NEURAL GEOMETRY CONTRASTIVE LEARNING: TOWARDS BRAIN-MACHINE INTELLIGENCE ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern artificial neural networks (ANNs) have exhibited remarkable capabilities in learning representations from large-scale data, yet improving their complex cognitive abilities (*e.g.*, understanding abstract concepts) remains a major challenge. One potential reason behind this limitation is that there are still many discrepancies between the representational structure formed by ANNs and that of the human brain. Therefore, we propose neural geometry contrastive learning (**NeuroGeo**), a novel method that learns representations from human brain activity and transfers their representational geometric structures to ANNs. Unlike most existing methods, NeuroGeo explicitly aligns geometric structures by preserving how representations vary relative to one another, thereby avoiding strict alignment constraints across heterogeneous modality spaces. This formulation yields rich combinatorial supervision from limited brain activity. Moreover, by introducing learnable virtual anchors and representing each embedding as an anchor-relative displacement, NeuroGeo establishes adaptive reference origins for different modalities and subjects, unifying cross-subject representation learning and relational structure alignment within a contrastive learning framework. We find that NeuroGeo effectively transfers human representational structures to ANNs, significantly enhancing their ability to represent abstract concepts. Results highlight that NeuroGeo yields highly interpretable concept representations, thus leading to substantial performance gains in downstream tasks such as few-shot learning and neural semantic decoding. Code will be released upon acceptance.

1 INTRODUCTION

The evolution of deep learning has been driven by the availability of large-scale data and increasingly powerful architectures. However, current large models (LMs) still lag behind biological brains, especially in their abilities to understand (abstract) concepts (Chen et al., 2026a). For instance, while an LM can easily recognize that an *owl*, *sparrow*, and *penguin* all belong to the category *bird*, it remains challenging to capture the fine-grained semantic structure within this concept (Shani et al., 2026). By contrast, the human brain excel at organizing conceptual knowledge into hierarchical structures such as *owl* $\rightarrow$ *bird* $\rightarrow$ *animal*, which serve as an important foundation for human cognitive abilities (Lake & Baroni, 2023; Hebart et al., 2020).

Recent studies have demonstrated that human-like semantic structures extracted from human behavioral data can provide valuable priors for improving model capabilities (Muttenthaler et al., 2025) such as few-shot learning. However, behavioral data typically provides discrete and indirect supervision, which may only captures partial aspects of human conceptual structures (Chen et al., 2026a). For example, human responses based on triplet odd-one-out tasks reveal relative relationships among concepts, indicating which concept is more similar to another, while failing to quantify the semantic distances between concepts. In contrast, neural activity reflects the latent representational states associated with human cognition. Therefore, learning representations from brain responses, such as functional Magnetic Resonance Imaging (fMRI), offers a promising fashion to model the human semantic structures.

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.

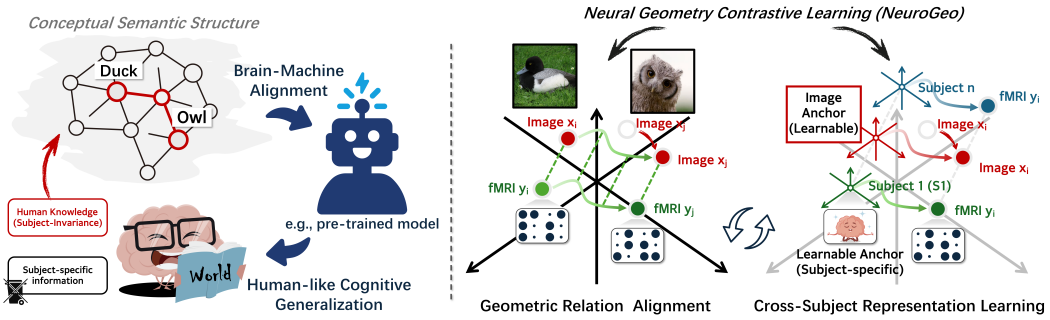

Figure 1: **Left:** Extracting human knowledge (*e.g.*, semantic structure) from the biological brain and transferring them to ANNs offers a promising paradigm for brain–machine intelligence enhancement. **Right:** Unlike existing methods, NeuroGeo aligns relational geometry rather than individual embedding locations, avoiding explicit cross-domain similarity-based matching computation. It further introduces learnable virtual anchors to unify node-level representation learning and pairwise relational alignment across subjects and modalities.

In fact, incorporating neurophysiological constraints into ANNs is an active field of research (Kubilius et al., 2019; Dapello et al., 2020), but previous studies have primarily focused on neural-guided co-training strategies (Federer et al., 2020; Li et al., 2019; Safarani et al., 2021; Dapello et al., 2023), *i.e.*, introducing neural activity prediction as an auxiliary objective to regularize model training (Sinz et al., 2019). Nevertheless, how to transfer representational structures from biological to artificial systems remains largely unexplored (See Fig. 1 Left). Learning representational structures from human brain activity and transferring them to ANNs remain challenging due to the heterogeneity between biological and artificial latent spaces. The low signal-to-noise ratio of neural signals and inter-subject variability further exacerbate this challenge.

To address the challenges, we propose neural geometry contrastive learning (**NeuroGeo**), which aims to directly decode the semantic structures from brain signals and transfer them to ANNs within a unified contrastive framework (See Fig. 1 Right). Our insight is that the useful knowledge of human brain activity lies not in the absolute locations of individual representations, but in their relational geometry, which describes how concepts are organized relative to one another. To this end, NeuroGeo constructs pairs of brain responses and their corresponding stimuli (*i.e.*, image), and guides ANNs to capture how concepts are organized in the human brain by learning pairwise transitions from brain representations. This formulation not only avoids the need to directly calculate embedding similarity between heterogeneous spaces, but also yields $\mathcal{O}(N^2)$ combinatorial supervision from N brain signals.

NeuroGeo also introduces learnable virtual anchors as adaptive reference origins and represents each embedding as an anchor-relative displacement (See Fig. 1 Right). By constructing anchor–brain (or anchor–image) pairs, NeuroGeo unifies node-level representation learning and pairwise relational alignment. More importantly, subject-specific anchors can be used to disentangle individual neural variability from shared semantic geometry, naturally enabling cross-subject learning within a unified framework. This design is motivated by the fact that conceptual knowledge such as *owl* $\rightarrow$ *bird* $\rightarrow$ *animal* is generally shared across individuals, whereas the neural patterns encoding these concepts may exhibit substantial subject-specific variability. In brief, our contributions can be summarized as follows:

- We propose NeuroGeo, a novel learning paradigm that transfers semantic structures from neural signals to ANNs by aligning the relational geometry of the brain and ANN representations rather than the absolute locations of individual embeddings, thereby avoiding direct similarity-based matching across heterogeneous spaces.

- NeuroGeo introduces learnable virtual anchors as adaptive coordinate origins to model subject- and modality-specific variations, thereby enabling node-level representation learning and pairwise relational alignment across subjects and modalities within a unified model.

- We demonstrate that NeuroGeo-enhanced ANNs acquire more human-like conceptual representation structures, achieving substantial improvements across various downstream tasks (such as few-shot image classification) and state-of-the-art performance in neural semantic decoding.

2 Related Work

As NeuroGeo involves both learning neural representations from brain activity (*i.e.*, fMRI) and leveraging them to guide ANNs, here we briefly review representative studies most relevant to our work from these two perspectives.

Neural Representation Learning. In recent years, non-invasive visual neural decoding (e.g., fMRI-to-image reconstruction) has made substantial progress (Scotti et al., 2024; Chen et al., 2025b; Takagi & Nishimoto, 2023a; Chen et al., 2023; 2024a; Lu et al., 2023; Scotti et al.; Wang et al., 2024; Qiu et al., 2025), driven by advances in pre-trained and generative models, such as Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021) and Stable Diffusion (Rombach et al., 2022). These visual reconstruction works have also facilitated the neural representation learning methods, which designed to capture visual perceptual or semantic information from brain activity.

In the era of deep learning, visual reconstruction generally follow a two-stage decoding paradigm (Chen et al., 2026b). The first stage learns neural representations by mapping brain activity into the latent spaces of pretrained models such as CLIP, while the second stage feeds the predicted representations into generative models to reconstruct visual stimuli or generate textual descriptions. Obviously, neural representation learning as a critical component of this decoding paradigm, as it bridges brain activity and generative models. For example, MinD-Vis (Chen et al., 2023) proposed sparse masked brain modeling for self-supervised fMRI representation learning, following a mask-and-reconstruct paradigm similar to masked autoencoders (MAEs) (He et al., 2022). MindGPT (Chen et al., 2025a) leveraged data augmentation and cross-attention mechanisms to learn robust neural semantic representations from fMRI signals. More recently, several studies have investigated cross-subject reconstruction, primarily by learning a shared representational space across subjects.

Brain Activity-Guided ANN Enhancement. Historically, the biological brain has inspired many innovations in deep learning. For example, ANNs typically adopt multi-stage hierarchical architectures inspired by the primate visual cortex (Yamins & DiCarlo, 2016). ANNs have witnessed great success in the computer vision community. However, important behavioral discrepancies remain, particularly in generalization and robustness. For instance, current ANNs are easily fooled by pixel-level perturbations, while biological visual systems are robust to these kinds of attacks.

A prevailing view is that task-optimized ANNs (e.g., object recognition) cannot capture a more physical and causal representation of the world, and representational discrepancies between modern ANNs and biological brains limit the emergence of human-like intelligence. Under the circumstances, Sinz et al. (2019) proposed neural co-training hypothesis. The idea behind it is that although the mechanisms of biological intelligence are not yet fully understood, we can use brain representations to transfer good inductive biases from biological to artificial systems. Large research efforts (Kubilius et al., 2019; Dapello et al., 2020; Federer et al., 2020; Li et al., 2019; Safarani et al., 2021; Dapello et al., 2023) have been devoted to transferring inductive biases from the brain representation to ANNs during the past few years. These studies have generated a remarkable array of new findings. For example, improving the representational similarity between models and the cerebral cortex led to gains in the adversarial robustness of ANNs (Dapello et al., 2023).

More recently, researchers have begun to leverage brain representation learning to extract human conceptual structures and then transfer decoded human knowledge to ANNs through alignment techniques. To the best of our knowledge, Chen et al. (2026a) proposed a novel mental representation-guided supervision method, which first combined neural representation learning with alignment techniques to extract human conceptual structures from brain signals and transfer them to ANNs. They demonstrated that the human-like conceptual structures significantly enhanced pre-trained models' cognitive generalization. However, the mental representation-guided supervision method lacks the ability to generalize neural representations across subjects, and requires explicit cross-domain computation to construct matching correspondences between brain and model representations. Our work expands on this line of research by introducing the neural geometry contrastive

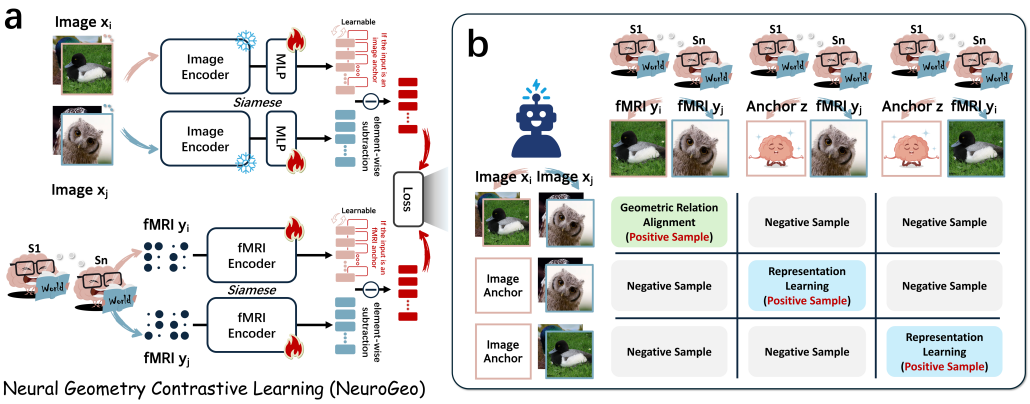

Figure 2: Architecture of the proposed NeuroGeo. **(a)** We employ a Siamese network architecture to reshape the latent space of a pre-trained foundation model (e.g., CLIP) toward a more brain-like structure. **(b)** During training, for each randomly sampled fMRI–fMRI relation, we additionally construct two anchor–fMRI relations. The fMRI–fMRI relation supervises relational geometry alignment, while the two anchor–fMRI relations enable node-level representation learning. Mismatched relations serve as negative samples for contrastive learning.

learning framework that enables cross-subject alignment without explicit cross-domain similarity-based matching computation.

3 Neural Geometry Contrastive Learning

In this section, we first introduce relative geometric relation construction (Sec. 3.1) and learnable virtual anchors (Sec. 3.2), which together enable relational geometry alignment and representation learning across subjects and modalities within a unified framework. We then present a concrete instance of NeuroGeo using a Siamese network architecture and a contrastive loss (Sec. 3.3), as shown in Fig 2. Finally, Sec. 3.4 describes how the NeuroGeo-enhanced ANNs are used for downstream evaluation.

3.1 Relative Geometric Relation Construction

To bridge the gap between brain and model representations, most existing methods adopt a point-wise matching strategy, *i.e.*, directly aligning individual brain representations with their corresponding model representations through cross-domain similarity calculation. In contrast, we hypothesize that the useful prior encoded in brain representations lies not in their absolute embedding locations, but in their conceptual structure, which reflects how concepts are organized relative to one another. Therefore, NeuroGeo is designed to align relative geometric relations between pairs of representations.

Formally, given a set of image-fMRI pairs $\mathcal{D} = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{N}$, where $\mathbf{y}_i$ denotes the Blood Oxygen Level-Dependent (BOLD) signal elicited by visual stimulus $\mathbf{x}_i$, we first encode them into their respective latent spaces:

$$\hat{\mathbf{x}}_i = f_{image}(\mathbf{x}_i), \quad \hat{\mathbf{y}}_i = f_{brain}(\mathbf{y}_i), \tag{1}$$

where $f_{brain}(\cdot)$ and $f_{image}(\cdot)$ is the fMRI and image encoders, respectively. Traditional point-to-point practice directly aligns $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{y}}_i$ using a distance metric, such as $\|\hat{\mathbf{x}}_i - \hat{\mathbf{y}}_i\|_2^2$. Instead, NeuroGeo randomly samples two brain recordings ($\mathbf{y}_i$, $\mathbf{y}_j$) and models their pairwise geometric relation as $d(\mathbf{y}_i, \mathbf{y}_j; f_{brain})$.

In this work, we instantiate $d(\cdot, \cdot)$ using simple element-wise subtraction, defined as

$$\hat{\mathbf{y}}_{i\to j} = d(\mathbf{y}_i, \mathbf{y}_j; f_{brain}) = f_{brain}(\mathbf{y}_j) - f_{brain}(\mathbf{y}_i) \in \mathbb{R}^D. \tag{2}$$

The corresponding geometric relation in the image latent space is obtained using the same operation and denoted as $\hat{\mathbf{x}}_{i\to j}$. Under this formulation, we transfer the relative semantic structure encoded

in brain representations to ANNs by aligning $\hat{\mathbf{y}}_{i\to j}$ and $\hat{\mathbf{x}}_{i\to j}$. Moreover, relative geometric relation construction provides combinatorial relational supervision, expanding the effective training signals from N paired image–fMRI samples to up to $\mathcal{O}(N^2)$ relational pairs.

3.2 Learnable Virtual Anchors

Even though neural representations exhibit substantial inter-subject variability, previous studies have suggested that humans share similar semantic organizations of concepts. For example, individuals generally agree that concepts such as *owl* and *bird* are semantically closer to each other than *lion* and *bird*. Therefore, we hypothesize that neural representations can be decomposed as

$$\hat{\mathbf{y}}_i^s = \mathbf{c}_i + \boldsymbol{\delta}_s, \tag{3}$$

where $\hat{\mathbf{y}}_i^s$ denotes the neural representation from subject s, $\mathbf{c}_i$ represents the shared semantic embedding, and $\boldsymbol{\delta}_s$ denotes the subject-specific variation. Under this assumption, the relative geometric relation between two concepts remains invariant across subjects, *i.e.*,

$$\hat{\mathbf{y}}_{i\to j} = \hat{\mathbf{y}}_j^{s1} - \hat{\mathbf{y}}_i^{s1} = \cdots = \hat{\mathbf{y}}_j^{sn} - \hat{\mathbf{y}}_i^{sn} = \mathbf{c}_j - \mathbf{c}_i. \tag{4}$$

To incorporate Eq. 3 into the NeuroGeo, we introduces a learnable virtual anchor $\mathbf{z}_s$ as an adaptive coordinate origin to absorbs subject-specific variability while preserving the original relational geometry among concepts, *i.e.*, $\hat{\mathbf{y}}_i^s = \mathbf{c}_i + \mathbf{z}_s$.

Since the ground-truth subject-specific variation $\boldsymbol{\delta}_s$ is unavailable during training, we do not explicitly supervise the anchor $\mathbf{z}_s$. However, if the virtual anchor captures the subject-specific variation, the anchor-relative representations should recover the shared semantic component. For the subjects $(s_1, \cdots, s_n)$, this can be written as

$$\hat{\mathbf{y}}_{*\to i} = \hat{\mathbf{y}}_i^s - \boldsymbol{\delta}_s = \mathbf{c}_i \approx \hat{\mathbf{y}}_i^{s1} - \mathbf{z}_{s1} \approx \cdots \approx \hat{\mathbf{y}}_i^{sn} - \mathbf{z}_{sn}. \tag{5}$$

Thus, we only need to maximize the similarity between anchor-relative representations during training, allowing the virtual anchors to be implicitly optimized without requiring explicit supervision of subject-specific variations.

In this process, anchor-relative representations can be regarded as a special case of geometric relations, where the virtual anchor serves as one endpoint and the semantic representation serves as the other. Therefore, they can be seamlessly integrated into NeuroGeo, enabling node-level representation learning and pairwise geometric alignment within a unified learning framework.

3.3 Architecture and Contrastive Loss

NeuroGeo adopts a Siamese-style architecture consisting of two modality branches for fMRI responses and visual stimuli. Each branch includes two weight-sharing encoders that process a pair of inputs and extract their corresponding representations for relative geometric relation construction. Specifically, the image branch consists of a pre-trained vision encoder (e.g., CLIP image encoder) followed by an MLP projection head. Note that the vision encoder remains frozen during training, while the MLP is optimized to reshape the image latent space toward a human-like semantic structure. Following Chen et al. (2026a), the fMRI branch employs a trainable ViT-based encoder.

During the training phase, for each randomly sampled fMRI-fMRI pair ($\mathbf{y}_i$, $\mathbf{y}_j$), we construct two anchor-fMRI pairs ($\mathbf{z}_s$, $\mathbf{y}_i$) and ($\mathbf{z}_s$, $\mathbf{y}_j$). NeuroGeo then jointly optimizes the MLP and fMRI encoder using a contrastive learning loss, which maximizes the similarity between matched relations while separating mismatched relations within a batch, as shown in Fig 2 Right. In particular, relations with matched indices denote positive pairs such as ($\mathbf{y}_i$, $\mathbf{y}_j$) and ($\mathbf{x}_i$, $\mathbf{x}_j$), while relations with mismatched indices are treated as negative pairs such as ($\mathbf{y}_i$, $\mathbf{y}_j$) and ($\mathbf{x}_i$, $\mathbf{x}_k$).

Formally, given a batch of N relation pairs, the loss function of NeuroGeo can be written as:

$$\mathcal{L} = -\frac{1}{N}\sum_{(i,j)} \log \frac{\exp\left(\mathrm{sim}\left(\hat{\mathbf{y}}_{i\to j}, \hat{\mathbf{x}}_{i\to j}\right)/\tau\right)}{\sum_{(p,q)} \exp\left(\mathrm{sim}\left(\hat{\mathbf{y}}_{i\to j}, \hat{\mathbf{x}}_{p\to q}\right)/\tau\right)}, \tag{6}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity and τ is a temperature parameter. NeuroGeo converts both inter-concept relations and anchor-to-concept relations into relational representations, and optimizes them with a unified InfoNCE loss.

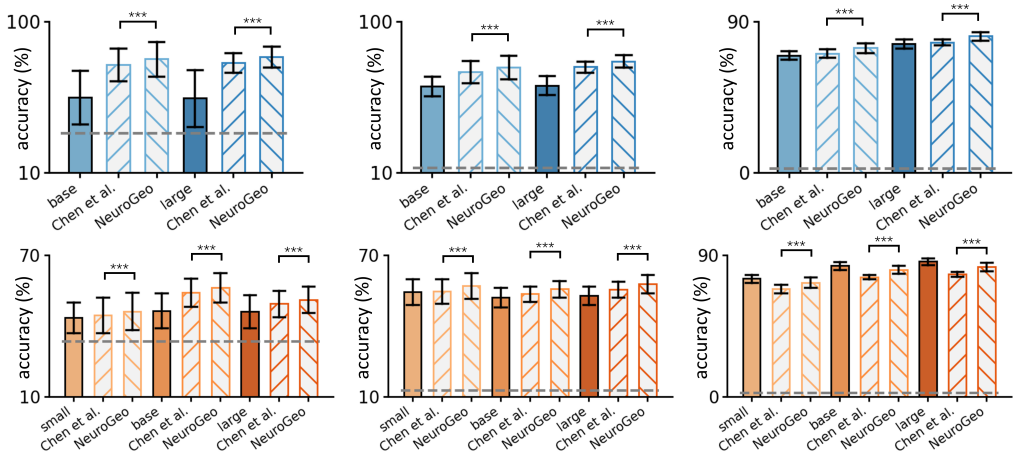

Figure 3: One-shot accuracy across three abstraction levels on different architectures. **Top**: One-shot accuracy of NeuroGeo-enhanced CLIP. **Bottom**: One-shot accuracy of NeuroGeo-enhanced DINOv2. From left to right: 3-, 8-, and 50-class tasks (two-tailed paired t-test, ***: $P < 0.001$).

3.4 Using NeuroGeo

NeuroGeo is a brain-guided learning paradigm designed to reshape the latent space of pre-trained foundation models such as CLIP. Therefore, it cannot be directly used to perform downstream tasks. Nevertheless, NeuroGeo-enhanced representations can be flexibly integrated with various models to perform diverse downstream tasks. For discriminative tasks such as image classification, we can adopt a linear probe on top of the NeuroGeo-enhanced representations, where the parameters of the enhanced ANN are frozen, and only a linear classifier is optimized. For generative tasks such as brain-to-text decoding, it can be seamlessly combined with existing decoding models by replacing the original supervision targets with NeuroGeo-enhanced representations. This allows us to evaluate whether brain-guided representations facilitate the semantic reconstruction from neural signals.

4 Experiments

Datasets. Following Chen et al. (2026a), fMRI signals used to train NeuroGeo are obtained from the publicly available DIR dataset (Shen et al., 2019), which were acquired using a Siemens MAGNETOM Verio 3-Tesla scanner. In image presentation experiments, three healthy subjects viewed 1,250 visual stimuli drawn from 200 representative object categories in the ImageNet (Deng et al., 2009). In DIR, the training set contains 150 categories (8 images per category), and the test set includes 50 categories (1 image per category). More details can be found in the original paper. Since the DIR only contains 50 test visual stimuli, we further collected a large-scale test set from ImageNet-21K, including 50 categories, totaling 66,067 object images (275–2,115 images per category). Note that these categories are identical to those used in the DIR test set and do not overlap with the training categories. This setting ensures that few-shot learning evaluations are conducted on unseen object concepts.

To evaluate whether NeuroGeo-enhanced ANNs can facilitate neural decoding tasks, we further conduct experiments on the Natural Scenes Dataset (NSD) (Allen et al., 2022), which is a large-scale fMRI-image dataset containing high-quality BOLD recordings collected from 8 subjects over 30-40 sessions using a 7-Tesla scanner. Following common practice, our neural decoding experiments employ 27,750 trials from subjects 1, 2, 5, and 7, where 2,770 trials (including 982 image stimuli) are used as the test set, while the remaining trials are used for training. The visual stimuli in NSD are drawn from the COCO dataset (Lin et al., 2014), and the 982 test stimuli are shared across subjects 1, 2, 5, and 7.

Implementation Details. NeuroGeo can integrate with any pre-trained foundation model. In this work, we employ two types of pre-trained foundation models: 1) multimodal vision-language models (Radford et al., 2021), including CLIP-base (ViT-B/16) and CLIP-large (ViT-L/14) trained by

Model	Dataset	Subject	Language Similarity Metrics ↑							
			B@1	B@2	B@3	B@4	METEOR	ROUGE	CIDEr	SPICE
UniBrain (Mai & Zhang, 2023)	NSD	1	-	-	-	-	16.90	22.20	-	-
Brain Captioning (Ferrante et al., 2023)	NSD	1, 2, 5, 7	54.82	35.14	21.96	14.04	16.24	40.39	38.86	8.84
Takagi *et al.* (Takagi & Nishimoto, 2023b)	NSD	1, 2, 5, 7	35.22	16.37	7.22	3.30	9.77	24.61	13.52	4.88
MindGPT (Chen et al., 2025a)	NSD	1, 2, 5, 7	56.91	36.95	23.41	15.87	20.04	38.41	43.56	10.97
Neuro2Language (Shen et al., 2024)	NSD	1, 2, 5, 7	57.19	37.17	23.78	15.85	18.60	36.67	49.51	12.39
UMBRAE* (Xia et al., 2024)	NSD	1, 2, 5, 7	57.38	37.98	25.18	17.11	18.27	41.99	51.68	11.82
UMBRAE (Xia et al., 2024)	NSD	1, 2, 5, 7	59.09	40.13	27.05	18.40	19.24	43.64	57.76	12.42
MindArt (Chen et al., 2024b)	NSD	1, 2, 5, 7	57.26	38.62	25.17	17.71	21.93	42.86	45.95	12.27
Mind-SA (Chen et al., 2025b)	NSD	1, 2, 5, 7	59.76	39.81	27.05	18.86	38.08	42.72	54.03	12.25
Yo'Mind (Chen et al., 2026b)	NSD	1, 2, 5, 7	61.81	42.75	29.76	20.88	38.40	43.63	56.07	12.42
NeuroGeo + MindGPT	NSD	1, 2, 5, 7	62.15	42.63	29.27	20.93	38.01	43.56	**58.89**	**12.62**
NeuroGeo + Yo'Mind	NSD	1, 2, 5, 7	**63.62**	**44.65**	**30.52**	**21.75**	**38.51**	**45.47**	58.42	12.61
MindArt† (Chen et al., 2024b)	NSD	1, 2, 5, 7	64.56	51.57	46.35	31.90	37.54	50.01	78.35	18.21
Mind-SA† (Chen et al., 2025b)	NSD	1, 2, 5, 7	64.78	53.54	46.38	32.46	36.25	54.26	79.48	18.72
Yo'Mind† (Chen et al., 2026b)	NSD	1, 2, 5, 7	65.19	53.53	48.46	33.36	39.25	55.19	81.16	19.25
NeuroGeo + MindGPT †	NSD	1, 2, 5, 7	69.12	54.12	49.24	34.35	39.26	54.78	81.11	19.28
NeuroGeo + Yo'Mind †	NSD	1, 2, 5, 7	**69.98**	**55.03**	**49.97**	**35.01**	**39.89**	**55.70**	**82.45**	**19.37**

Table 1: Brain-to-text reconstruction results. UMBRAE* denotes trained on data from a single subject, whereas UMBRAE is trained using data from all four subjects. † indicate results obtained with Harmon (Qwen2.5)-generated (Wu et al., 2025) captions as ground truth. Some results are derived from Chen et al. (2026b). The best and second-best results are highlighted using **Bold** and Underline, respectively.

OpenAI; 2) self-supervised vision models (Oquab et al., 2024), including DINOv2-small (ViT-S/14), DINOv2-base (ViT-B/14), and DINOv2-large (ViT-L/14) trained by Meta AI. We adopt the same fMRI encoder architecture as Chen et al. (2026a), and used voxels from brain higher visual areas including LOC, FFA, and PPA to guide pre-trained foundation models. The projection MLP consists of a hidden layer with ReLU activation (embedding size of 256) and an output linear layer. NeuroGeo is implemented in PyTorch, and we train it using the Adam (Kingma & Ba, 2014) optimizer (learning rate is 2e-5) with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Following Chen et al. (2026a; 2024b;a), our training data consists of 206,848 object images from ImageNet-21K and 6,000 fMRI signals, with both modalities covering the same 150 object categories.

One-Shot Classification Results. We first investigate how NeuroGeo affects the capacity of ANNs to generalize across different levels of abstraction. A key task to evaluate such generalization ability is one-shot classification, where models are required to recognize novel concepts from only one labeled observation. In the adaptation phase, we train a linear classifier using multinomial logistic regression with cross-entropy loss and an L_2 penalty term. Following Chen et al. (2026a), we adopt the same evaluation pipeline, which uses 66,067 object images from 50 object categories in ImageNet-21K, and construct classification tasks at three levels of semantic abstraction based on the WordNet (Fellbaum, 1998) hierarchy: the original 50 ImageNet categories, their corresponding 8 higher-level semantic categories (*i.e., arthropod, bird, ungulate, home appliance, equipment, musical instrument, vehicle, clothing*), and 3 more abstract super-categories (*i.e., invertebrate, vertebrate, and artifact*).

For each classification task, we report the average accuracy over 100 independent trials for NeuroGeo and Chen et al. (2026a). In each trial, one labeled image per class is randomly sampled for adaptation, while the remaining images are used for evaluation. As shown in Fig. 3, NeuroGeo consistently improves few-shot classification performance across different levels of semantic abstraction, with larger gains observed on higher-level semantic categories, demonstrating that brain representations can provide good inductive bias for enhancing the abstract semantic generalization ability of ANNs. Compared with the latest approaches, NeuroGeo achieves further improvements on the classification of 50 concrete concepts. This may be attributed to the fact that NeuroGeo aligns relative semantic structures rather than enforcing strict point-wise alignment of individual representations, thus better retaining the inherent capabilities of the pre-trained foundation model.

Semantic Decoding Results. In this part, we quantitatively evaluate whether NeuroGeo-enhanced representations can improve brain-to-text decoding performance by comparing against a wide range of baselines, including Mai & Zhang (2023); Ferrante et al. (2023); Takagi & Nishimoto (2023b); Chen et al. (2025a); Shen et al. (2024); Xia et al. (2024); Chen et al. (2025b; 2026b). Since Neuro-

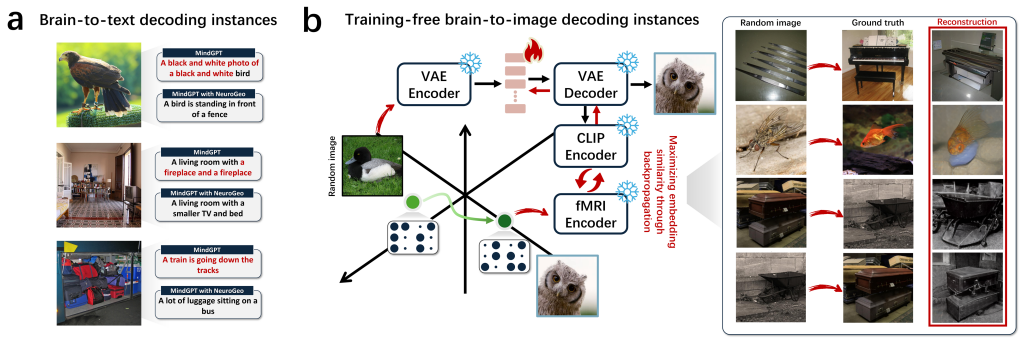

Figure 4: Qualitative results of neural decoding. **(a)** Some brain-to-text reconstruction cases. We use MindGPT as a language decoder. **(b)** Some training-free brain-to-image reconstruction cases. To reconstruct visual stimuli, we optimize the latent representation of the VAE through backpropagation to minimize the discrepancy between its CLIP embedding and the predicted fMRI representation.

Geo is designed for ANN enhancement rather than neural decoding, it does not include a generative decoder and cannot directly perform brain-to-text generation. Therefore, our strategy is to incorporate NeuroGeo into existing decoding frameworks by using NeuroGeo-enhanced representations as semantic supervision signals to guide the decoding process.

Specifically, we integrate NeuroGeo with two existing decoding approaches, namely MindGPT (Chen et al., 2025a) and Yo'Mind (Chen et al., 2026b). The decoding results on NSD (S1, S2, S5, and S7) are reported in Tab. 1. Quantitative results demonstrate that NeuroGeo can be used to improve the decoding performance of MindGPT and Yo'Mind, achieving the SOTA performance on most evaluation metrics. In particular, NeuroGeo-enhanced MindGPT exhibits remarkable improvements over the original MindGPT. For example, it improves the performance by 31.9% and 89.7% on BLEU-4 (B@4) and METEOR, respectively. Interestingly, although MindGPT exhibits inferior performance compared with Yo'Mind, incorporating NeuroGeo substantially narrows the performance gap between the two approaches. Qualitative examples are shown in Fig. 4a. The reason behind this phenomenon may be that NeuroGeo provides brain-aligned supervision signals; meanwhile, the core idea of Yo'Mind is to bridge the discrepancy between machine and brain representations to provide more effective decoding supervision, thereby enhancing brain-to-text reconstruction performance.

Beyond the supervised brain-to-text decoding, we also demonstrate a training-free fMRI-to-image reconstruction approach (Fig. 4b). Given a random image initialization, we optimize the latent representation of a pre-trained VAE through backpropagation to minimize the discrepancy between its CLIP embedding and the predicted fMRI representation. Qualitative cases only illustrates a potential training-free application of our NeruoGeo.

Low-Dimensional Semantic Manifold Analysis. There is considerable evidence that human behaviour or mental representations can be mapped to a low-dimensional, navigable manifold. Therefore, if NeuroGeo-guided ANNs generates human-like representational structures, it should be able to produce a more efficient low-dimensional embedding space. To intuitively verify this, here we visualize the embeddings using principal component analysis (PCA) at different levels of semantic granularity, as shown in Fig. 5a. Each color in the Fig. 5a represents a different concepts, and the position of each concept name is determined based on the average coordinates of its samples.

Through PC visualization, we can find that NeuroGeo-guided representations become more principled and logical, which naturally forms clusters related to high-level abstract concepts such as living and artificial. NeuroGeo produces a more efficient low-dimensional latent space, which is reflected in the reconstruction error and the amount of variance explained by the first few principal components. Quantitative results (See Fig. 5b) show that NeuroGeo achieves lower reconstruction error and a higher proportion of total variance explained by the first few principal components. In contrast, the low-dimensional embedding of CLIP does not clearly organize concepts according to high-level semantic properties, resulting in less discriminative features for abstract concepts. This improved semantic organization may explain why NeuroGeo enhances few-shot learning performance, particularly for abstract concepts that require stronger semantic generalization.

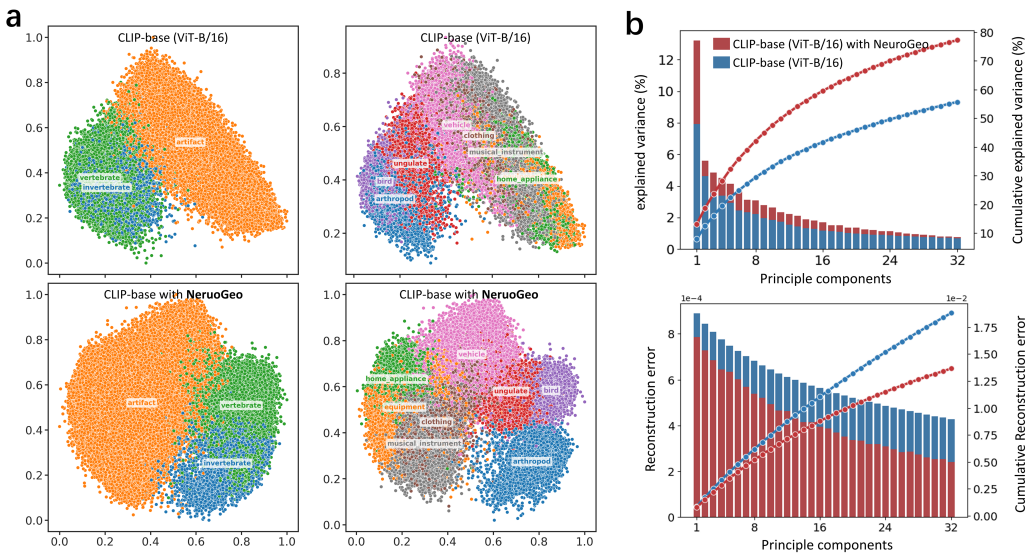

Figure 5: Analysis of embedding spaces using PCA. **(a)** Projection of the representations onto their first two principal components both before (top) and after (bottom) alignment. The representations of the NeuroGeo-enhanced model tend to group themselves based on concepts. **(b)** Comparison of explained variance ratio and reconstruction error between CLIP-base with and without NeuroGeo.

Alignment	Anchors	One-Shot (%) (3-, 8-, and 50-class)	Brain-to-Text Reconstruction				
			B@4	METEOR	ROUGE	CIDEr	SPICE
Geometric Relation	×	(75.9, 71.2, 73.7), (76.0, 72.8, 74.3)	19.48	36.22	42.89	57.96	12.10
Geometric Relation	✓	(76.9, 72.7, 75.1), (78.2, 74.2, 77.2)	20.93	38.01	43.56	58.89	12.62
Embedding	×	(74.9, 70.9, 71.8), (74.8, 71.1, 72.0)	16.42	36.12	41.98	55.88	11.53
Embedding	✓	(75.3, 71.1, 72.8), (75.8, 72.5, 74.4)	17.21	36.79	42.18	56.81	11.79

Table 2: Ablation Studies of NeuroGeo. In the classification tasks, underlined values indicate that NeuroGeo is trained on subjects 1-3, whereas non-underlined values correspond to NeuroGeo trained on subject 3 only. For the decoding tasks, we use MindGPT as a language decoder.

Ablation Studies. To investigate the contribution of each component in NeuroGeo, here we evaluate the effects of relative geometric relation alignment and learnable virtual anchors by comparing the full NeuroGeo model with its corresponding variants (based on CLIP-base). We repeat the brain-to-text decoding and one-shot classification experiments under the same experimental settings. The results are reported in Tab. 2. We find that geometric relation-based alignment achieves better classification and decoding performance than embedding alignment, and the incorporation of learnable anchors further improves one-shot classification accuracy when NeuroGeo is trained on subjects 1-3.

5 Conclusion

In this paper, we introduce NeuroGeo, a brain-guided supervised learning paradigm that designed to transfer the human knowledge encoded in brain activity to ANNs. Unlike existing approaches, NeuroGeo focuses on the relative geometric structures among conceptual representations, which capture how concepts are organized in latent space. Extensive experiments demonstrate that NeuroGeo can effectively reshape the representation spaces of pre-trained ANNs toward human-like semantic structures. More importantly, our results suggest that brain-aligned latent spaces provide valuable structural priors for ANNs, leading to substantial performance gains on downstream tasks such as few-shot learning and brain-to-text decoding. Overall, this work offers new insights into advancing deep learning beyond a scale-driven paradigm toward a brain-machine fusion paradigm that leverages biological intelligence as a source of inductive bias. We hope this work will inspire further exploration of brain-machine intelligence.

AI USE STATEMENT

In this work, we used generative AI tools only for language polishing and grammar checking of the manuscript. We have reviewed all AI-assisted work. Specifically, all AI-assisted language modifications were manually checked by the authors to ensure that they accurately reflect the intended scientific meaning and maintain the integrity of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available fMRI datasets collected in previous studies. We do not collect new human data or conduct additional human subject experiments. All experiments are performed in accordance with the original data usage policies and ethical guidelines of the corresponding datasets. We take responsibility for the integrity and ethical conduct of this research.

REPRODUCIBILITY STATEMENT

Our experiments and analysis are based on the datasets and basic experimental settings introduced in Chen et al. (2026a). We also provide detailed descriptions of NeuroGeo, including the model architecture, optimization objectives, training configurations, and evaluation protocols, in the main paper. These details cover the key components required to reproduce our experiments. The source code and implementation will be made publicly available upon acceptance.

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.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Bridging the semantic latent space between brain and machine: Similarity is all you need. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11302–11310, 2024a.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Mind artist: Creating artistic snapshots with human thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27207–27217, 2024b.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Mindgpt: Interpreting what you see with non-invasive brain recordings. *IEEE Transactions on Image Processing*, 34:3281–3293, 2025a.

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*, 2025b.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Human-like cognitive generalization for large models via mental representation-guided supervision. *Nature Communications*, 2026a.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Reducing semantic mismatch in brain-to-text decoding through personalized multimodal masking. In *International Conference on Learning Representations*, volume 2026, pp. 68777–68793, 2026b.

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22710–22720, 2023.

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. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13073–13087, 2020.

Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Baldwin Geary, Michael Ferguson, David Daniel Cox, and James J. DiCarlo. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. In *The Eleventh International Conference on Learning Representations*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Callie Federer, Haoyan Xu, Alona Fyshe, and Joel Zylberberg. Improved object recognition using neural networks trained to mimic the brain's statistical properties. *Neural Networks*, 131:103–114, 2020.

Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.

Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L Yamins, and James J DiCarlo. Brain-like object recognition with high-performing shallow recurrent anns. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Brenden M Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.

Zhe Li, Wieland Brendel, Edgar Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian Sinz, Zachary Pitkow, and Andreas Tolias. Learning from brains how to regularize machines. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5899–5908, 2023.

Weijian Mai and Zhijun Zhang. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. *arXiv preprint arXiv:2308.07428*, 2023.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Weikang Qiu, Zheng Huang, Haoyu Hu, Aosong Feng, Yujun Yan, and Rex Ying. MindLLM: A subject-agnostic and versatile model for fMRI-to-text decoding. In *Forty-second International Conference on Machine Learning*, 2025.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Shahd Safarani, Arne Nix, Konstantin Willeke, Santiago Cadena, Kelli Restivo, George Denfield, Andreas Tolias, and Fabian Sinz. Towards robust vision by multi-task learning on monkey visual cortex. In *Advances in Neural Information Processing Systems*, volume 34, pp. 739–751, 2021.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36, 2024.

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *Forty-first International Conference on Machine Learning*.

Chen Shani, Liron Soffer, Dan Jurafsky, Yann LeCun, and Ravid Shwartz-Ziv. From tokens to thoughts: How llms and humans trade compression for meaning. In *International Conference on Learning Representations*, volume 2026, pp. 83019–83055, 2026.

Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. *Advances in Neural Information Processing Systems*, 37:98083–98110, 2024.

Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, pp. 21, 2019.

Fabian H Sinz, Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S Tolias. Engineering a less artificial intelligence. *Neuron*, 103(6):967–979, 2019.

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, 2023a.

Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. *arXiv preprint arXiv:2306.11536*, 2023b.

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, 2024.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025.

Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision*, pp. 242–259. Springer, 2024.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.