RETHINKING BRAIN DECODING WITH CLIP: THE ROLE OF ADVERSARIAL ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain decoding aims to uncover neural mechanisms by inferring stimulus-related representations from brain signals. In fMRI studies, this is typically achieved by mapping fMRI responses to the latent representations of computational models. Recently, CLIP has become a popular choice for brain decoding due to its rich vision–language embedding space. However, aligning fMRI signals with CLIP representations remains challenging. As CLIP is not explicitly optimized for neural alignment, its representations may capture statistically predictive cues that are only partially reflected in brain activity, limiting decoding performance. In this paper, we investigate whether adversarially robust representations improve neural decoding with CLIP. Adversarial training suppresses reliance on non-robust features and can promote more stable, perceptually structured representations, which may better align with brain activity. We evaluate this by fixing the architecture of fMRI decoders and varying only the target representation (standard CLIP vs. robust variants) on fMRI-image retrieval and zero-shot classification tasks across NSD and GOD datasets. Empirical results show that this simple change consistently improves task performance and yields stronger alignment across multiple metrics. Attribution analysis further reveals consistently low agreement between standard CLIP and its robust variants, suggesting that adversarial robustness reorganizes feature importance in the visual representation. These findings suggest that the choice of target representation influences neural decoding performance and that adversarial robustness may serve as a useful criterion for brain decoding.

1 INTRODUCTION

With the recent advances of Artificial Neural Networks (ANNs), brain decoding using non-invasive neural modalities such as Functional Magnetic Resonance Imaging (fMRI) has gained significant attention. The goal of brain decoding is to uncover the neural mechanisms by inferring stimulus-related representations from brain activity; to this end, the typical pipeline requires a mapping from fMRI to the human-interpretable semantic representations, such as latent embeddings of ANNs, which serve as proxies for the underlying stimuli.

For this semantic proxy, early approaches utilized engineered image features like Gabor wavelet patches (Miyawaki et al., 2008). With the advent of deep learning, end-to-end feature extractors became widely used (Horikawa & Kamitani, 2017; Shen et al., 2019). Recently, large-scale vision–language pretrained models such as CLIP (Radford et al., 2021) have emerged as effective target representations due to their semantically rich embedding space, and are now broadly adopted in brain decoding (Liu et al., 2025; Wang et al., 2025; Li et al., 2025; Wang et al., 2024b; Scotti et al., 2023; Gong et al., 2024; Ma et al., 2025; Efird et al., 2025; Wu et al., 2025).

Despite its popularity, aligning fMRI signals with CLIP representations remains challenging. Recent works (Chen et al., 2025; Vafaei et al., 2026) have shown that neural decoding performance with CLIP can be further improved via a biologically constrained pipeline, such as adding a visual attention mechanism (Desimone et al., 1995) or enforcing CLIP embeddings with geometric similarity to fMRI responses (Vafaei et al., 2026), suggesting that the native CLIP representation space may be suboptimal for neural decoding. Moreover, Shirakawa et al. (2025) indicate that visually plausible reconstructions can arise without faithfully capturing underlying neural representations, highlighting that the decoding pipeline can rely on visual features that are predictive yet less aligned with human

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

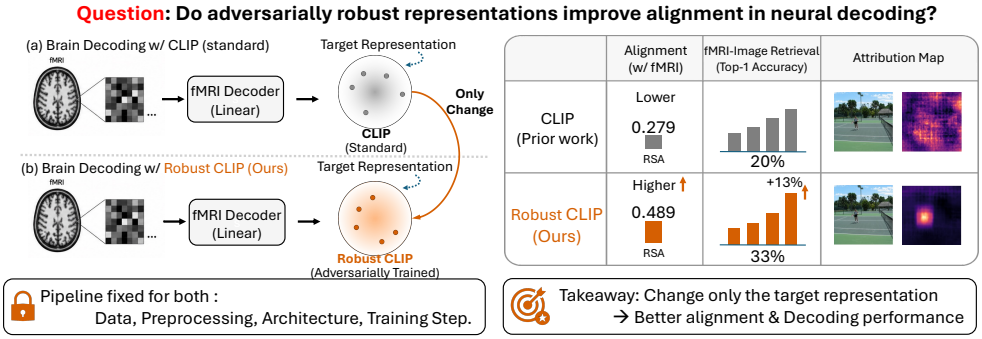


Figure 1: A summary of our work: changing target representation to an adversarially robust one improves neural decoding with CLIP. Empirical results support increased neural alignment and fMRI-image retrieval performance, while attribution maps further reveal qualitatively different image feature usage.

neural representations. At the representation level, since CLIP is optimized for web-scale image–text contrastive learning rather than neural alignment, its embeddings may capture non-robust (Ilyas et al., 2019; Cui et al., 2024) or shortcut-driven features (Geirhos et al., 2020; Wang et al., 2024a) that are predictive for image–text matching but not necessarily well-aligned with brain activity. Together, these findings suggest that improving brain decoding may require revisiting the properties of the target representation itself, beyond architectural or pipeline-level modifications.

Therefore, we hypothesize that adversarially robust CLIP variants can improve neural alignment in brain decoding. Although adversarial perturbations that preserve the perceptual identity of the original image can influence human perception (Veerabadran et al., 2023), human neural representations have been reported to remain relatively invariant to such perturbations (Zhang et al., 2021). This suggests that features exploited by such perturbations may be less reflected in fMRI responses. Adversarial training has been shown to reduce reliance on non-robust features (Ilyas et al., 2019), while adversarially robust models have also been reported to exhibit visual features more closely aligned with human perceptual properties (Tsipras et al., 2019; Geirhos et al., 2021). As such, these observations motivate the central question: "Do adversarially robust representations improve alignment in neural decoding?"

In this paper, we propose adversarial robustness as a criterion for selecting target representations in brain decoding. Importantly, prior studies have shown that adversarially robust vision models better predict neural responses in primate neural data (Kong et al., 2022; Gokce & Schrimpf, 2024). While these studies examined the relationship between adversarial robustness and neural responses from an encoding perspective, we further test whether the benefits of adversarial robustness extend to the complementary decoding setting, where neural responses are mapped to target representations. Empirically, we observe that adversarially trained CLIP variants consistently achieve stronger alignment across multiple metrics and improved fMRI-image retrieval accuracy. Here, we summarize our contribution as follows:

- We show that adversarially trained CLIP variants, FARE (Schlarmann et al., 2024) and TeCoA (Mao et al., 2022), improve alignment between fMRI responses and model representations as measured by Pearson correlation, cosine similarity, and Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008), yielding significantly higher fMRI-image retrieval accuracy on the NSD dataset.
- We further show that FARE and TeCoA improve zero-shot performance on fMRI-image retrieval and fMRI-text classification on the GOD dataset, a dataset carefully designed for zero-shot settings with non-overlapping train-test semantics (Horikawa & Kamitani, 2017).
- Finally, using an attribution-based analysis framework introduced in Palazzo et al. (2020), we show that adversarially trained models exhibit low agreement with standard CLIP in attribution patterns, yielding spatially concentrated and distinct patterns.

2 Background and Setup

2.1 Related Works

Brain Decoding with CLIP. Since its remarkable zero-shot generalization capability (Radford et al., 2021), both image (Ma et al., 2025; Efird et al., 2025; Chen et al., 2025; Gong et al., 2024; Lin et al., 2022) and text latent embeddings (Li et al., 2025; Scotti et al., 2023; Takagi & Nishimoto, 2023) of CLIP have been widely adopted as target representation spaces for brain decoding. Despite its popularity, Shirakawa et al. (2025) show that CLIP-driven reconstructions can be spurious under distribution shifts, raising concerns on reliability. Beyond pipeline-wise concern, several works explored modifying the CLIP representation space itself. Chen et al. (2025) showed that simulating visual attentions to extract selective semantics from CLIP embeddings can improve decoding quality, while Vafaei et al. (2026) demonstrated that aligning the geometry of pretrained embeddings, including CLIP, with fMRI responses improves decoding performance. These works exhibited current limitations on brain decoding with native CLIP representation and presented engineering solutions to modify the representation space. In contrast, our work investigates how intrinsic properties of the representation, specifically adversarial robustness, affect brain decoding performance.

Adversarial Robustness. Adversarial training (Goodfellow et al., 2014; Madry et al., 2019) has been extensively studied for improving adversarial robustness by suppressing non-robust features (Ilyas et al., 2019; Tsipras et al., 2019) which are associated with shortcut-like behavior and part of spurious correlations in previous works (Li et al., 2023; Zhang et al., 2022). Beyond robustness, previous works have shown that adversarially trained models exhibit more structured representations, including human-interpretable gradients (Tsipras et al., 2019; Ross & Doshi-Velez, 2018; Kim et al., 2019), concise and spatially-concentrated attribution maps (Chalasani et al., 2020; Etmann et al., 2019; Gong et al., 2025), and increased reliance on more structured features like shape-bias (Geirhos et al., 2021; Subramanian et al., 2023; Gavrikov et al., 2023). Several studies have also shown that incorporating biological constraints into models can improve robustness and perceptual alignment (Dapello et al., 2020; Li et al., 2019; Guo et al., 2024), connecting adversarial robustness with brain-like representations (Nonaka et al., 2021; Mineault et al., 2025). From an encoding perspective, recent works demonstrated that adversarially robust models can better predict neural responses (Kong et al., 2022; Gokce & Schrimpf, 2024). However, it remains unclear whether the complementary holds—namely, whether adversarially robust representations can be better predicted from neural signals in brain decoding.

2.2 Experiment Setups

2.2.1 Dataset

NSD dataset. Natural Scene Dataset (NSD) (Allen et al., 2022) is currently the largest open-source vision–fMRI dataset. NSD contains 73,000 static images cropped from the MSCOCO dataset (Lin et al., 2015) with corresponding fMRI activation data attained from 8 different subjects viewing these images during multiple fMRI trials. In our work, we followed the data preprocessing pipeline introduced in Efird et al. (2025), using denoised beta responses and selecting voxels based on noise ceiling criteria. Following prior work, we evaluate on the standard NSD test split (Allen et al., 2022; Wang et al., 2024b; Ma et al., 2025) and report results on subjects 1, 2, 5, and 7, which have full trial repetitions. More detailed preprocessing and data handling procedures are provided in the **Appendix A.2**.

GOD dataset. Generic Object Decoding Dataset (GOD) (Horikawa & Kamitani, 2017) was introduced with the motivation for testing the generalization ability of models to novel stimuli (zero-shot). GOD contains 1250 images collected from ImageNet (Deng et al., 2009) and corresponding beta data from 5 different subjects. The last 50 images were used for test data, and 35 repeated trial betas for these test data were averaged to generate a single beta file in our experiment.

2.2.2 Problem Setting

The typical fMRI-based brain decoding pipelines include: 1) a **decoder** that maps fMRI responses to the latent representations of a computational model. 2) a **generator** that reconstructs the stimuli (e.g., images) from these representations. While recent works have often focused on improving reconstruction quality by enhancing the generator component of the pipeline (Takagi & Nishimoto,

2023; Li et al., 2025; Scotti et al., 2023; Liu et al., 2025), our approach isolates the decoder, aiming to study the fidelity of the mapping between fMRI responses and target representation without confounding effects from downstream generation.

To this end, we adopt a controlled setting in which only the target representation is varied while other components remain fixed, including the decoder architecture and training procedure. Implementation details are summarized in the **Appendix A.1.1**.

Training Objective. For the main experiments, we train a linear decoder that maps fMRI voxel data to the predicted embeddings of the target model, using the InfoNCE objective (van den Oord et al., 2019), following Efird et al. (2025). While contrastive training has been shown to achieve higher fMRI-image retrieval accuracy (Ma et al., 2025; Efird et al., 2025), we additionally report results using ridge regression, a standard baseline in brain decoding (Kriegeskorte & Douglas, 2019). The full objectives are provided in **Appendix A.1.2**.

Target Representation Models. While many existing works (Chen et al., 2025; Vafaei et al., 2026) improve brain decoding by engineering target representations, we instead study how the adversarial robustness of representations affects decoding performance.

We consider two adversarially robust CLIP variants, TeCoA (Mao et al., 2022) and FARE (Schlarmann et al., 2024), enabling us to examine how adversarial robustness in the target space influences decoding performance through a simple substitution of target representations. TeCoA is a robust CLIP variant in which the visual encoder is adversarially fine-tuned under text-guided supervision. In contrast, FARE employs an unsupervised adversarial fine-tuning scheme that does not rely on text supervision, while maintaining strong zero-shot capabilities. Their training objectives are given in the **Appendix A.1.3**.

For each robust variant, we evaluate two levels of adversarial strength: FARE-2 / FARE-4 and TeCoA-2 / TeCoA-4, where the suffix indicates the perturbation budget ϵ used during adversarial training (e.g., $\epsilon = 2/255$ for FARE-2). We refer to decoders trained to predict CLIP, TeCoA, and FARE representations as the CLIP decoder, TeCoA decoders, and FARE decoders, respectively.

Evaluation. We used standard metrics from brain decoding literature. The main metric is top-k accuracy (Recall@k) following Efird et al. (2025); Ma et al. (2025) for the fMRI-image retrieval task, both in the NSD and GOD datasets. In GOD, we additionally provide fMRI-text classification accuracy.

We also provide the measures of neural alignment between fMRI signals and target model representations, including Pearson correlation, cosine similarity, and Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) on the NSD test dataset. We evaluate this alignment at two levels. First, decoder-mediated alignment measures Pearson correlation and cosine similarity between predicted and target embeddings. Second, direct representational alignment uses RSA between fMRI responses and target representations.

Statistical testing. Statistical significance is assessed using paired Wilcoxon signed-rank tests (Wilcoxon, 1945) across subjects, comparing each robust variants to CLIP. The results are summarized in **Appendix B.3**.

3 Results

To examine whether adversarially robust representations improve alignment in neural decoding, in this section, we provide the experimental results based on the training and evaluation scheme described in section 2.2.2. We observe a consistent trend that robust models achieve improved decoding performance both on the NSD and GOD datasets.

3.1 Decoders with Adversarially Trained Models Improve Brain Decoding on the NSD Dataset

Decoders with Adversarially Trained Models Show Stronger Alignment. Although the decoder is trained on a sample-wise objective, we observe consistent differences in global alignment across the dataset. Figure 2 summarizes both decoder-based alignment (Pearson correlation and cosine similarity) and representation-level alignment (RSA). Across all metrics, adversarially trained models

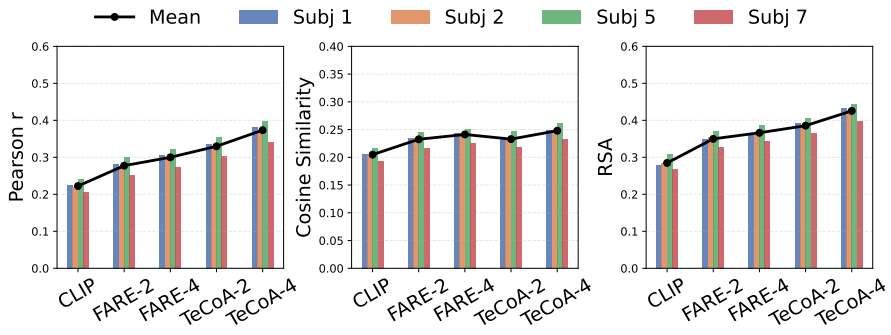
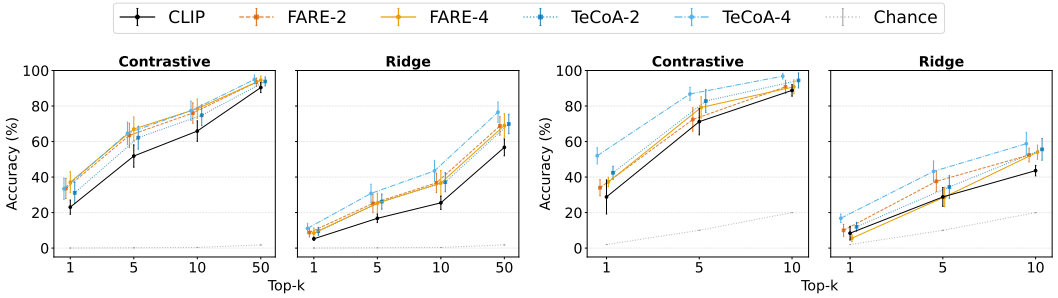


Figure 2: Comparison of alignment metrics between different models on the NSD dataset. All the measures are visualized with subjects 1, 2, 5, 7, and their mean. Adversarially trained variants achieved higher scores in all the metrics.



(a) Top-k accuracy on NSD dataset

(b) Top-k accuracy on GOD dataset

Figure 3: fMRI-image retrieval accuracy of different models. Robust CLIP improves retrieval both on NSD and GOD. Error bars indicate variability across subjects. See **Appendix B**, Table 4 for the complete subject-wise result.

(FARE and TeCoA) achieve higher alignment scores than standard CLIP, with TeCoA-4 consistently being the highest.

Decoders with Adversarially Trained Models Show Higher Retrieval Accuracy. We first re-implemented the contrastive decoding pipeline from Efird et al. (2025). Despite the differences in data split, we confirmed that the standard CLIP decoder is comparable to the original results. For fMRI-image retrieval, TeCoA and FARE decoders achieve consistently higher accuracy than the standard CLIP decoder, with the best-performing model (FARE-4) improving Top-1 accuracy by 13% relative to the CLIP baseline. Figure 3-(a) shows this trend across 4 subjects and their mean. Under ridge regression—which yields lower overall performance than the contrastive decoding (Efird et al., 2025)—the TeCoA and FARE decoders still outperform the CLIP decoder, while the best-performing model changed from FARE-4 to TeCoA-4. We demonstrate consistent improvements under both contrastive learning and ridge regression, suggesting that the choice of target representation plays a more significant role than the training objective. Detailed results are provided in **Appendix B**, Table 3.

Effect of Adversarial Strength on Decoding Performance. We further observe that, under the contrastive decoding setting, both FARE and TeCoA decoders achieve higher retrieval accuracy with strongly adversarially trained models (FARE-4 and TeCoA-4) compared to moderately robust models (FARE-2 and TeCoA-2). Interestingly, this trend does not consistently hold under ridge regression. In this setting, FARE-2 and FARE-4 decoders exhibit closely overlapping performance, whereas the TeCoA-4 decoder continues to outperform the TeCoA-2 decoder.

TeCoA and FARE Exhibit Metric-Dependent Performance Difference. While both consistently outperform the standard CLIP decoder, we further observe that TeCoA and FARE exhibit different

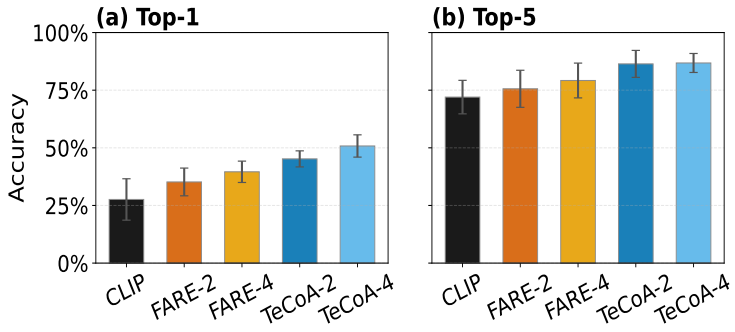


Figure 4: Top-k fMRI-text zero-shot classification on the GOD dataset. Error bars indicate variability across subjects.

performance tendencies, such as relative ranking, depending on metrics. Specifically, TeCoA decoders tend to achieve higher scores on representation-level alignment metrics, including Pearson correlation, cosine similarity, and RSA. In contrast, FARE decoders achieve higher fMRI-image retrieval performance under the contrastive decoding setting. This indicates that no single adversarial training strategy uniformly dominates across all evaluation criteria. Instead, despite both being derived from adversarial training, TeCoA and FARE induce distinct representational properties that are emphasized differently depending on the metric.

3.2 Decoders with Adversarially Trained Models Show Better Zero-Shot Capacity on the GOD Dataset

Recent criticism (Shirakawa et al., 2025) has raised concerns that some CLIP-based pipelines (Li et al., 2025; Scotti et al., 2023) evaluated on the NSD dataset may exhibit limited cross-dataset generalizability, indicating potential evaluation bias, such as double dipping (Kriegeskorte et al., 2009). In this section, we rely on the GOD dataset, which enforces zero-shot settings with non-overlapping train-test semantics, to investigate whether the observed improvements on the NSD dataset are reproducible and generalize under stricter semantic separation.

Decoders with Adversarially Trained Models Maintain Generalizability in Zero-Shot fMRI-image retrieval. Figure 3-(b) shows fMRI-image retrieval accuracy under contrastive decoding and ridge regression on the GOD dataset. Consistent with the NSD results, decoders with adversarially trained variants show higher performance than the standard CLIP decoder. In terms of zero-shot accuracy, the contrastive decoding setting shows TeCoA decoders consistently outperform FARE decoders, with TeCoA-4 being the best-performing model, while ridge regression shows this trend clearly only between strongly adversarially trained models (TeCoA-4 and FARE-4). Notably, FARE-4 exhibits low Top-1 accuracy, but recovers at higher Top-K values.

Decoders with Adversarially Trained Models also Show Higher Zero-Shot fMRI-Text Classification Performance. We further evaluate text-prompt-based zero-shot classification (Radford et al., 2021) using each model's text encoder. Following standard practice in machine learning (Radford et al., 2021; Dhillon et al., 2020) and brain decoding literature (Ma et al., 2025; Yu et al., 2025), we construct prompts of the form "a photo of {cls}" and encode them to text embeddings, which are subsequently compared with predicted vision embeddings from decoders via cosine similarity for classification. This yields an fMRI-text zero-shot classifier. Consistent with retrieval results, both TeCoA and FARE decoders outperform the standard CLIP decoder, while TeCoA decoders also consistently outperform FARE decoders. The results are summarized in Figure 4.

4 Attribution Analysis

Despite the observed improvements in neural decoding performance, it remains unclear why decoders targeting adversarially robust representations exhibit such improvement and how this performance increase connects to the specific image features that decoders may rely on differently. To investigate this, we conduct an attribution-based analysis of feature utilization across decoders. Our findings

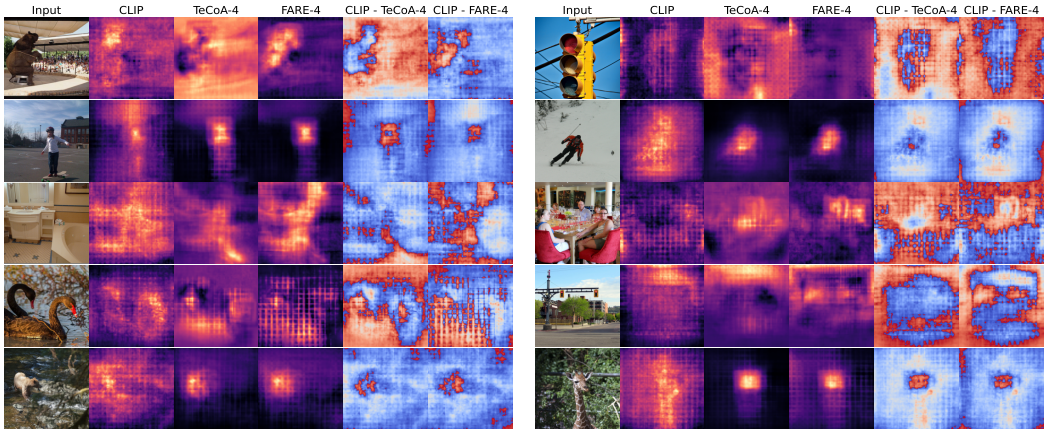


Figure 5: Attribution maps of 10 randomly selected images from the test dataset. From left to right: input image, CLIP, TeCoA-4, FARE-4, difference between CLIP and TeCoA-4, and difference between CLIP and FARE-4. All the attribution maps are generated from Subject 01.

are twofold: (1) decoders trained on adversarially robust CLIP variants exhibit less spatially diffuse and more concentrated attribution patterns than the standard CLIP decoder; (2) these differences extend beyond simple refinement, with robust representations inducing systematic shifts in feature importance, leading to structurally distinct attribution patterns.

4.1 Methodology

We adopt the attribution framework of Palazzo et al. (2020), which estimates pixel-level contributions to neural–visual compatibility without requiring an additional generative model. Full methodological details are provided in the **Appendix C.1**. We adapt the original framework by replacing the neural encoder with the trained fMRI decoder. Specifically, given fMRI data x and an image v, we define the compatibility score as:

$$F(x, v) = d(x)^\top \phi(v), \tag{1}$$

where $d(\cdot)$ is the decoder and $\phi(\cdot)$ is the vision encoder (CLIP, TeCoA, or FARE).

Pixel-wise attribution is computed via occlusion by masking local regions of the image and measuring the change in compatibility:

$$S(p, x, v) = F(x, v) - F(x, m(p) \odot v), \tag{2}$$

where $m(p)$ masks a patch centered at pixel $p = (i, j)$. We apply this procedure across multiple scales and aggregate the results.

Both $d(x)$ and $\phi(v)$ are L_2-normalized, making $F(x, v)$ equivalent to cosine similarity. The resulting attribution map reflects the contribution of each image region to the alignment between fMRI responses and the target representation. We interpret these maps as **importance maps**, indicating which visual features each decoder relies on. Example attribution maps are shown in Figure 5.

4.2 Results

Both Standard CLIP and Robust Variants Exhibit High Global Spatial Diffusion. To quantify the spatial distribution of attribution, we convert each importance map into a probability distribution over pixels and compute entropy over the flattened spatial domain. Let s_i denote the attribution score at pixel i, and define

$$p_i = \frac{s_i}{\sum_j s_j}. \tag{3}$$

We measure entropy as:

$$H(p) = -\sum_i p_i \log p_i. \tag{4}$$

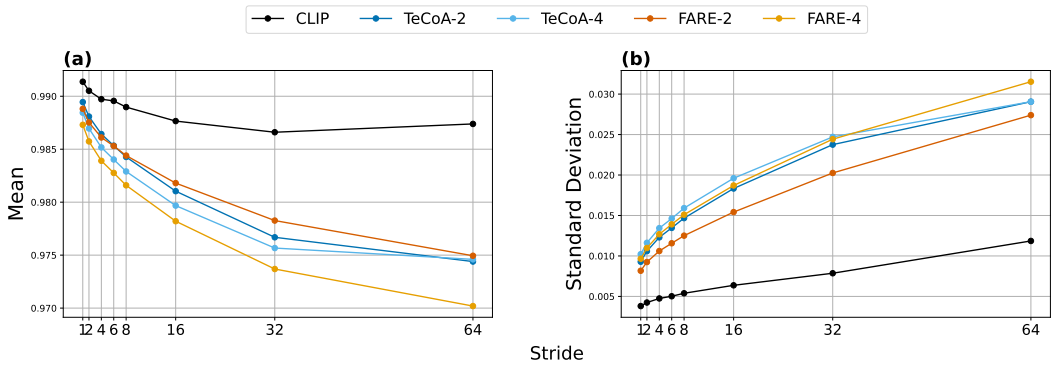


Figure 6: (a) Mean of Normalized Entropy measured by stride. (b) Standard Deviation of Normalized Entropy measured by stride.

To enable comparison across images, we normalize entropy by its maximum value $\log(HW)$ where H and W denote the width and height of the attribution map:

$$\tilde{H}(p) = \frac{H(p)}{\log(HW)} \in [0, 1]. \tag{5}$$

We observe that all models, including CLIP and adversarially robust variants (TeCoA, FARE), exhibit consistently high normalized entropy values, indicating broadly distributed spatial attribution across pixels (see **Appendix C.2**, Table 11). This suggests that global diffusion alone does not distinguish the models.

Robust Variants Exhibit Stronger Scale-Dependent Structural Variation in Attribution Maps. While global entropy is saturated, it does not capture differences in spatial structure across scales. To analyze this, we compute multi-scale attribution maps by applying max pooling with varying strides, producing coarse-grained versions of the original importance maps. Entropy is then computed for each scale.

Figure 6-(a) shows the mean entropy across scales. We observe that CLIP maintains relatively stable entropy across spatial resolutions, indicating limited sensitivity to changes in spatial scale. In contrast, both TeCoA and FARE exhibit a consistent decrease in entropy as spatial resolution becomes coarser, suggesting more structured and scale-dependent attribution patterns. We further examine variability across samples in Figure 6-(b). CLIP shows relatively stable entropy variance across scales, whereas robust models exhibit larger changes in variance under coarse-graining, indicating greater variability in spatial attribution patterns across images. We observe similar trends when using the Gini index (Hurley & Rickard, 2009), as reported in the **Appendix C.2**, Figure 8-(a).

Overall, these results suggest that while global diffusion is similar across models, robust variants exhibit stronger scale-dependent structure in attribution behavior.

Decoders with Adversarially Trained Models Rely on Spatially Distinct Image Features from Standard CLIP. As a follow-up analysis, we examine whether decoders trained with FARE and TeCoA rely on the same image features as the CLIP-based decoder. To this end, we quantify both spatial overlap and rank consistency between attribution maps using top-K Intersection-over-Union (IoU) and Spearman rank correlation, respectively.

Figure 7 summarizes the results. We observe consistently low IoU across all top-K thresholds, including at larger regions (e.g., 20%), indicating that the spatial locations of highly attributed pixels differ substantially between CLIP and robust variants. Notably, IoU at the top-1% level remains extremely low (below 0.1 for all models), suggesting minimal overlap in the most salient regions.

Spearman rank correlation further reveals moderate but highly variable agreement across images. While some samples exhibit high correlation, many show weak or even negative correlation, indicating substantial reordering of feature importance. Together, these results suggest that adversarially trained

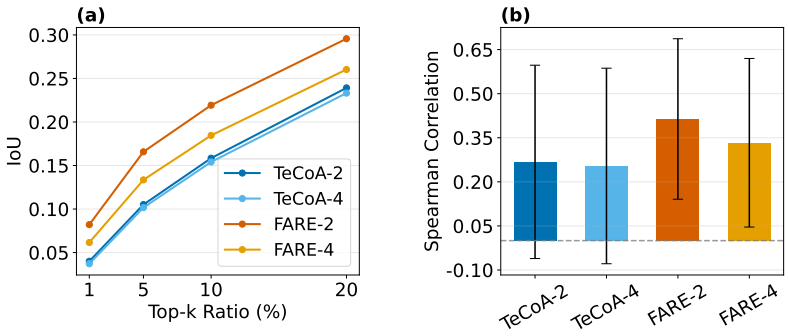


Figure 7: (a) Average Top-k IoU with CLIP (b) Average Spearman Correlation with CLIP. Both plots are generated from attribution maps. The plots are made from Subject 1, while the same trend is observed in subjects 1, 2, 5, and 7 (See **Appendix C.2**, Table 12, Table 13, and Table 14).

models do not simply refine CLIP's feature usage, but instead rely on structurally distinct attribution patterns for alignment.

Interestingly, TeCoA and FARE exhibit different ranges of IoU and correlation values, suggesting that different adversarial training strategies may induce distinct modes of feature reorganization.

5 Discussion

Limitations. While we observe that adversarially robust models improve brain decoding performance and exhibit spatially distinct attribution structures from the standard CLIP decoder, the interpretability of these differences remains limited. In particular, attribution maps do not provide ground-truth correspondence to human visual saliency, making it difficult to determine whether highlighted regions reflect semantically meaningful features for human vision or model-specific predictive cues. We additionally observe qualitative cases where the standard CLIP decoder seems to assign non-negligible importance to background regions (see **Appendix C.3**, Figure 9). While such behavior has been associated with shortcut-like feature usage in vision-language models (Wang et al., 2024a; Li et al., 2024), it is unclear whether these cues are irrelevant for neural processing, as human perception may also leverage contextual information. We therefore interpret these results as differences in representational usage rather than definitive evidence of spurious feature reliance.

Scope and Future direction. Our study does not investigate causal mechanisms, but provides empirical evidence that adversarially robust CLIP representations yield stronger correspondence with human fMRI responses across multiple datasets and evaluation metrics in brain decoding. Recent studies have highlighted limitations of purely correlation-based evaluations, showing highly predictive neural response or representation similarity does not necessarily reflect biological fidelity (Bowers et al., 2022; Vila-Tomás et al., 2025; Martinez-Garcia et al., 2019). Future work could use carefully designed, controlled stimuli to further investigate the underlying mechanisms and factors contributing to the observed correspondence.

6 Conclusion

In this work, we investigated whether adversarial robustness in CLIP representations improves neural decoding performance and alignment with fMRI signals. Across two datasets (NSD and GOD), we found that decoders targeting adversarially trained representations consistently outperform the standard CLIP decoder in fMRI-image retrieval and zero-shot classification, while also achieving higher alignment scores measured by Pearson correlation, cosine similarity, and RSA. Beyond performance gains, our attribution analysis reveals that adversarially trained representations induce structurally distinct feature usage patterns compared to standard CLIP, including differences in spatial concentration and scale-dependent attribution behavior. These findings suggest that robustness not only affects accuracy but also reorganizes how visual features are utilized in brain decoding. Overall, our results highlight that the choice of target representation is a critical factor in neural decoding pipelines, and that adversarial robustness provides a meaningful inductive bias for improving alignment between computational models and brain activity.

AI USE STATEMENT

In this work, we did not use generative AI tools for any of the tasks requiring disclosure that are addressed in the ICLR AI policy guidance. Therefore, the tasks registered as the required disclosure are not applicable to this work. However, we used generative AI tools for other tasks with the recommended disclosure: 1. draft parts of a research paper 2. summarize or analyse existing literature 3. discover research topics or identify gaps. We have reviewed all AI-assisted work manually and further elaborated on it where necessary. For example, "we have reviewed polished parts of writing, confirming their contents and wordings", "we have manually checked the original source of the paper or summary suggested by LLM", etc. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

For reproducibility, we provide detailed information on the implementation, computational resources and costs, hyperparameters, and original references for the main experiments in **Appendix A.1.1**. Details of the datasets used in our experiments and the corresponding data-processing procedures are provided in **Appendix A.2**. Details of the methodology for attribution map generation, including specific hyperparameters used, are provided in **Appendix C.1**. We will also **release the source code** and relevant implementation materials to facilitate reproduction of our results.

REFERENCES

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

Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, et al. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46, 2022.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Somesh Jha, and Xi Wu. Concise explanations of neural networks using adversarial training, 2020. URL https://arxiv.org/abs/1810.06583.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Bridging the gap between brain and machine in interpreting visual semantics: Towards self-adaptive brain-to-text decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21938–21948, 2025.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24625–24634, 2024.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, 33:13073–13087, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Robert Desimone, John Duncan, et al. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

Guneet S. Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification, 2020. URL https://arxiv.org/abs/1909.02729.

Cory Daniel Efird, Alex Murphy, Joel Zylberberg, and Alona Fyshe. Finding shared decodable concepts and their negations in the brain. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=L07zWidgdW.

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability, 2019. URL https://arxiv.org/abs/1905.04172.

Paul Gavrikov, Janis Keuper, and Margret Keuper. An extended study of human-like behavior under adversarial training, 2023. URL https://arxiv.org/abs/2303.12669.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision, 2021. URL https://arxiv.org/abs/2106.07411.

A. T. Gifford, B. Lahner, S. Saba-Sadiya, M. G. Vilas, A. Lascelles, A. Oliva, K. Kay, G. Roig, and R. M. Cichy. The algonauts project 2023 challenge: How the human brain makes sense of natural scenes, 2023. URL https://arxiv.org/abs/2301.03198.

Abdulkadir Gokce and Martin Schrimpf. Scaling laws for task-optimized models of the primate visual ventral stream. *arXiv preprint arXiv:2411.05712*, 2024.

Shizhan Gong, LEI Haoyu, Qi Dou, and Farzan Farnia. Boosting the visual interpretability of clip via adversarial fine-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zixuan Gong, Guangyin Bao, Qi Zhang, Zhongwei Wan, Duoqian Miao, Shoujin Wang, Lei Zhu, Changwei Wang, Rongtao Xu, Liang Hu, et al. Neuroclips: Towards high-fidelity and smooth fmri-to-video reconstruction. *Advances in Neural Information Processing Systems*, 37:51655–51683, 2024.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Manshan Guo, Bhavin Choksi, Sari Sadiya, Alessandro T. Gifford, Martina G. Vilas, Radoslaw M. Cichy, and Gemma Roig. Limited but consistent gains in adversarial robustness by co-training object recognition models with human eeg, 2024. URL https://arxiv.org/abs/2409.03646.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017.

Niall P. Hurley and Scott T. Rickard. Comparing measures of sparsity, 2009. URL https://arxiv.org/abs/0811.4706.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features, 2019. URL https://arxiv.org/abs/1905.02175.

Beomsu Kim, Junghoon Seo, and Taegyun Jeon. Bridging adversarial robustness and gradient interpretability, 2019. URL https://arxiv.org/abs/1903.11626.

Nathan CL Kong, Eshed Margalit, Justin L Gardner, and Anthony M Norcia. Increasing neural network robustness improves match to macaque v1 eigenspectrum, spatial frequency preference and predictivity. *PLOS Computational Biology*, 18(1):e1009739, 2022.

Nikolaus Kriegeskorte and Pamela K. Douglas. Interpreting encoding and decoding models, 2019. URL https://arxiv.org/abs/1812.00278.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540, 2009.

Ang Li, Yifei Wang, Yiwen Guo, and Yisen Wang. Adversarial examples are not real features. *Advances in Neural Information Processing Systems*, 36:17222–17237, 2023.

Haoyu Li, Hao Wu, and Badong Chen. Neuraldiffuser: Neuroscience-inspired diffusion guidance for fmri visual reconstruction. *IEEE Transactions on Image Processing*, 34:552–565, 2025. ISSN 1941-0042. doi: 10.1109/tip.2025.3526051. URL http://dx.doi.org/10.1109/TIP.2025.3526051.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training, 2024. URL https://arxiv.org/abs/2304.05653.

Zhe Li, Wieland Brendel, Edgar Y. Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian H. Sinz, Xaq Pitkow, and Andreas S. Tolias. Learning from brains how to regularize machines, 2019. URL https://arxiv.org/abs/1911.05072.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities, 2022. URL https://arxiv.org/abs/2210.01769.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Pengyu Liu, Guohua Dong, Dan Guo, Kun Li, Fengling Li, Xun Yang, Meng Wang, and Xiaomin Ying. A survey on fmri-based brain decoding for reconstructing multimodal stimuli, 2025. URL https://arxiv.org/abs/2503.15978.

Yongqiang Ma, Yulong Liu, Liangjun Chen, Guibo Zhu, Badong Chen, and Nanning Zheng. Brainclip: Brain representation via clip for generic natural visual stimulus decoding. *IEEE Transactions on Medical Imaging*, 44(10):3962–3972, 2025. doi: 10.1109/TMI.2025.3537287.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. URL https://arxiv.org/abs/1706.06083.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022.

Marina Martinez-Garcia, Marcelo Bertalmío, and Jesús Malo. In praise of artifice reloaded: Caution with natural image databases in modeling vision. *Frontiers in neuroscience*, 13:8, 2019.

Patrick Mineault, Niccolò Zanichelli, Joanne Zichen Peng, Anton Arkhipov, Eli Bingham, Julian Jara-Ettinger, Emily Mackevicius, Adam Marblestone, Marcelo Mattar, Andrew Payne, Sophia Sanborn, Karen Schroeder, Zenna Tavares, Andreas Tolias, and Anthony Zador. Neuroai for ai safety, 2025. URL https://arxiv.org/abs/2411.18526.

Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.

Soma Nonaka, Kei Majima, Shuntaro C Aoki, and Yukiyasu Kamitani. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *IScience*, 24(9), 2021.

Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3833–3849, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024.

Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors, 2023. URL https://arxiv.org/abs/2305.18274.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019.

Ken Shirakawa, Yoshihiro Nagano, Misato Tanaka, Shuntaro C Aoki, Yusuke Muraki, Kei Majima, and Yukiyasu Kamitani. Spurious reconstruction from brain activity. *Neural Networks*, 190: 107515, 2025.

Ajay Subramanian, Elena Sizikova, Najib Majaj, and Denis Pelli. Spatial-frequency channels, shape bias, and adversarial robustness. *Advances in neural information processing systems*, 36: 4137–4149, 2023.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14453–14463, 2023.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2019. URL https://arxiv.org/abs/1805.12152.

Shirin Vafaei, Ryohei Fukuma, Takufumi Yanagisawa, Huixiang Yang, Satoru Oshino, Naoki Tani, Hui Ming Khoo, Hidenori Sugano, Yasushi Iimura, Hiroharu Suzuki, et al. Brain-aligning of semantic vectors improves neural decoding of visual stimuli. *Communications Biology*, 2026.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Vijay Veerabadran, Josh Goldman, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, Jonathon Shlens, Jascha Sohl-Dickstein, Michael C Mozer, et al. Subtle adversarial image manipulations influence both human and machine perception. *Nature communications*, 14 (1):4933, 2023.

Jorge Vila-Tomás, Pablo Hernández-Cámara, Qiang Li, Valero Laparra, and Jesús Malo. A turing test for artificial nets devoted to model human vision. *arXiv preprint arXiv:2502.00721*, 2025.

Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. A sober look at the robustness of clips to spurious features, 2024a. URL https://arxiv.org/abs/2403.11497.

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11333–11342, 2024b.

Yifan Wang, Shaonan Wang, Yunhao Zhang, Changde Du, Cunhang Fan, Dongyang Li, Hongpeng Zhou, Hongyu Zhang, Jixing Li, Quanying Liu, et al. Representation, alignment, and generation: A comprehensive survey of foundation models for non-invasive brain decoding. *bioRxiv*, pp. 2025–11, 2025.

Futa Waseda, Saku Sugawara, and Isao Echizen. Quality text, robust vision: The role of language in enhancing visual robustness of vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4808–4816, 2025.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior, 2025. URL https://arxiv.org/abs/2503.04207.

Muquan Yu, Mu Nan, Hossein Adeli, Jacob S. Prince, John A. Pyles, Leila Wehbe, Margaret M. Henderson, Michael J. Tarr, and Andrew F. Luo. Meta-learning an in-context transformer model of human higher visual cortex, 2025. URL https://arxiv.org/abs/2505.15813.

Chi Zhang, Xiao-Han Duan, Lin-Yuan Wang, Yong-Li Li, Bin Yan, Guo-En Hu, Ru-Yuan Zhang, and Li Tong. Dissociable neural representations of adversarially perturbed images in convolutional neural networks and the human brain. *Frontiers in Neuroinformatics*, 15:677925, 2021.

Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Causaladv: Adversarial robustness through the lens of causality, 2022. URL https://arxiv.org/abs/2106.06196.

A Experiment Details

A.1 Training Scheme

A.1.1 Implementation Details

We provide additional implementation details for training the linear brain decoder. Our training protocol follows prior work (Efird et al., 2025). Specifically, we train a linear contrastive decoder for 5000 iterations. Data augmentation is applied to the fMRI inputs by adding normalized Gaussian noise scaled by a factor of 0.1. We use a learning rate of 1×10^{-3}, a temperature parameter of 0.03, and a batch size of 128. Training is performed without a learning rate scheduler. The same hyperparameter configuration is used for all GOD decoders.

All models were implemented using publicly available HuggingFace model checkpoints. We use a consistent ViT-L/14 backbone across CLIP, TeCoA, and FARE to ensure a controlled comparison.

For ridge regression, we perform a grid search over the regularization parameter λ and select the model with the best validation performance.

A.1.2 Decoder Training Objective

For the decoder, we train a linear model $d(X;\theta) = \hat{Y}$ following Efird et al. (2025). The model maps fMRI voxel data $X = [x_i, ..., x_n]$, $x_i \in \mathbb{R}^v$ to predicted embeddings $\hat{Y} = [\hat{y}_1, ..., \hat{y}_n]$, $\hat{y}_i \in \mathbb{R}^{768}$, where n and v denote the number of samples and subject-specific voxels, respectively. Here, $\hat{Y}$ approximates the target embeddings Y_M from model M.

For the main experiments, we optimize the decoder using the InfoNCE loss (van den Oord et al., 2019):

$$Contrast(A, B) = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp(\boldsymbol{a}_i \cdot \boldsymbol{b}_i/\tau)}{\sum_{j=1}^{N}\exp(\boldsymbol{a}_i \cdot \boldsymbol{b}_j/\tau)}\right) \tag{6}$$

$$L_{\text{infoNCE}}(A, B) = \frac{1}{2}[Contrast(A, B) + Contrast(B, A)] \tag{7}$$

where $A = \{\boldsymbol{a}_i\}_{i=1}^N$ and $B = \{\boldsymbol{b}_i\}_{i=1}^N$ denote batches of paired embeddings, $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are the i-th embeddings in each batch, and τ is a temperature hyperparameter. In our setting, the loss is applied to the predicted embeddings $\hat{Y}$ and target embeddings Y_M, i.e., $L(\hat{Y}, Y_M) = L(d(X;\theta), Y_M)$ where L denotes L_{infoNCE}.

While contrastive training has been shown to improve fMRI-image retrieval accuracy (Ma et al., 2025; Efird et al., 2025), we additionally report results using ridge regression, a standard baseline in brain decoding (Kriegeskorte & Douglas, 2019):

$$\hat{B} = \arg\min_{B} \left\{ \sum_{i=1}^{n} \|y_i - \beta_0 - x_i^\top B\|_2^2 + \lambda \|B\|_F^2 \right\} \tag{8}$$

where y_i denotes the i-th target embedding from Y_M, while $B \in \mathbb{R}^{v \times 768}$ and $\beta_0 \in \mathbb{R}^{768}$ denote the regression weights and intercept, respectively. The intercept is not regularized.

A.1.3 Adversarial Training Objectives

TeCoA (Mao et al., 2022) introduces a text-supervised adversarial fine-tuning objective. Let $\{(x_i, y_i)\}_{i=1}^n$ denote the training dataset. The objective is defined as:

$$L_{\mathrm{TeCoA}}(y, f(\phi, x)) = -\log\left(\frac{e^{f_y(\phi,x)}}{\sum_{k=1}^{K} e^{f_k(\phi,x)}}\right), \tag{9}$$

where $f(\phi, x)$ denotes the classifier logits produced by the model.

While text supervision has been reported to degrade CLIP's zero-shot performance, Schlarmann et al. (2024) proposed FARE, a self-supervised adversarial fine-tuning method that preserves zero-shot capability. Let ϕ_{Org} denote the original CLIP encoder. The FARE objective is given by:

$$L_{\mathrm{FARE}}(\phi, x) = \max_{\|z-x\|_\infty \leq \varepsilon} \|\phi(z) - \phi_{\mathrm{Org}}(x)\|_2^2. \tag{10}$$

A.1.4 Computing Resources

All linear decoder training experiments were conducted on a single NVIDIA RTX 6000 Ada Generation GPU. For the NSD dataset, a single training run required approximately 10 minutes, while training on the GOD dataset completed in under one minute.

Attribution map generation was performed on a high-performance computing cluster using a single NVIDIA H100 SXM5 GPU (94GB HBM). This process required approximately 4–5 days per subject for 982 images.

A.2 Datasets

In this section, we further report the details of the dataset settings

A.2.1 NSD Dataset

In the NSD dataset, each subject has 1–3 times repeated betas on each image, depending on original experiment completion and held-out data for the Algonaut project (Gifford et al., 2023). We followed the fMRI preprocessing settings of the previous work (Efird et al., 2025), which selected denoised beta voxels (*betas_fithrf_GLMdenoise_RR*) by thresholding based on noise ceiling. To avoid possible double dipping (Kriegeskorte et al., 2009), this noise ceiling was computed using only the train dataset split in advance. Although Efird et al. (2025) introduced a data split to make sure 3 repeated betas for the test dataset, we arranged the test set back to the original NSD test set (Allen et al., 2022; Ma et al., 2025), which contains subject-shared 982 image stimuli and corresponding fMRI betas trials to test the consistency of the results on the same stimuli. For this setting, subjects 1, 2, 5, and 7 have full fMRI trials (3 repeated).

Preprocessing Parameters. In our experiment, voxel selection was done by threshold 8.0. After voxel selection, we did not apply further parcellations, such as via Region of Interest (ROI). Lastly, we normalized the selected voxels by session. This procedure and the parameter values followed previous work's setting (Efird et al., 2025).

Table 1: Summary of the Representational Similarity Analysis (RSA) results.

	RSA (Subj01)	RSA (Subj02)	RSA (Subj05)	RSA (Subj07)	Mean
CLIP	0.279	0.283	0.308	0.269	0.285
FARE-2	0.348	0.354	0.371	0.327	0.350
FARE-4	0.365	0.369	0.388	0.344	0.366
TeCoA-2	0.392	0.382	0.405	0.364	0.386
TeCoA-4	0.433	0.428	0.444	0.397	0.425

Table 2: Summary of the Centered Kernel Alignment (CKA) results.

	CKA (Subj01)	CKA (Subj02)	CKA (Subj05)	CKA (Subj07)	Mean
CLIP	0.408	0.416	0.448	0.393	0.416
FARE-2	0.446	0.460	0.478	0.420	0.451
FARE-4	0.461	0.472	0.493	0.432	0.464
TeCoA-2	0.461	0.460	0.484	0.427	0.458
TeCoA-4	0.489	0.492	0.508	0.444	0.483

A.2.2 GOD DATASET

In the GOD dataset, test images are semantically disjoint from training images, with 100 categories used for training and 50 for testing. We use the preprocessed beta signals provided by the dataset authors, which are parcellated into Regions of Interest (ROI). In our experiments, we use all voxels within the visual cortex without additional preprocessing or feature selection.

B NEURAL ALIGNMENT AND FMRI-IMAGE RETRIEVAL.

In this section, we provide the inclusive results of our experiments.

B.1 NEURAL ALIGNMENT BETWEEN MODELS AND FMRI RESPONSES.

Pearson correlation and cosine similarity were measured between the latent embeddings of models and predicted embeddings from trained decoders, while Representation Similarity Analysis (RSA) was computed from the two Representation Difference Matrices (RDMs) between latent embeddings of models and fMRI responses, directly. All the metrics are measured subject-wise on the test dataset to observe inter-subject variability. We additionally report the Centered Kernel Alignment (CKA) results as well in Table 2, which was measured between fMRI responses and latent embeddings of models on the test dataset, following the same procedure as RSA.

B.2 FMRI-IMAGE RETRIEVAL.

We provide additional fMRI-image retrieval results, including TeCoA-0 and QT-AFT (Waseda et al., 2025) alongside the main experimental models (CLIP, TeCoA, and FARE). Table 3 reports the mean and variation of Top-K retrieval accuracy, while Table 4 reports subject-wise Top-1 accuracy on NSD of the main experimental models.

We include TeCoA-0 as an additional baseline because both TeCoA and FARE are fine-tuned on ImageNet (Deng et al., 2009). This allows us to distinguish the effect of ImageNet fine-tuning from that of adversarial robustness. As shown in Table 3, TeCoA-0 improves retrieval performance over CLIP on both GOD and NSD. However, the improvement is smaller than that obtained with the robust variants. On GOD, for example, Top-1 accuracy increases from 0.29 for CLIP to 0.31 for TeCoA-0 (+0.02), compared with 0.37 for FARE-4 (+0.08) and 0.52 for TeCoA-4 (+0.23). Similar trends are observed on NSD.

Therefore, these results suggest that ImageNet domain familiarity likely contributes to the observed improvements, but it cannot fully explain the performance gain. Robust training provides substantial additional gains beyond those obtained by ImageNet fine-tuning alone.

QT-AFT was introduced as an additional adversarially robust variant of CLIP. Consistent with TeCoA and FARE, QT-AFT-4 shows noticeable performance gains across GOD and NSD.

B.3 Statistical significance test

In our graphs of neural alignment, fMRI-Image retrieval, and fMRI-text classification, the reported standard deviations mainly reflect substantial differences in overall decoding performance across subjects. When the models are compared within each subject, however, the relative ordering between standard CLIP and the robust CLIP variants is highly consistent. Thus, the overlapping error bars mainly reflect between-subject variability, rather than inconsistent effects of robust representations. We report the full subject-wise results in the Table 4.

To further quantify these paired subject-level improvements, we conducted two-sided Wilcoxon signed-rank tests (Wilcoxon, 1945) across subjects. We evaluated robust CLIP variants against standard CLIP for: (1) Pearson/RSA/ Cosine Distance on Table 5. (for Figure 2) (2) fMRI-Image retrieval on NSD and GOD on Table 6 and Table 7. (for Figure 3), and (3) fMRI-Text zero-shot classification on GOD on Table 8. (for Figure 4).

Additionally, we offer identical Wilcoxon signed-rank test on fMRI-image retrieval with cross-family setting (TeCoA versus FARE) and whitin-family setting with different adversarial budget ($\varepsilon = 2/255$ versus $\varepsilon = 4/255$) across the dataset. Table 9 presents NSD dataset results, while Table 10 presents GOD dataset results.

Table 5: Wilcoxon test for alignment metrics. μ, σ, p and Δ indicate mean, standard deviation, p-value, and median difference, respectively. σ reflects subject-wise variance.

Model	Pearson r			CosDis			RSA		
	$\mu \pm \sigma$	p	Δ	$\mu \pm \sigma$	p	Δ	$\mu \pm \sigma$	p	Δ
CLIP	$.208 \pm .023$	–	–	$.805 \pm .014$	–	–	$.285 \pm .017$	–	–
TeCoA-2	$.311 \pm .031$	.008	.105	$.777 \pm .016$	.008	.029	$.386 \pm .017$	.125	.098
TeCoA-4	$.353 \pm .033$	.008	.148	$.761 \pm .016$	.008	.044	$.425 \pm .020$	.125	.140
FARE-2	$.260 \pm .028$	.008	.052	$.778 \pm .016$	.008	.026	$.350 \pm .018$	.125	.066
FARE-4	$.281 \pm .030$	.008	.074	$.769 \pm .015$	.008	.034	$.366 \pm .018$	.125	.082

Table 6: Wilcoxon test for NSD fMRI-Image retrieval. μ, σ, p and Δ indicate mean, standard deviation, p-value, and median difference, respectively. σ reflects subject-wise variance.

Model	$\mu \pm \sigma$ (top-1)	p (top-1)	Δ (top-1)	$\mu \pm \sigma$ (top-5)	p (top-5)	Δ (top-5)
CLIP	0.20 ± 0.05			0.46 ± 0.09		
TeCoA-2	0.27 ± 0.07	0.008	0.076	0.57 ± 0.09	0.008	0.11
TeCoA-4	0.30 ± 0.08	0.008	0.11	0.60 ± 0.09	0.008	0.14
FARE-2	0.30 ± 0.07	0.008	0.1	0.58 ± 0.09	0.008	0.12
FARE-4	0.33 ± 0.08	0.008	0.13	0.62 ± 0.10	0.008	0.15

Table 3: Summary results of fMRI-image retrieval by Top-K Accuracy on the GOD and NSD datasets. TeCoA-0, an ImageNet-finetuned CLIP, is given as another baseline, and QT-AFT, another adversarially robust model, is given as an additional robust CLIP variant. The highest number in each setting is highlighted in **bold**.

Dataset	Model	k / Method	1	5	10	50
GOD	CLIP	Contrastive	0.29 ± 0.11	0.71 ± 0.08	0.89 ± 0.04	–
		Ridge	0.08 ± 0.04	0.29 ± 0.06	0.44 ± 0.03	–
	FARE-2	Contrastive	0.34 ± 0.05	0.72 ± 0.07	0.91 ± 0.04	–
		Ridge	0.10 ± 0.04	0.38 ± 0.06	0.52 ± 0.04	–
	FARE-4	Contrastive	0.37 ± 0.03	0.79 ± 0.07	0.91 ± 0.04	–
		Ridge	0.06 ± 0.02	0.29 ± 0.06	0.54 ± 0.04	–
	QT-AFT	Contrastive	0.38 ± 0.06	0.80 ± 0.05	0.90 ± 0.03	–
		Ridge	0.07 ± 0.02	0.30 ± 0.05	0.54 ± 0.04	–
	TeCoA-0	Contrastive	0.31 ± 0.08	0.72 ± 0.05	0.88 ± 0.05	–
		Ridge	0.10 ± 0.05	0.33 ± 0.06	0.53 ± 0.06	–
	TeCoA-2	Contrastive	0.42 ± 0.04	0.83 ± 0.07	0.94 ± 0.05	–
		Ridge	0.12 ± 0.03	0.34 ± 0.07	0.56 ± 0.07	–
	TeCoA-4	Contrastive	$\mathbf{0.52} \pm 0.05$	$\mathbf{0.87} \pm 0.04$	$\mathbf{0.97} \pm 0.02$	–
		Ridge	$\mathbf{0.17} \pm 0.03$	$\mathbf{0.43} \pm 0.07$	$\mathbf{0.59} \pm 0.07$	–
NSD	CLIP	Contrastive	0.20 ± 0.05	0.46 ± 0.09	0.60 ± 0.09	0.88 ± 0.05
		Ridge	0.04 ± 0.01	0.15 ± 0.04	0.23 ± 0.05	0.53 ± 0.07
	FARE-2	Contrastive	0.30 ± 0.07	0.58 ± 0.09	0.71 ± 0.09	0.92 ± 0.04
		Ridge	0.07 ± 0.03	0.22 ± 0.06	0.33 ± 0.07	0.65 ± 0.07
	FARE-4	Contrastive	$\mathbf{0.33} \pm 0.08$	$\mathbf{0.62} \pm 0.10$	$\mathbf{0.74} \pm 0.08$	$\mathbf{0.93} \pm 0.04$
		Ridge	0.07 ± 0.03	0.22 ± 0.06	0.33 ± 0.08	0.66 ± 0.08
	QT-AFT	Contrastive	$\mathbf{0.33} \pm 0.08$	$\mathbf{0.62} \pm 0.10$	$\mathbf{0.74} \pm 0.09$	$\mathbf{0.93} \pm 0.04$
		Ridge	0.08 ± 0.03	0.24 ± 0.07	0.35 ± 0.08	0.67 ± 0.08
	TeCoA-0	Contrastive	0.24 ± 0.06	0.51 ± 0.09	0.64 ± 0.09	0.89 ± 0.05
		Ridge	0.06 ± 0.02	0.19 ± 0.05	0.28 ± 0.06	0.60 ± 0.07
	TeCoA-2	Contrastive	0.27 ± 0.07	0.57 ± 0.09	0.70 ± 0.08	0.92 ± 0.05
		Ridge	0.08 ± 0.03	0.23 ± 0.05	0.34 ± 0.07	0.66 ± 0.07
	TeCoA-4	Contrastive	0.30 ± 0.08	0.60 ± 0.09	0.73 ± 0.08	0.93 ± 0.04
		Ridge	$\mathbf{0.10} \pm 0.03$	$\mathbf{0.28} \pm 0.06$	$\mathbf{0.40} \pm 0.07$	$\mathbf{0.73} \pm 0.07$

Table 4: Subject-wise Top-1 accuracy on the NSD dataset. Across subject, decoders with adversarially trained CLIP variants consistently improve Top-1 accuracy relative to CLIP.

Model	Subject / Method	Subj01	Subj02	Subj03	Subj04	Subj05	Subj06	Subj07	Subj08	Mean
CLIP	Contrastive	0.25	0.21	0.16	0.17	0.28	0.22	0.18	0.12	0.20
	Ridge	0.06	0.05	0.04	0.04	0.06	0.05	0.04	0.02	0.04
FARE-2	Contrastive	0.37	0.32	0.24	0.27	0.40	0.33	0.25	0.18	0.30
	Ridge	0.09	0.08	0.05	0.06	0.12	0.08	0.06	0.04	0.07
FARE-4	Contrastive	0.41	0.36	0.28	0.29	0.43	0.36	0.28	0.20	0.33
	Ridge	0.09	0.08	0.05	0.07	0.13	0.07	0.04	0.04	0.07
TeCoA-2	Contrastive	0.33	0.29	0.23	0.25	0.39	0.31	0.23	0.17	0.27
	Ridge	0.09	0.09	0.07	0.07	0.13	0.10	0.07	0.05	0.08
TeCoA-4	Contrastive	0.36	0.32	0.26	0.27	0.41	0.34	0.24	0.17	0.30
	Ridge	0.10	0.11	0.08	0.08	0.15	0.12	0.08	0.06	0.10

Table 7: Wilcoxon test: fMRI-Image retrieval on GOD. μ, σ, p and Δ indicate mean, standard deviation, p-value, and median difference, respectively. σ reflects subject-wise variance.

Model	$\mu \pm \sigma$ (top-1)	p (top-1)	Δ (top-1)	$\mu \pm \sigma$ (top-5)	p (top-5)	Δ (top-5)
CLIP	0.29 ± 0.11			0.71 ± 0.08		
TeCoA-2	0.42 ± 0.04	0.062	0.18	0.83 ± 0.07	0.062	0.12
TeCoA-4	0.52 ± 0.05	0.062	0.26	0.87 ± 0.04	0.062	0.14
FARE-2	0.34 ± 0.05	0.25	0.04	0.72 ± 0.07	1	0
FARE-4	0.37 ± 0.03	0.19	0.1	0.79 ± 0.07	0.062	0.08

Table 8: Wilcoxon test for GOD fMRI-Text Classification. μ, σ, p and Δ indicate mean, standard deviation, p-value, and median difference, respectively. σ reflects subject-wise variance.

Model	$\mu \pm \sigma$ (top-1)	p (top-1)	Δ (top-1)	$\mu \pm \sigma$ (top-5)	p (top-5)	Δ (top-5)
CLIP	0.29 ± 0.11			0.71 ± 0.08		
TeCoA-2	0.43 ± 0.05	0.062	0.18	0.83 ± 0.07	0.062	0.12
TeCoA-4	0.52 ± 0.05	0.062	0.26	0.86 ± 0.05	0.062	0.14
FARE-2	0.34 ± 0.06	0.38	0.02	0.72 ± 0.07	0.88	0
FARE-4	0.37 ± 0.03	0.19	0.1	0.79 ± 0.07	0.062	0.08

Table 9: Wilcoxon test for NSD fMRI-image retrieval with Cross-family and whitin-family pairwise results among robust CLIP variants. p and Δ indicate p-value and median difference, respectively.

Model comparison	p (top-1)	Δ (top-1)	p (top-5)	Δ (top-5)
FARE-2 − TeCoA-2	0.008	0.018	0.008	0.0109
FARE-4 − TeCoA-4	0.008	0.0287	0.031	0.0165
FARE-4 − FARE-2	0.008	0.0319	0.008	0.032
TeCoA-4 − TeCoA-2	0.008	0.0222	0.008	0.029

Table 10: Wilcoxon test for GOD fMRI-image retrieval with Cross-family and whitin-family pairwise results among robust CLIP variants. p and Δ indicate p-value and median difference, respectively.

Model comparison	p (top-1)	Δ (top-1)	p (top-5)	Δ (top-5)
FARE-2 − TeCoA-2	0.0625	-0.08	0.0625	-0.12
FARE-4 − TeCoA-4	0.0625	-0.16	0.125	-0.08
FARE-4 − FARE-2	0.188	0.04	0.0625	0.06
TeCoA-4 − TeCoA-2	0.0625	0.08	0.125	0.04

C Attribution Maps

C.1 Methodology

Given neural representation e and image v, compatibility is defined as:

$$F(e, v) = \varphi(e)^T \theta(v), \tag{11}$$

where φ and θ are modality-specific encoders.

Pixel-wise attribution is computed via occlusion:

$$S(p, \sigma, e, v) = F(e, v) - F(e, m_\sigma(p) \odot v), \tag{12}$$

where $m_\sigma(p)$ masks a patch of size σ, centered at pixel coordinate $p = (i, j)$. The final attribution map aggregates across scales:

$$S(p, e, v) = \sum_\sigma S(p, \sigma, e, v). \tag{13}$$

We adapt this framework by replacing the neural encoder $\varphi(\cdot)$ with an fMRI decoder $d(\cdot)$ and the visual encoder θ with pretrained CLIP, TeCoA, and FARE. The compatibility score becomes:

$$F(x, v) = d(x)^T \phi(v), \tag{14}$$

where x denotes fMRI voxel data and ϕ denotes the vision encoder (CLIP, TeCoA, and FARE).

Pixel-wise attribution is defined analogously:

$$S(p, \sigma, x, v) = F(x, v) - F(x, m_\sigma(p) \odot v), \tag{15}$$

with final aggregation:

$$S(p, x, v) = \sum_\sigma S(p, \sigma, x, v). \tag{16}$$

We further apply L_2-normalization to embeddings from d and ϕ, making the compatibility equivalent to cosine similarity as noted in the main paper.

Hyperparameters. For the scale parameter σ, we used 3, 5, 9, 17, 33, and 65. For each pixel location p, attribution scores $S(p, \sigma, x, v)$ computed across different scales were summed and then normalized to obtain the final attribution map, following the implementation details from prior work (Palazzo et al., 2020). To reduce computational cost, we additionally applied a spatial stride of 3 when iterating perturbation locations. Concretely, attribution scores were evaluated every three pixels along both spatial dimensions rather than exhaustively at every pixel location.

C.2 Aggregation and Statistics

Here, we provide comprehensive measurement for the metrics introduced in the main paper for attribution maps, Entropy, Top-K IoU, and Spearman correlation. We additionally introduce the summary of Gini index (Hurley & Rickard, 2009) in Figure 8-(a) as supplementary statistics of the scale-dependent sparsity measure. Gini index is broadly used for sparsity measure of the saliency maps Chalasani et al. (2020); Gong et al. (2025). We show that our results show consistent trends where coarse-grained statistics between CLIP and adversarially robust CLIP variants constantly diverge, indicating a structural difference in attribution patterns.

We also offer average pooling implemented version of scale-dependent entropy measured by stride in Figure 8-(b). As max-pool operation may induce bias in sampling of pixels, we alternatively employ average pooling to check consistency of the results. Although mean and standard deviation of the original CLIP show relative changes in magnitude, Figure 8-(b) shows the trend and contrast between CLIP and robust CLIP variants are consistent through strides.

In terms of Top-K IoU and Spearman correlation, the results are summarized from Table 13 and Table 12 respectively. The results are grouped by subject as our attribution maps were generated from these subjects for consistent comparison across shared image sets.

For the comprehensive view, we also present Top-K IoU and Spearman correlation for the all combinations on Table 14. the lower triangle indicates Spearman correlation while the upper triangle indicates Top-K IoU for each k values being 1%, 5%, 10%, 20% in order from left to right. Variability among samples is shown together for Spearman correlation.

Within-family comparisons (FARE-2 versus FARE-4 and TeCoA-2 versus TeCoA-4) show the strongest agreement, with consistently higher IoU across all top-k thresholds, higher Spearman correlations, and lower standard deviation than the cross-family FARE–TeCoA comparisons. Cross-family comparisons remain more similar than comparisons with native CLIP as shown in the Figure 7, but are clearly less aligned than variants trained within the same framework. These results suggest that adversarial robustness induces a shared reorganization of feature attribution, while the specific robust training objective further shapes distinct attribution patterns.

Table 11: Average Entropy and Gini Index of the generated attribution maps per model. The results are grouped by subjects.

		Entropy			Gini		
Subject	Model	Mean	Min	Max	Mean	Min	Max
Subj01	FARE-2	0.989	0.949	0.999	0.252	0.056	0.559
	TeCoA-2	0.989	0.944	0.999	0.236	0.052	0.589
	FARE-4	0.987	0.939	0.999	0.265	0.050	0.612
	TeCoA-4	0.988	0.934	0.999	0.247	0.050	0.628
	CLIP	0.991	0.973	0.999	0.233	0.078	0.420
Subj02	FARE-2	0.989	0.948	0.999	0.246	0.051	0.564
	TeCoA-2	0.990	0.920	0.999	0.232	0.033	0.659
	FARE-4	0.988	0.931	0.999	0.259	0.056	0.627
	TeCoA-4	0.989	0.910	0.999	0.242	0.045	0.694
	CLIP	0.991	0.968	0.999	0.235	0.084	0.460
Subj05	FARE-2	0.988	0.936	0.999	0.262	0.060	0.569
	TeCoA-2	0.989	0.941	1.000	0.241	0.039	0.599
	FARE-4	0.986	0.916	0.999	0.277	0.061	0.687
	TeCoA-4	0.988	0.921	1.000	0.255	0.052	0.660
	CLIP	0.991	0.976	0.999	0.234	0.086	0.401
Subj07	FARE-2	0.989	0.943	0.999	0.248	0.049	0.590
	TeCoA-2	0.990	0.944	1.000	0.233	0.035	0.587
	FARE-4	0.988	0.931	0.999	0.262	0.056	0.632
	TeCoA-4	0.989	0.936	1.000	0.245	0.033	0.620
	CLIP	0.991	0.971	0.999	0.232	0.063	0.431

Table 12: Average Spearman correlation between attribution maps from the standard CLIP and robust CLIP variants. Results are grouped by subject 1, 2, 5, and 7. Variability among samples are shown together.

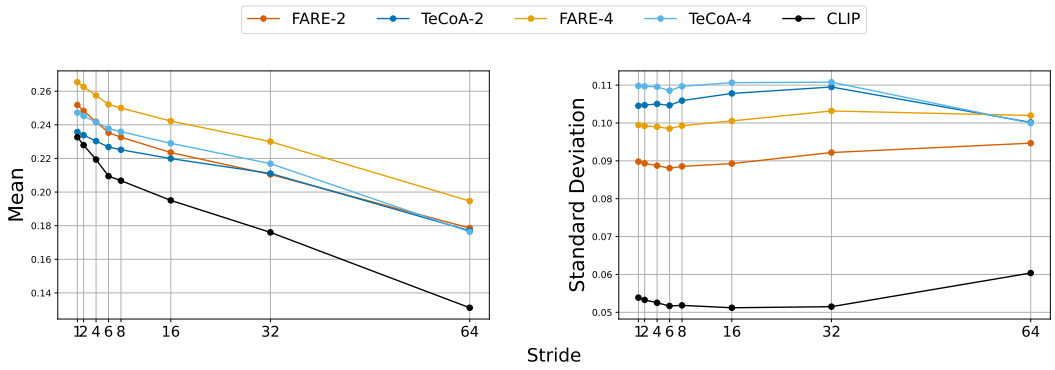
Subject	Model	Spearman Correlation
Subj01	TeCoA-2	0.268 ± 0.329
	TeCoA-4	0.254 ± 0.333
	FARE-2	0.414 ± 0.273
	FARE-4	0.333 ± 0.287
Subj02	TeCoA-2	0.267 ± 0.338
	TeCoA-4	0.257 ± 0.338
	FARE-2	0.416 ± 0.277
	FARE-4	0.335 ± 0.296
Subj05	TeCoA-2	0.284 ± 0.326
	TeCoA-4	0.268 ± 0.326
	FARE-2	0.429 ± 0.275
	FARE-4	0.350 ± 0.287
Subj07	TeCoA-2	0.273 ± 0.326
	TeCoA-4	0.258 ± 0.335
	FARE-2	0.419 ± 0.273
	FARE-4	0.330 ± 0.294

Table 13: Average Top-K IoU between attribution maps from standard CLIP and robust CLIP variants. The results are grouped by subjects. Variability among samples are shown together.

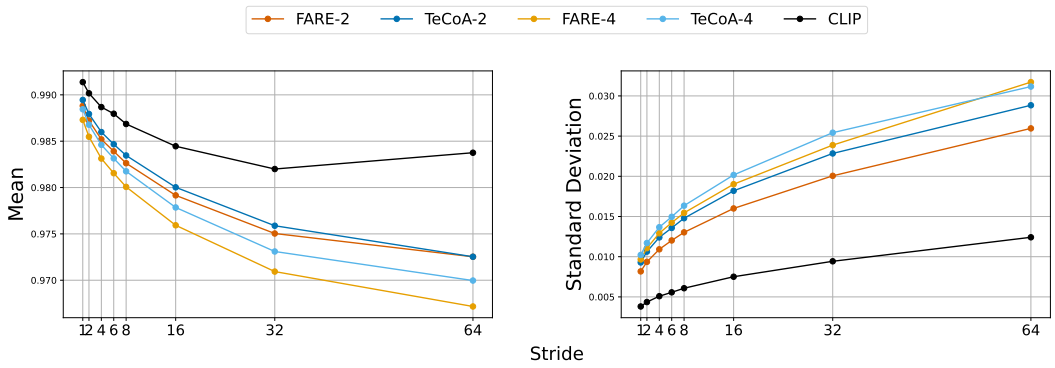
Subject	Model	Top-1%	Top-5%	Top-10%	Top-20%
Subj01	TeCoA-2	0.040 ± 0.084	0.105 ± 0.126	0.158 ± 0.147	0.239 ± 0.157
	TeCoA-4	0.037 ± 0.082	0.102 ± 0.125	0.154 ± 0.145	0.233 ± 0.156
	FARE-2	0.082 ± 0.127	0.166 ± 0.160	0.219 ± 0.159	0.296 ± 0.149
	FARE-4	0.062 ± 0.110	0.133 ± 0.145	0.185 ± 0.149	0.260 ± 0.145
Subj02	TeCoA-2	0.038 ± 0.083	0.105 ± 0.125	0.161 ± 0.146	0.241 ± 0.159
	TeCoA-4	0.034 ± 0.075	0.100 ± 0.126	0.156 ± 0.147	0.236 ± 0.161
	FARE-2	0.080 ± 0.120	0.166 ± 0.152	0.221 ± 0.154	0.301 ± 0.153
	FARE-4	0.057 ± 0.103	0.133 ± 0.142	0.188 ± 0.148	0.266 ± 0.150
Subj05	TeCoA-2	0.043 ± 0.088	0.112 ± 0.131	0.165 ± 0.146	0.248 ± 0.157
	TeCoA-4	0.040 ± 0.084	0.108 ± 0.130	0.161 ± 0.145	0.243 ± 0.156
	FARE-2	0.094 ± 0.140	0.179 ± 0.164	0.233 ± 0.159	0.310 ± 0.153
	FARE-4	0.071 ± 0.122	0.147 ± 0.155	0.200 ± 0.156	0.278 ± 0.153
Subj07	TeCoA-2	0.037 ± 0.076	0.106 ± 0.127	0.161 ± 0.145	0.245 ± 0.160
	TeCoA-4	0.036 ± 0.077	0.102 ± 0.131	0.156 ± 0.148	0.239 ± 0.161
	FARE-2	0.077 ± 0.119	0.162 ± 0.152	0.218 ± 0.152	0.302 ± 0.152
	FARE-4	0.054 ± 0.097	0.130 ± 0.143	0.184 ± 0.146	0.265 ± 0.150

Table 14: Average Top-K IoU and Average Spearman correlation for all the model combinations. Upper triangle: Top-k IoU in order $1\%/5\%/10\%/20\%$ from left to right. Lower Triangle: Spearman correlation with std.

	CLIP	TeCoA-2	TeCoA-4	FARE-2	FARE-4
CLIP	–	.04/.11/.16/.24	.04/.10/.15/.23	.08/.17/.22/.30	.06/.13/.18/.26
TeCoA-2	$.27 \pm .33$	–	.34/.48/.53/.59	.16/.26/.31/.37	.20/.31/.36/.43
TeCoA-4	$.25 \pm .33$	$.81 \pm .15$	–	.12/.22/.27/.34	.18/.30/.35/.41
FARE-2	$.41 \pm .27$	$.54 \pm .26$	$.45 \pm .30$	–	.36/.48/.53/.58
FARE-4	$.33 \pm .29$	$.61 \pm .25$	$.57 \pm .28$	$.82 \pm .13$	–



(a) Mean and standard deviation of the Gini index across spatial scales. CLIP maintains relatively stable sparsity statistics across resolutions, whereas robust models exhibit larger changes under coarse-graining.



(b) Scale-dependent entropy mean and standard deviation measured with average pooling. Similar to the original max-pooling implementation, consistent contrast is observed between CLIP and robust variants.

Figure 8: Statistics of the scale-dependent Gini index (Hurley & Rickard, 2009) and average-pooled entropy across models. Robust variants exhibit larger changes in attribution distributions under spatial coarse-graining compared to CLIP.

C.3 Further Visualizations

C.3.1 Contrast of Attribution Maps

In this section, we provide additional qualitative examples of attribution maps. While the examples of the Figure 5 were randomly sampled from the FARE-4 and TeCoA-4 decoders, we further explore qualitative differences using an alternative sampling strategy.

Specifically, Figure 9 presents a subset of test images selected based on the Jensen–Shannon divergence (JSD), which we use to identify samples exhibiting highly divergent attribution distributions. Let $(\mathcal{X}, \mathcal{F})$ be a measurable space, and let P and Q be two probability distributions on $\mathcal{X}$ such that $P \ll M$ and $Q \ll M$, where $M = \frac{1}{2}(P + Q)$ denotes their mixture distribution. The *Jensen–Shannon divergence* between P and Q is defined as

$$\mathrm{JSD}(P \,\|\, Q) \;=\; \frac{1}{2} D_{\mathrm{KL}}(P \,\|\, M) \;+\; \frac{1}{2} D_{\mathrm{KL}}(Q \,\|\, M), \tag{17}$$

where D_{KL} denotes the Kullback–Leibler divergence:

$$D_{\mathrm{KL}}(P \,\|\, Q) \;=\; \int_{\mathcal{X}} \log\!\left(\frac{dP}{dQ}\right) dP. \tag{18}$$

In our setting, P and Q correspond to normalized attribution maps produced by different models for the same input. By selecting samples with high JSD, we emphasize cases where the spatial distribution of attribution differs most substantially across models.

Qualitatively, these samples reveal a consistent pattern. The standard CLIP decoder tends to produce spatially diffuse attribution maps, whereas TeCoA and FARE decoders exhibit more spatially concentrated responses aligned with objects in the scene. Importantly, although no explicit constraint was imposed to encourage either spatial concentration or diffusion, a consistent qualitative pattern nevertheless emerges across models.

Furthermore, consistent with the observations discussed in the main text, the standard CLIP decoder often assigns non-negligible attribution mass to background regions, while robust CLIP variants yield more localized activation patterns. These differences become particularly pronounced in the high-divergence samples selected by JSD.

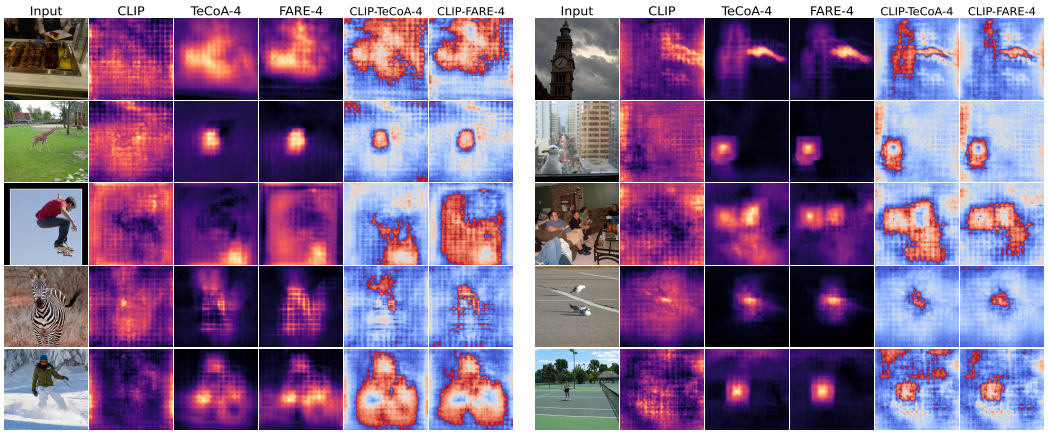


Figure 9: Attribution maps of 10 selected images from the test dataset by high distributional differences measured by JSD (Jensen-Shannon-Divergence). From left to right: input image, CLIP, TeCoA-4, FARE-4, difference between CLIP and TeCoA-4, and difference between CLIP and FARE-4. All the attribution maps are generated from Subject 01.

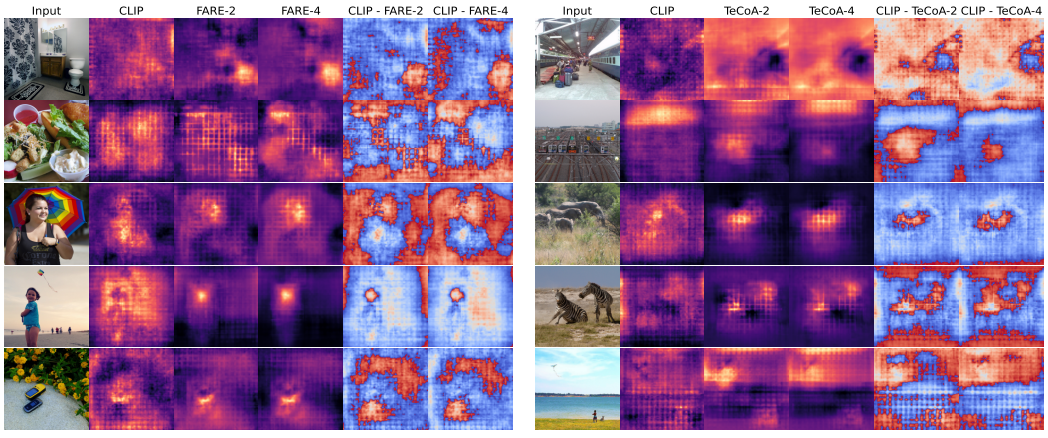


Figure 10: (Left) Attribution maps of 5 randomly selected images from the test dataset. FARE-2 and FARE-4 were selected for the comparison between different adversarial training strengths. From left to right: input image, CLIP, FARE-2, FARE-4, difference between CLIP and FARE-2, and difference between CLIP and FARE-4. (Right) Attribution maps of 5 randomly selected images from the test dataset. TeCoA-2 and TeCoA-4 were selected for the comparison between different adversarial training strengths. From left to right: input image, CLIP, TeCoA-2, TeCoA-4, difference between CLIP and TeCoA-2, and difference between CLIP and TeCoA-4. All the attribution maps are generated from Subject 01.

C.3.2 COMPARISON BETWEEN ADVERSARIAL STRENGTHS

Meanwhile, another interesting question is about the effect of adversarial training strengths (eps) used for adversarial training of the target representations. Therefore, we present the attribution map comparison among CLIP, moderately robust models ($\epsilon = 2$), and strongly robust models ($\epsilon = 4$) in the Figure 10. While the contrast between CLIP and moderately robust models is more noticeable, we observe the salient activation regions between moderately robust models and strongly robust models remain qualitatively similar. The examples are selected randomly from test dataset.

D FURTHER ANALYSIS

Table 15: Effective rank of model representations on NSD and GOD.

Model	Effective Rank (NSD)	Effective Rank (GOD)
CLIP	235.50	201.63
TeCoA-2	115.95	137.01
TeCoA-4	89.64	116.44
FARE-2	164.75	155.41
FARE-4	133.04	135.14

Table 16: Correlation between effective rank and decoding performance (Top-1).

Dataset	Decoder	Pearson r
NSD	Contrastive	-0.7672
NSD	Ridge	-0.9653
GOD	Contrastive	-0.8787
GOD	Ridge	-0.4935

Effective Rank and Decoding Performance. Here, we also provide an analysis of each model's effective dimensionality (Kong et al., 2022). we measured the effective dimensionality of each target space using Effective Rank. As shown in Table 15, the robust CLIP variants generally exhibit lower effective rank than the original CLIP. Table 16 also demonstrates that effective rank is negatively correlated with Top-1 retrieval accuracy across CLIP and its robust variants, suggesting that lower-dimensional target representations may be easier to decode.

However, effective dimensionality does not fully explain the improvements. On NSD, FARE outperforms the corresponding TeCoA variants under contrastive decoding despite having higher effective rank. Moreover, RSA and CKA show similar improvements without relying on model's regression but on the representational geometry, suggesting that the gains are not solely due to lower-dimensional representations being easier to decode.

E BROADER IMPACT

This research aims to contribute to the intersection of neuroscience and AI security, particularly in brain decoding and the adversarial robustness of deep learning models. While brain decoding has potential applications in brain–computer interfaces (BCIs), our work focuses on improving the reliability and performance of brain decoding pipelines. At the same time, brain decoding technologies raise important ethical considerations. Neural data may contain sensitive and personal information, introducing risks related to privacy, consent, and potential misuse. It is therefore critical that research in this area is conducted with appropriate safeguards. In this work, we use only publicly available, de-identified datasets and do not involve any new data collection from human participants. Although our study focuses on methodological improvements rather than direct real-world deployment, we emphasize that future progress in this field should be accompanied by careful consideration of ethical, legal, and societal implications, including strong standards for data protection and responsible use.