Simple Temporal Attention Beats Complex Decoders for Neural-to-Visual Mapping from Primate Spiking Data

Matteo Ciferri
University of Rome, Tor Vergata
Department of Biomedicine and Prevention
matteo.ciferri@students.uniroma2.eu

Matteo Ferrante
University of Rome, Tor Vergata
Department of Biomedicine and Prevention
matteo.ferrante@uniroma2.it

Nicola Toschi
University of Rome, Tor Vergata
Department of Biomedicine and Prevention
A.A. Martinos Center for Biomedical Imaging
Harvard Medical School/MGH, Boston (US)

Abstract

Understanding how neural activity gives rise to perception remains a fundamental challenge in neuroscience. Here, we address the problem of visual decoding from high-density intracortical recordings in primates using the THINGS Ventral Stream Spiking Dataset. We systematically evaluate the effects of model architecture, loss function, and temporal aggregation, showing that decoding accuracy is primarily driven by temporal dynamics rather than architectural complexity. A lightweight model combining temporal attention with a shallow MLP achieves up to 70% top-1 image retrieval accuracy, outperforming linear and recurrent baselines. Building on this, we introduce a modular generative pipeline that combines low-resolution latent reconstruction with semantically guided diffusion. By generating and ranking multiple candidate images via rejection sampling, our approach enables photo-realistic reconstructions from 200 ms of brain activity. These results provide actionable insights for neural decoding and establish a flexible framework for future brain–computer interfaces and semantic reconstruction from brain signals.

1 Introduction

A complete account of perception and behavior must bridge *neural representations* with *mental states*, linking spikes and field potentials to the contents of subjective experience and overt action. Recent progress in cognitive science and computational neuroscience has been catalyzed by three intertwined developments. First, community-driven efforts now release large, meticulously curated datasets that pair rich sensory stimulation with high-resolution neural recordings [1, 15, 5, 14]. Second, advances in machine learning—particularly deep generative modeling and scalable optimization—provide expressive function classes capable of capturing the complex structure of brain–world mappings [2, 22, 3]. Third, experimental practice is changing from *wide* surveys of many individuals to *deep*, longitudinal studies that expose a few subjects to tens of thousands of stimuli, drastically increasing statistical power [19]. These factors have revived bidirectional modeling of the stimulus–brain relationship. *Encoding* models predict neural responses from sensory features, helping to understand the functional organization of the cortex. In contrast, *decoding* models seek to reconstruct stimuli, or latent variables relevant to a task, from brain activity, a line of work central to both basic science

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

and emerging brain-computer interfaces. Successes span multiple modalities (EEG, MEG, fMRI, ECoG, and Utah array recordings) and cognitive domains, including language comprehension, speech production, music, and vision [4, 22].

Recent advances in visual decoding have largely focused on non-invasive data such as fMRI [22, 4, 2, 12, 16, 8, 10, 11], utilizing pretrained vision-language models such as CLIP combined with linear regression or contrastive learning. These approaches enable retrieval-based decoding and, when integrated with diffusion models, increasingly realistic image reconstruction [23, 9, 21, 29, 30, 6, 33]. Invasive approaches (e.g., ECoG) offer higher-resolution decoding, though they have historically been limited by the dataset scale. The release of the THINGS Ventral-Stream Spiking Dataset (TVSD) [24] marked a turning point, providing large-scale recordings from primate V1, V4, and IT paired with a diverse set of visual stimuli. Based on this, the MonkeySee study [20] introduced a CNN-based decoder with inverse retinotopic mapping, adversarial training, and VGG-based perceptual losses to achieve high-fidelity reconstructions.

Even with invasive data, key questions persist: What properties of intracortical spike trains carry the information necessary for high-fidelity decoding? How do architectural choices, such as linear versus non-linear models, temporal aggregation windows, loss functions, shape performance limits? And how do these factors interact with scale, both in terms of training data and in terms of the dimensionality of neural input? To address these questions, we adopt a decoding approach grounded in semantic understanding.

Our study begins with a zero-shot retrieval setup to assess the quality of the brain-to-vision mapping. Instead of relying on generative models—known to suffer from confounds due to strong image priors [31]—we use frozen CLIP embeddings and retrieval from a fixed candidate set. This approach enables controlled evaluation of model class, temporal structure, and data scaling on decoding performance. However, retrieval has inherent limitations: its reliance on a predefined candidate set restricts generalization. For this reason, we turn to the more ambitious goal of generative decoding. We introduce a two-stage generative decoder in which candidate images are sampled from a frozen generative prior and filtered via rejection sampling guided by a learned neural likelihood.

2 Material & Methods

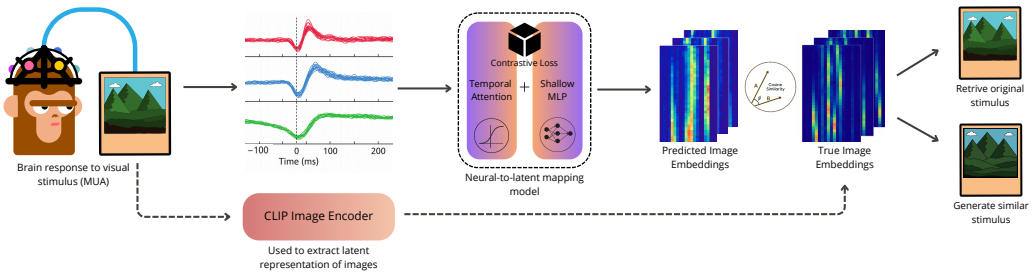

Figure 1: Architecture of our neural-to-semantic visual decoding pipeline. Brain responses recorded via intracortical multi-unit activity (MUA) during visual stimulation are passed through a temporal attention module followed by a shallow MLP to predict image embeddings in the CLIP space.

In this work, we propose a brain decoding framework to estimate the embedding of a visual stimulus directly from primate neural recordings (Figure 1). The goal is to reconstruct a meaningful representation of the perceived image from intracortical data, allowing two downstream applications: (i) *stimulation retrieval*, where the estimated embedding is compared against a set of candidate stimuli, and (ii) *image generation*, where the predicted embedding is used as input to a generative model to synthesize a recognizable version of the original visual input.

2.1 Data

We perform our analysis on both **Monkey F** and **Monkey N**, from the THINGS Ventral Stream Spiking Dataset (TVSD) [24]. This data set comprises intracortical multi-unit activity (MUA) recorded from 15 Utah arrays implanted in visual areas V1, V4, and IT of two macaque monkeys.

Neural responses were collected while the monkeys passively viewed 22,248 unique natural images from the THINGS database, covering a wide distribution of object categories. Each image was presented once for 200 ms, interleaved with 200 ms of a gray screen. The recordings were sampled at 30 kHz and temporally aligned with stimulus onset. For decoding, we used a 200 ms post-stimulus window. A subset of 100 images was presented 30 times each to the same subject and held out for evaluation, while 10% of the training set samples were held out as the validation set for hyperparameter optimization. Retrieval during testing was always performed among these 100 unique images, without including repeated trials. This design enables both training on a large-scale dataset and reliable, low-noise testing. The neural data was pre-processed as follows. Let $X_{\text{raw}} \in \mathbb{R}^{N \times T \times C}$ denote the raw MUA, with N the number of samples, $T = 200$ timepoints, and $C = 1024$ channels. The MUA data was standardized by z-scoring each channel across all timepoints and samples in the training set. The same normalization parameters were then applied to the test set.

2.2 Neural Model

In this section, we describe our proposed decoding model. The target representations $Y \in \mathbb{R}^{N \times D}$, with $D = 512$, consist of high-level visual embeddings corresponding to the presented images, computed from the pre-trained CLIP visual encoder [27]. In order to learn a mapping from MUA signals to the corresponding image embedding, we propose a neural architecture that can take into account the time evolution of the neural response. The model is designed to account for the temporal dimension of the neural sequence and project the aggregated representation to the 512-dimensional target space.

Given an input tensor $X \in \mathbb{R}^{N \times T \times C}$, the model computes a soft attention over time points: $\alpha = \sigma(W_{\text{attn}} X + b) \in \mathbb{R}^{N \times T \times 1}$ where W_{attn} denotes a linear layer and σ is the activation of the sigmoid. The attended representation is calculated as: $z = \frac{1}{T} \sum_{t=1}^{T} \alpha_t X_t \in \mathbb{R}^{N \times C}$ The vector z is then projected into the output space via a multilayer perceptron consisting of two fully connected layers with GELU activation and dropout: $\hat{Y} = W_2 \cdot \text{Dropout}(\text{GELU}(W_1 z + b_1)) + b_2 \in \mathbb{R}^{N \times D}$.

The training objective is a *contrastive loss* based on cosine similarity between predicted and ground-truth embeddings. Let $S \in \mathbb{R}^{N \times N}$ be the cosine similarity matrix between $\hat{Y}$ predicted outputs and Y targets: $S_{ij} = \frac{\hat{y}_i^\top y_j}{\|\hat{y}_i\| \, \|y_j\|}$ The loss function is a variant of the NT-Xent loss with temperature scaling τ, learned during training: $\mathcal{L}_{\text{contrastive}} = -\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N} \exp(S_{ij}/\tau)} \right)$.

We benchmark our model against: (i) linear model with attention, (ii) linear model with temporal averaging, (iii) MLP with averaging, (iv) recurrent network, and (v) temporal convolutional network. The overall architecture of each benchmark is described in Appendix A.1. Hyperparameters, including learning rate, batch size, network depth for deep models, and regularization strength, were selected based on the performance of the validation set.

2.3 Decoding Tasks

In order to assess the quality of the predicted embeddings, we performed a retrieval task, where each predicted embedding $\hat{y}_i$ from the test set is matched against all ground truth embeddings $\{y_j\}_{j=1}^{N}$, and the nearest neighbors are retrieved based on cosine distance. After generating predictions for all test samples in evaluation mode, we collected both the predicted embeddings $\hat{Y} \in \mathbb{R}^{N \times D}$ and the corresponding ground truth embeddings $Y \in \mathbb{R}^{N \times D}$. The cosine distance was used to compute nearest neighbors in the embedding space: $\text{dist}(\hat{y}_i, y_j) = 1 - S_{ij} = 1 - \frac{\hat{y}_i^\top y_j}{\|\hat{y}_i\| \|y_j\|}$. For each test sample i, we identified the top-k most similar ground truth embeddings over the test set. We computed the proportion of test samples for which the nearest neighbor is the ground truth target embedding, i.e., when the index of the closest neighbor matches the sample index (Top-1 Accuracy). We also evaluated the proportion of test samples for which the ground truth embedding appears within the top-5 retrieved neighbors (denoted as Top-5 Accuracy). This retrieval setup provides a quantitative measure of the semantic similarity between predicted and target embeddings, and acts as an indirect proxy of brain to image representation mapping quality in form of a decoding metric.

In addition to retrieval-based evaluation, we explore a generative setting in which predicted neural embeddings are used to synthesize images. We use the Stable Diffusion model as the generative

backbone [26]. In order to condition the model on high-level visual embeddings, we incorporate the IP-Adapter module [35] into the diffusion pipeline. This adapter enables conditioning via learned visual representations rather than textual prompts. We first train our neural model with output dimensionality matching that of the IP-Adapter input (e.g., 1280) to predict visual embeddings from MUA data. In a second setup, we train the same model to directly predict the flattened latent representation of the image expected by the Stable Diffusion VAE (i.e., a tensor of shape $4 \times 32 \times 32$). Given the predicted visual embedding and structural latents, we generate different images for each test sample using classifier-free guidance. The predicted structural latents are only used to evaluate the low-resolution version of the image estimated from neural signals. Our approach is inspired by the recent trend to increase test time computation in AI systems [7, 17]. Using the strong mapping between brain activity and semantic content, and the ability of generative models to reconstruct semantically coherent images, we first generate N candidate semantic images for each trial. Simultaneously, we evaluate the low-resolution image reconstruction that preserves the primary structure (i.e., overall shape and color) from brain activity. We further compare the generated images with the low-resolution preview decoded by the VAE using the predicted latents and computing Structural Similarity (SSIM) [32] to rank the outputs.

2.4 Scaling Laws

In order to investigate how the performance of the decoding model scales with different properties of the input data, we conducted two different sets of controlled experiments evaluating the impact of (i) the input dimensionality, and (ii) the size of the training set (i.e., number of available samples). We applied Principal Component Analysis (PCA) to the MUA signals across channels to analyze input dimensionality impact on performance. We fitted PCA models with different components, and projected both the training and test sets into the reduced channel subspace. The resulting PCA-reduced data had shape $\mathbb{R}^{N \times T \times C'}$, where $C' < C$ is the selected number of components. This allowed us to test the model under different values of C' while keeping the temporal resolution constant. To evaluate the effect of training data size on model performance, we subsampled the training set in a different scenario using a random selection of N' samples, with $N' < N$. Specifically, we randomly selected a fixed number of training samples, using a controlled random seed for reproducibility. The test set remained fixed across all experimental conditions.

3 Results

We evaluated the decoding performance using a retrieval task in which the predicted visual embeddings were matched against the ground-truth embeddings of all test samples. The model was considered successful if the correct target appeared within the top-k nearest neighbors, with Top-1 and Top-5 accuracy used as metrics. Table 1 reports the retrieval accuracy across different decoding models and feature processing strategies. Our best-performing model combines temporal attention with a shallow MLP, achieving the highest retrieval performance. Sequence-based models such as LSTMs did not outperform simpler baselines, despite dedicated hyperparameter tuning (see Appendix A.2), indicating that complexity alone does not improve decoding.

Table 1: Retrieval performance averaged over five seeds and two primates with different decoding models (using all channels). Best results in bold.

Decoding Model	Top-1 Accuracy		Top-5 Accuracy	
	MSE Loss	NT-Xent Loss	MSE Loss	NT-Xent Loss
Linear/TimeFlat	10.1% ± 2.10%	41.3% ± 3.92%	27.3% ± 2.89%	72.6% ± 5.66%
Linear/AvgTime	21.4% ± 1.36%	54.9% ± 1.53%	45.0% ± 2.97%	86.0% ± 1.50%
LSTM	11.0% ± 1.41%	37.7% ± 2.06%	30.2% ± 1.60%	73.7% ± 2.37%
Binning LSTM	16.1% ± 2.01%	56.1% ± 1.88%	40.8% ± 1.96%	85.3% ± 2.37%
TCN	17.0% ± 2.39%	58.3% ± 3.11%	44.1% ± 2.56%	86.6% ± 1.77%
Linear/TimeAtt	19.4% ± 3.11%	62.7% ± 2.79%	42.8% ± 2.70%	89.4% ± 1.34%
MLP/AvgTime	24.0% ± 1.10%	65.5% ± 2.08%	50.6% ± 2.42%	90.1% ± 2.48%
MLP/TimeAtt	22.4% ± 3.14%	**69.3% ± 2.38%**	48.8% ± 3.54%	**93.6% ± 1.03%**

Attention Weights

In order to gain insight into the temporal focus of the decoding model, we extracted and visualized the attention weights produced by our neural model over the entire test set. For each input sample $x \in \mathbb{R}^{T \times C}$, the model outputs a sequence of attention scores $\alpha \in \mathbb{R}^{T}$ reflecting the relative importance of each time point in the final prediction. During inference, we aggregated the attention weights for all test samples and constructed a matrix $A \in \mathbb{R}^{N \times T}$, where each row corresponds to a test trial and each column to a time point. The analysis reveals interpretable patterns in the temporal sensitivity of the model (see Figure 2). The resulting attention maps align with known neurophysiological dynamics (50–100 ms post-stimulus), showing that selective temporal integration is more effective than modeling temporal dependencies end-to-end.

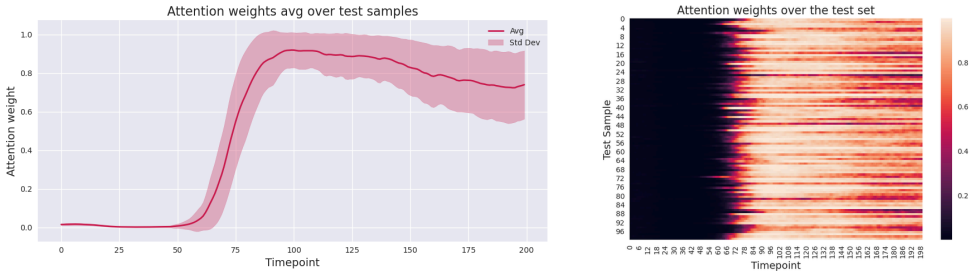

Figure 2: Heatmap of attention weights for the test set (right side) and average weights plot over the entire set (left side). Warmer colors of the heatmap indicate higher attention weights.

Scaling Laws

We investigated how model performance scales with neural feature dimensionality and the size of the training dataset. As illustrated in Figure 3 (left), increasing the number of principal components retained after applying PCA to the neural data consistently improves top-1 and top-5 classification accuracy. The trends are approximately logarithmic, as confirmed by a log-fit model, with large initial gains followed by diminishing returns beyond 256 dimensions. This supports the idea that while semantic information spans a high-dimensional neural space, a low-rank subspace can capture most of the information required for coarse-grained decoding. On the right (Fig 3), we observe a strong scaling trend with respect to training set size. Performance improves rapidly as the number of training examples increases, particularly between 100 and 5000 samples. Although gains persist beyond 10,000 samples, they begin to taper off, indicating a regime of diminishing returns. Both plots exhibit classic scaling law behavior [28, 2, 3, 18], where more data or higher-capacity representations improve performance predictably, albeit with sublinear returns.

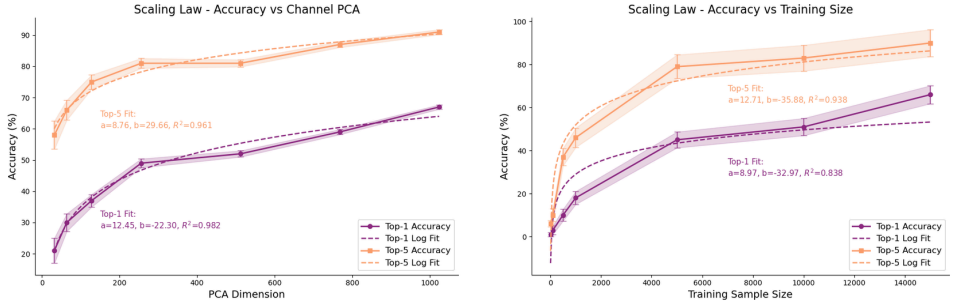

Figure 3: Left: Top-1 and Top-5 classification accuracy as a function of PCA dimensionality applied to neural channels. Accuracy increases logarithmically with dimensionality, as shown by high R^2 in the log-fit curves. Right: Top-1 and Top-5 accuracy as a function of training set size. Performance scales log-linearly with data, underscoring the importance of dataset size in brain-based visual decoding.

Image Reconstruction

We evaluated the generative potential of our brain decoding model by estimating the latent representations required to guide the pre-trained diffusion model. Figure 4 shows representative examples of the image generation results. For each test sample, we display the original stimulus, the low-resolution

preview decoded from the brain-inferred latents (using the VAE component of the model) and the high-resolution reconstruction generated by the diffusion model conditioned on the estimated visual embedding. Reconstructions are ranked on the basis of the Structural Similarity index.

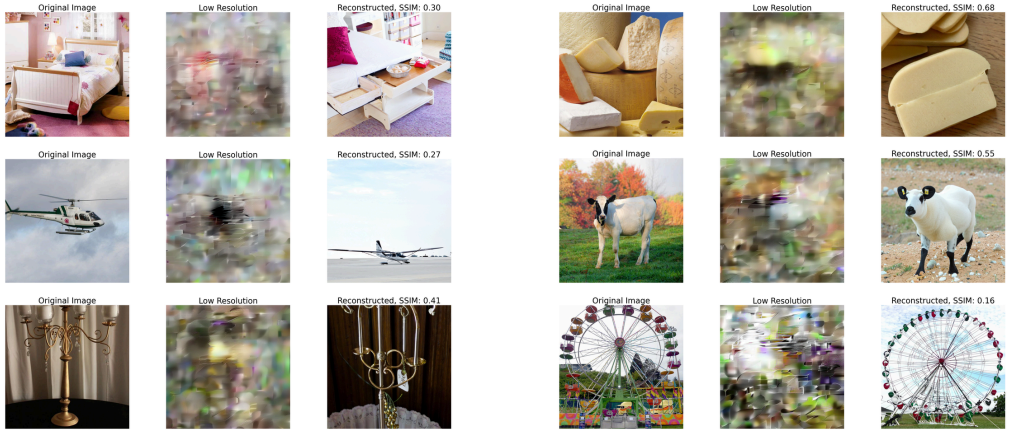

Figure 4: Examples of image reconstructions from neural activity. Each triplet shows the original image (left), a low-resolution baseline (middle), and our reconstruction (right) with corresponding SSIM score. The generations capture essential attributes such as object structure, color distribution, and general content.

4 Discussion & Conclusion

We explored the main factors influencing visual decoding accuracy from primate intracortical recordings, focusing on temporal modeling, model complexity and loss functions. We also examined whether semantic decoding can support high-fidelity image reconstruction.

A soft-attention aggregator combined with a shallow MLP consistently outperformed both linear and recurrent baselines, supporting our core claim that selective temporal integration matters more than modeling complex temporal dependencies end-to-end. High-resolution neural data contain many low-informative timepoints: sequence models propagate all of them, which increases noise sensitivity. In contrast, our mechanism learns to upweight the most informative intervals, effectively filtering out noisy or irrelevant fluctuations. Our model that explicitly retain temporal resolution consistently outperformed temporally averaged baselines. This finding emphasizes that the MUA signal carries rich semantic content that shows a peak from 50 to 100 ms after stimulus onset and slowly decays.

To ensure interpretability, we prioritized retrieval-based evaluation over direct image generation. By mapping neural activity to a fixed, semantically grounded embedding space (CLIP), we bypassed the confounding influence of strong generative priors that often obscure whether high-quality outputs really reflect accurate brain decoding [31]. Based on the reliability of our retrieval results, we extended the framework to generative decoding using a modular two-stage pipeline. We first sampled candidate images from a frozen diffusion prior and then applied rejection sampling based on SSIM, enabling reconstructions that combine semantic accuracy with structural faithfulness.

The scaling trends provide practical guidance for experimental design: to optimize decoding, the most effective thing is to acquire more diverse trials. Increasing input dimensionality—such as through higher channel count—also contributes to performance gains. Next-generation BCI systems could prioritize large-scale data collection and ultra-high-density recordings.

Despite these advances, several limitations remain. Retrieval is restricted by a fixed candidate pool, and CLIP embeddings may not fully align with non-human representations [34]. Pre-trained generative models still introduce priors, and more expressive non-linear models may further improve decoding. Our passive viewing task omits top-down effects such as attention or memory. Finally, translating this work to human settings raises ethical concerns around neural privacy and informed consent [36].

References

[1] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, J. B. Hutchinson, T. Naselaris, and K. Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, Jan 2022.

[2] R. Antonello, A. Vaidya, and A. G. Huth. Scaling laws for language encoding models in fmri, 2023.

[3] H. Banville, Y. Benchetrit, S. d'Ascoli, J. Rapin, and J.-R. King. Scaling laws for decoding images from brain activity, 2025.

[4] T. Bazeille, E. DuPre, H. Richard, J.-B. Poline, and B. Thirion. An empirical evaluation of functional alignment using inter-subject decoding. *NeuroImage*, 245:118683, 2021.

[5] N. Chang, J. A. Pyles, A. Marcus, A. Gupta, M. J. Tarr, and E. M. Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific Data*, 6(1):49, May 2019.

[6] Z. Chen, J. Qing, T. Xiang, W. L. Yue, and J. H. Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding, 2022.

[7] A. Damian, E. Nichani, and J. D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. *arXiv preprint arXiv:2209.15594*, 2022.

[8] M. Ferrante, T. Boccato, F. Ozcelik, R. VanRullen, and N. Toschi. Through their eyes: Multi-subject brain decoding with simple alignment techniques. *Imaging Neuroscience*, 2:1–21, 05 2024.

[9] M. Ferrante, T. Boccato, L. Passamonti, and N. Toschi. Retrieving and reconstructing conceptually similar images from fmri with latent diffusion models and a neuro-inspired brain decoding model. *Journal of Neural Engineering*, 21(4):046001, 2024.

[10] M. Ferrante, T. Boccato, G. Rashkov, and N. Toschi. Towards neural foundation models for vision: Aligning eeg, meg, and fmri representations for decoding, encoding, and modality conversion, 2024.

[11] M. Ferrante, M. Ciferri, and N. Toschi. R&b – rhythm and brain: Cross-subject decoding of music from human brain activity, 2024.

[12] J. L. Gallant, S. Nishimoto, and T. Naselaris. The brain's eye: Decoding mental images from the human brain. *Frontiers in Human Neuroscience*, 6:68, 2012.

[13] J. I. Glaser, A. S. Benjamin, R. H. Chowdhury, M. G. Perich, L. E. Miller, and K. P. Kording. Machine learning for neural decoding. *eneuro*, 7(4), 2020.

[14] M. N. Hebart, O. Contier, L. Teichmann, A. H. Rockter, C. Y. Zheng, A. Kidder, A. Corriveau, M. Vaziri-Pashkam, and C. I. Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, feb 2023.

[15] T. Horikawa and Y. Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):15037, Aug. 2017.

[16] A. G. Huth, W. A. de Heer, T. L. Griffiths, et al. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

[17] Y. Ji, J. Li, H. Ye, K. Wu, J. Xu, L. Mo, and M. Zhang. Test-time computing: from system-1 thinking to system-2 thinking. *arXiv preprint arXiv:2501.02497*, 2025.

[18] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020.

[19] E. R. Kupers, T. Knapen, E. P. Merriam, and K. N. Kay. Principles of intensive human neuroimaging. *Trends in Neurosciences*, 47(11):856–864, 2024.

[20] L. Le, P. Papale, K. Seeliger, A. Lozano, T. Dado, F. Wang, P. Roelfsema, M. van Gerven, Y. Güçlütürk, and U. Güçlü. Monkeysee: Space-time-resolved reconstructions of natural images from macaque multi-unit activity. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 93826–93848. Curran Associates, Inc., 2024.

[21] S. Lin, T. Sprague, and A. K. Singh. Mind reader: Reconstructing complex images from brain activities, 2022.

[22] S. R. Oota, M. Gupta, R. S. Bapi, G. Jobard, F. Alexandre, and X. Hinaut. Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey), July 2023. arXiv:2307.10246 [cs, q-bio].

[23] F. Ozcelik and R. VanRullen. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion, 2023.

[24] P. Papale, F. Wang, M. W. Self, and P. R. Roelfsema. An extensive dataset of spiking activity to reveal the syntax of the ventral stream. *Neuron*, 113(4):539–553.e5, February 2025. NeuroResource.

[25] S. T. Piantadosi and et al. Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, 28(9):844–856, 2024. Trends in Cognitive Sciences, Volume 28, Issue 9.

[26] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[28] M. Sato, K. Tomeoka, I. Horiguchi, K. Arulkumaran, R. Kanai, and S. Sasai. Scaling law in neural data: Non-invasive speech decoding with 175 hours of eeg data, 2024.

[29] P. S. Scotti, A. Banerjee, J. Goode, S. Shabalin, A. Nguyen, E. Cohen, A. J. Dempster, N. Verlinde, E. Yundler, D. Weisberg, K. A. Norman, and T. M. Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors, 2023.

[30] P. S. Scotti, M. Tripathy, C. K. T. Villanueva, R. Kneeland, T. Chen, A. Narang, C. Santhirasegaran, J. Xu, T. Naselaris, K. A. Norman, and T. M. Abraham. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data, 2024.

[31] K. Shirakawa, Y. Nagano, M. Tanaka, S. C. Aoki, K. Majima, Y. Muraki, and Y. Kamitani. Spurious reconstruction from brain activity, 2024.

[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[33] W. Xia, R. de Charette, C. Öztireli, and J.-H. Xue. Dream: Visual decoding from reversing human visual system, 2023.

[34] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

[35] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[36] R. Yuste, S. Goering, B. A. Y. Arcas, G. Bi, J. M. Carmena, A. Carter, J. J. Fins, P. Friesen, J. Gallant, J. E. Huggins, et al. Four ethical priorities for neurotechnologies and ai. *Nature*, 551(7679):159–163, 2017.

A Technical Appendices and Supplementary Material

This section provides additional qualitative and quantitative results to support the main findings of the article, particularly with regard to the advantages of contrastive learning in the brain decoding task. Code is available at this repository: `https://github.com/fidelioc55/monkeys-ieeg-decode`.

In order to qualitatively assess the decoding performance, we select some random images and visualize the retrieved nearest neighbors in the image space based on the predicted CLIP embeddings with our model. For each test image, we display the original stimulus along with its top-5 neighbors retrieved from the test set (Figure 5).

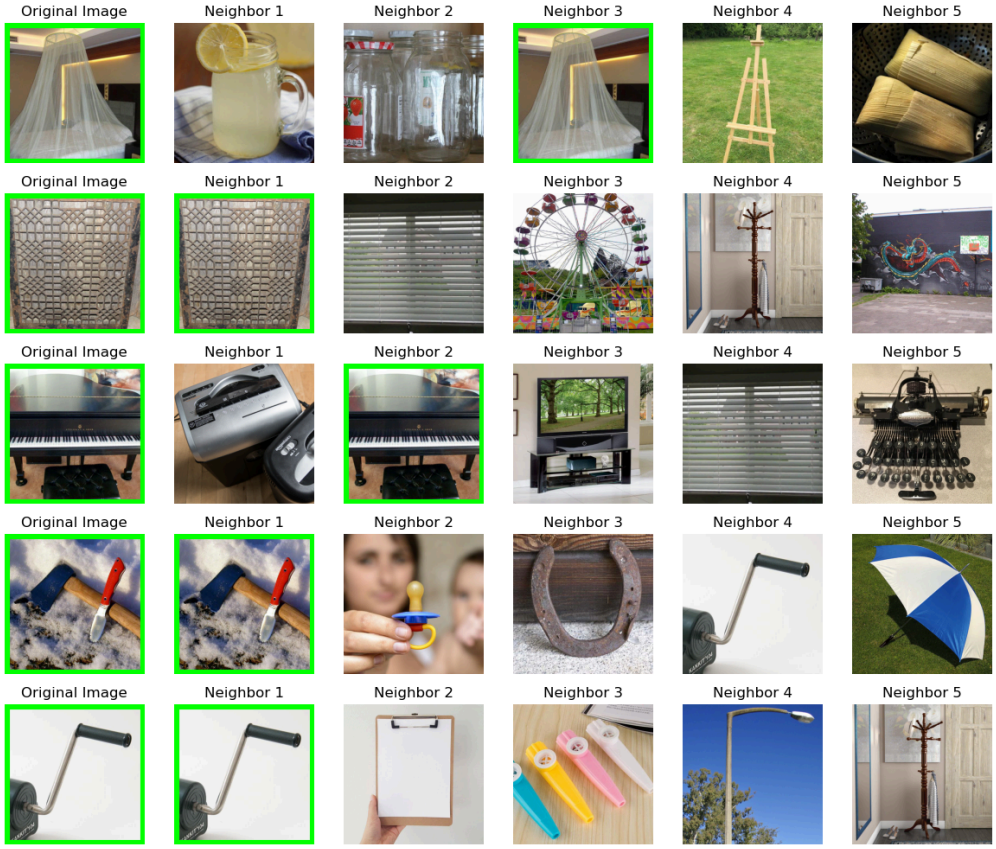

Figure 5: Top-5 image retrieval examples based on predicted embeddings. Each row shows one test sample: the original image (left) and the five nearest neighbors retrieved from the test set.

A.1 Evaluation of Decoding Models

In order to contextualize the performance of our proposed model, we compare it against several standard baselines. The following baseline models were considered:

- Linear Model with Temporal Attention: similar to the proposed model, this variant uses the same temporal attention mechanism, but replaces the final MLP with a single linear layer to project the representation to the CLIP embedding space.
- Linear Model with Temporal Averaging: a linear regression trained on neural features obtained by averaging the MUA signal over the temporal dimension, i.e., reducing each trial from $\mathbb{R}^{T \times C}$ to $\mathbb{R}^{C}$.
- Linear Model on Flattened Input: a linear regression trained on the fully flattened MUA, reshaped from $\mathbb{R}^{T \times C}$ to a 1D-vector $\mathbb{R}^{T \cdot C}$.

- MLP with Temporal Averaging: a feedforward neural network trained on the same time-averaged representation as above, introducing non-linearity over the input features at channel-wise level.
- Recurrent neural network: an LSTM processes the MUA sequence over time. The last hidden state of the LSTM is extracted and passed through a projection layer to obtain the predicted embedding.
- LSTM with Temporal Binning: following standard practice in spike-based decoding [13], we apply temporal binning to aggregate the neural sequence into coarser windows before feeding it to the LSTM. The binning step aggregates consecutive timepoints into non-overlapping windows by averaging neural activity within each window (20 ms as the best performing).
- Temporal Convolutional Network: a model composed of stacked 1D convolutional layers, followed by adaptive average pooling to reduce the temporal dimension. The pooled representation is then passed through an MLP to produce the final embedding.

Metric	MLP/TimeAtt	Linear/TimeAtt	MLP/AvgTime
Pixel Correlation ↑	0.140 ± 0.163	0.151 ± 0.167	**0.156 ± 0.163**
SSIM ↑	**0.364 ± 0.199**	0.356 ± 0.202	0.342 ± 0.206
MSE ↓	0.110 ± 0.020	0.111 ± 0.024	**0.107 ± 0.018**
Cosine Similarity ↑	0.811 ± 0.108	**0.815 ± 0.123**	0.801 ± 0.100
InceptionV3 ↑	**0.868 ± 0.225**	0.836 ± 0.236	0.808 ± 0.230
CLIP ↑	**0.879 ± 0.201**	0.842 ± 0.229	0.827 ± 0.245
EffNet Distance ↓	**0.792 ± 0.144**	0.822 ± 0.139	0.827 ± 0.137
SwAV Distance ↓	**0.492 ± 0.111**	0.516 ± 0.114	0.528 ± 0.117

Table 2: Quantitative evaluation of three decoder variants across multiple metrics. Metrics marked with ↑ are better when higher (e.g., similarity or structural alignment), while those with ↓ are better when lower (e.g., error or distance). Bold values indicate the best score per row.

Table 2 reports the evaluation metrics in image reconstruction for the best three model configurations. Across all metrics, the three variants exhibit similar performance, with subtle differences depending on the evaluation criterion. MLP with time attention shows better results in perceptual and semantic metrics (e.g., CLIP), indicating stronger high-level feature alignment. Linear model or averaging the time information performs well on loss-like metrics and pixel correlation, suggesting effective representation matching and favoring low-level reconstruction. These results underscore the importance of evaluating reconstruction not just with pixel-wise losses but also with perceptual and embedding-based metrics, which better reflect the semantic fidelity of the decoded stimuli.

Layer	**MonkeySee Spatial**	**MonkeySee ST**	**MLP/TimeAtt (Gen)**	**MLP/TimeAtt (Retr)**
conv1	0.358	0.372	0.528	0.874
conv2	0.320	0.334	0.368	0.819
conv3	0.429	0.443	0.335	0.808
conv4	0.385	0.401	0.319	0.804
conv5	0.292	0.318	0.305	0.798
FC6	0.344	0.377	0.522	0.839
FC7	0.534	0.579	0.500	0.827
FC8	0.579	0.610	0.712	0.884

Table 3: Feature correlation (mean Pearson) between AlexNet features extracted from reconstructed and original images, across different decoding models (MonkeySee as a baseline).

Table 3 compares feature correlations across AlexNet layers between our MLP/TimeAtt model (generative and retrieval frameworks) and the MonkeySee baselines. Focusing on the generative variant (third column), our model consistently outperforms both MonkeySee Spatial and Spatiotemporal in the early convolutional layers (conv1 and conv2). This suggests that the rejection sampling procedure, which selects generated images based on low-level SSIM from a pool of candidates, effectively

Model	Hyperparameter	Best Value
TCN	Learning Rate	1e-3
	Conv. Channels	256
	Hidden Dim.	256
	Kernel Size	7
	Num. Layers	2
	Loss Type	CL
LSTM	Learning Rate	1e-4
	LSTM Hidden Dim.	512
	MLP Hidden Dim.	512
	Num. Layers	2
	Loss Type	CL
Linear (TimeFlat)	Learning Rate	1e-3
	Loss Type	CL
Linear (AvgTime)	Learning Rate	1e-3
	Loss Type	CL
MLP (AvgTime)	Learning Rate	1e-3
	Num. Layers	2
	Hidden Dim.	768
	Loss Type	CL
MLP (TimeAtt)	Learning Rate	1e-3
	Num. Layers	2
	Hidden Dim.	768
	Loss Type	CL

Table 4: Best hyperparameter configuration per model.

enhances alignment with early visual features encoded in the brain. In the deeper fully connected layers (FC6–FC8), our model maintains strong performance, surpassing MonkeySee. This indicates that the generative decoder is also capable of capturing high-level semantic content, highlighting the model's ability to represent both perceptual structure and abstract semantics from neural activity.

A.2 Hyperparameter Tuning

We performed an hyperparameter search for each model architecture to identify the optimal configuration (in term of retrieval accuracy), as reported in Table 4. Across all models, the best results were consistently obtained using a contrastive loss.

All experiments were carried out on a high performance server equipped with eight NVIDIA A100 GPUs (80 GB each, interconnected via NVLINK), 256 CPU threads and 2 TB of system memory.

A.3 Contrastive Loss vs MSE Loss

Across multiple decoding architectures, NT-Xent-trained models consistently achieve higher Top-1 and Top-5 accuracy. Although MSE focuses on bringing the predicted and target embeddings as close as possible in Euclidean terms, it does not account for the relative positioning of other embeddings. In contrast, contrastive learning explicitly pushes embeddings of mismatched pairs apart and pulls the matched pairs together, preserving semantic relationships and improving the model's ability to generalize in retrieval tasks. This highlights the benefit of contrastive learning for aligning neural and sensory representations, ultimately improving retrieval and decoding performance.

Beyond empirical performance, our choice is also grounded in theoretical work: recent findings [25] argue that conceptual and semantic representations are best modeled as directional vectors in high-dimensional spaces. In this view, the direction of a vector encodes most of the meaningful structure, making cosine similarity—sensitive to direction but not to scale—particularly well-suited for tasks involving semantic embeddings such as ours.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims: the importance of temporal dynamics over architectural complexity, the value of contrastive learning, and the effectiveness of a two-stage generative decoding pipeline. These claims are consistently supported by the experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper discusses several limitations in Section 4 ("Discussion and Conclusions"), including reliance on a few subjects, the constraints of retrieval-based evaluation, potential mismatches between CLIP embeddings and non-human visual representations, and limitations related to generalizability and generative priors.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results requiring formal proofs. Instead, it provides empirical results and experimental validations to support findings as answers to research questions posed in the introduction.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All data are from a publicly available dataset (THINGS Ventral Stream Spiking Dataset) `https://gin.g-node.org/paolo_papale/TVSD`. The Methods section provides sufficient details about preprocessing, model architectures, training procedures, baselines, and evaluation metrics. The code is released anonymously (in the Appendix).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes] ,

Justification: The data are publicly available [24], and the code is released in anonymized form for NeurIPS revision phase to support reproducibility and freely accessible upon publication for everyone.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The "Material and Methods" section specifies the dataset, training/test splits, loss functions, baselines, and evaluation protocols. Compute details (hyper-parameters tuning) are also provided in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports retrieval accuracies with standard errors, computed over five random seeds to ensure reproducibility and statistical robustness (Table 1 and Figure 5).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer:[Yes]

   Justification: The paper specifies the compute resources used: eight NVIDIA A100 GPUs, 256 GPU threads, and 2 TB RAM (also mentioned in Appendix).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study follows the NeurIPS Code of Ethics. It uses publicly available data collected under approved animal research protocols and considers neuroethical implications in the Discussion (Section 4).

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive impacts (advancing brain-computer interfaces and neuroscience research) and potential negative impacts, including ethical concerns about neural privacy and consent (Section 4, Discussion).

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The generative model (Stable Diffusion XL) includes its own safeguards. The work does not introduce new models or data that inherently pose high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits and cites all datasets (TVSD [24]), models (CLIP [27], Stable Diffusion XL [26], IP-Adapter [35]) and assets. Usage complies with the respective licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: While no new datasets are released, the code and processing scripts will be released with appropriate documentation to reproduce the reported experiments.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The research does not involve crowdsourcing or experiments with human subjects. It uses pre-existing non-human primate datasets.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The study uses data from non-human primates collected under prior ethical approvals. No new human or animal experiments were conducted.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Large language models were not used in the core methods or experiments of this work. Use of LLM was limited to text editing and did not affect the scientific content or methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.