refines its selection, ultimately retrieving an image (outlined in red) that closely matches the semantic representation of the reference image.

4.3 Regulation of intensity of neural activity

We implemented a closed-loop stimulus image generation framework using the 200 $\times$ 1 = 200 image space of THINGS-EEG2 as initialization. We set the crossover rate α to 0.6, the mutation rate β to 0.2, and randomly select 10 images from 200 images during initialization. We used StableDiffusion XL-turbo (Rombach et al., 2022) integrated by IP-Adapter (Ye et al., 2023) to generate new samples each time based on the new stimulus images obtained after crossover and mutation, and randomly selected 2 samples from the image feature space, calculated the similarity of EEG activity representation, and selected the next step of stimulation according to the roulette method of cumulative probability.

The results of our stimulus generation experiments are shown in Figure 4. Figure 4(A) shows the similarity and mean square error between the EEG features generated by the step stimulation image at different iterations and the target EEG features. In addition, we calculated the explained variance of different channels and selected the three channels O_1, O_z, and O_2 with the largest variance for regulation. Figure 4(B) shows the comparison of the PSD of the EEG predicted by the random and step-best samples relative to the target EEG representation. Figure 4(DEF) plots the synthetic EEG

ICLR_2025/figures/fig3.pdf

Figure 3: **Results of our framework in the retrieval task.** (A) Similarity between the neural representation obtained by our framework at different iteration steps (i.e., step-1, step-2, step-second-last, step-last) and the target neural representation compared to random stimulus (i.e., random). (B) The evolution of EEG representation similarity (blue) and loss curves (yellow) on Subject 8 at different iteration steps. (C) The t-SNE visualization of Subject 8's latent trajectories within the feature space across all iterations. (D) The images retrieved by our framework at different iteration steps. Note that only the neural activity representation evoked by the reference image is known during the iteration process.

of three different channels obtained by step-best, random and target stimulation images respectively. All three channels show that the EEG corresponding to step-best, random and target images is quite different before 100 data points (corresponding to 0.4s). After 0.4s, due to the limitations of the encoding model itself, the synthetic EEG of the target image is not much different from the synthetic EEG of the optimal stimulation and the synthetic EEG of the random image. This corresponds to Fig.4 in (Gifford et al., 2022). Using the tick image as an example, Figure 4(C) shows the image and its corresponding time-frequency features, as well as the generated image and corresponding features at each iteration. The image enclosed by a red border represents the image synthesized by the generator, while the unbordered image is a sample selected from the original dataset.

ICLR_2025/figures/fig4.pdf

Figure 4: **Results of our framework in the generation task.** (A) Similarity and loss curves of EEG neural representations for Subject 8. (B) The difference of PSD between the neural activity representations evoked by the final step of generated and random stimulus, with the target neural representations used as the relative baseline. (C) For a given target EEG semantic representation, our framework iteratively predicts synthetic data, extract feature and synthesizes images at each iteration. (D) EEG timing diagram generated by our stimulus images for O_1 channel. (E) EEG timing diagram generated by our stimulus images for O_z channel. (F) EEG timing diagram generated by our stimulus images for O_2 channel.

4.4 Regulation of Individual Variability

Table 1 summarizes the results in the retrieval setting (corresponding to the representation score, SS) and the generation model setting (corresponding to the intensity score, IS), highlighting the results of our framework in achieving the optimal number of iterations in a given search space. The data show that for different target EEG features, our method has a good improvement in feature similarity across different subjects. For instance, the similarity score (SS) of the semantic feature of Subject 7 is improved from 0.874 in step-1 to 0.974, with an improvement of 10.04%. Similarly, the feature similarity score (IS) of the channel intensity of Subject 8 is improved from 0.913 in step-1 to 0.990, accompanied by a 7.744% improvement. Even on the subjects with poor performance, our framework achieves a positive performance, which shows that our framework has a generalized improvement effect across different subjects, highlighting its potential in practical applications.

Table 1: **Performance (EEG semantic representation and intensity) of brain responses.** We provide two metrics: EEG semantic representation score (i.e., SS) and EEG response intensity score (i.e., IS) to measure the difference between the neural activity generated by the optimal stimulation image we obtained and the target EEG neural activity.

	Step-1		Step-Best		Improvement	
Subject	SS	IS	SS	IS	ΔSS (%)	ΔIS (%)
1	0.871	0.989	0.967	0.997	9.593	0.801
7	0.874	0.960	0.974	0.995	10.040	3.444
8	0.904	0.913	0.976	0.990	7.162	7.744
10	0.915	0.986	0.961	0.998	4.587	1.163

5 Discussion and conclusion

In this study, we propose a novel and feasible EEG-based closed-loop visual stimulation framework for controllable generation. To our knowledge, this is the first framework that successfully implements closed-loop stimulus generation to modulate brain activity using natural priors on EEG.

Technical Impact: We present a closed-loop iterative strategy that samples new random stimuli each time a new round of stimulus images is generated. By passing the gradient of the target neural activity representation to the diffusion model in a proxy manner, we eliminate the need to train the generative model or update its weights. In both the feature space interactive retrieval task and the image stimulus generation task, we obtained effective stimulus images that are closest to the target EEG activity features. This demonstrates that our framework is an efficient and optimal closed-loop stimulus generation method that does not require any model parameter updates.

Neuroscience Insights: Our results demonstrate that closed-loop, controllable generation of visual stimuli based on EEG signals is not only feasible but also effective in two distinct contexts. First, we successfully modulated the activity of specific electrode channels, indicating that fine-tuning neural activity in targeted brain regions can be achieved through controlled visual stimulation. Second, we showcased our framework's capability to guide the brain in generating specific neural representations, which is crucial for understanding how different regions of the brain process visual information and respond to external stimuli.

This study provides significant insights into the neural mechanisms underlying visual perception and stimulus processing. The successful implementation of EEG-driven closed-loop generation offers a novel real-time approach for manipulating brain activity, revealing the possibility of targeted modulation of cognitive functions. Additionally, by linking specific EEG patterns to visual representations, our work contributes to a broader understanding of how neural signatures correlate with perceptual experiences. This opens new avenues for applications in brain-computer interfaces, neurofeedback systems, and therapeutic interventions for neurological disorders where precise regulation of brain activity is needed (Jang et al., 2021; Alamia et al., 2023).

Interesting Phenomena and Future Directions: Since different stimulus images in our framework can produce the same or similar EEG features, this suggests the existence of Metamers (Feather et al., 2023), which may not be unique. The presence of Metamers complicates feature discrimination in our analysis, and these Metamers vary across different subjects. Future research should delve deeper into understanding the underlying neural mechanisms that lead to the generation of similar EEG features from different stimuli. This can be approached by incorporating real-time EEG data from a diverse set of subjects, facilitating more individualized and precise modulation of neural activity. Another promising direction involves integrating more sophisticated models that account for inter-individual variability in neural responses, aiming to fine-tune the stimulus generation process for enhanced brain-computer interaction (Alamia et al., 2021). For instance, it is expected to combine other modes of stimulation to regulate the electrical characteristics of the brain similar to gamma oscillations, and realize a new idea of depression treatment (Li et al., 2023). Further exploration could involve combining this closed-loop framework with other brain imaging modalities, such as fMRI or MEG, to gain a more comprehensive understanding of the multimodal neural representations of visual stimuli.

REFERENCES

Andrea Alamia, Milad Mozafari, Bhavin Choksi, and Rufin VanRullen. On the role of feedback in visual processing: a predictive coding perspective. *arXiv preprint arXiv:2106.04225*, 2021.

Andrea Alamia, Milad Mozafari, Bhavin Choksi, and Rufin VanRullen. On the role of feedback in image recognition under noise and adversarial attacks: A predictive coding perspective. *Neural Networks*, 157:280–287, 2023.

Luca Baroni, Mohammad Bashiri, Konstantin F Willeke, Ján Antolík, and Fabian H Sinz. Learning invariance manifolds of visual sensory neurons. In *NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, pp. 301–326. PMLR, 2023.

Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

Jenelle Feather, Guillaume Leclerc, Aleksander Madry, and Josh H McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034, 2023.

Marin Ferecatu and Donald Geman. Interactive search for image categories by mental matching. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. IEEE, 2007.

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.

Tijl Grootswagers, Ivy Zhou, Amanda K Robinson, Martin N Hebart, and Thomas A Carlson. Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1):3, 2022.

Zijin Gu, Keith Jamison, Mert R Sabuncu, and Amy Kuceyeski. Modulating human brain responses via optimal natural image selection and synthetic image generation. *ArXiv*, 2023.

Matthias Guggenmos, Philipp Sterzer, and Radoslaw Martin Cichy. Multivariate pattern analysis for meg: A comparison of dissimilarity measures. *Neuroimage*, 173:434–447, 2018.

Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10):e0223792, 2019.

Hojin Jang, Devin McCormack, and Frank Tong. Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. *PLoS biology*, 19(12): e3001418, 2021.

Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019.

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

Qun Li, Yuichi Takeuchi, Jiale Wang, Levente Gellért, Livia Barcsai, Lizeth K Pedraza, Anett J Nagy, Gábor Kozák, Shinya Nakai, Shigeki Kato, et al. Reinstating olfactory bulb-derived limbic gamma oscillations alleviates depression-like behavioral deficits in rodents. *Neuron*, 111(13): 2065–2075, 2023.

Andrew Luo, Maggie Henderson, Leila Wehbe, and Michael Tarr. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *Advances in Neural Information Processing Systems*, 36, 2024a.

Junwen Luo, Chengyong Jiang, Qingyuan Chen, Dongqi Han, Yansen Wang, Biao Yan, Dongsheng Li, and Jiayi Zhang. The vep booster: A closed-loop ai system for visual eeg biomarker autogeneration. *arXiv preprint arXiv:2407.15167*, 2024b.

Carlos R Ponce, Will Xiao, Peter F Schade, Till S Hartmann, Gabriel Kreiman, and Margaret S Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009, 2019.

Yongrong Qiu, David A Klindt, Klaudia P Szatko, Dominic Gonschorek, Larissa Hoefling, Timm Schubert, Laura Busse, Matthias Bethge, and Thomas Euler. Efficient coding of natural scenes improves neural system identification. *PLoS computational biology*, 19(4):e1011037, 2023.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Tiberiu Tesileanu, Eugenio Piasini, and Vijay Balasubramanian. Efficient processing of natural scenes in visual cortex. *Frontiers in Cellular Neuroscience*, 16:1006703, 2022.

Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12): 2060–2065, 2019.

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

A APPENDIX

A.1 VALIDITY VERIFICATION OF SYNTHETIC EEG

To evaluate the performance of our EEG encoding models, we compare the synthetic EEG signals generated by two deep neural networks (DNNs)—AlexNet and CORnet-S—with real EEG data. Here's a step-by-step breakdown of how we processed and compared the data.

We selected 17 specific channels from the original 63-channel EEG dataset. These channels were chosen based on their relevance to visual processing, ensuring that we focused on neural regions most closely related to the visual stimuli. For each stimulus, we averaged the EEG signals across all trials, resulting in a representative dataset for each stimulus. This reduced the dimensionality of the data, making it easier to compare with synthetic data. We used a pretrained end-to-end encoding model to generate synthetic EEG signals based on the visual stimuli. The model captures the mapping between the visual input and the resulting EEG signals using deep neural networks. These synthetic signals represent the neural responses that the model predicts based on the stimuli.

Table 2: MSE Values for synthesized EEG

Subject	**Pretrained**		**Random Init**		**Average**
	AlexNet	**CORnet-S**	**AlexNet**	**CORnet-S**	
Sub-01	0.1095	0.1126	0.1161	0.0994	0.1094
Sub-02	0.0764	0.0788	0.0840	0.0994	0.0847
Sub-03	0.0787	0.0806	0.0816	0.0910	0.0830
Sub-04	0.0652	0.0664	0.0662	0.1011	0.0747
Sub-05	0.0493	0.0515	0.0704	0.0975	0.0672
Sub-06	0.0690	0.0719	0.0498	0.0966	0.0718
Sub-07	0.1267	0.1300	0.0914	0.1312	0.1198
Sub-08	0.0718	0.0727	0.1038	0.1165	0.0912
Sub-09	0.0529	0.0563	0.0781	0.0756	0.0657
Sub-10	0.1122	0.1151	0.0961	0.1149	0.1096
Average	0.0810	0.0832	0.0838	0.1023	0.0876

Table 2 presents the mean squared error (MSE) between the synthetic EEG signals generated by AlexNet and CORnet-S, and the real EEG signals for 10 subjects. The MSE was computed for each individual test sample and then averaged across the entire test set. Lower MSE values indicate better alignment between the synthetic and real EEG signals.

From the comparison shown in the Figure 5, the retrieval accuracy for S-S (both training and testing sets consist of generated signals) is significantly higher than other categories, including T-T (both training and testing sets consist of real signals), T-S (training set consists of real signals, testing set consists of generated signals), and S-T (training set consists of generated signals, testing set consists of real signals), under both AlexNet and CORnet-S models. This indicates:

Advantages of generated signals Supported by black-box ANN models (e.g., AlexNet and CORnet-S), generated signals perform significantly better in retrieval tasks compared to real signals. In particular, the highest retrieval accuracy for S-S demonstrates the consistency and model adaptability of generated signals in this retrieval task.

Model adaptability: Different ANN models (e.g., AlexNet and CORnet-S) show consistent superiority in the retrieval tasks for generated signals, indicating that generated signals are more easily captured and distinguished by black-box models.

In Figure 6, we computed the variance across all samples and time points for each channel. This allows us to quantify the overall variability of the EEG signals for different visual stimuli and their temporal dynamics. The variance analysis provides insights into the spatial distribution of neural

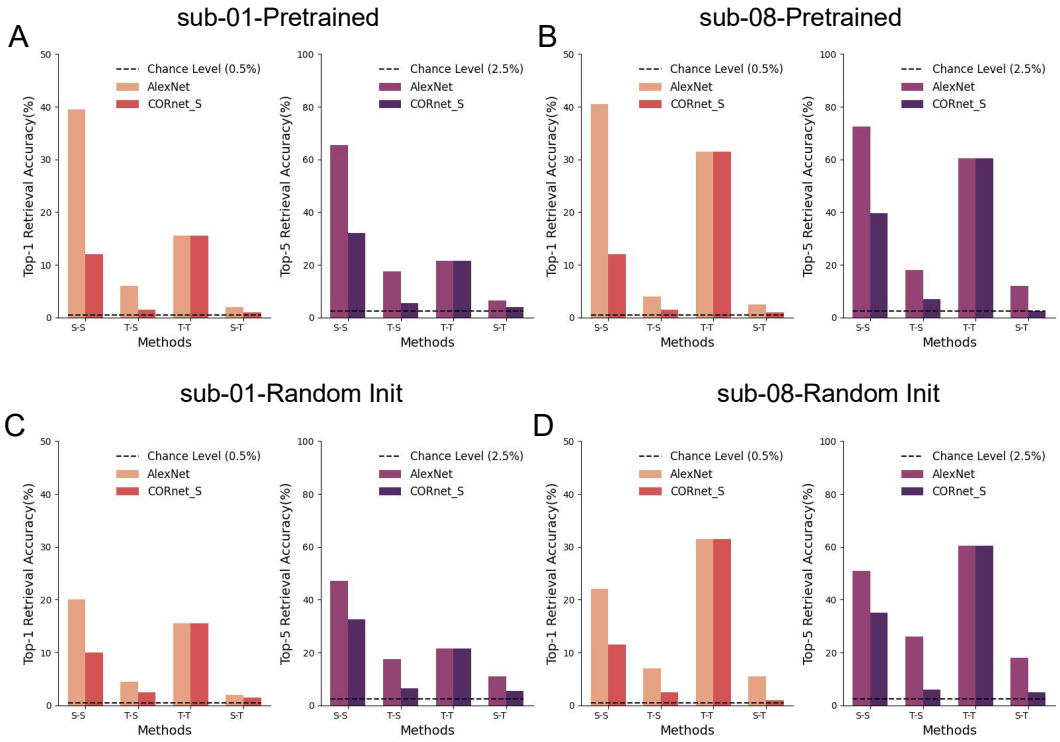

Figure 5: Retrieval accuracy under different training and test datasets. Zero-shot retrieval performance of EEG data from different sources in Subject 1 and Subject 8 using ATM-S in different Settings. AlexNet and CORnet-S used in the first row were both pre-trained end-to-end models, and the second row was randomly initialized end-to-end.

responses to stimuli, highlighting how different channels vary in their responsiveness. This can guide the selection of specific channels for further analysis or modulation.

In Figure 7, showing the variance and standard deviation of the EEG signals computed across samples (stimuli) for each time point, and then averaged across channels. This analysis allows us to assess how signal variability changes over time. By comparing the real EEG data with synthetic data (generated by AlexNet and CORnet-S), we can evaluate how well each model captures the temporal variability of the real EEG signals.

In Figure8, we computed the Pearson correlation coefficient between the averaged real EEG data and the synthetic data for each stimulus. This gives a measure of how well the synthetic data matches the real EEG data on a per-sample basis. The resulting histogram shows the distribution of Pearson correlation coefficients across all samples for both AlexNet and CORnet-S. Higher correlation values indicate better alignment between the synthetic EEG signals and the real EEG data. The comparison of distributions for both models provides insights into which model better replicates the real neural activity, with higher peaks in the histogram representing better performance.truth data.

In Figure9, we analyze the time-resolved Pearson correlation between real and synthetic EEG signals over time. For each time point (from 1 to 250), we compute the Pearson correlation between the real EEG signal and the synthetic signals from both AlexNet and CORnet-S. This time-resolved analysis allows us to visualize how well each model replicates the temporal structure of real neural responses to visual stimuli. Shaded regions in the plot represent the standard deviation across samples, showing the variability in model performance over time. The results provide a detailed view of how each model performs at different time points, highlighting which model more accurately captures the temporal dynamics of EEG signals.stimuli.

From the above analysis, we observe that the synthetic EEG signals generated by AlexNet and CORnet-S closely replicate the variability patterns of real EEG data. Both models perform well, showing comparable results in terms of MSE, spatial (channel-wise) variability, and temporal (time-

resolved) variability. The Pearson correlation analysis further confirms that both models are able to generate synthetic EEG signals that align well with real data, with subtle differences in performance across models. These findings highlight the robustness of our EEG encoding models, demonstrating their ability to generate synthetic EEG signals that not only mimic the structural features of real EEG data but also capture the realistic variability seen in neural responses to visual stimuli. This suggests that our models are effective in approximating the neural representations underlying visual processing.

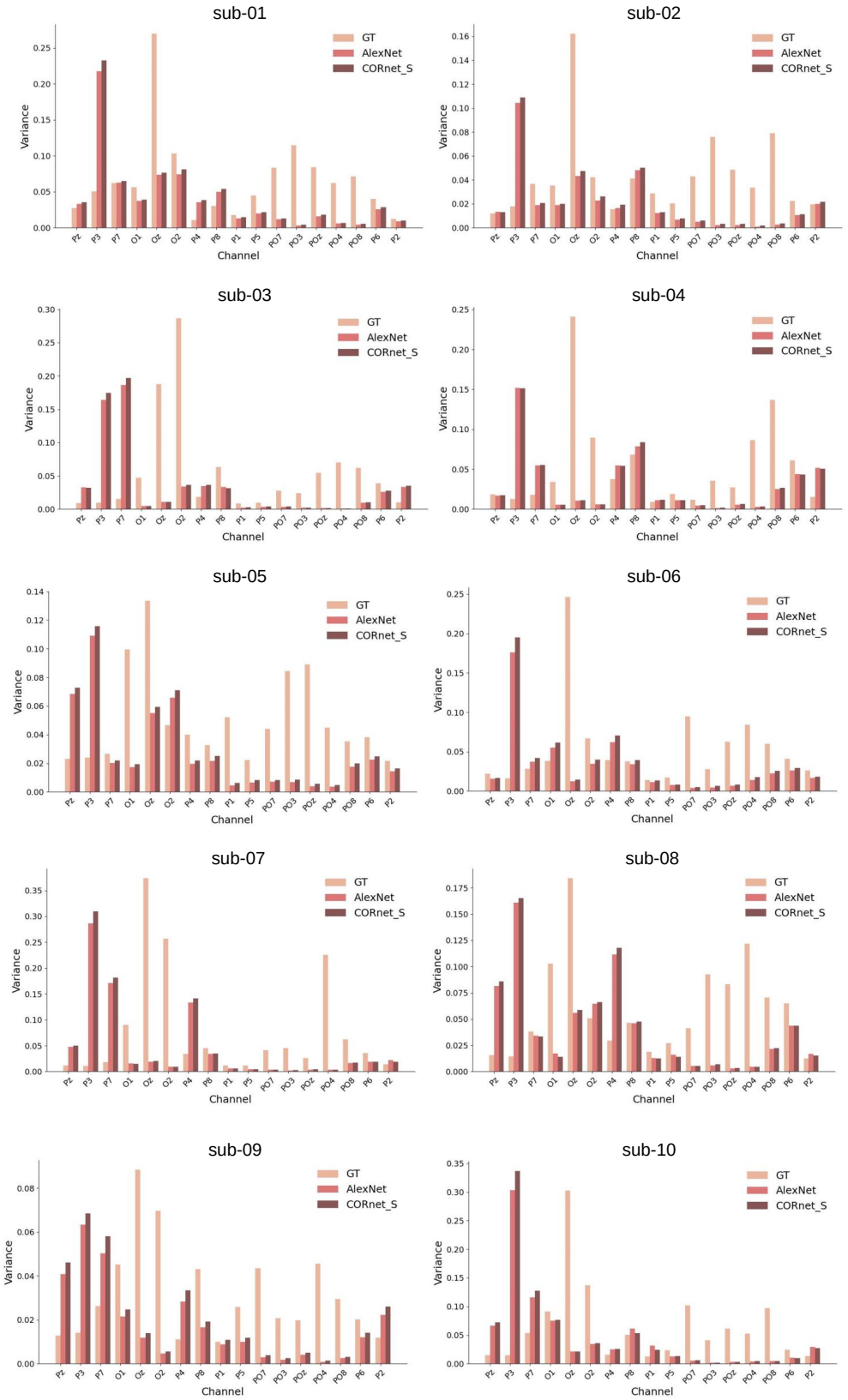

Figure 6: Variance across different channels for different visual stimulus and temporal dynamics

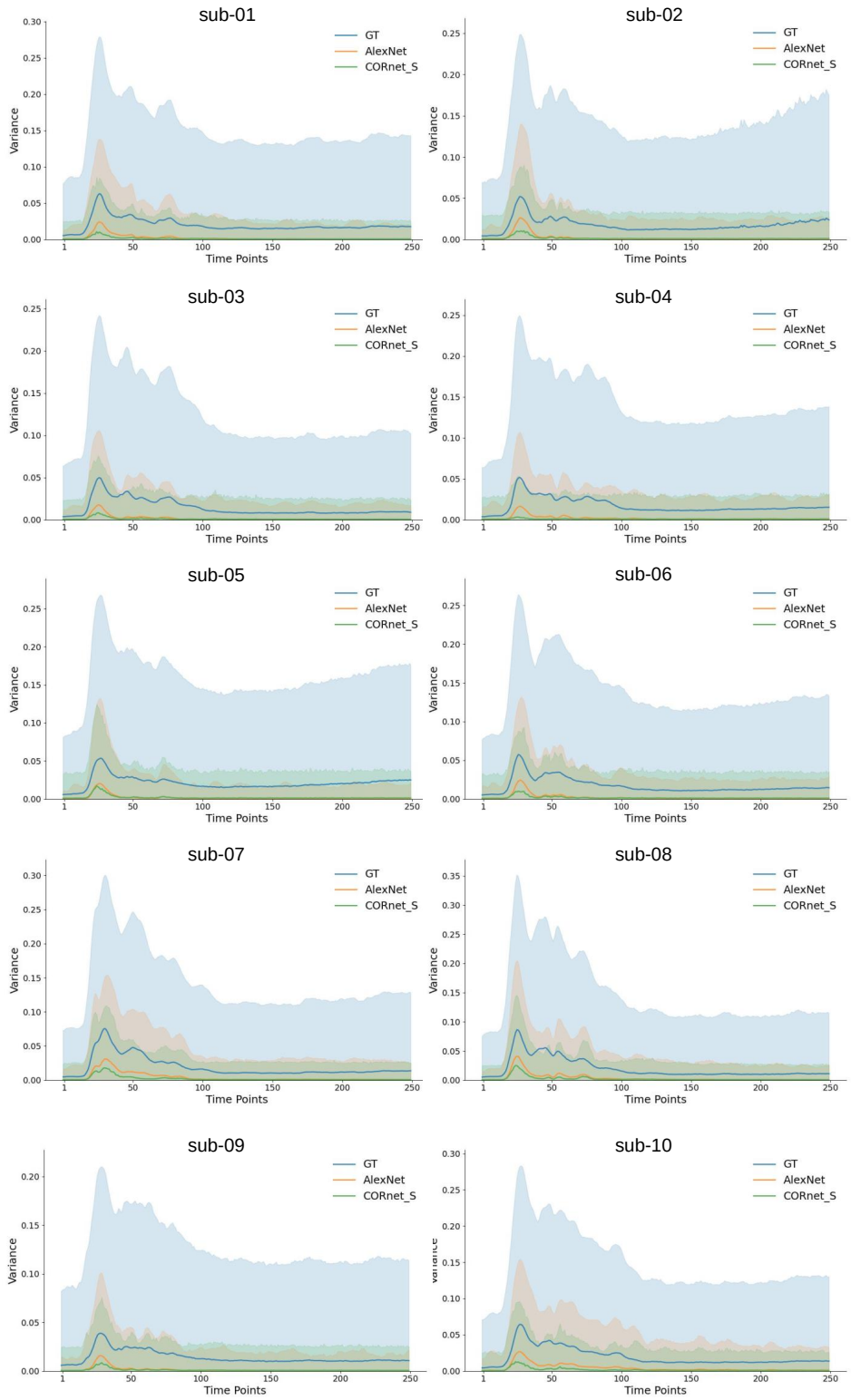

Figure 7: Variance across different time points for different visual stimuli and channels.

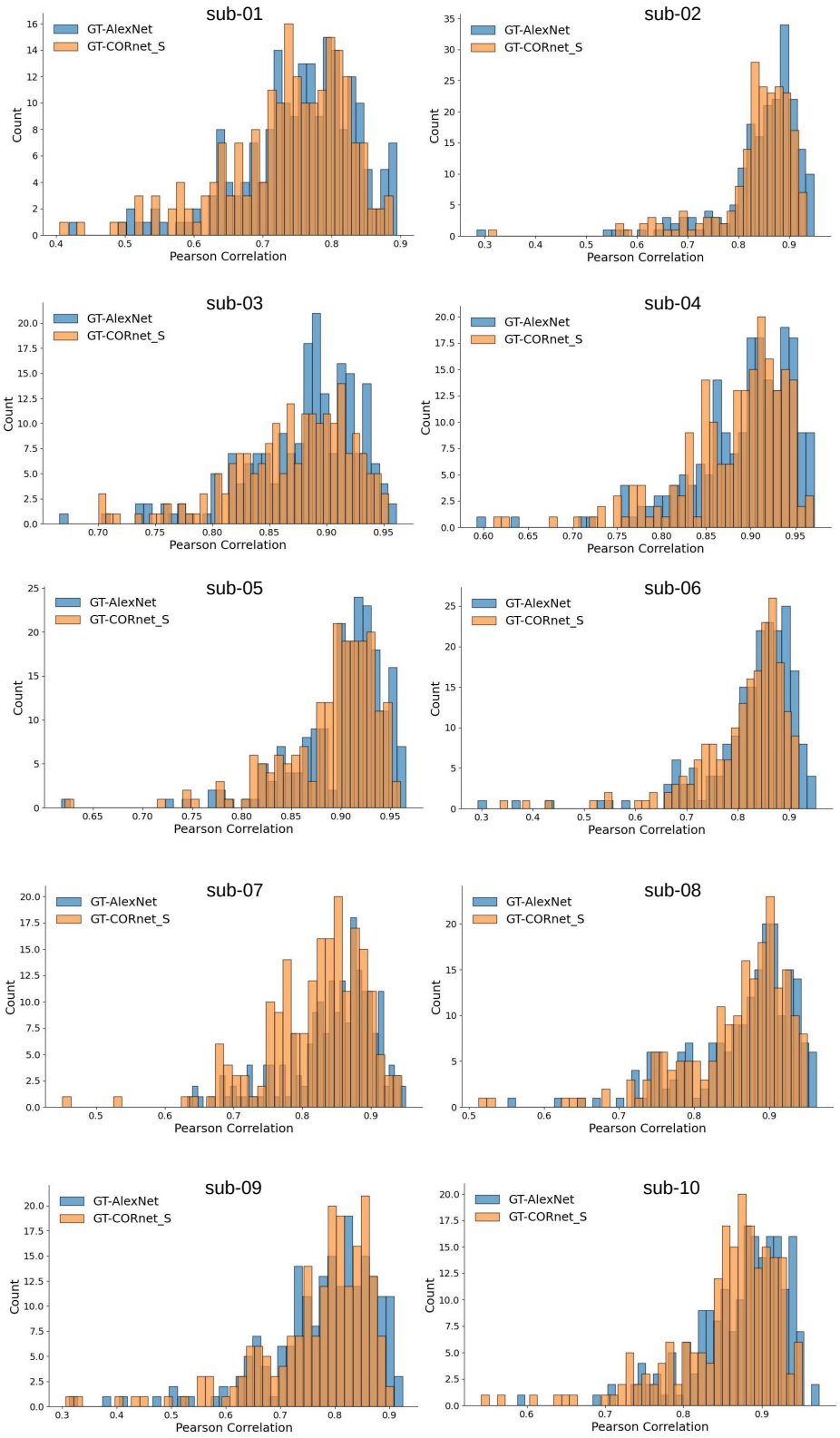

Figure 8: Distribution of Pearson correlation coefficients across all sample pairs.

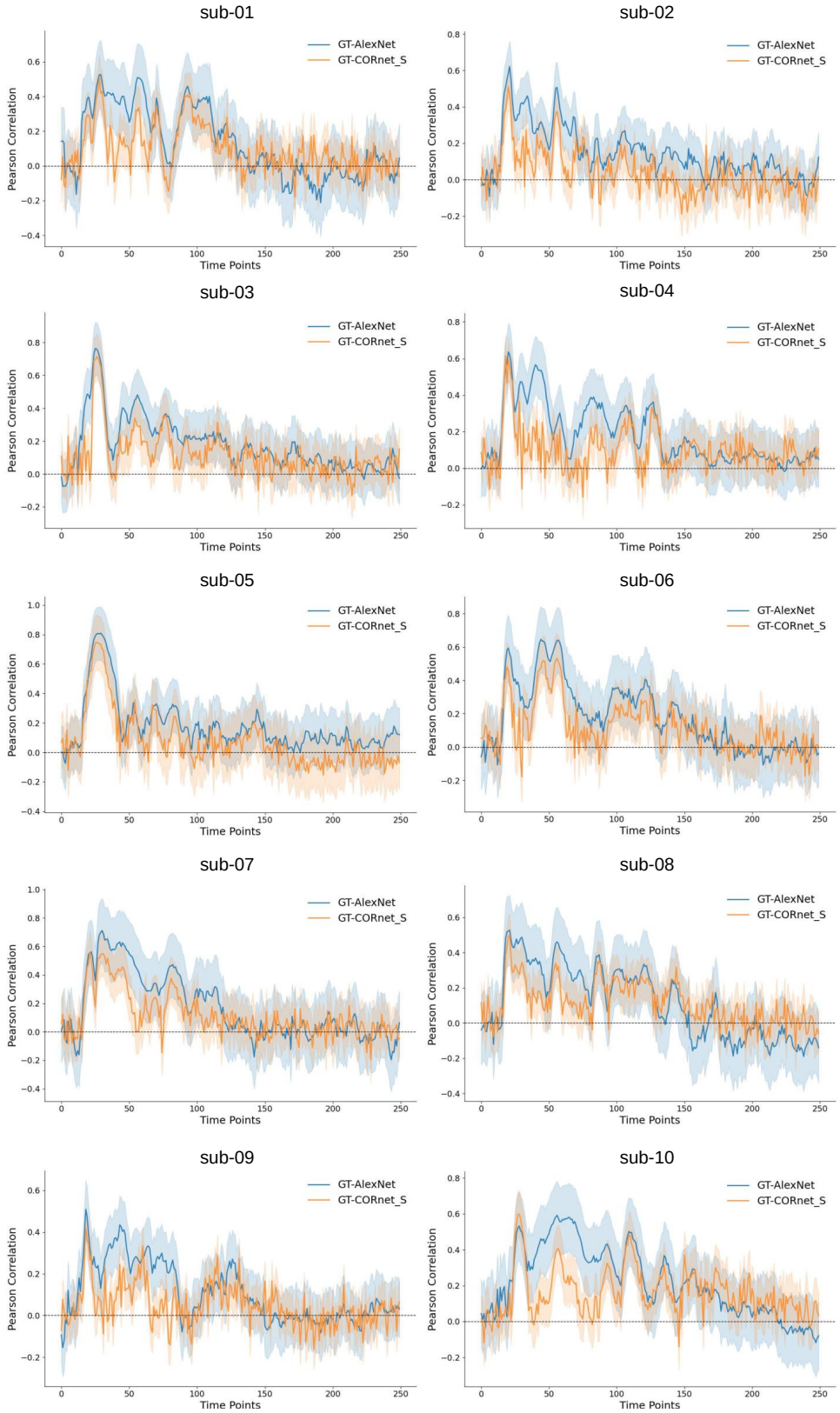

Figure 9: Time-resolved Pearson correlation between ground truth EEG signals and synthetic EEG signals predicted by two neural network models (AlexNet and CORnet-S).

A.1.1 Additional Retrieval Examples of Semantic Representation

A.1.2 Some Failure Examples of Retrieval

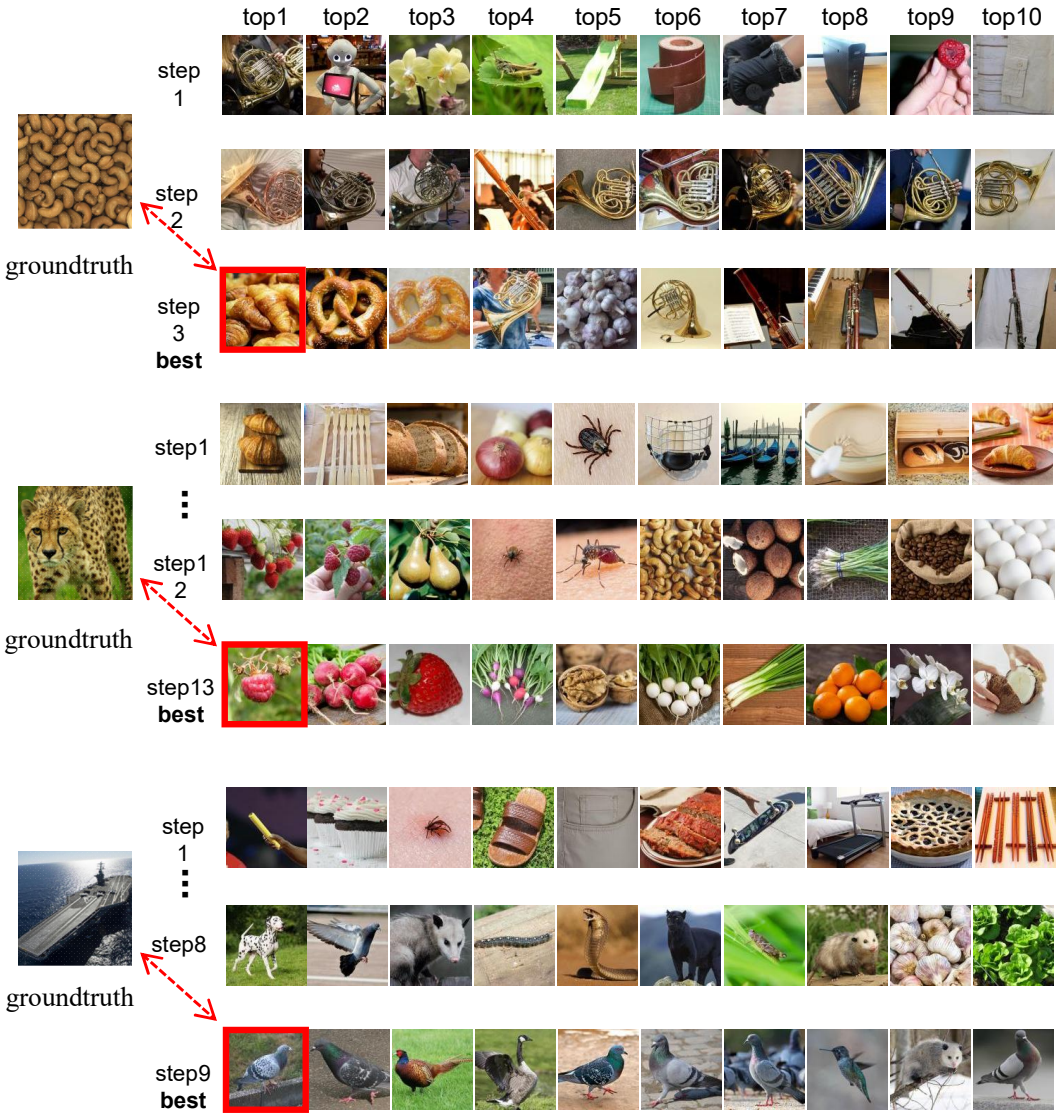

Figure 10: Some retrieval failure examples. By setting different targets, we show examples where the stimulus retrieved at the end of the iteration is far from the true category.