NEURO-SYMBOLIC DECODING OF NEURAL ACTIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose NEURONA, a modular neuro-symbolic framework for fMRI decoding and concept grounding in neural activity. Leveraging image- and video-based fMRI question-answering datasets, NEURONA learns to decode interacting concepts from visual stimuli from patterns of fMRI signals, integrating symbolic reasoning and compositional execution with fMRI grounding across brain regions. We demonstrate that incorporating structure into the decoding pipeline improves both decoding accuracy and generalization performance. NEURONA shows that modeling the compositional structure of concepts through hierarchical predicate-argument dependencies enables more precise decoding from fMRI, highlighting neuro-symbolic frameworks as promising tools for neural decoding.

1 INTRODUCTION

A long-standing hypothesis in cognitive science, the Language of Thought (LoT) hypothesis (Fodor, 1975), proposes that human cognition operates over structured, symbolic representations that compose systematically. Rather than storing concepts as isolated units, the brain is thought to organize knowledge into compositional structures—such as predicates and their arguments—that enable flexible and generalizable reasoning. To test whether such structures can improve neural decoding, we study concept grounding in functional magnetic resonance imaging (fMRI), with the goal of predicting symbolic concepts (e.g., `person`, `baseball-bat`, and `holding`) from patterns of neural activity (Mitchell et al., 2008). This alignment offers insights into whether incorporating structural priors can enable more accurate, precise, and generalizable neural decoding.

There has been vast literature on concept grounding in the past decades, with several influential works studying how concepts are organized across the cortex (Mitchell et al., 2008; Palatucci et al., 2009; Huth et al., 2016; Pereira et al., 2018). Recent advances in machine learning has enabled growing efforts toward data-driven approaches to concept grounding. However, most large-scale fMRI decoding studies focus on isolated concepts or holistic stimulus reconstruction (Nishimoto et al., 2011; Naselaris et al., 2011; Chen et al., 2023a; Takagi & Nishimoto, 2023; Scotti et al., 2023; Chen et al., 2023b), leaving open the question of how the to decode relational meaning between interacting visual concepts. Specifically, we ask, does the decoding of relational concepts (e.g., `holding`) improve by accounting for the systematic combination of their constituent arguments (e.g., `person` and `baseball-bat`) across multiple brain regions?

To explore these questions, we leverage rich data from image- and video-based fMRI datasets, which naturally encode complex semantic and compositional structure. Naturalistic stimuli such as images and videos often involve multiple interacting concepts (e.g., a person holding a baseball-bat), making them well-suited for probing how to decode entities and their relations from neural activity. Hence, we build challenging fMRI-question-answering (fMRI-QA) datasets based on BOLD5000 (Chang et al., 2019) and CNeuroMod (Gifford et al., 2024; Boyle et al., 2023), with the goal of learning concept grounding and improving decoding accuracy based on computational hypotheses on composition.

However, neither simple linear models nor purely end-to-end neural decoding models are sufficient for solving this task. Linear models lack the capacity to capture interactions between multiple interacting components, while large neural decoders (e.g., those with language model backbones) tend to encode stimuli holistically, without explicitly modeling modular concepts or their relationships. To address these limitations, we adopt a neuro-symbolic approach that integrates the compositionality of symbolic systems with the expressivity of neural networks for fMRI-QA: each query is decomposed into a symbolic expression composing concepts, and neural activities in the brain are routed through corresponding concept modules (implemented as neural networks) to answer the given query.

Specifically, we build upon the general neuro-symbolic framework of the Logic-Enhanced Foundation Model (LEFT) (Hsu et al., 2023), and adapt it for the domain of fMRI-based question answering, enabling the use of QA supervision to learn disentangled concept groundings. Crucially, we introduce the incorporation of various *compositional priors* into the model, by defining candidate entities in neural activity and specifying how they are composed based on the symbolic expressions. From this paradigm, we propose **NEURONA**, a **NEURO**-symbolic framework for decoding in **N**eural **A**ctivity, which integrate symbolic reasoning and compositional execution with fMRI grounding, and significantly improves decoding accuracy compared to prior works.

With NEURONA, we find that incorporating structural priors to explicitly *guide* the concept grounding process, such as enforcing hierarchical predicate-argument dependencies (e.g., grounding for `holding` conditioned on the grounding of `baseball-bat`)—notably improves decoding accuracy on fMRI-QA tasks. These priors guide the model to compose high-level relational concepts from their constituent entity groundings, showing that relational meaning is better predicted across multiple co-activated brain networks via its arguments, rather than localized to a single region or to multiple regions without guidance.

Evaluating concept grounding in fMRIs is inherently challenging due to the lack of direct supervision: there is no ground truth mapping from abstract concepts to specific brain regions. Instead, we report experiments on BOLD5000 and CNeuroMod fMRI-QA datasets, which demonstrate that our neuro-symbolic framework significantly outperforms baseline neural decoding methods, and importantly, exhibits strong generalization to unseen compositional queries. Notably, ablation studies with NEURONA highlight the importance of encoding hierarchical structure: conditioning predicate grounding modules on the regions associated with their subject and object arguments consistently yields large performance gains in decoding of neural activity.

2 Related Works

Visual decoding from fMRI. Reconstructing visual content from fMRI signals has become a central research focus of works in the field, with many approaches leveraging state-of-the-art generative backbones for the task, following early studies (Thirion et al., 2006; Miyawaki et al., 2008; Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011). Takagi et al. demonstrated that a pre-trained diffusion model can reconstruct high-resolution images from fMRI (Takagi & Nishimoto, 2023). MinD-Vis uses masked brain modeling with a diffusion model for semantically faithful image generation (Chen et al., 2023a). MindEye projects fMRI into a CLIP embedding space and applies a diffusion prior for pixel-level synthesis (Scotti et al., 2023). Extending to video, MinD-Video and NeuroCLIP incorporate spatiotemporal masked modeling and keyframe-perception flow cues, respectively, into diffusion-based reconstruction (Chen et al., 2023b; Gong et al., 2024). These visual reconstruction works focus on recovering stimulus appearance from neural data; in contrast, our work addresses a distinct goal of concept grounding: rather than generating pixel-level images or videos, we aim to predict modular concepts and their relationships from neural activity.

Concept grounding. Several influential works have focused on how semantic information is organized across the cortex. As a representative work, Huth et al. used voxel-wise encoding models with natural narrative stimuli to construct a semantic atlas, showing that different semantic domains selectively ground to distinct brain regions (Huth et al., 2016). Mitchell et al. predicted fMRI patterns for concrete nouns using corpus-derived semantic features, showing generalization to unseen words (Mitchell et al., 2008). SOC enables zero-shot decoding by mapping fMRI to semantic codes and recognizing novel object categories (Palatucci et al., 2009). Pereira et al. introduced a general decoder that maps fMRI into a shared semantic space, enabling generalization from limited data (Pereira et al., 2018). Beyond semantic mapping, several studies also explored how concepts are organized in the brain (Frankland & Greene, 2015; Eichenbaum, 2001). There is converging evidence that certain brain regions support invariant concept and rule representations. For example, the prefrontal cortex—spanning networks such as the dorsal attention and default mode networks—and the medial temporal lobe have been implicated in abstract concept and rule processing (Quiroga et al., 2005; Rey et al., 2015; Tian et al., 2024; Dijksterhuis et al., 2024). Additionally, single-neuron recordings in the human prefrontal cortex have revealed neurons that encode abstract task rules independently of sensory or motor details (Mian et al., 2014). In contrast to these prior works, our approach emphasizes compositional grounding in the neural decoding process, with focus on functional instead of representational compositionality, where we explicitly model not only individual concepts, but also the relationships between them. This modeling allows us to investigate how relational concepts can be more accurately decoded through guided grounding across brain networks.

fMRI-question answering. Recent works have explored using fMRI data for question-answering by integrating large vision-language models (VLMs). These methods typically map neural activity to visual embeddings, then generate answers using pre-trained VLMs. For example, SDRecon (Takagi & Nishimoto, 2023) projects fMRI signals into BLIP (Li et al., 2022) embeddings for captioning; BrainCap (Ferrante et al., 2023) maps fMRI to GIT (Wang et al., 2022) features for visual description; and UMBRAE (Xia et al., 2024) aligns fMRI to multimodal embeddings with subject-specific tokenization and answers questions via LLaVA (Liu et al., 2023). These methods commonly use BLEU scores (Papineni et al., 2002) to measure alignment with ground-truth text, but they do not explicitly verify whether the predicted answer captures exact concepts or relational structure. In contrast, our framework grounds fMRI signals to modular concepts before performing structured reasoning, enabling precise, accurate, and generalizable question answering.

3 Method

3.1 Neuro-symbolic framework

We introduce NEURONA as a neuro-symbolic framework for concept grounding and decoding in neural activity. Neuro-symbolic models are a class of methods that decompose queries into symbolic expressions containing concepts, and then differentiably execute those expressions over input data, using learned concept grounding modules to perform a variety of downstream tasks (Yi et al., 2018; Mao et al., 2019; Hsu et al., 2023; Mao et al., 2025). Each symbolic concept (e.g., `person`, `holding`) is associated with a small neural network that maps entity-centric representations from input data to a predicted semantic signal, enabling the learning of intermediate concept grounding from weak supervision of reasoning tasks. Execution is conducted via differentiable functions combining concept grounding modules, which enables end-to-end training.

In this work, we build our model, NEURONA, based on the Logic-Enhanced Foundation Model (LEFT) (Hsu et al., 2023). LEFT is a general neuro-symbolic framework that unifies grounding and reasoning via a differentiable executor for logic programs. It is designed to support concept grounding across various visual domains (e.g., 2D images, 3D scenes), notably, where the relevant entities are known a priori, such as objects in a room. In contrast, our setting introduces a unique challenge: concept grounding from fMRI signals, where the relevant neural "entities" (i.e., brain regions) are not predefined, and must instead be inferred as part of the learning process. This makes our setting significantly more challenging than prior works: we are testing hypotheses for what these entities should be, as well as learning how they best compose to precisely decode neural activity.

Concretely, we frame concept grounding in neural activity as a weakly supervised fMRI-question answering (fMRI-QA) task. Each input consists of two elements (See Figure 1). The first is a visual stimuli v that is parsed into a symbolic query that describe possible concepts in v. The second is an fMRI recording $f \in \mathbb{R}^{N \times T}$ of the neural activity recorded when the stimuli was viewed, where N is the number of fine-grained brain networks, and T is the number of time steps. The goal is to predict an answer a, either as a Boolean label (e.g., True/False) or as a classification over a concept vocabulary. Crucially, no supervision is provided on which brain regions are relevant to each concept.

To enable modular concept grounding to neural activity, we first map the fine-grained networks N to P functional networks defined by an atlas. We experiment with Yeo-7, Yeo-17 (Yeo et al., 2011), DiFuMo-64, DiFuMo-128 (Dadi et al., 2020), and Schaefer-100 (Schaefer et al., 2018) atlases, where P is 7, 17, 64, 128, 100 respectively. While all atlases yield consistently strong performance with NEURONA, we report decoding accuracy with the Yeo-17 atlas in the main text, as performance is highest. Experiment results across all atlases can be found in Appendix A. Hence, we have $P = 17$ resulting network-specific fMRI signals, which we then encode to form a unified set $\mathcal{E}$ of embeddings $\{e_1, \ldots, e_P\}$, to form parcellation embeddings. From these base embeddings, we propose hypotheses of candidate entities from which concepts can be mapped to. Here, neural entities precisely are

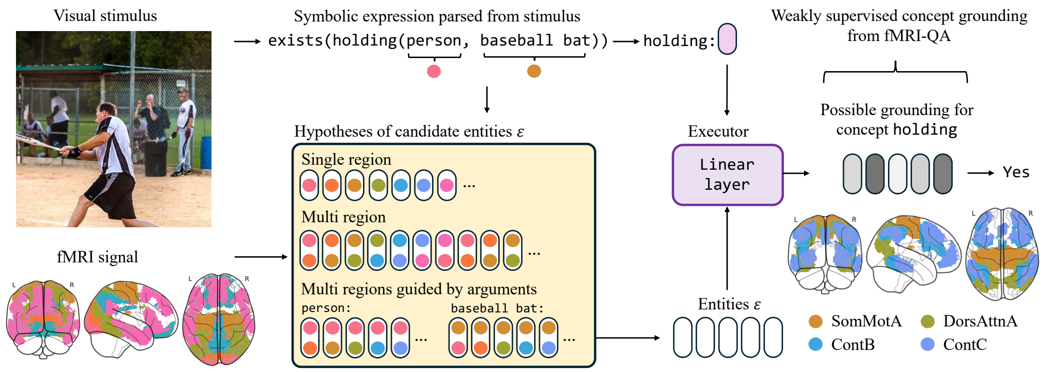

Figure 1: NEURONA is a neuro-symbolic framework for neural decoding that parses visual stimuli into symbolic expressions and fMRI signals into candidate entities, optionally guided by the concepts in the expression. These composed concepts are then grounded to the entities via a linear layer for question answering; the predicted answer provides weak supervision for the grounding process.

region-specific fMRI signals derived from functional parcellations (e.g., Yeo-17, DiFuMo-128 atlases) that serve as possible groundings for symbolic concepts.

In NEURONA, each concept is associated with a grounding module that maps between the concepts and such candidate neural entity embeddings—intuitively, selecting the brain regions that best predict the modular concept. Then, NEURONA uses a differentiable executor to compose these concept groundings according to the structure of the symbolic expression, yielding a final output a of either a Boolean value or a distribution over concept labels. Together, these components allow NEURONA to accurately decode concepts from fMRI, and test hypotheses on how to guide the grounding process to patterns of neural activity without direct supervision. In the following sections, we describe our method for grounding concepts in fMRI data, our hypotheses, and NEURONA's training objective.

3.2 Grounding concepts onto neural activities

In concept grounding, we aim to identify which neural entities predict modular concepts or compositions of concepts, over C total concepts. Let us consider a base set of candidate neural entities, which we extract by aggregating fMRI signals into functional parcellations. This yields a set of entity embeddings $\mathcal{E} = [e_1, \ldots, e_P]$ for each network p, where each entity is embedded to dimension d.

To model unary concepts (e.g., subject and objects such as `person` and `baseball-bat`), we formulate grounding as a C-way classification problem over $\mathcal{E}$ using a linear classifier with weight matrix $\mathcal{W}_{\text{unary}} \in \mathbb{R}^{d \times C}$ and bias $\mathbf{b}_{\text{unary}} \in \mathbb{R}^C$. For each entity e_p, the predicted logits are computed via a linear layer $\mathbf{z}_p = \mathcal{W}_{\text{unary}}^\top e_p + \mathbf{b}_{\text{unary}} \in \mathbb{R}^C$. The overall predicted logits $\mathcal{Z}_{\text{unary}} \in \mathbb{R}^{P \times C}$ is the grounding probability of all concepts, and the *grounding score* for concept c across all entities is $G_{\text{unary}}(c) = [z_{1c}, \ldots, z_{Pc}]^\top \in \mathbb{R}^P$.

To model high-arity relational concepts (e.g., predicates such as `holding`), we augment the neural entity embeddings $\mathcal{E}$ with learnable embeddings $\mathcal{E}_b \in \mathbb{R}^{P \times d}$ and concatenate them to form $\mathcal{E}_c \in \mathbb{R}^{P \times 2d}$, where $\mathcal{E}_c = \mathcal{E} \oplus \mathcal{E}_b$ and $\oplus$ denotes feature concatenation. $\mathcal{E}_b$ provides features that represent each brain network. For each pair of entities (i, j), we concatenate their embeddings $e_{c_i} \oplus e_{c_j}$ and apply a learnable transformation $\mathcal{W}_{\text{pair}} \in \mathbb{R}^{4d \times d}$ to obtain a pairwise representation $\mathcal{E}'_{ij} \in \mathbb{R}^d$. We then apply a linear classifier to compute the logits $\mathbf{z}_{ij}$ for each pair. The overall logits of high-arity concept are $\mathcal{Z}_{\text{binary}} \in \mathbb{R}^{P \times P \times C}$, which represents the grounding probability of all concepts. The grounding score for a concept c is $G_{\text{binary}}(c) = [z_{ij,c}]_{1 \le i,j \le P} \in \mathbb{R}^{P \times P}$.

These grounded concepts are then used (and optionally composed) to answer queries from our weakly supervised fMRI-question answering task. Let c_p, c_s, and c_o be the concepts for predicate, subject, and object, respectively. For Boolean queries, we first optionally condition $G_{\text{binary}}(c_p)$ on the unary groundings $G_{\text{unary}}(c_s)$ and $G_{\text{unary}}(c_o)$, and aggregate the resulting scores. Then, we apply a sigmoid over the scores, and threshold the result to produce a binary decision in inference. For concept classification queries, given a concept vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$, we compute grounding scores for each $v \in \mathcal{V}$. Since the grounding logits $\mathcal{Z}_{\text{unary}}$ and $\mathcal{Z}_{\text{binary}}$ have already been computed for all concepts, we can directly extract the grounding scores for any vocabulary concept v, treating them as the unary similarity $\mathcal{S}_{\text{unary}}$ and relational similarity $\mathcal{S}_{\text{binary}}$ between each neural candidate and v. Unary and relational similarity scores are computed as

$$S_{\text{unary}} = [z_{iv}]_{1 \le i \le P, v \in \mathcal{V}} \in \mathbb{R}^{P \times |\mathcal{V}|}, \quad S_{\text{binary}} = [z_{ij,v}]_{1 \le i,j \le P, v \in \mathcal{V}} \in \mathbb{R}^{P \times P \times |\mathcal{V}|}. \tag{1}$$

When subject and object groundings are available, we compute guided scores

$$S_{\text{unary}}^{\text{guided}} = G_{\text{unary}}(c_s)^\top S_{\text{unary}} + G_{\text{unary}}(c_o)^\top S_{\text{unary}} \in \mathbb{R}^{|\mathcal{V}|}, \tag{2}$$

$$S_{\text{binary}}^{\text{guided}} = G_{\text{unary}}(c_s)^\top \left(G_{\text{unary}}(c_o)^\top S_{\text{binary}}\right) \in \mathbb{R}^{|\mathcal{V}|}, \tag{3}$$

and combine them as $S^{\text{final}} = S_{\text{unary}}^{\text{guided}} + S_{\text{binary}}^{\text{guided}} \in \mathbb{R}^{|\mathcal{V}|}$, with the final predicted concept selected by $\hat{v} = \arg\max_{v \in \mathcal{V}} S_v^{\text{final}}$. This formulation enables unified, differentiable concept grounding for unary and high-arity concepts, over Boolean and vocabulary-based queries. Additionally, NEURONA's role-conditioned aggregation mechanism enables flexible hypothesis testing.

3.3 Testing hypotheses of grounding structures

Notably, we propose and evaluate five hypotheses on how concepts can be composed during grounding in brain networks. Due to NEURONA's modular structure, we can conduct guided grounding via conditioning the representation of a relational concept on the activations of its constituent arguments, rather than modeling each concept independently, which we see in the latter three hypotheses.

H1: Single-region localized. Concepts are localized to a single brain network, where the grounding score is defined as $G_{\mathrm{H1}}(c) = G_{\mathrm{unary}}(c)$.

H2: Multi-region co-activation. Concepts are represented by co-activation across region pairs, where the grounding score is defined as $G_{\mathrm{H2}}(c_p) = G_{\mathrm{binary}}(c_p)$.

H3: Subject-guided multi-region. Predicate representations are guided by subject region activation, with the grounding score defined as $G_{\mathrm{H3}}(c_p) = G_{\mathrm{binary}}(c_p) + G_{\mathrm{unary}}(c_s)$.

H4: Object-guided multi-region. Predicate representations are guided by object region activation, with the grounding score defined as $G_{\mathrm{H4}}(c_p) = G_{\mathrm{binary}}(c_p) + G_{\mathrm{unary}}(c_o)$.

H5: Full argument-guided multi-region. Multi-region groundings are combined with argument-specific guidance. The predicate grounding score is computed as $G_{\mathrm{H5}}(c_p) = G(c_p) + G(c_s) + G(c_o)$, where $G(c)$ aggregates unary and binary grounding as $G(c) = G_{\mathrm{unary}}(c) + \frac{1}{P}\sum_{i=1}^{P} G_{\mathrm{binary}}(c_i)$. The subject and object scores are implemented as above.

Together, these hypotheses define different subspaces and ways of grounding $\mathcal{E}$, allowing us to test with NEURONA whether structure-guided grounding leads to better decoding. Full derivations and experimental details are provided in Appendix D.

3.4 Training objective

Our model is trained in a weakly supervised fMRI-QA setting, where ground-truth answers are provided for each query, but no supervision is given on intermediate concept groundings, as none are available. We optimize a standard cross-entropy loss over the model's predicted output distribution $\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{K} a_i \log(\hat{a}_i)$. Here, $\hat{a} \in \mathbb{R}^K$ is the predicted probability distribution over K possible answer classes, and a is the ground-truth label. The prediction $\hat{a}$ is the model's prediction for the full symbolic expression after composing the neural groundings of its component concepts.

4 Datasets

To train NEURONA, we create fMRI-question-answering (fMRI-QA) datasets by adapting existing large-scale fMRI-vision datasets. We first extract structure defining multiple interacting concepts from the rich visual data. Each visual stimuli is processed with a pre-trained vision-language model, which outputs a scene graph that captures both object-level and relational semantics (e.g., unary

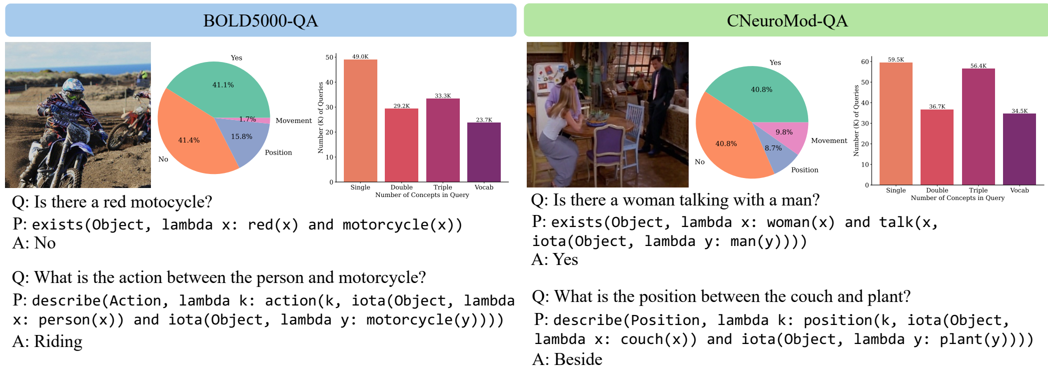

Figure 2: We include example queries and dataset distribution overviews for BOLD5000-QA and CNeuroMod-QA; both datasets span diverse queries and tasks.

objects such as `person`, `baseball-bat`, and high-arity predicates that capture their interaction such as `holding (person, baseball-bat)`). We then convert the scene graph into a set of structured question-answer pairs, where each question corresponds to a symbolic query and each answer serves as a weak supervision signal for concept grounding. For example, given the above scene graph, we construct questions such as: "What is the relation between person and baseball-bat?". We additionally generate negative samples by randomly sampling concepts from other stimuli in the dataset. This process produces diverse fMRI-QA examples covering both unary entities and high-arity relations, and notably, with precise answers. Specifically, we apply our pipeline to two datasets: BOLD5000 (Chang et al., 2019) and CNeuroMod (Gifford et al., 2024; Boyle et al., 2023) (See Figure 2). More dataset details can be found in Appendix C.

BOLD5000-QA. The BOLD5000 dataset is a large-scale fMRI dataset collected while participants viewed naturalistic images. The dataset consists of approximately $5,000$ distinct images from three datasets: Scene Images (Xiao et al., 2010), COCO (Lin et al., 2014), and ImageNet (Deng et al., 2009). Four participants viewed these images while undergoing whole-brain fMRI scanning. To create BOLD5000-QA, we follow the process above, and generate queries containing $4,258$ unary and 135 relational concepts, with $133,146$ train and $2,095$ test examples.

CNeuroMod-QA. The CNeuroMod dataset is a large-scale fMRI dataset that includes recordings of participants watching full-length naturalistic videos. Specifically, we build upon the Friends dataset, and set season 1 to 5 to be the train samples, and unseen season 6 to be the test. To create CNeuroMod-QA, we sample video frames based on motion energy, defined as the absolute difference between consecutive frames. We then extract scene graphs for each sampled frame, aggregate the changes, and construct corresponding symbolic queries. The CNeuroMod-QA dataset includes $1,966$ unary and 106 relational concepts, with $157,046$ train and $30,059$ test examples.

5 Results

Our goal is precise neural decoding and consistent concept grounding in neural activity. We present quantitative metrics in Section 5.1, qualitative analyses in Section 5.2, and discussion in Section 5.3.

5.1 Quantitative performance

We test quantitative performance of NEURONA on both the BOLD5000-QA and CNeuroMod-QA datasets. We compare against baseline fMRI decoding methods, evaluate generalization to unseen compositional queries, conduct ablation studies to test hypotheses about grounding structures, and report quantitative consistency metrics of concept grounding. Results on performance across atlases, effect sizes between atlases, evaluation over all subjects, statistical tests across hypotheses, cross-dataset transfer experiments, network ablations, detailed concept accuracy, predicate argument binding, fine-grained generalization analyses, additional null models, and fMRI retrieval results are provided in Appendix A.

Comparison to prior works. We first evaluate NEURONA on the fMRI-question-answering (fMRI-QA) task, compared against existing decoding methods: a linear baseline, UMBRAE (Xia et al., 2024),

Table 1: We evaluate NEURONA on BOLD5000-QA (subject-CSI1) and CNeuroMod-QA (subject-01), comparing its performance to prior fMRI language decoding models and a linear baseline.

	BOLD5000				CNeuroMod			
Method	Overall	Action	Position	T/F	Overall	Action	Position	T/F
Linear	0.4692	0.2069	0.1778	0.5260	0.4638	0.3043	0.1285	0.5192
UMBRAE	0.4668	0.2069	0.1238	0.5328	0.4614	0.0549	0.1439	0.5442
SDRecon	0.4711	0.2414	0.1937	0.5248	0.4430	0.1350	0.1481	0.5238
BrainCap	0.4773	0.1937	0.1724	0.5551	0.4417	0.1257	0.1477	0.5112
NEURONA (Ours)	**0.7041**	**0.6207**	**0.5079**	**0.7407**	**0.7046**	**0.6514**	**0.5746**	**0.7250**

Table 2: Generalization results of NEURONA and prior work on unseen queries.

	BOLD5000				CNeuroMod			
Method	Overall	Action	Position	T/F	Overall	Action	Position	T/F
Linear	0.4587	0.0690	0.0794	0.5231	0.4143	0.0398	0.0323	0.5003
UMBRAE	0.4162	0.1724	0.0540	0.4854	0.4306	0.1315	0.1022	0.5018
SDRecon	0.4702	0.2414	0.1937	0.5237	0.4341	0.1236	0.1473	0.5008
BrainCap	0.4754	0.1724	0.1937	0.5311	0.4347	0.1198	0.1402	0.5042
NEURONA (Ours)	**0.6840**	**0.6207**	**0.4984**	**0.7184**	**0.6583**	**0.4365**	**0.5261**	**0.6991**

SDRecon (Takagi & Nishimoto, 2023), and BrainCap (Ferrante et al., 2023). All models are trained on subject CSI1 (BOLD5000) and subject sub-01 (CNeuroMod). As shown in Table 1, NEURONA consistently outperforms prior approaches across both the BOLD5000-QA and CNeuroMod-QA datasets. Compared to prior linear or language model-based approaches, NEURONA achieves a 47% relative improvement on the top performing prior work. These results show that linear methods lack in expressivity, and purely end-to-end decoding pipelines struggle to learn fMRI embeddings that capture the detailed concepts required for fMRI-QA. In particular, NEURONA demonstrates significant gains on queries about actions and positions, which involve precise relational reasoning over subject and object roles. This highlights the strength of our neuro-symbolic framework, which explicitly grounds neural activity guided by hierarchical structures to answer compositional queries.

We further evaluate NEURONA's ability to generalize to unseen compositions of concepts, which robustly tests whether learned concept groundings are semantically meaningful. Specifically, we construct evaluation splits where all training and testing queries are disjoint, with no overlapping combinations of entities and relations, ensuring that the model must generalize. For example, the training set may contain only queries such as `in_front_of(person,baseball-bat)`, while the test set includes novel compositions such as `in_front_of(baseball-bat,person)`. As shown in Table 2, NEURONA achieves the strongest performance across both datasets, substantially outperforming all baselines. Notably, prior methods suffer significant performance degradation, often falling to near-random levels when generalizing to unseen compositions. The language model-based baselines retain slightly better performance due to their use of pre-trained language models, which carry general linguistic priors. However, since they lack explicit concept grounding, their performance remains well below NEURONA. Overall, these results show that NEURONA's neuro-symbolic framework learns meaningful decoding from neural activity and generalizes to new compositional queries for robust fMRI-QA.

Ablations & hypotheses testing. Notably, we conduct ablation studies on different hypotheses about the candidate neural entity space $\mathcal{E}$, and analyze how they affect decoding performance. We summarize results in Table 3 and Table 4: we compare single-region grounding, multi-region grounding without guidance, and forms of guided grounding based on subject and object arguments. We report standard deviation across 4 subjects in BOLD5000-QA and 3 subjects in CNeuroMod-QA, and our analyses evaluate how different grounding assumptions affect QA accuracy. With NEURONA, we answer the following questions.

DOES GROUNDING TO MULTIPLE REGIONS IMPROVE PERFORMANCE? We first test whether allowing concepts to ground to combinations of brain regions improves decoding performance relative to grounding to a single region. As shown in Table 3 and Table 4, we observe that multi-region grounding alone does not yield substantial gains over single-region grounding in terms of overall accuracy. This is especially evident in vocabulary classification tasks, where both approaches tend to overfit to the most frequent vocabulary labels without capturing finer variations.

Table 3: We ablate our hypotheses about the structure of concept grounding on BOLD5000-QA.

	BOLD5000			
Method	Overall	Action	Position	T/F
Single region	0.6451 ± 0.0161	0.2973 ± 0.0361	0.2005 ± 0.0047	0.7293 ± 0.0191
Multi-region (MR)	0.6476 ± 0.0048	0.2973 ± 0.0361	0.2005 ± 0.0047	0.7324 ± 0.0056
Subject-guided MR	0.6678 ± 0.0029	0.2881 ± 0.0299	0.3469 ± 0.0134	0.7308 ± 0.0024
Object-guided MR	0.6733 ± 0.0051	0.3892 ± 0.0752	0.3429 ± 0.0164	0.7363 ± 0.0045
Full argument-guided MR	$\mathbf{0.7102 \pm 0.0053}$	$\mathbf{0.5965 \pm 0.0322}$	$\mathbf{0.5378 \pm 0.0135}$	$\mathbf{0.7425 \pm 0.0057}$

Table 4: We evaluate our hypotheses about compositional priors in neural activity on CNeuroMod.

	CNeuroMod			
Method	Overall	Action	Position	T/F
Single region	0.6429 ± 0.0013	0.2165 ± 0.0193	0.2445 ± 0.0009	0.7400 ± 0.0016
Multi-region (MR)	0.6162 ± 0.0027	0.2165 ± 0.0193	0.2445 ± 0.0009	0.7042 ± 0.0023
Subject-guided MR	0.6265 ± 0.0042	0.2339 ± 0.0043	0.3722 ± 0.0045	0.7008 ± 0.0051
Object-guided MR	0.6872 ± 0.0040	0.6320 ± 0.0019	0.4933 ± 0.0172	0.7149 ± 0.0045
Full argument-guided MR	$\mathbf{0.7189 \pm 0.0009}$	$\mathbf{0.6422 \pm 0.0072}$	$\mathbf{0.5931 \pm 0.0084}$	$\mathbf{0.7417 \pm 0.0005}$

DOES GUIDED GROUNDING IMPROVE PERFORMANCE? Next, we evaluate whether guiding multi-region grounding based on the subject and object arguments of relational concepts improves performance. As shown, models that incorporate guided grounding significantly outperform both single-region and unguided multi-region baselines. We find that guiding NEURONA by object improves performance over by subject. In particular, grounding based on both subject and object regions achieves the highest accuracy, notably on action and position queries that require precise relational reasoning. With NEURONA, we demonstrate the importance of argument-conditioned composition for interpreting relational semantics in neural activity. Overall, these results demonstrate that while multi-region grounding provides a more flexible representation space, explicit structural guidance based on predicate-argument relationships is crucial. The consistent results on both natural image and video datasets validate NEURONA ability to conduct complex neural decoding.

Concept grounding consistency. As there is no ground truth to evaluate the reliability of intermediate concept grounding, we introduce a consistency metric to test whether the same concept grounds to consistent brain regions across different fMRI-QA instances. We calculate consistency as follows. Let a concept c appear in N QA examples, and let the predicted grounding in the i-th example be a set of brain regions $B^{(i)}(c) \subseteq \{1, \dots, P\}$, where P is the total number of regions, and each $B^{(i)}(c)$ is the set of regions selected as the grounding for concept c in that example. We compute the frequency count for each region r as $\text{Count}(r) = \sum_{i=1}^{N} \mathbf{1}[r \in B^{(i)}(c)]$. Then, the score for concept c is defined as $\text{Consistency}(c) = \frac{1}{|R|} \sum_{r \in R} \frac{\text{Count}(r)}{N}$, where R is the set of all regions that appear in any grounding of c, and $\frac{\text{Count}(r)}{N}$ is the fraction of times region r was selected.

Our proposed metric captures how concentrated the concepts groundings are: a score of 1.0 implies perfect consistency (all instances of the concept used the same region set), while lower scores indicate more variability in the grounding of concept c. We report these consistency scores over all concepts in BOLD5000 and CNeuroMod. As a baseline, we define a null model that randomly assigns each concept to a region subset via uniform sampling. In Table 5, we see that NEURONA significantly outperforms the null baseline. These results highlight that NEURONA not only improves QA accuracy but also grounds concepts in a structured and reproducible way across different stimuli. Consistency results over all other atlases can be found in Appendix A.

5.2 QUALITATIVE CONCEPT GROUNDING ANALYSES

Table 5: Consistency of concept grounding between NEURONA and a null baseline.

	BOLD5000	CNeuroMod
Null	0.5357	0.5358
NEURONA	**0.8220**	**0.8700**

Finally, we qualitatively examine how NEURONA learns to decode high-level relational concepts from neural activity using intermediate concept groundings, focusing on how this grounding varies with different subject-object pairs. Figure 3 shows representative examples from both BOLD5000-QA and CNeuroMod-QA, where we project the learned grounding scores onto network parcellations that define our neural entities. We observe that the same relational predicate, such as `hold` or `look`, is best decoded from different brain networks depending on the object. For example, in BOLD5000,

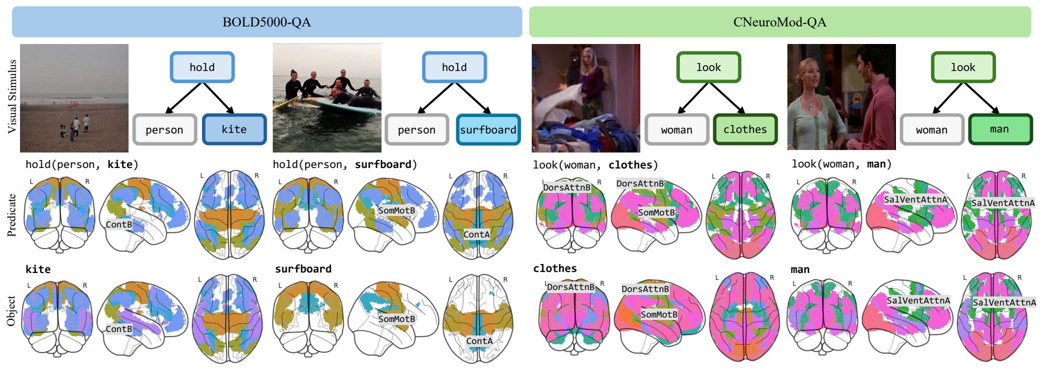

Figure 3: We show examples of learned concept grounding from NEURONA. On both BOLD5000-QA and CNeuroMod-QA, we see that predicate concepts ground to the brain regions that their constituent objects are grounded to, following hierarchical predicate-argument structure.

`holding (person, kite)` is best decoded using the Control B network, while `holding (person, surfboard)` is decoded using both the Somatomotor B and Control A networks. This suggests that decoding accuracy improves when the grounding of a relational concept is modulated by its arguments, supporting the importance of argument-dependent composition in neural decoding.

Interestingly, the model-inferred concept groundings are not confined to early visual areas, despite the task involving visual stimuli. Instead, objects such as `baseball-bat` and `surfboard` receive high grounding scores in motor-related regions, including the Somatomotor network. This qualitative pattern resembles prior findings that perceiving action-related objects is linked to motor and premotor areas (Martin, 2007; Gallese et al., 1996). Additionally, both `hold` and `look` are often decoded using prefrontal networks, including the Dorsal Attention and Salience/Ventral Attention networks, which have been associated with high-level cognitive control and abstract rule processing (Miller & Cohen, 2001; Quiroga et al., 2005; Tian et al., 2024). We emphasize that these analyses reflect model-dependent decoding patterns rather than direct estimates of underlying encoding representations. Additional concept grounding visualizations are provided in Appendix B.

5.3 Discussion

Our findings demonstrate that incorporating compositional structure into the decoding pipeline significantly improves performance. While NEURONA does not establish representational compositionality in neural patterns, the substantial gains from modeling hierarchical predicate–argument structure provide proof-of-concept that compositional principles can inform neural decoding. However, our study also has several limitations. First, participants in our datasets engaged only in passive viewing, and the ground truth symbolic expressions were extracted through automated scene-graph parsing rather than participant-driven reasoning, limiting the cognitive conclusions that can be drawn. Second, we restrict candidate neural entities to predefined parcellations from established atlases (e.g., Yeo-7, Yeo-17, DiFuMo-64, DiFuMo-128, and Schaefer-100), which, while widely used, provide coarser representations of the brain from which we build upon. In the Appendix, we provide performance on BOLD5000 and CNeuroMod across atlases, effect sizes between atlases, and concept grounding consistency across atlases. These results demonstrate robust and modular concept grounding across different levels of spatial granularity, suggesting that our results are not tied to specific parcellation choices. However, we believe that scaling NEURONA to incorporate whole-brain voxel-level data is a promising next step.

6 Conclusion

We propose NEURONA, a neuro-symbolic framework for concept grounding and decoding in neural activity. By leveraging symbolic reasoning and compositional execution with fMRI grounding, NEURONA enables precise and generalizable decoding. Experiments on BOLD5000-QA and CNeuroMod-QA demonstrate that NEURONA outperforms baseline decoding methods and generalizes to novel compositions, with explicit grounding guidance significantly improving performance. Our findings show that incorporating predicate-argument structure improves decoding, and highlight neuro-symbolic modeling as a promising approach for interpreting and structuring fMRI decoding models.

Reproducibility statement. We refer readers to Section 3 for details on the grounding process and Appendix D for train settings, and note that our work builds off the public codebase of LEFT. We will release code upon acceptance. We also describe our dataset processing steps in detail in Appendix C.

REFERENCES

Julie Boyle, Basile Pinsard, Valentina Borghesani, Francois Paugam, Elizabeth DuPre, and Pierre Bellec. The Courtois NeuroMod project: Quality Assessment of the Initial Data Release (2020). In *2023 Conference on Cognitive Computational Neuroscience*, pp. 2023–1602, 2023.

Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. BOLD5000, A Public fMRI Dataset While Viewing 5000 Visual Images. *Scientific data*, 6(1):49, 2019.

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, 2023a.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity. *Advances in Neural Information Processing Systems*, 36: 24841–24858, 2023b.

Kamalaker Dadi, Gaël Varoquaux, Antonia Machlouzarides-Shalit, Krzysztof J Gorgolewski, Demian Wassermann, Bertrand Thirion, and Arthur Mensch. Fine-grain Atlases of Functional Modes for fMRI Analysis. *NeuroImage*, 221:117126, 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Kingma Diederik. Adam: A Method for Stochastic Optimization. *(No Title)*, 2014.

Doris E Dijksterhuis, Matthew W Self, Jessy K Possel, Judith C Peters, ECW van Straaten, Sander Idema, Johannes C Baaijen, Sandra MA van der Salm, Erik J Aarnoutse, Nicole CE van Klink, et al. Pronouns Reactivate Conceptual Representations in Human Hippocampal Neurons. *Science*, 385(6716):1478–1484, 2024.

Howard Eichenbaum. The Hippocampus and Declarative Memory: Cognitive Mechanisms and Neural Codes. *Behavioural brain research*, 127(1-2):199–207, 2001.

Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Multimodal Decoding of Human Brain activity into Images and Text. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023.

Jerry Fodor. *The Language of Thought*. Harvard University Press, 1975.

Steven M Frankland and Joshua D Greene. An Architecture for Encoding Sentence Meaning in Left Mid-superior Temporal Cortex. *Proceedings of the National Academy of Sciences*, 112(37): 11732–11737, 2015.

Vittorio Gallese, Luciano Fadiga, Leonardo Fogassi, and Giacomo Rizzolatti. Action Recognition in the Premotor Cortex. *Brain*, 119(2):593–609, 1996.

Alessandro T Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M Cichy. The Algonauts Project 2025 Challenge: How the Human Brain Makes Sense of Multimodal Movies. *arXiv preprint arXiv:2501.00504*, 2024.

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.

Joy Hsu, Jiayuan Mao, Josh Tenenbaum, and Jiajun Wu. What's Left? Concept Grounding with Logic-Enhanced Foundation Models. *Advances in Neural Information Processing Systems*, 36: 38798–38814, 2023.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural Speech Reveals the Semantic Maps that Tile Human Cerebral Cortex. *Nature*, 532(7600):453–458, 2016.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying Natural Images from Human Brain Activity. *Nature*, 452(7185):352–355, 2008.

Aaron Kucyi, Amy Daitch, Omri Raccah, Baotian Zhao, Chao Zhang, Michael Esterman, Michael Zeineh, Casey H Halpern, Kai Zhang, Jianguo Zhang, et al. Electrophysiological Dynamics of Antagonistic Brain Networks Reflect Attentional Fluctuations. *Nature communications*, 11(1):325, 2020.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping Language-image Pre-training for Unified Vision-language Understanding and Generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision. *arXiv preprint arXiv:1904.12584*, 2019.

Jiayuan Mao, Joshua B Tenenbaum, and Jiajun Wu. Neuro-Symbolic Concepts. *arXiv preprint arXiv:2505.06191*, 2025.

Alex Martin. The Representation of Object Concepts in the Brain. *Annu. Rev. Psychol.*, 58(1):25–45, 2007.

Vinod Menon. 20 Years of the Default Mode Network: A Review and Synthesis. *Neuron*, 111(16): 2469–2487, 2023.

Matthew K Mian, Sameer A Sheth, Shaun R Patel, Konstantinos Spiliopoulos, Emad N Eskandar, and Ziv M Williams. Encoding of Rules by Neurons in the Human Dorsolateral Prefrontal Cortex. *Cerebral cortex*, 24(3):807–816, 2014.

Earl K Miller and Jonathan D Cohen. An Integrative Theory of Prefrontal Cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting Human Brain Activity Associated with the Meanings of Nouns. *science*, 320(5880):1191–1195, 2008.

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.

Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, 63(6):902–915, 2009.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and Decoding in fMRI. *Neuroimage*, 56(2):400–410, 2011.

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current biology*, 21(19):1641–1646, 2011.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot Learning with Semantic Output Codes. *Advances in neural information processing systems*, 22, 2009.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a Universal Decoder of Linguistic Meaning from Brain Activation. *Nature communications*, 9(1):963, 2018.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant Visual Representation by Single Neurons in the Human Brain. *Nature*, 435(7045):1102–1107, 2005.

Hernan G Rey, Matias J Ison, Carlos Pedreira, Antonio Valentin, Gonzalo Alarcon, Richard Selway, Mark P Richardson, and Rodrigo Quian Quiroga. Single-cell Recordings in the Human Medial Temporal Lobe. *Journal of anatomy*, 227(4):394–408, 2015.

Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral cortex*, 28(9):3095–3114, 2018.

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*, 36:24705–24728, 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.

Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. Inverse Retinotopy: Inferring the Visual Content of Images from Brain Activation Patterns. *Neuroimage*, 33(4):1104–1116, 2006.

Zhenghe Tian, Jingwen Chen, Cong Zhang, Bin Min, Bo Xu, and Liping Wang. Mental Programming of Spatial Sequences in Working Memory in the Macaque Frontal Cortex. *Science*, 385(6716): eadp6091, 2024.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A Generative Image-to-text Transformer for Vision and Language. *arXiv preprint arXiv:2205.14100*, 2022.

Yanchen Wang, Adam Turnbull, Tiange Xiang, Yunlong Xu, Sa Zhou, Adnan Masoud, Shekoofeh Azizi, Feng Vankee Lin, and Ehsan Adeli. Decoding visual experience and mapping semantics through whole-brain analysis using fmri foundation models. *arXiv preprint arXiv:2411.07121*, 2024.

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.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. SUN Database: Large-scale Scene Recognition from Abbey to Zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The Organization of the Human Cerebral Cortex Estimated by Intrinsic Functional Connectivity. *Journal of neurophysiology*, 2011.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. *Advances in neural information processing systems*, 31, 2018.

Supplementary for: Neuro-Symbolic Decoding of Neural Activity

The appendix is organized into five main sections. Appendix A includes additional experiment results: performance across atlases, effect sizes between atlases, concept grounding consistency across atlases, evaluation over all subjects, statistical tests across hypotheses, cross-dataset transfer experiments, network ablations, detailed concept accuracy, predicate argument binding, fine-grained generalization analyses, additional null models, and fMRI retrieval results. Appendix B provides additional visualizations and analyses of NEURONA's concept grounding performance. Appendix D describes the training procedure of NEURONA, the implementation of baseline methods, and the setup of our hypothesis ablation experiments. Appendix C presents more examples illustrating our fMRI-QA datasets and detail the data generation process. Appendix E details our ethics statement. Here, we also note that we use large language models to make minor improvements to writing.

A Additional results

A.1 Performance across atlases

We report performance from NEURONA across atlases to show robustness of decoding. We map parcellated fMRI signals (1024 regions for BOLD5000, 1000 for CNeuroMod) to multiple atlases, including Yeo-7, Yeo-17 (Yeo et al., 2011), DiFuMo-64, DiFuMo-128 (Dadi et al., 2020), and Schaefer-100 (Schaefer et al., 2018), then train NEURONA on these neural entities. Results in Table 6 and Table 7 show that NEURONA consistently learns across these atlases, and still significantly outperforms prior works in decoding accuracy. Yeo-17 yields the highest accuracy among all tested atlases, followed by DiFuMo-128 and Schaefer-100.

Table 6: NEURONA's performance with coarse and fine-grained atlases on BOLD5000.

BOLD5000	Overall	Action	Position	T/F
Yeo-7	0.6864	0.5517	0.4730	0.7270
Yeo-17	**0.7041**	**0.6207**	0.5079	**0.7407**
DiFuMo-64	0.6992	0.5517	**0.5524**	0.7282
DiFuMo-128	0.7026	0.5862	0.5460	0.7327

Table 7: NEURONA's performance with coarse and fine-grained atlases on CNeuroMod.

CNeuroMod	Overall	Action	Position	T/F
Yeo-7	0.6969	0.6459	0.5577	0.7180
Yeo-17	**0.7046**	0.6514	**0.5746**	0.7250
Schaefer-100	0.7043	**0.6549**	0.5614	**0.7258**

A.2 Effect sizes between atlases

To additionally evaluate the robustness of NEURONA to different brain parcellations, we compute Cohen's d effect sizes between QA predictions from different atlases. For each atlas pair, we compute paired effect sizes using QA predictions across the test set. As seen in Table 8 and Table 9, effect sizes are consistently small, showing that NEURONA is robust to the choice of atlas and performs reliably across a range of parcellations.

Table 8: Effect sizes between atlases in BOLD5000.

BOLD5000	Yeo-7	Yeo-17	Difumo-64	Difumo-128
Yeo-7	-	-0.017	-0.133	-0.012
Yeo-17	0.017	-	-0.116	0.006
Difumo-64	0.133	0.116	-	0.121
Difumo-128	0.012	-0.006	-0.121	-

Table 9: Effect sizes between atlases in CNeuroMod.

CNeuroMod	Yeo-7	Yeo-17	Schaefer-100
Yeo-7	-	0.094	0.089
Yeo-17	-0.094	-	-0.006
Schaefer-100	-0.089	0.006	-

A.3 Concept grounding consistency across atlases

In Table 10 and Table 11, we report additional consistency scores averaged over all concepts in BOLD5000 and CNeuroMod, under multiple atlas configurations: Yeo-7, Yeo-17, DiFuMo64, DiFuMo128, and Schaefer100. NEURONA achieves consistently high grounding consistency across all atlases, significantly above the null baseline. Unary concepts show higher consistency than relational ones, as expected due to their simpler structure. Across the atlases, all show consistent results, with DiFuMo-64 best for BOLD5000 and Yeo-17 best for CNeuroMod. NEURONA's concept grounding is reproducible way across different stimuli and parcellations.

Table 10: Concept grounding consistency in BOLD5000.

BOLD5000	Overall	Unary Concept	Relational Concept
Yeo-7 Null	0.5738	–	–
Yeo-7 NEURONA	0.8207	0.8283	0.7343
Yeo-17 Null	0.5357	–	–
Yeo-17 NEURONA	0.8220	0.8351	0.6646
DiFuMo-64 Null	0.5075	–	–
DiFuMo-64 NEURONA	0.8462	0.8644	0.6064
DiFuMo-128 Null	0.5039	–	–
DiFuMo-128 NEURONA	0.8224	0.8380	0.6241

A.4 Evaluation over all subjects

We include evaluation over all subjects on both BOLD5000 (4 subjects) and CNeuroMod (3 subjects). For each dataset, we train an individual model for each subject and report the mean ± standard deviation across subjects. We evaluate the in-distribution neural decoding performance in Table 12 and our generalization split in Table 13. We see that NEURONA continues to significantly outperform all prior works.

In addition, we evaluate cross-subject consistency in concept groundings. To evaluate whether concepts are grounded similarly across individuals, we aggregate grounding scores for each concept across all subjects and queries (for concept c appearing N times per subject, we obtain $N \times S$ samples where S is the number of subjects). We then compute a cross-subject consistency metric that measures the similarity of the spatial grounding patterns for each concept across individuals. We evaluate NEURONA against a random assignment null model and a multinomial null model that preserves each concept's distribution across regions. Our results in Table 14 show that NEURONA's cross-subject grounding consistency scores, averaged across all concepts, are significantly higher than both null models ($p < 0.001$), demonstrating that NEURONA's concept groundings converge to more similar sets of regions across participants.

Table 11: Concept grounding consistency in CNeuroMod.

CNeuroMod	Overall	Unary Concept	Relational Concept
Yeo-7 Null	0.5740	–	–
Yeo-7 NEURONA	0.8437	0.8564	0.7563
Yeo-17 Null	0.5358	–	–
Yeo-17 NEURONA	0.8700	0.8967	0.6812
Schaefer-100 Null	0.5029	–	–
Schaefer-100 NEURONA	0.8346	0.8695	0.5838

Table 12: Results across subjects for BOLD5000 and CNeuroMod.

BOLD5000	Overall	Action	Position	T/F
Linear	0.4702 ± 0.0073	0.0677 ± 0.0815	0.1817 ± 0.0289	0.5280 ± 0.0068
UMBRAE	0.4948 ± 0.0098	0.2488 ± 0.0572	0.2092 ± 0.0401	0.5401 ± 0.0097
SDRecon	0.4751 ± 0.0064	0.2887 ± 0.0499	0.2005 ± 0.0047	0.5266 ± 0.0075
BrainCap	0.4842 ± 0.0052	0.2412 ± 0.0470	0.2005 ± 0.0047	0.5383 ± 0.0061
NEURONA	**0.7102 ± 0.0053**	**0.5965 ± 0.0322**	**0.5378 ± 0.0135**	**0.7425 ± 0.0057**
CNeuroMod	Overall	Action	Position	T/F
Linear	0.4579 ± 0.0142	0.2078 ± 0.0400	0.1680 ± 0.0338	0.5184 ± 0.0153
UMBRAE	0.4588 ± 0.0425	0.1588 ± 0.0102	0.1705 ± 0.0220	0.5225 ± 0.0248
SDRecon	0.4368 ± 0.0012	0.1428 ± 0.0010	0.1554 ± 0.0010	0.5012 ± 0.0024
BrainCap	0.4422 ± 0.0026	0.1245 ± 0.0099	0.1497 ± 0.0074	0.5110 ± 0.0024
NEURONA	**0.7189 ± 0.0009**	**0.6422 ± 0.0072**	**0.5931 ± 0.0084**	**0.7417 ± 0.0005**

A.5 STATISTICAL TESTS ACROSS HYPOTHESES

We conduct additional statistical analyses for our hypotheses ablations to more rigorously evaluate the effects of the five hypotheses across subjects. Specifically, we use the Wilcoxon signed-rank test to assess statistical significance and Cohen's d to estimate effect size. For each subject, we collect accuracies across all metrics, and perform comparisons between hypotheses using these values. The BOLD5000 and CNeuroMod results are summarized in Figure 4 and Figure 5. Across both datasets, NEURONA's full argument-guided multi-region hypothesis (H5) consistently shows statistically significant improvements over alternative hypotheses.

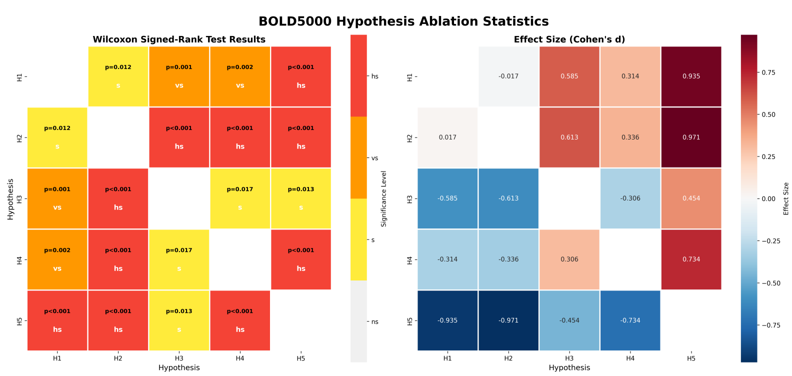

Figure 4: Statistical tests evaluating the effects of the five hypotheses across subjects in BOLD5000.

A.6 CROSS-DATASET TRANSFER EXPERIMENTS

Here, we include cross-dataset generalization experiments by training our model on BOLD5000-QA and evaluating it on the CNeuroMod-QA test set. Since BOLD5000 spans a broader concept space,

Table 13: Generalization results across subjects for BOLD5000 and CNeuroMod.

BOLD5000	Overall	Action	Position	T/F
Linear	0.4595 ± 0.0093	0.0308 ± 0.0312	0.1291 ± 0.0545	0.5250 ± 0.0059
UMBRAE	0.4886 ± 0.0068	0.1214 ± 0.0217	0.0914 ± 0.0212	0.5392 ± 0.0080
SDRecon	0.4752 ± 0.0064	0.2973 ± 0.0361	0.2005 ± 0.0047	0.5266 ± 0.0075
BrainCap	0.4838 ± 0.0055	0.2412 ± 0.0470	0.1996 ± 0.0040	0.5380 ± 0.0064
NEURONA	**0.6812 ± 0.0055**	**0.4952 ± 0.0682**	**0.4696 ± 0.0190**	**0.7217 ± 0.0057**
CNeuroMod	Overall	Action	Position	T/F
Linear	0.4206 ± 0.0129	0.0462 ± 0.0654	0.1045 ± 0.0667	0.4986 ± 0.0052
UMBRAE	0.4487 ± 0.0040	0.1107 ± 0.0156	0.1386 ± 0.0047	0.5247 ± 0.0035
SDRecon	0.4365 ± 0.0014	0.1407 ± 0.0008	0.1560 ± 0.0012	0.5012 ± 0.0019
BrainCap	0.4382 ± 0.0012	0.1217 ± 0.0070	0.1455 ± 0.0060	0.5069 ± 0.0003
NEURONA	**0.6676 ± 0.0119**	**0.2916 ± 0.0878**	**0.5377 ± 0.0215**	**0.7260 ± 0.0072**

Table 14: Cross-subject consistency of concept grounding.

	BOLD5000	CNeuroMod
Null (random)	0.5276 ± 0.0629	0.5267 ± 0.0683
NULL (multinomial)	0.6045 ± 0.1685	0.6481 ± 0.1644
NEURONA	0.7000 ± 0.2514	0.7027 ± 0.2165

we selected overlapping queries across datasets. In the CNeuroMod test set, this includes $1,169$ queries for the action task, $2,600$ for the position task, and $23,038$ for the T/F task (out of full test set sizes of $2,912$, $2,661$, and $24,486$, respectively).

In Table 15, we compare NEURONA to UMBRAE (Xia et al., 2024), the top performing baseline model. NEURONA significantly outperforms UMBRAE across all queries, demonstrating stronger cross-dataset robustness and generalization. Notably, while overall performance of NEURONA drops, largely due to a performance gap on T/F queries, accuracy on action and position tasks remains high, indicating some degree of cross-dataset transfer. This drop is expected, as our model is trained as a subject-specific model and there is substantial variance across subjects. Additionally, the two datasets differ in preprocessing pipelines: BOLD5000 uses the DiFuMo-1024 parcellation, while CNeuroMod uses the Schaefer-1000 atlas. This difference requires us to apply padding to align the feature dimensions when evaluating on CNeuroMod. Furthermore, some concepts in CNeuroMod, such as `telephone`, occur infrequently in BOLD5000, which limits NEURONA's ability to generalize to them. Nonetheless, we find that NEURONA maintains strong performance on queries such as action decoding, suggesting meaningful transfer of motor-related neural representations across datasets.

Table 15: Cross-dataset generalization results, where models are trained on BOLD5000 and tested on CNeuroMod.

	Overall	Action	Position	T/F
UMBRAE	0.4494	0.0106	0.0485	0.5036
Ours	**0.5535**	**0.7237**	**0.5246**	**0.5481**

A.7 Network ablations

To evaluate how each functional network contributes to decoding performance, we perform a network-level ablation in which NEURONA was trained using only a constrained subset of Yeo-7 networks at a time, isolating the contribution of each network. In Table 16, we see that the top performing networks are the Default Mode, Control, Dorsal Attention, and Visual networks, all of which have been closely linked to visual processing and high-level perceptual representations in prior works Menon (2023); Kucyi et al. (2020); Miyawaki et al. (2008).

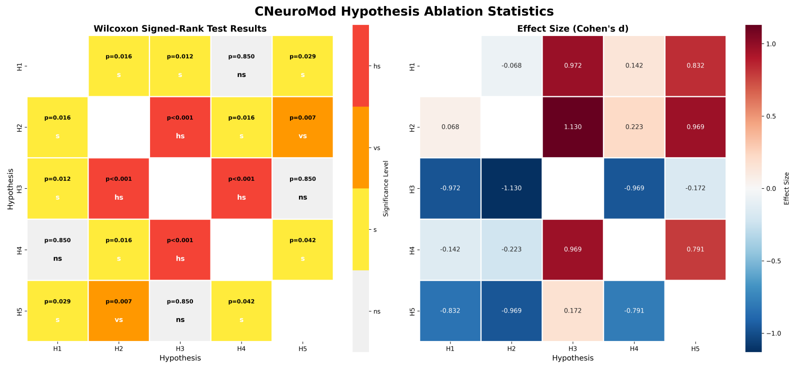

Figure 5: Statistical tests evaluating the effects of the five hypotheses across subjects in CNeuroMod.

Table 16: Network ablations to evaluate the contributions of each network to decoding performance.

Selected network	Overall	Action	Position	T/F
VisCent VisPeri	0.6750 ± 0.0067	0.3577 ± 0.0459	0.3765 ± 0.0159	0.7330 ± 0.0072
SomMotA/B	0.6724 ± 0.0069	0.3480 ± 0.0317	0.3620 ± 0.0131	0.7326 ± 0.0079
DorsAttnA/B	0.6753 ± 0.0073	0.3281 ± 0.0786	0.3850 ± 0.0282	0.7324 ± 0.0076
SalVentAttnA/B	0.6709 ± 0.0069	0.2910 ± 0.0655	0.3839 ± 0.0267	0.7281 ± 0.0060
LimbicA/B	0.6669 ± 0.0101	0.3379 ± 0.0918	0.3512 ± 0.0603	0.7282 ± 0.0044
ContA/B/C	0.6785 ± 0.0038	0.3380 ± 0.0808	0.4129 ± 0.0173	0.7310 ± 0.0035
DefaultA/B/C TempPar	0.6804 ± 0.0080	0.3248 ± 0.0752	0.4357 ± 0.0324	0.7297 ± 0.0066
All	0.7102 ± 0.0053	0.5965 ± 0.0322	0.5378 ± 0.0135	0.7425 ± 0.0057

A.8 Detailed concept accuracy

In Table 17 and Table 18, we report QA accuracy for unary and relational concepts separately across BOLD5000 and CNeuroMod, to analyze whether query structure affects performance. We see that that performance is generally stable across concept types, and across multiple atlases.

In Figure 6 and Figure 7, we illustrate confusion matrices for both BOLD5000 and CNeuroMod. We see that, as expected, while many concepts are reliably decoded (e.g., visually distinctive actions), some still cause confusion (e.g., spatial relations that are semantically similar).

Table 17: Accuracy breakdown between unary and relational concepts in BOLD5000.

BOLD5000	Overall	Unary	Relation
Yeo-7	0.727	0.717	0.751
Yeo-17	0.740	0.732	0.760
DiFuMo-64	0.728	0.727	0.728
DiFuMo-128	0.732	0.733	0.730

A.9 Predicate argument binding

We investigate whether unguided predicate representations contain decodable information about their arguments, by computing pairwise correlations between concept groundings across brain regions. For each relational query (e.g., `hold(person, baseball)`), we extract grounding score vectors across functional networks for the subject, object, unguided predicate, and guided predicate. We then compute the correlation matrix between these grounding patterns across all relational queries in both BOLD5000 (4 subjects) and CNeuroMod (3 subjects).

In Table 19, we report the mean correlations, and see minimal correlation between unguided predicate groundings and their subject (-0.0940) or object (0.0048) arguments in BOLD5000, with similarly

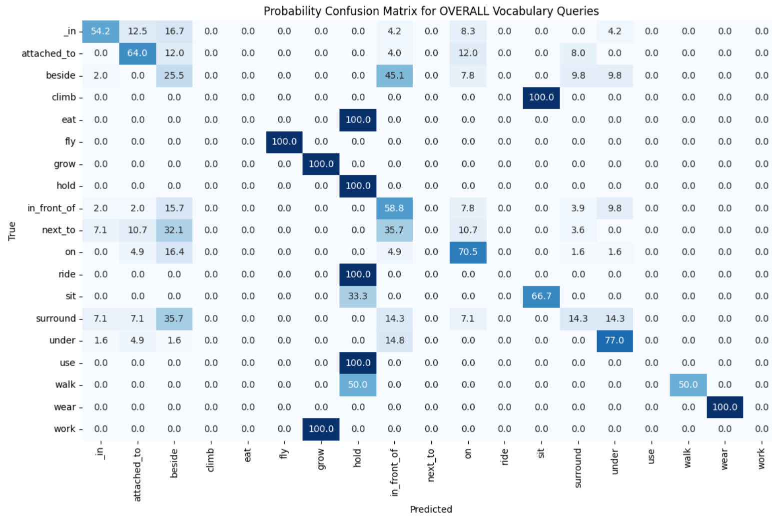

Figure 6: Accuracy confusion matrix across concepts in BOLD5000.

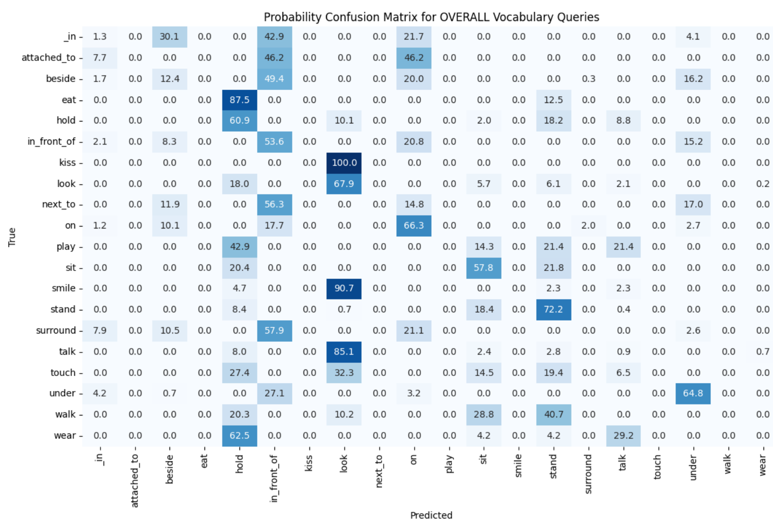

Figure 7: Accuracy confusion matrix across concepts in CNeuroMod.

low correlations in CNeuroMod (0.0230 for subject, 0.0019 for object). This suggests that when no structural guidance is provided, predicate representations do not bind to specific arguments.

In contrast, guided predicate representations showed substantially higher correlations with both subjects (0.2020 in BOLD5000, 0.1095 in CNeuroMod) and objects (0.2975 in BOLD5000, 0.2552 in CNeuroMod). This significant increase in correlation indicates that NEURONA's explicit guidance notably helps decode relations from predicate-argument interactions.

A.10 Fine-grained generalization analyses

Our generalization experiments on both BOLD5000 (subject-CSI1) and CNeuroMod (subject-01) include argument swapping, predicate transfer, and role systematicity. In Table 20, we report fine-grained systematicity tests across each of these settings. We see that NEURONA shows strong performance on each category, with relatively small drops from all to unseen combinations.

Table 18: Accuracy breakdown between unary and relational concepts in CNeuroMod.

CNeuroMod	Overall	Unary	Relation
Yeo-7	0.718	0.696	0.754
Yeo-17	0.725	0.707	0.754
Schaefer-100	0.725	0.708	0.754

Table 19: Results of predicate argument binding.

	BOLD5000	CNeuroMod
Unguided-predicate / subject	-0.0940	0.0230
Unguided-predicate / object	0.0048	0.0019
Unguided-predicate / guided-predicate	0.0336	0.1488
Guided-predicate / subject	0.2020	0.1095
Guided-predicate / object	0.2975	0.2552

Table 20: Fine-grained generalization analyses in BOLD5000 and CNeuroMod.

Bold5000	Argument swapping	Predicate transfer	Role systematicity	Other	All (T/F)
Unseen	0.7451	0.75	0.7391	0.7096	0.7184
All	0.7562	0.6667	0.8030	0.7393	0.7407
CNeuroMod	Argument swapping	Predicate transfer	Role systematicity	Other	All (T/F)
Unseen	0.7526	0.7333	0.7390	0.6842	0.6991
All	0.7011	0.5088	0.7020	0.7307	0.7250

A.11 Additional null model

For our consistency metric, we include a stronger null model that uses a multinomial distribution to randomly assign concept groundings across brain networks, preserving the exact total sample count for each concept while distributing samples uniformly across all networks. This ensures that the null model retains the same data structure as the observed data (i.e., same total grounding counts per concept) while randomizing only the network assignments. We run this null model 10 times and report the mean and standard deviation in Table 21. NEURONA similarly significantly outperforms the null model in grounding consistency.

Table 21: Consistency metric comparison with a multinomial null model.

	BOLD5000	CNeuroMod
Null (multinomial)	0.6267 ± 0.0067	0.6424 ± 0.0057
NEURONA	0.8475 ± 0.0186	0.8592 ± 0.0084

A.12 fMRI retrieval results

Here, we detail results from an fMRI retrieval task, where we adapt NEURONA to retrieve the corresponding fMRI given a symbolic query. We use positive queries to ensure that the concept and fMRI are well aligned. Each concept is represented as a one-hot vector over the full concept vocabulary, from which a concept embedding is obtained using a small MLP. We then follow the structure of the symbolic expression by applying an aggregation operation to combine the specified concepts in the query. A parallel MLP encodes the fMRI input into an fMRI embedding, and we train the embedding spaces jointly using a CLIP-based contrastive loss. In this setting, we achieve a test top-1 retrieval accuracy of 0.1325 and a top-5 accuracy of 0.3012, substantially outperforming a random-choice baseline (top-1: 0.0120, top-5: 0.0602).

B Concept Grounding Visualizations

In Figure 8, we present concept grounding examples from the BOLD5000 (Chang et al., 2019) and CNeuroMod (Gifford et al., 2024; Boyle et al., 2023) datasets.

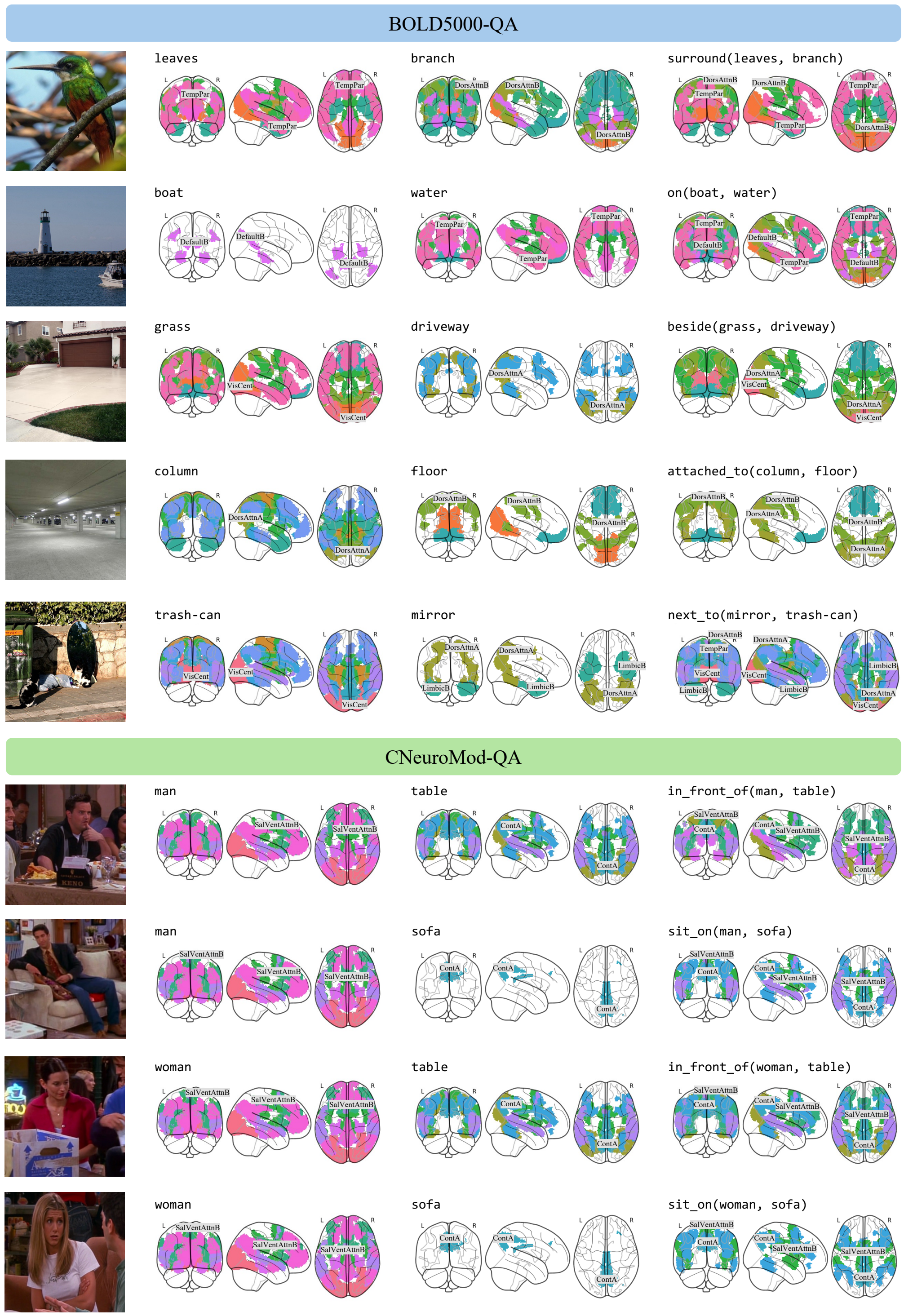

Figure 8: We show examples of learned concept grounding by NEURONA on BOLD5000-QA and CNeuroMod-QA, across subject, object, and predicate concepts.

C Datasets

C.1 License for existing datasets

We train NEURONA on the BOLD5000 (Chang et al., 2019) and CNeuroMod (Gifford et al., 2024; Boyle et al., 2023) datasets, which are both licensed under the Creative Commons 0 License. More information can be found on their websites: BOLD5000 and CNeuroMod.

C.2 fMRI-QA Datasets

BOLD5000. We utilize the BOLD5000 dataset, which has been preprocessed and aligned with image stimuli following WAVE (Wang et al., 2024). The fMRI data has a shape of $[5, 1024]$, representing 5 TRs and 1024 brain regions. We use preprocessed, image-aligned fMRI data provided here. Each TR (repetition time) is 2 seconds, resulting in a chunk duration of 10 seconds (5×2s). We account for a hemodynamic lag of 2 TRs. All four subject pairs are included, following the same train-test split as in previous studies.

CNeuroMod. We use the CNeuroMod dataset preprocessed by the Algonauts Challenge (Gifford et al., 2024; Boyle et al., 2023). The fMRI data has a TR of 1.49 seconds and a shape of $[5, 1000]$, representing 5 TRs and 1000 brain regions based on the Schaefer-1000 atlas (Schaefer et al., 2018). This yields a chunk duration of 7.45 seconds (5×1.49s), with a hemodynamic lag of 3 TRs. For each chunk, we select the most motion-informative video frame by computing motion energy as the absolute difference between consecutive frames. The chunks are extracted from Friends episodes, with seasons 1–5 used for training and the unseen season 6 reserved for testing. We use $1,000$ fMRI-video chunks per season, resulting in $5,000$ training samples and $1,000$ testing samples.

We provide additional examples from our datasets in Figure 9.

Figure 9: Examples of queries in BOLD5000-QA and CNeuroMod-QA.

D EXPERIMENT DETAILS

D.1 TRAIN SETTINGS

We train and evaluate NEURONA on the specified training and test sets for both BOLD5000 (Chang et al., 2019) and CNeuroMod (Gifford et al., 2024; Boyle et al., 2023) datasets. Training is conducted for 100 epochs using the Adam optimizer (Diederik, 2014), with learning rate 0.001 and batch size 32.

D.2 COMPUTE RESOURCES

Since NEURONA consists of a lightweight convolutional neural network for fMRI feature extraction followed by a linear classifier for concept grounding and execution, its computational requirements are minimal. All experiments are conducted on a single NVIDIA A100 GPU, with training completing in approximately 30 minutes. Data loading is parallelized using 16 CPU workers, and the system uses 64 GB of RAM.

D.3 BASELINE IMPLEMENTATIONS

We describe the implementation details of the baseline models compared in our study below. In all methods, we treat the fMRI input as a sequence of length 5, with each time step as a token.

Linear We tokenize the input query using the BERT tokenizer (Devlin et al., 2019) and pad all sequences to a fixed length. The tokens are then encoded using a linear layer. We concatenate the fMRI token sequence and the query token sequence, and apply a final linear classification layer to predict the output (either a binary T/F answer or a vocabulary token).

SDRecon We implement SDRecon (Takagi & Nishimoto, 2023) following the official repository*. A ridge regression model aligns fMRI features with image embeddings, which are then passed to a VQA-GIT language model (Wang et al., 2022) to generate answers. We set the ridge regularization parameter to $\lambda = 20$. A custom parser (described below) is used to map the generated language response to a valid prediction.

BrainCap BrainCap similarly uses a linear encoder to align fMRI features with visual embeddings (Ferrante et al., 2023). The aligned embeddings are passed to a BLIP language model (Li et al., 2022) to generate answers. We apply the same parser as in SDRecon to extract final predictions from the language output. We follow the implementation of the official repository†.

UMBRAE UMBRAE leverages a transformer-based encoder to map fMRI features to image embeddings (Xia et al., 2024). These embeddings are then passed to LLaVA (Liu et al., 2023) for language-based prediction, followed by response parsing. We follow the implementation of the official repository‡.

We implement a rule-based parser to convert language model outputs into structured predictions. The parser first cleans the text by removing punctuation, digits, and formatting inconsistencies. It identifies binary answers (“yes” or “no”) when the query requires. For other queries, it extracts the first valid word from a predefined vocabulary. If no valid word is found, it defaults to the most common answers of “on” for spatial queries or “hold” otherwise. For all image-grounded baselines, the ground-truth image embeddings are derived from the visual encoder of a vision-language model for BOLD5000. We use the embedding of the first selected video frame as the ground truth for CNeuroMod.

*`https://github.com/yu-takagi/StableDiffusionReconstruction`
†`https://github.com/enomodnara/BrainCaptioning`
‡`https://github.com/weihaox/UMBRAE`

D.4 Ablation details

In this section, we provide the full definitions of the entities and grounding hypotheses introduced in the main paper.

D.4.1 Entity processing

To enable concept grounding to neural activity $\mathrm{f} \in \mathbb{R}^{N \times T}$, we first map the fine-grained networks $N = 1024$ to P functional networks defined by the given atlas. This results in P network-specific fMRI signals $\{f_1, \ldots, f_P\}$, where each $f_p \in \mathbb{R}^{m_p \times T}$ represents the aggregated signal from m_p fine-grained regions assigned to network p. Since the number of regions m_p vary across networks, we apply a linear stitcher to project each $f_p \in \mathbb{R}^{m_p \times T}$ to a fixed-dimensional representation $e_p \in \mathbb{R}^{d \times T}$, where $d = 256$, using network-specific linear projections $W_p \in \mathbb{R}^{m_p \times d}$, such that $e_p = W_p^\top f_p$. This produces a unified set $\mathcal{E}$ of P embeddings $\{e_1, \ldots, e_P\}$, which are then processed by a small 1-D convolutional encoder to form parcellation embeddings. From these base embeddings, we propose hypotheses of candidate entities from which concepts can be grounded.

D.4.2 General grounding formulation

We define single-region (unary) grounding for a concept c as $G_{\text{unary}}(c)$ and multi-region (binary) grounding as $G_{\text{binary}}(c)$. We further define a fused grounding score combining unary and binary components:

$$G(c) = G_{\text{unary}}(c) + \frac{1}{P} \sum_{i=1}^{P} G_{\text{binary}}(c_i). \tag{4}$$

D.4.3 Hypotheses definitions

Let c_p, c_s, and c_o be the concepts for predicate (e.g., `holding`), subject (e.g., `person`), and object (e.g., `baseball-bat`), respectively.

H1: Single-region grounding. Concepts are grounded to a single brain region:

$$\begin{aligned} G_{\text{H1}}(c_p) &= G_{\text{unary}}(c_p), \\ G_{\text{H1}}(c_s) &= G_{\text{unary}}(c_s), \\ G_{\text{H1}}(c_o) &= G_{\text{unary}}(c_o). \end{aligned} \tag{5}$$

H2: Multi-region co-activation. Concepts are grounded through co-activation across region pairs:

$$\begin{aligned} G_{\text{H2}}(c_p) &= G_{\text{binary}}(c_p), \\ G_{\text{H2}}(c_s) &= G_{\text{unary}}(c_s), \\ G_{\text{H2}}(c_o) &= G_{\text{unary}}(c_o). \end{aligned} \tag{6}$$

H3: Predicate conditioned on subject. Predicate representations are guided by the activation of the subject region:

$$\begin{aligned} G_{\text{H3}}(c_p) &= G_{\text{binary}}(c_p) + G_{\text{unary}}(c_s), \\ G_{\text{H3}}(c_s) &= G_{\text{unary}}(c_s), \\ G_{\text{H3}}(c_o) &= G_{\text{unary}}(c_o). \end{aligned} \tag{7}$$

H4: Predicate conditioned on object. Predicate representations are guided by the activation of the object region:

$$\begin{aligned} G_{\text{H4}}(c_p) &= G_{\text{binary}}(c_p) + G_{\text{unary}}(c_o), \\ G_{\text{H4}}(c_s) &= G_{\text{unary}}(c_s), \\ G_{\text{H4}}(c_o) &= G_{\text{unary}}(c_o). \end{aligned} \tag{8}$$

H5: Full argument-guided grounding. Our proposed method combines multi-region grounding with subject and object guidance. The grounding scores are defined as:

$$\begin{aligned} G_{\mathrm{H5}}(c_p) &= G(c_p) + G(c_s) + G(c_o), \\ G_{\mathrm{H5}}(c_s) &= G(c_s), \\ G_{\mathrm{H5}}(c_o) &= G(c_o). \end{aligned} \tag{9}$$

These formulations enable systematic evaluation of how structural priors affect downstream neural decoding accuracy.

E ETHICS STATEMENT

This work uses only publicly released fMRI datasets, and focuses on decoding concepts from visual stimuli rather than personal traits or identity, which reduces the risk of individual fingerprinting. However, we acknowledge the inherent risks in building architectures that enable neural decoding from participants. We emphasize that our method is intended solely for scientific analysis, and should not be used for identification or inference of sensitive personal attributes from neural activity.