BRAIN-IT-VQA: FROM BRAIN SIGNALS TO ANSWERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding visual content from fMRI signals recorded while a person views images, and specifically answering questions about the seen images, is a long-standing challenge. While significant progress has been made in recent years in visual question answering (VQA) from fMRI, performance remains limited. Moreover, although recent models can make increasingly accurate predictions, they have rarely been used as tools for understanding the structure of visual representations in the brain. We present ***Brain-IT-VQA***, a framework for visual question answering from fMRI. Unlike previous methods, which extract a fixed representation from the fMRI signal, our extraction is conditioned on the question itself, so what is decoded from brain activity depends on what is being asked. Our model substantially outperforms previous fMRI-based captioning and VQA approaches. We further introduce ***NSD-VQA***, a new large-scale dataset and benchmark for visual question answering from fMRI. Unlike existing image-fMRI VQA datasets, which typically provide only a few broad and weakly controlled questions per image, *NSD-VQA* provides on average 20 question-answer pairs per image, across a wide range of controlled question categories that disentangle multiple levels of visual understanding. This enables more reliable and interpretable evaluation despite limited fMRI test data. Together, *Brain-IT-VQA* and *NSD-VQA* provide both a strong predictive framework and a tool for studying brain representations. Using this benchmark, we quantify which forms of visual and semantic information can be reliably decoded from fMRI responses to natural images. We further analyze the contributions of different brain regions across question types.

1 INTRODUCTION

Understanding what visual information is represented in the human brain, and how different aspects of a visual scene are encoded across cortex, is a long-standing challenge in neuroscience. One approach to this question is to examine what can be decoded from functional MRI signals recorded while a person views images. Decoding visual information from fMRI can reveal information at multiple levels of abstraction, from broad semantic content, such as object and scene categories, to more specific visual properties, such as color, shape, spatial layout, and object attributes.

Recent advances in machine learning and generative models have substantially improved the ability to decode visual information from fMRI. These efforts include methods for reconstructing perceived images from brain activity (Beliy et al., 2026; Takagi & Nishimoto, 2023a; Ozcelik & VanRullen, 2023) mapping brain activity to fixed visual representations (Shen et al., 2019; Wang et al., 2024a), and decoding fMRI into language or multimodal representations (Xia et al., 2024; Huang et al., 2025; Qiu et al., 2025). While reconstructing images or decoding specific visual properties provides an important way to study visual representations in the brain, decoding language-related representations offers a more flexible and interpretable interface for probing specific concepts and attributes. In particular, generating captions to perceived images from fMRI and visual question answering (VQA) provide direct ways to ask what information about an image can be extracted from brain activity. However, existing fMRI-based captioning and VQA models remain limited in performance, and are not necessarily aligned with the kinds of controlled aspects that are useful for neuroscientific analysis. Since these models are typically trained and evaluated using existing vision-oriented captioning or VQA datasets, the available questions often do not target controlled neuroscientific distinctions. Together with the limited analysis of model behavior across question types and brain regions, this leaves the field still insufficiently explored.

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.

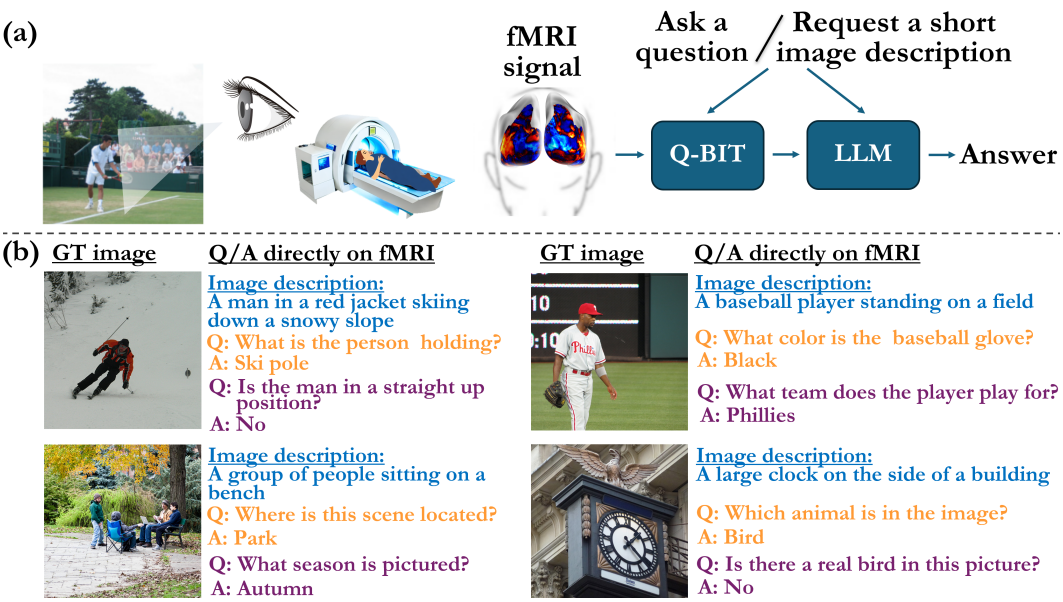


Figure 1: **fMRI-to-Language decoding: Captioning & VQA directly from fMRI brain activity.** ***(a)*** *Overview of the pipeline; fMRI signals recorded while a subject views an image are used to generate a caption or answer questions about the image.* ***(b)*** *Example captions on question-answer pairs generated by Brain-IT-VQA directly from fMRI signals.*

To address these limitations, we propose ***Brain-IT-VQA*** (Fig. 1), a new framework for visual question answering from fMRI, together with ***NSD-VQA***, a new dataset and benchmark designed for controlled evaluation of question answering from brain activity. *Brain-IT-VQA* provides a strong predictive model for answering questions about perceived images directly from fMRI, while NSD-VQA enables a more detailed analysis of which visual and semantic information can be decoded from the brain. *Brain-IT-VQA* builds on the Brain Interaction Transformer introduced in Brain-IT (Beliy et al., 2026), a method for reconstructing seen images from fMRI. It represents fMRI signals through shared functional voxel [1] groups and integrates distributed neural information across subjects. We adapt this architecture to decode *language-conditioned* representations from brain activity and combine them with a pretrained language model. Importantly, unlike prior brain-to-language methods the question guides our decoding, allowing different questions to extract different information from the same fMRI signal. This allows the model to answer natural-language questions about perceived images from brain activity. *Brain-IT-VQA* achieves state-of-the-art performance on fMRI-based captioning and visual question answering.

In addition to the model, we introduce ***NSD-VQA***, a large-scale benchmark for fine-grained evaluation of visual question answering from fMRI. Existing fMRI-VQA benchmarks typically contain only a small number of broad or weakly controlled questions per image, making it difficult to determine which types of information are actually decoded. ***NSD-VQA*** addresses this by providing, on average, 20 question-answer pairs per image in the NSD dataset (Allen et al., 2022). These questions are automatically generated from a wide range of controlled question categories that target different aspects of visual understanding, including objects, attributes, spatial relations, counting, actions, and scene-level information. This dense and structured annotation makes each measured fMRI response substantially more informative, enabling more reliable evaluation under limited test data.

NSD-VQA allows us to move beyond a single overall VQA score and evaluate which forms of visual and semantic information can be reliably inferred from fMRI recordings of image viewing. Using this benchmark, we evaluate and analyze Brain-IT-VQA across question categories, and can further relate its predictions to brain regions. Together, Brain-IT-VQA and NSD-VQA provide a framework for using visual question answering not only as a brain-decoding task, but also as a tool for probing the organization of visual representations in the human brain.

[1] A 'brain-voxel' is a small 3D cube of brain tissue, whose activity is estimated via fMRI and represented by a single scalar value.

Our contributions are therefore as follows:

- We introduce ***Brain-IT-VQA***, a framework for end-to-end visual question answering from fMRI, yielding state-of-the-art results.
- We introduce ***NSD-VQA***, a new extensive fMRI-VQA dataset specifically tailored for fMRI analysis, enabling quantitative evaluation of distinct types of visual and semantic information.
- We conduct a systematic empirical study of decodable information from fMRI responses to natural images, identifying which types of visual and semantic content can be reliably inferred.
- We provide an interpretable analysis of brain regions contributions, quantifying how different brain regions support distinct types of questions.

2 Related Work

Vision-Language Models for VQA (on images): Recent advances in vision-language models (VLMs) have significantly improved performance on visual question answering and multimodal reasoning tasks. However, VLMs differ fundamentally in how they bridge visual and language modalities. One line of work, including Flamingo (Alayrac et al., 2022) and the LLaVA family (Liu et al., 2023; 2024), connects visual encoders to LLMs via cross-attention gating or direct MLP projection, with the LLM fine-tuned on visual instruction data. Large proprietary systems such as GPT-4V (OpenAI, 2023) and Gemini (Gemini Team et al., 2023) train vision and language jointly from scratch, yielding strong performance but limiting modularity and research accessibility. BLIP-2 (Li et al., 2023) introduced an alternative: a lightweight Q-Former that distils image encoder outputs into a fixed set of query token embeddings fed as soft prompts to a fully *frozen* LLM. InstructBLIP (Dai et al., 2023) extends this with instruction-aware feature extraction, conditioning the Q-Former on the task prompt. In our work, we draw on InstructBLIP's idea of instruction-aware extraction and adapt it to operate on brain tokens. Its Q-Former, conditioned on the question, provides the supervision target for our first training stage, guiding the brain-derived tokens toward soft prompts compatible with the frozen LLM. In the second stage, our module is then adapted end-to-end through language-model supervision, without modifying any LLM weights. Direct-projection models require LLM fine-tuning on image-specific features, making this substitution significantly harder.

Vision-Based Brain Decoding: Decoding visual information from brain activity (fMRI) into perceptual and semantic representations has seen rapid progress in recent years. Early work focused on mapping fMRI signals to handcrafted or low-level visual features (Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011), followed by approaches leveraging deep neural network representations (Güçlü & Van Gerven, 2015; Shen et al., 2019; Zhang et al., 2018). End-to-end methods were later introduced (Seeliger et al., 2018; St-Yves & Naselaris, 2018; Beliy et al., 2019), followed by approaches predicting latent codes of generative models such as VAEs and GANs (Han et al., 2019; Lin et al., 2019; Mozafari et al., 2020; Qiao et al., 2020; Ren et al., 2021). More recent methods employ diffusion models to reconstruct images with increasing fidelity (Chen et al., 2023b; Takagi & Nishimoto, 2023a; Ozcelik & VanRullen, 2023). In parallel, there has been growing focus on leveraging shared structure across subjects to improve generalization under limited data (Scotti et al., 2024; Gong et al., 2025; Ferrante et al., 2024; Liu et al., 2025; Beliy et al., 2024).

Multimodal and Language-Based Decoding: Recent works extend brain decoding beyond reconstruction by mapping fMRI signals to natural language or multimodal representations. Methods such as MindGPT (Chen et al., 2023a), UniBrain (Wang et al., 2024b), BrainCap (Ferrante et al., 2023), BrainChat (Huang et al., 2025) align fMRI with visual and textual embeddings and decode language using pretrained models, typically compressing brain activity into a single shared representation. Other approaches predict intermediate stimulus representations and apply off-the-shelf visual-language models for downstream tasks (Xia et al., 2024). Contrastive alignment with vision-language models has also been explored to improve captioning quality and enable region-level interpretability (Shen et al., 2025). Recent work explores end-to-end fMRI-to-text decoding with large language models across different language-based tasks (Qiu et al., 2025). MindLLM (Qiu et al., 2025) is the most closely related and highest performing prior method, sharing transformer-based fMRI processing and multi-subject parcellations.

Our approach differs from all previous methods in that fMRI processing is question-dependent: the information extracted from the brain signal depends on the question being asked, rather than being a fixed readout subsequently conditioned on by the LLM. Additionally, our approach differs in its processing of brain activity. Following Brain-IT, we use data-driven functional clusters rather than

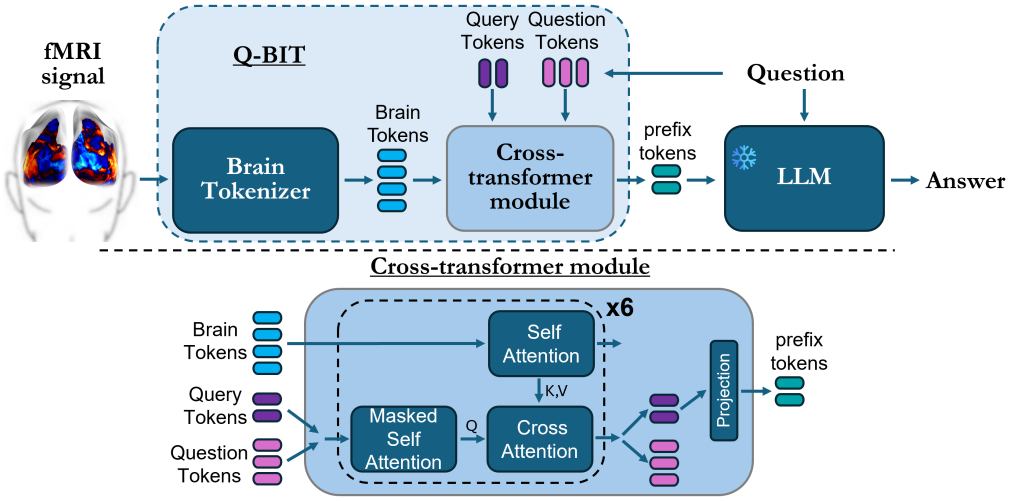


Figure 2: **Overview of the *Brain-IT-VQA* architecture.**

anatomical parcellations, which was shown to be superior for image decoding (Beliy et al., 2026), and dedicated cross-attention blocks to distill task-relevant representations rather than directly prepending Brain tokens to the LLM. Brain-IT reconstructs a single image from the fMRI signal, so its output is fixed and independent of any query, whereas we condition the extraction on the question, so that different questions draw different information from the same brain activity.

Datasets and Evaluation for fMRI Decoding: The dominant fMRI datasets for brain decoding, including NSD (Allen et al., 2022) and BOLD5000 (Chang et al., 2019), provide image–fMRI pairs, and are primarily designed for image reconstruction tasks. Evaluation in this setting often relies on pixel- and feature-level similarity metrics that capture global visual fidelity. These evaluations do not necessarily distinguish which types of visual or semantic information are recoverable from neural signals. Recent work has broadened the evaluation scope: BrainHub (Xia et al., 2024) extends NSD with captioning and grounding tasks, and BrainChat (Huang et al., 2025) introduces fMRI question answering evaluated by classification accuracy on broad COCO VQA questions. However, neither provides controlled question categories that systematically disentangle different levels of visual understanding such as; object identity, attribute, spatial relation, or scene-level semantics. Our newly designed NSD-VQA dataset fills this gap by providing structured question types designed to probe distinct aspects of visual cognition. It enables fine-grained, interpretable evaluation of what can and cannot be decoded from fMRI responses to natural images.

Interpretability and Brain-Region Contributions: Understanding how different brain regions contribute to decoded representations remains a central challenge. Many existing approaches compress fMRI signals into global embeddings (Takagi & Nishimoto, 2023a; Scotti et al., 2024; Wang et al., 2024a), obscuring the contribution of individual voxels or functional regions. Some methods attempt to preserve spatial structure through voxel grouping or attention mechanisms (Wang et al., 2024a; Huo et al., 2024), while others explore voxel-level or cross-subject representations (Beliy et al., 2024). Recently, BrainExplore (Wasserman et al., 2025) explores data-driven discovery of interpretable concepts from fMRI activity. However, these approaches do not provide systematic analysis linking brain regions to specific decoded information. In contrast, our approach allows for direct quantification of contributions from chosen brain regions across different question categories, providing insight into how brain regions support different forms of visual and semantic processing.

3 Method

3.1 Overview of our approach

We present ***Brain-IT-VQA*** (Fig. 1), a framework for decoding fMRI brain activity into natural language, supporting both image captioning and visual question answering (VQA). Given fMRI brain activity recorded while a subject views an image, the model generates either a descriptive caption or an answer to a textual query about the image from the fMRI signal. Our approach extends the Brain Interaction Transformer (BIT) of (Beliy et al., 2026) to predict language tokens from fMRI signals. We denote this extension as ***Q-BIT***, a question-conditioned Brain Interaction Transformer that produces soft prompts for an LLM. Q-BIT organizes voxel-level fMRI signals into clusters

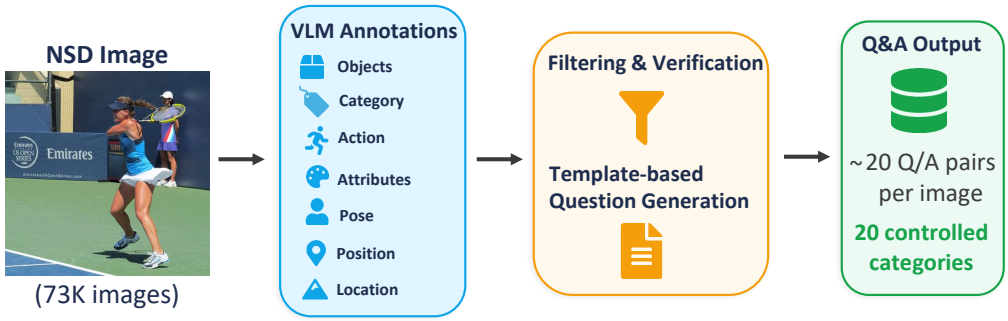


Figure 3: ***NSD-VQA* Dataset construction pipeline.** *Starting from NSD images, we generate structured annotations using a VLM, followed by filtering and verification. Template-based question generation then produces multiple question–answer pairs per image across controlled question categories.*

of *functionally* similar brain-voxels shared across multiple subjects. Each cluster is summarized into a compact *Brain Token*, and the question itself guides the attention that extracts task-relevant representations from these tokens. As such, the information extracted from brain activity depends on what is being asked, thus allowing different questions to extract different information from the same fMRI response. This enables direct generation of captions and answers from brain activity. For assumptions and limitations of our method, see App.A.

3.2 Model Architecture

Q-BIT transforms fMRI activations into a set of *"Brain Tokens"* – structured representations that summarize the activity of clusters of functionally similar brain-voxels. Following BIT (Beliy et al., 2026), the Brain Tokens interact through self-attention layers, while a cross-attention mechanism with learnable query tokens extracts task-relevant information from them. In our Brain-IT-VQA (see Fig. 2), the question guides this extraction. The question is embedded into a set of *question tokens*, and these are concatenated with the learnable query tokens. Together they pass through masked self-attention, where the mask prevents attention to padded positions in the question, and then serve as queries in the cross-attention over the Brain Tokens. This process is repeated across blocks, with the Brain Tokens themselves updated by self-attention in parallel, so that what is extracted from the fMRI signal depends on the question at every stage. The query token outputs are then projected into *soft prompts* which, together with the textual query, condition the frozen LLM of InstructBLIP to generate answers. Further architectural and implementation details are provided in App. B.

3.3 Training Setup

Training proceeds in two stages. In the first stage (**Q-BIT Pretraining**), InstructBLIP's Q-Former serves as a teacher. Given an image and a question, it produces a question-conditioned representation, and Q-BIT is trained with an MSE loss to reproduce it from the fMRI signal alone. This aligns the brain-derived representation with the question-conditioned representation the language model expects, before any language supervision is applied. In the second stage (**End-to-End Fine-Tuning**), Q-BIT is adapted using LoRA and trained end-to-end on either caption generation or VQA with an autoregressive cross-entropy loss over the target tokens, while the LLM remains frozen. Further training details, including computational resources are provided in App. B.

Enriching the Training Data. Both stages use fMRI recordings from the NSD dataset (Allen et al., 2022) paired with their corresponding images. Since subject-specific data is limited, we augment the training set with additional natural images which do not have fMRI recordings, by predicting their fMRI responses using the Image-to-fMRI encoder (Beliy et al., 2024). In the first stage, ~120k images from the unlabeled portion of COCO are used this way. In the second stage, all COCO images which are not part of the validation or test sets are used, as fMRI responses can be predicted for any subject, regardless of whether the image was actually observed.

4 The NSD-VQA Dataset

We introduce ***NSD-VQA***, a large-scale dataset tailored for fMRI-based visual question answering. NSD-VQA is generated from the NSD dataset (Allen et al., 2022), which comprises 73K images with corresponding fMRI brain recordings from 8 subjects, as well as ~123K unlabeled COCO images (Lin et al., 2014). We generate structured annotations and question-answer pairs for both sets of images, while fMRI recordings are available only for the NSD images.

Annotation pipeline. Starting from NSD images, we generate structured annotations using the vision-language model `Qwen3-VL-8B` (Bai et al., 2023). For each image, we prompt the model to extract the most salient objects together with attributes relevant for downstream questioning. These include the following controlled categories: object identity, counts, semantic categories (e.g. animal, vehicle, food), color, person actions, simple interactions (e.g. holding), and scene-level descriptors (type and location). These categories capture distinct aspects of visual and semantic processing represented in the natural-scene images and enable controlled evaluation across question types.

Filtering and verification. To improve annotation reliability, we perform a verification step focusing on object counts and presence. Counts are estimated using both `Qwen3-VL-8B` and `Gemma-4-31B-it` (Gemma Team, 2024), and retained only when both models agree; otherwise, the corresponding annotations are discarded. We additionally verify consistent object presence by ensuring that predicted counts are non-zero across both model predictions. Finally, a post-processing step using a BGE text encoder (Xiao et al., 2023) removes semantic redundancies by merging labels with high embedding similarity into a unified vocabulary (e.g. merging *laptop* and *notebook* into *computer*). We additionally assess the quality of the generated question-answer pairs through an independent human evaluation C.1.

Question & Answer generation. From the structured annotations, we generate VQA-style question-answer pairs using templates aligned with the annotated attributes. The dataset covers object-level properties (presence, counting, color), and scene understanding (indoor/outdoor and location), as well as category-specific queries (e.g. animals, vehicles) and human-centric attributes (actions and interactions). Question types are instantiated conditionally based on the presence of relevant annotations (e.g. animal, person, or object categories), ensuring that questions are grounded in visible content. For example, an annotation specifying *carrot* under the food category with an associated count of 2 would generate "What food is in the image?" and "How many carrots are in the image?", with answers "carrot" and "2", respectively. To ensure meaningful evaluation, we augment the dataset with targeted negative examples, particularly for presence and counting questions, and enforce minimum support constraints by discarding question types with fewer than 50 instances. We further balance answer distributions across categories by filtering questions with highly skewed answers (single class dominating more than 70% of instances). Default answers are short-form (e.g. single words or short expressions), enabling controlled evaluation. We additionally construct a full-sentence variant, **NSD-VQA-FS**, by prompting a large-language model (`Llama-3.2-3B` (Grattafiori et al., 2024)) to rewrite each question-answer pair into a full-sentence response.

Summary. NSD-VQA provides a structured benchmark for evaluating how different types of visual and semantic information are represented in fMRI signals. By decomposing the task into targeted question categories grounded in explicit annotations, it enables systematic analysis of what information can be reliably inferred from brain activity. All used prompts and additional details, including the question category design and human evaluation, are provided in App. C.

5 Experiments

5.1 Experimental Setup

Datasets. We train and evaluate on the Natural Scenes Dataset (NSD), a large-scale 7-Tesla fMRI dataset recording brain activity of 8 subjects viewing images drawn from COCO (Lin et al., 2014). Following standard practice, we use the 1,000 images shared across all subjects as the test set. We consider subjects 1, 2, 5, and 7, in line with prior NSD-based brain decoding works. For voxel selection, we adopt the post-processed version provided by Gifford et al. (Gifford et al., 2023), which includes $\sim$40k voxels, mainly from vision-related cortical areas. COCO provides image captions, which we use for captioning evaluation following the standard brain captioning benchmark (Xia et al., 2024). For VQA, we evaluate on four benchmarks that extend COCO with question-answer annotations: VQA-v2 (Goyal et al., 2017) provides multiple questions per image with 10 human answers each, covering general visual understanding. FSVQA (Shin et al., 2016) extends this setting to full-sentence answers, requiring richer language generation. Further, we evaluate on NSD-VQA, our proposed benchmark introduced in Sec. 4, and its full-sentence variant NSD-VQA-FS.

Evaluation Setup. For short-answer tasks (VQA-v2 & NSD-VQA), we report accuracy following standard evaluation protocol (Goyal et al., 2017; Antol et al., 2015). For captioning and full-sentence generation settings (FSVQA and NSD-VQA-FS), we report standard text generation metrics, including BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004),

Table 1: **Results on Brain captioning** *evaluated against COCO captions (subject 1).*

Method	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	METEOR ↑	ROUGE ↑	CIDEr ↑	SPICE ↑
SDRecon (Takagi & Nishimoto, 2023b)	36.21	17.11	7.22	3.43	10.03	25.13	0.138	5.02
OneLLM (Han et al., 2024)	47.04	26.97	15.49	9.51	13.55	35.05	0.230	6.26
UniBrain (Mai & Zhang, 2023)	–	–	–	–	16.90	22.20	–	–
BrainCap (Ferrante et al., 2023)	55.96	36.21	22.70	14.51	16.68	40.69	0.413	9.06
BrainChat (Huang et al., 2025)	52.30	29.20	17.10	10.70	14.30	45.70	0.261	–
UMBRAE (Xia et al., 2024)	59.44	40.48	27.66	19.03	19.45	43.71	0.611	12.79
UniBrain (Wang et al., 2024b)	59.08	39.64	26.36	17.68	17.49	43.48	0.482	9.38
MindLLM (Qiu et al., 2025)	61.75	42.84	29.86	21.24	19.54	45.82	0.610	–
Brain-language fusion (Bosch et al., 2025)	62.1	43.7	29.8	20.4	–	46.0	0.659	–
BRAIN-IT VQA (Ours)	**68.11**	**49.30**	**35.08**	**24.81**	**24.82**	**47.97**	**0.683**	**16.00**

Table 2: **Results on VQA-v2 and FSVQA datasets for subject 1.**

Dataset	Metric	OneLLM	UMBRAE	BrainChat	MindBridge	UniBrain	MindLLM	BRAIN-IT VQA (Ours)
VQA-v2	Accuracy ↑	33.68	51.23	40.02	47.91	48.58	52.14	**57.61**
	VQA Acc. ↑	31.44	40.67	36.30	45.95	44.58	48.03	**55.10**
	FSVQA Acc. ↑	21.02	0.00	30.22	40.97	37.87	43.00	**52.00**
	BLEU-1 ↑	37.42	23.11	83.99	86.52	85.10	87.10	**89.15**
	BLEU-2 ↑	31.72	5.86	78.50	82.28	80.01	83.03	**86.14**
FSVQA	BLEU-3 ↑	26.95	2.10	73.00	78.34	75.49	79.27	**83.19**
	BLEU-4 ↑	22.48	1.04	69.73	74.35	70.73	75.50	**80.17**
	METEOR ↑	26.35	8.93	44.76	48.63	46.89	49.05	**52.40**
	CIDEr ↑	0.313	0.004	0.600	0.657	0.629	0.666	**0.722**

CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016), which respectively measure n-gram precision, semantic similarity, recall via longest common subsequence, consensus with human references, and structured semantic content. Following prior works (Wang et al., 2024a; Xia et al., 2024; Qiu et al., 2025), results on VQA-v2, FSVQA, and captioning are reported for subject 1. For NSD-VQA and NSD-VQA-FS, we average across subjects 1, 2, 5 & 7.

5.2 Results

Captioning. Table 1 shows the captioning results for COCO captioning evaluated on subject 1. Brain-IT-VQA achieves state-of-the-art performance across all captioning metrics in our full setting, outperforming all prior methods by a substantial margin. Compared to the strongest prior methods, our model improves BLEU-4 by **+3.57** and METEOR by **+5.28** over MindLLM (Qiu et al., 2025), indicating improved semantic fidelity in the decoded captions.

Visual Question Answering. Table 2 shows the results for VQA for subject 1. Our method achieves the best performance across all benchmarks, improving over the strongest baseline (MindLLM) by **+5.47** accuracy on VQA-v2, with consistent gains also observed on FSVQA across both accuracy and generative metrics. Results on NSD-VQA and its full-sentence variant NSD-VQA-FS are shown in Table 3 (average across subjects). Our model outperforms the previous state-of-the-art (MindLLM) across all metrics. Improvements across both short-form and full-sentence settings demonstrate that our model captures richer and more detailed semantic information from fMRI signals. Additional quantitative results are provided in App. D, including captioning metrics, VQA performance, and per-category breakdowns, all reported per subject, as well as NSD-VQA performance by category against MindLLM and a question-only sanity check. Qualitative examples are shown in Appendix F.

5.3 Decoding Performance by Question Category

We leverage NSD-VQA to analyze which types of visual and semantic information can be reliably decoded from fMRI. By organizing questions into controlled categories, the benchmark enables fine-grained evaluation beyond aggregate accuracy. This setup allows us to examine systematic differences across object recognition, attributes, and relational reasoning. Results averaged across subjects 1, 2, 5, and 7 are shown in Table 4, with per-subject results provided in App. D.3.

We observe a clear dependence on question type. Binary (Y/N) questions consistently achieve high accuracy (typically 83–95%), reflecting the relative simplicity of binary decision tasks and indicating that fMRI signals support robust decoding of coarse object presence and categorical distinctions. In contrast, open-ended questions that require selecting among multiple semantic alternatives are substantially more challenging, with lower performance for categories such as *color* (43.14%), *food* (50.28%) and *action* (58.26%).

Overall, these results suggest that fMRI-based decoding preferentially captures coarse, high-level visual and categorical information, while remaining limited in resolving fine-grained semantic attributes. This pattern is consistent across question categories and highlights a gap between coarse recognition and detailed semantic discrimination. Failure-mode analyses are provided in App. D.5.

Table 3: **Results on NSD-VQA**. *Values are reported as mean ± std across subjects 1,2,5 and 7. NSD-VQA-FS denotes the full-sentence variant.*

Dataset	Metric	MindLLM	BRAIN-IT VQA (Ours)
NSD-VQA	Accuracy ↑	73.43 ± 0.97	**78.23 ± 1.08**
NSD-VQA-FS	FSVQA Acc. ↑	73.83 ± 0.66	**78.18 ± 1.05**
	BLEU-1 ↑	93.18 ± 0.18	**94.42 ± 0.29**
	BLEU-2 ↑	91.21 ± 0.24	**92.80 ± 0.38**
	BLEU-3 ↑	89.29 ± 0.30	**91.22 ± 0.48**
	BLEU-4 ↑	87.03 ± 0.37	**89.42 ± 0.58**
	METEOR ↑	60.13 ± 0.31	**62.33 ± 0.55**
	ROUGE-L ↑	92.41 ± 0.20	**93.74 ± 0.31**
	CIDEr ↑	0.801 ± 0.005	**0.836 ± 0.008**

Table 4: **NSD-VQA accuracy by category** *reported as mean ± std across subjects 1, 2, 5, and 7.*

Category	Acc (%)	Category	Acc (%)	Category	Acc (%)
action	58.26 ± 3.21	electronic Y/N	89.35 ± 3.51	person Y/N	93.36 ± 0.67
animal	64.56 ± 3.57	food	50.28 ± 6.57	scene	93.78 ± 0.88
animal Y/N	91.13 ± 2.29	food Y/N	94.42 ± 0.17	sport Y/N	93.46 ± 1.63
appliance Y/N	94.24 ± 2.56	holding	54.47 ± 4.90	street Y/N	87.86 ± 1.54
building Y/N	82.61 ± 6.15	household Y/N	90.06 ± 0.72	structure Y/N	88.83 ± 1.06
clothing Y/N	94.74 ± 3.04	location	54.82 ± 2.59	vehicle	67.23 ± 4.92
color	43.14 ± 0.91	nature Y/N	83.76 ± 0.38	vehicle Y/N	89.91 ± 1.86
counting	79.44 ± 1.01	objects Y/N	89.21 ± 2.82		

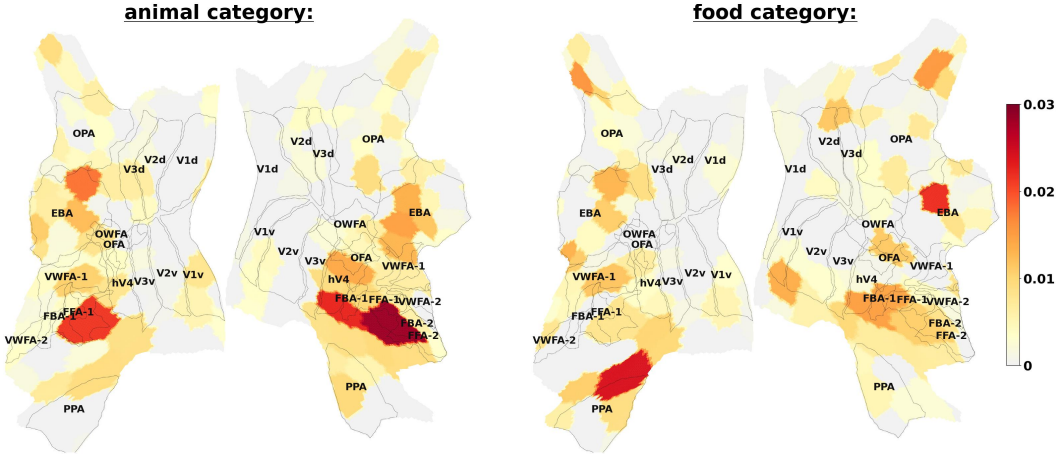


Figure 4: **Visualization of Schaefer-1000 parcel marginal contributions across question categories.** *The marginal contribution of a region refers to the change in decoding performance when that region is added to a coalition (subset) of other regions. Different parcels vary in importance depending on the question type, highlighting how distinct brain regions support different aspects of visual and semantic processing.*

5.4 Ablations

We conduct an ablation study to evaluate the contribution of Q-BIT and of the external data augmentation (Table 5). We consider the following ablations: replacing Q-BIT with the original BIT architecture of Brain-IT (Beliy et al., 2026), designed for image reconstruction, which extracts a single representation per fMRI signal regardless of the question (w/o Q-BIT), and removing external data augmentation from predicted fMRI responses (w/o external data). Both contribute meaningful improvements (**+1.54** and **+2.47** respectively on VQAv2 Acc.), with larger gains on FSVQA Acc. (**+6.17** and **+4.07**), where full-sentence answers are required. Additional ablations are provided in App.E, where we show that even without external data augmentation we outperform all previous methods on VQA, and lead on 7 out of 8 captioning metrics.

6 Decoding Contribution Analysis

Overview: We leverage Brain-IT-VQA to analyze which brain regions encode information relevant to specific types of visual and semantic understanding. We estimate the marginal contribution of each

Table 5: **Ablation study.** Effect of predicted-fMRI data enrichment (Ext. Data) and question conditioning (Q-BIT, vs. the question-independent BIT readout). ✓/× indicate whether each is used.

Ext. Data	Q-BIT	VQAv2 Acc.	FSVQA Acc.
✓	✓	**57.61**	**52.00**
×	✓	55.14	47.93
✓	×	56.07	45.83
×	×	53.75	39.7

brain region to decoding performance across the controlled question categories of NSD-VQA. The marginal contribution of a region refers to the change in decoding performance when that region is added to a coalition (subset) of other regions, compared to the prediction without it. Masking any single region is likely to underestimate its contribution, as other regions can compensate due to the distributed and redundant nature of visual representations in the brain (Haxby et al., 2001; Norman et al., 2006). Instead, we adopt a randomized masking approach that accounts for this redundancy by estimating contributions across many masking configurations.

Technical Details: We estimate the marginal contribution of each brain region to each question category using a masking-based regression approach, inspired by perturbation-based attribution methods, specifically occlusion sensitivity (Zeiler & Fergus, 2014) and local surrogate modeling over perturbed inputs (Ribeiro et al., 2016). Regions are defined by the 1000 parcels of the Schaefer atlas (Schaefer et al., 2018), of which 198 fall within our global voxel mask. At each trial, a random subset of these parcels is masked G.1, and the model generates predictions on 200 stimuli drawn from the NSD test set. This is repeated for 5,000 trials, yielding a dataset of masking configurations and their corresponding per-category VQA accuracy. We then fit a ridge regression model with the binary masking vector over parcels as input and the per-category score as output. The resulting regression coefficients provide an estimate of each parcel's marginal contribution to decoding performance for that category, while controlling for the contributions of all other parcels.

Analysis: Fig. 4 shows the estimated parcel contributions for the *animal* and *food* question categories for subject 1. We observe distinct contribution patterns across the two categories, suggesting that different types of visual and semantic information rely on partially different brain representations. The *animal* category exhibits contributions distributed across lateral and ventral occipitotemporal cortex, peaking near FBA and FFA, with additional contributions around EBA and OFA, consistent with the involvement of body- and form-selective regions in recognizing animate categories. In contrast, contributions for *food* questions are more spatially concentrated in a small number of parcels near PPA and EBA, broadly consistent with recent findings of food-selective representations in ventral visual cortex (Khosla et al., 2022). These results suggest that different question categories engage distinct brain representations, and that the nature of the information, whether localized or distributed, is reflected in the spatial structure of the contributions. This demonstrates the potential of our framework as a tool for probing visual-semantic organization in the brain. Additional visualization results across additional question categories, subjects, and functional ROIs, together with an extended discussion of the contribution patterns, are provided in App. G.

7 Conclusion

We present *Brain-IT-VQA*, a SotA framework for visual Captioning & VQA from fMRI brain recordings, which outperforms previous methods by a large margin. Unlike previous approaches, the extraction of information from the brain signal is conditioned on the question itself, so that different questions draw different information from the same recording. We further introduce *NSD-VQA*, a new extensive automatically-curated benchmark dataset, which provides $\sim$20 question-answer pairs per image, across 20 controlled question categories. This enables more reliable and interpretable evaluation of VQA from fMRI than ever before. Moreover, using this benchmark we can analyze the contributions of different brain regions across question types. We provide initial results for attributing information content to functional brain regions via a masking-based analysis, demonstrating that different regions contribute selectively to different types of visual and semantic understanding. Part of our future work is a systematic evaluation of these attribution results against known functional neuroimaging literature, to assess the extent to which the identified contributions align with established region-function mappings, and whether they discover new ones.

AI USE DISCLOSURE

Generative AI tools were used as part of the research pipeline for constructing the NSD-VQA dataset. Specifically, Qwen3-VL-8B was used to generate structured image annotations, and Qwen3-VL-8B and Gemma-4-31B-it were used to independently verify object counts and presence. Llama-3.2-3B was additionally used to generate the full-sentence variant of the dataset from the structured question-answer pairs. These uses are described in detail in Sec. 4 and App. C. We additionally used generative AI tools to assist with code development and debugging, and to assist with writing, including polishing and improving the clarity and readability of the manuscript. All AI-assisted work was reviewed by the authors, and generated annotations were subject to filtering, cross-model verification, and independent human evaluation. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY

We provide details throughout the main text and appendices to facilitate reproduction, including the model architecture and training procedure (Sec. 3, App. B), NSD-VQA construction and evaluation (Sec. 4, App. C), and additional experimental results and analyses (App. D–E).

REFERENCES

Hossein Adeli, Sun Minni, and Nikolaus Kriegeskorte. Transformer brain encoders explain human high-level visual responses. *Advances in neural information processing systems*, 38:56840–56871, 2025.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering . In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, Los Alamitos, CA, USA, December 2015. IEEE Computer Society. doi: 10.1109/ICCV.2015.279. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.279.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32:6514–6524, 2019.

Roman Beliy, Navve Wasserman, Amit Zalcher, and Michal Irani. The wisdom of a crowd of brains: A universal brain encoder. *arXiv preprint arXiv:2406.12179*, 2024.

Roman Beliy, Amit Zalcher, Jonathan Kogman, Navve Wasserman, and Michal Irani. Brain-it: Image reconstruction from fmri via brain-interaction transformer. In *International Conference on Learning Representations (ICLR)*, 2026.

Victoria Bosch, Daniel Anthes, Adrien Doerig, Sushrut Thorat, Peter König, and Tim Christian Kietzmann. Brain-language fusion enables interactive neural readout and in-silico experimentation, 2025. URL https://arxiv.org/abs/2509.23941.

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. doi: 10.1038/s41597-019-0052-3.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Mindgpt: Interpreting what you see with non-invasive brain recordings, 2023a. URL https://arxiv.org/abs/2309.15729.

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, 2023b.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, et al. Scaling instruction-finetuned language models, 2022.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Curran Associates Inc., 2023.

Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text, 2023. URL https://arxiv.org/abs/2305.11560.

Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Through their eyes: multi-subject brain decoding with simple alignment techniques. *Imaging Neuroscience*, 2:1–21, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team. Gemma: Open models based on gemini research and technology, 2024. URL https://arxiv.org/abs/2403.08295.

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

Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Rongtao Xu, Ke Liu, Liang Hu, and Duoqian Miao. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14247–14255, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, Los Alamitos, CA, USA, July 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.670. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.670.

Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Umut Güçlü and Marcel AJ Van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26584–26595, 2024. doi: 10.1109/CVPR52733.2024.02510.

Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136, 2019.

James V Haxby, M. Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

Wanqiu Huang, Ke Ma, Tingyu Xie, and Hongwei Wang. Brainchat: Interactive semantic information decoding from fmri using large-scale vision-language pretrained models. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889434.

Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision*, pp. 56–73. Springer, 2024.

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.

Meenakshi Khosla, N. Apurva Ratan Murty, and Nancy Kanwisher. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32(19):4159–4171.e9, 2022. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2022.08.009. URL https://www.sciencedirect.com/science/article/pii/S0960982222012866.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

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.

Yunfeng Lin, Jiangbei Li, and Hanjing Wang. Dcnn-gan: Reconstructing realistic image from fmri. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pp. 1–6. IEEE, 2019.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yulong Liu, Yongqiang Ma, Guibo Zhu, Haodong Jing, and Nanning Zheng. See through their minds: Learning transferable brain decoding models from cross-subject fmri. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5730–5738, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Weijian Mai and Zhijun Zhang. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity, 2023. URL https://arxiv.org/abs/2308.07428.

Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstructing natural scenes from fmri patterns using bigbigan. In *2020 International joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

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.

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.

Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in Cognitive Sciences*, 10(9):424–430, 2006.

OpenAI. GPT-4 technical report. Technical report, OpenAI, 2023. arXiv:2303.08774.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Kai Qiao, Jian Chen, Linyuan Wang, Chi Zhang, Li Tong, and Bin Yan. Biggan-based bayesian reconstruction of natural images from human brain activity. *Neuroscience*, 444:92–105, 2020.

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. URL https://openreview.net/forum?id=EiAQrilPYP.

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.

Ziqi Ren, Jie Li, Xuetong Xue, Xin Li, Fan Yang, Zhicheng Jiao, and Xinbo Gao. Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, 228:117602, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

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 Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/scotti24a.html.

Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel AJ Van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018.

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

Vyacheslav Shen, Kassymzhomart Kunanbayev, Donggon Jang, and Daeshik Kim. Interpretable fmri captioning via contrastive learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference, Daejeon, South Korea, September 23–27, 2025, Proceedings, Part XIV*, pp. 295–304, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-032-05184-4. doi: 10.1007/978-3-032-05185-1_29. URL https://doi.org/10.1007/978-3-032-05185-1_29.

Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. The color of the cat is gray: 1 million full-sentences visual question answering (fsvqa), 2016. URL https://arxiv.org/abs/1609.06657.

Ghislain St-Yves and Thomas Naselaris. Generative adversarial networks conditioned on brain activity reconstruct seen images. In *2018 IEEE international conference on systems, man, and cybernetics (SMC)*, pp. 1054–1061. IEEE, 2018.

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. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, 2023b.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation . In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, Los Alamitos, CA, USA, June 2015. IEEE Computer Society. doi: 10.1109/CVPR.2015.7299087. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7299087.

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, 2024a.

Zicheng Wang, Zhen Zhao, Luping Zhou, and Parashkev Nachev. Unibrain: A unified model for cross-subject brain decoding. *arXiv preprint arXiv:2412.19487*, 2024b.

Navve Wasserman, Matias Cosarinsky, Yuval Golbari, Aude Oliva, Antonio Torralba, Tamar Rott Shaham, and Michal Irani. Brainexplore: Large-scale discovery of interpretable visual representations in the human brain, 2025. URL https://arxiv.org/abs/2512.08560.

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.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, pp. 818–833. Springer, 2014.

Chi Zhang, Kai Qiao, Linyuan Wang, Li Tong, Ying Zeng, and Bin Yan. Constraint-free natural image reconstruction from fmri signals based on convolutional neural network. *Frontiers in human neuroscience*, 12:242, 2018.

Appendix

A Limitation

fMRI Assumptions. Following standard practice in the field, our model assumes that fMRI responses are memoryless and replicable. The memoryless assumption implies that prior stimuli do not influence the response to the current image, while replicability assumes that repeated presentations of the same image yield consistent responses. The latter is important for signal averaging, a common practice to mitigate the low SNR of fMRI. These assumptions may not hold in all settings, as they neglect effects such as representational drift over time.

Subject Variability. There is significant variability in signal quality across subjects. The interpretability analysis we present is most reliable for subjects with high SNR, where voxel functionality can be estimated more accurately. For subjects with poor signal quality, the estimated contributions of brain regions may be less reliable.

Interpretability Analysis. The masking-based analysis presented in this work is intended as a demonstration of the framework's potential for brain exploration rather than a definitive functional mapping. The attribution estimates depend on the quality of the underlying VQA model and the coverage of the NSD test set, and should be interpreted with appropriate caution.

Model versus Brain Attribution. Our pipeline relies not only on fMRI measurements, but also on many other components, including pretrained vision-language and language models, synthetic fMRI generated by an image-to-fMRI encoder, and engineering choices in clustering and training. Consequently, strong performance of our method may reflect both information present in the fMRI data and prior information introduced by other components of the pipeline. Our method can therefore be used to generate hypotheses about brain representations, but these hypotheses require targeted experiments to validate. We encourage readers to interpret Sec. G accordingly.

Scope of the Neuroscientific Analysis. This paper is primarily an ML contribution: our goal is a substantially stronger decoding model together with a benchmark that makes fine-grained evaluation possible. We regard building models that can decode brain activity as a necessary step toward understanding how information is represented in the brain. The accompanying analysis is intended to demonstrate what the framework makes possible, not to settle neuroscientific questions. A systematic, quantitative comparison of our attribution maps against established functional maps in the neuroimaging literature, including an assessment of whether any of the recovered patterns are genuinely novel, would constitute a substantial study on its own and is left to future work.

B Model Implementation Details

B.1 Architecture Details

We follow the BIT model proposed in (Beliy et al., 2026), using 128 voxel clusters. BIT consists of a Brain Tokenizer, which maps fMRI activations into 512-dimensional Brain Tokens via a single-head graph attention layer, and a Cross-Transformer Module with 5 self-attention blocks and 6 cross-attention blocks (8 heads each). We modify BIT by increasing the hidden dimensionality from 512 to 768, matching the dimension of the soft prompts passed to the language model. We add 6 masked self-attention blocks (8 heads each), which allow the question to shape the extraction: the 32 learnable query tokens are concatenated with the question tokens, embedded using InstructBLIP's frozen Q-Former embedding layer, and these blocks let the query tokens attend to the question before each cross-attention step over the Brain Tokens. The mask prevents attention to padded positions in the question. The query token outputs are projected to 32 soft prompts of dimension 768.

We use `Salesforce/instructblip-flan-t5-xl` (Dai et al., 2023), which consists of a frozen EVA-CLIP ViT-g/14 image encoder (1408-dimensional features), a Q-Former module, and Flan-T5-XL (Chung et al., 2022) as the frozen language model.

B.2 Training Details

We reserve 10% of the training data for validation, which is used for hyperparameter selection and early stopping.

For training, we use the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 5×10^{-4}. Both stages are trained for 50 epochs, with a warmup period of 15 epochs in stage 1. The learning rate is reduced by a factor of 0.1 on validation plateau with patience 5, and the best checkpoint on validation data is saved.

In stage 2, Q-BIT LoRA uses $r = \alpha = 2$.

B.3 Compute Resources

Brain-IT-VQA is trained on a single H200 GPU (joint-subject model). Stage 1 training completes in approximately 10 hours, and Stage 2 training requires approximately 18 hours, for a total of 28 hours per task/dataset. At inference, processing a single image with 20 queries takes 0.1s, demonstrating the practicality of the model for real-world applications.

C NSD-VQA dataset generation

C.1 Human Evaluation

To assess the quality of the automatically generated question-answer pairs, we conducted an independent human evaluation on a randomly sampled subset of NSD-VQA. Ten independent participants were presented with 50 randomly sampled images and five questions per image, balanced across the different question categories, resulting in 2,500 question-answer evaluations in total. For open-ended questions, participants were asked to select the correct answer from four multiple-choice options, with distractors sampled from other examples within the same semantic category. For every question, participants could additionally select an "N/A" option when they judged that the question was not applicable to the image or, for multiple-choice questions, that none of the provided answers was correct.

Across the evaluated cases, participants agreed with the generated annotations in 89.85% of cases on average, while the "N/A" option was selected in 6.06% of evaluations. The agreement rates across individual question categories are reported in Table S1. Overall, these results indicate that the automatically generated question-answer pairs have high agreement with human judgments, supporting the use of NSD-VQA as a benchmark for fMRI-based visual question answering.

Table S1: Human evaluation accuracy by question category. Accuracy is computed over valid responses, excluding "N/A" responses.

Category	**Accuracy (%)**
Action	85.71
Animal	96.63
Color	74.47
Counting	87.14
Food	95.12
Holding	92.86
Location	89.59
Scene	97.47
Vehicle	91.07
Yes/No	94.75

C.2 Dataset Statistics

Fig. S1 shows the distribution of question-answer pairs across NSD-VQA categories after filtering described in Sec. 4. While NSD-VQA is organized around 20 semantic question categories, several categories are additionally separated into open-ended and binary (Y/N) variants for evaluation (e.g., animal vs. animal Y/N), resulting in the 23 displayed categories shown in Fig. S1.

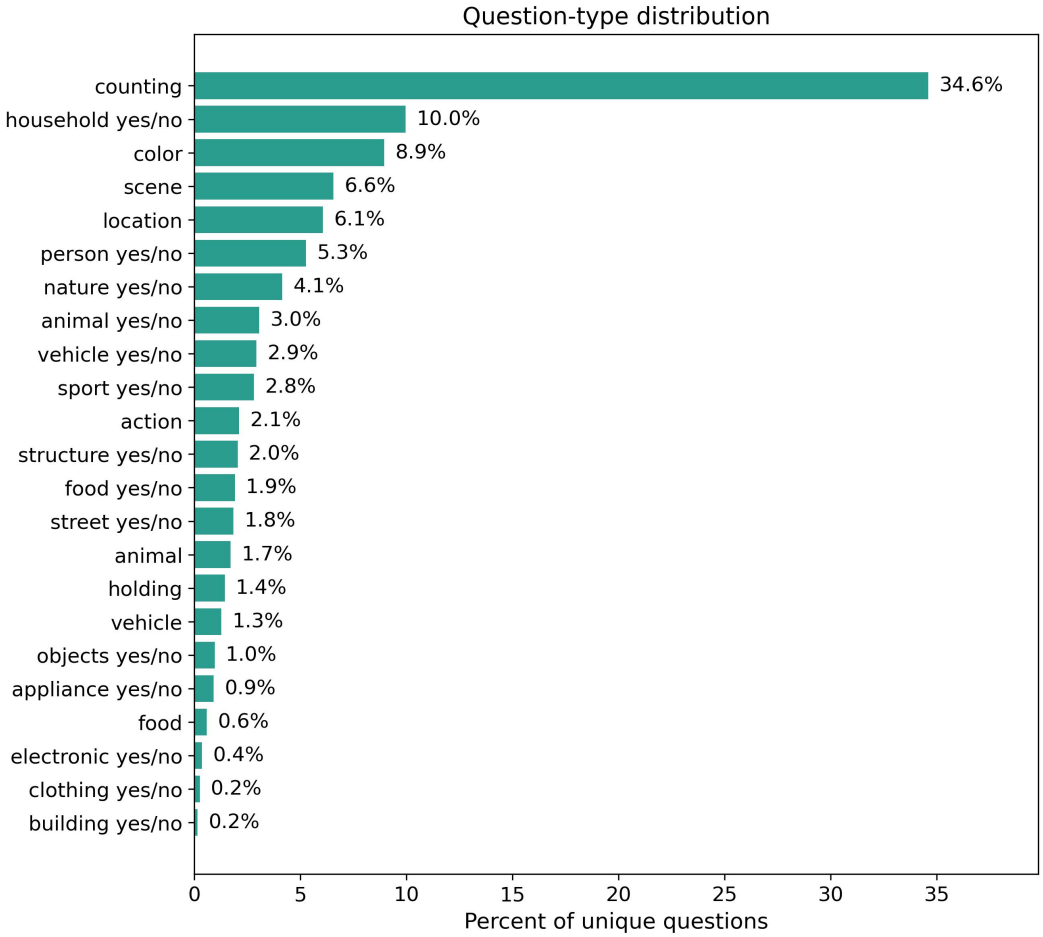


Figure S1: Distribution of question-answer pairs across NSD-VQA categories.

C.3 Question Category Design and Scope

NSD-VQA is not intended to provide an exhaustive taxonomy of neuroscientific questions. The Natural Scenes Dataset (NSD) was designed to capture a broad range of complex natural scenes rather than to probe any particular neuroscientific hypothesis. We therefore selected question categories based on what could be reliably annotated from the images and evaluated at sufficient scale. In particular, we retained distinctions that are explicitly observable in the images and occur frequently enough to support reliable evaluation, using a minimum support of 50 instances and filtering categories with highly imbalanced answer distributions.

The resulting categories cover several complementary aspects of visual understanding, including object identity and semantic category, counting, color and other attributes, actions and interactions, spatial position, and scene-level context. These distinctions are not claimed to be the uniquely most informative questions for neuroscience, nor to exhaust the space of visual processing. Rather, they provide a practical set of controlled dimensions that can be consistently grounded in the available NSD stimuli and decomposed for fine-grained analysis. The same annotation framework also allows the pipeline to construct additional question types when the corresponding visual evidence is present.

This design choice follows the goal of NSD-VQA: to provide a reliable and interpretable benchmark for studying which aspects of visual and semantic information can be decoded from fMRI.

C.4 Compute resources

The NSD-VQA dataset construction pipeline, including annotation and verification using `Qwen3-VL-8B` and `Gemma-4-31B-it`, required approximately 30 GPU hours on a single H200 GPU. The generation of the full-sentence variant (NSD-VQA-FS) using Llama-3.2-3B required an additional 8 GPU hours.

C.5 Dataset Prompts

Figs. S3 and S2 show the prompts used during NSD-VQA construction. The annotation prompt is used to extract structured object- and scene-level information from NSD images using `Qwen3-VL-8B`[2], including object identity, attributes, actions, spatial position, and scene context.

The counting verification prompt is used to validate object counts and object presence consistency during the filtering stage. Counts are independently estimated using both `Qwen3-VL-8B` and `Gemma-4-31B-it`[3], and annotations are retained only when both models agree. Object presence consistency is additionally verified by requiring non-zero predicted counts across both models.

The generated structured annotations are subsequently converted into question-answer pairs using template-based generation rules conditioned on the available attributes and semantic categories.

```
You are given ONE image.

Count the number of CLEARLY visible "{object_name}".

Rules:
- Count only MAIN, clearly visible instances; ignore small/background/unclear ones.
- If "{object_name}" is not a clear main object, output 0.
- If unsure, choose the lower number.

Counting scheme:
   0-3 -> exact number
   4-9 → "few"
   ≥10 → "many"

Answer with ONLY one of: 0 1 2 3 few many
```

Figure S2: Counting verification prompt used for validating object counts during dataset construction.

[2] https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct
[3] https://huggingface.co/google/gemma-4-31B-it

```
You are given ONE image.
Extract a structured representation of the MAIN visible entities and their
attributes.

IMPORTANT RULES:
- Include only MAIN, SALIENT objects (clearly visible).
- Include at least 1 and at most 5 objects.
- Ignore small, occluded, or irrelevant objects.
- Be STRICT and CONSERVATIVE: do NOT guess or infer beyond what is clearly visible.
- If unsure, omit the field.
- All fields must refer only to objects listed in "objects".

OUTPUT FORMAT (STRICT JSON ONLY):
{
  "objects": [
    {"name": "person", "count": "1", "category": "person",
     "position": "foreground","action": "cooking", "holding": "spoon",
     "body_pose": {"legs": "straight", "hands": "holding object"}},

    {"name": “carrot", "count": “3", "category": “food",
     "position": “background", "color": “orange"},
    ...
  ],
  "scene type": "indoor",
  "scene location": "kitchen"
}

ADDITIONAL RULES:
-  Object names: general (e.g., person, car, dog, table).
-  Each type appears EXACTLY ONCE, representing the main visible
   instance in the image.
-  "count": 1–3 → exact; 4–9 → "few"; ≥10 → "many“
-  "color": required for non-person (if visible).
    Use simple terms(e.g., "red", "dark"); omit if unclear.
-  "category": person, animal, vehicle, food, appliance, sport, household, etc
-  "position": foreground → main/large
               background → distant but visible
- "action": only for persons (e.g., cooking, running); otherwise omit
- "holding": include only if clearly visible
-  "body_pose" (if visible):
       legs → straight/bent/extended/resting
       hands → holding object/extended/resting/balancing/in action
-  "scene type": indoor/outdoor
-  "scene location": specific place if identifiable

Return ONLY the JSON.
```

Figure S3: Structured annotation prompt used for extracting object- and scene-level attributes from NSD images.

D ADDITIONAL RESULTS

D.1 CAPTIONING ACROSS SUBJECTS

Table S2: Captioning performance of **Brain-IT-VQA** across subjects (1, 2, 5, and 7).

Metric	S1	S2	S5	S7
CIDEr	0.683	0.646	0.722	0.598
BLEU-1	68.11	66.26	69.11	65.39
BLEU-2	49.30	47.54	50.51	46.00
BLEU-3	35.08	33.48	36.12	32.00
BLEU-4	24.81	23.37	25.71	22.19
ROUGE	47.97	47.10	48.67	46.51
METEOR	24.82	24.30	25.76	23.44
SPICE	16.00	15.23	16.78	14.78

D.2 VQA RESULTS ACROSS SUBJECTS

Table S3: Performance of **Brain-IT-VQA** across multiple datasets and subjects (1, 2, 5, and 7).

Dataset	**Metric**	**S1**	**S2**	**S5**	**S7**
VQA-v2	Accuracy ↑	57.61	57.04	57.71	56.01
FSVQA	VQA Acc. ↑	55.10	54.87	55.30	53.22
	FSVQA Acc. ↑	52.00	51.73	51.80	49.53
	BLEU-1 ↑	89.15	89.15	89.12	88.65
	BLEU-2 ↑	86.14	86.12	86.06	85.50
	BLEU-3 ↑	83.19	83.12	83.08	82.41
	BLEU-4 ↑	80.17	80.03	79.99	79.16
	METEOR ↑	52.40	52.28	52.38	51.69
	CIDEr ↑	0.722	0.720	0.721	0.713
NSD-VQA	Accuracy ↑	78.86	77.69	79.36	77.00
	Acc (per-category) ↑	79.91	78.83	81.05	78.29
	Acc (weighted) ↑	78.84	77.68	79.35	76.99
	Acc (grouped) ↑	67.40	64.35	67.86	62.85
NSD-VQA-FS	FSVQA Acc. ↑	78.31	78.24	79.36	76.80
	BLEU-1 ↑	94.47	94.42	94.75	94.04
	BLEU-2 ↑	92.87	92.79	93.23	92.30
	BLEU-3 ↑	91.30	91.21	91.76	90.61
	BLEU-4 ↑	89.52	89.40	90.09	88.67
	METEOR ↑	62.42	62.33	62.96	61.61
	ROUGE-L ↑	93.79	93.75	94.09	93.34
	CIDEr ↑	0.837	0.836	0.845	0.826

D.3 NSD-VQA RESULTS PER CATEGORY

Table S4 reports NSD-VQA performance by category across subjects 1, 2, 5, and 7. Table S5 compares Brain-IT-VQA against MindLLM on Subject 1 across categories. Statistical significance is evaluated using paired bootstrap testing with 10,000 bootstrap samples.

Table S4: NSD-VQA performance by category across subjects (1, 2, 5 and 7).

Category	S1	S2	S5	S7	Category	S1	S2	S5	S7
action	61.99	56.07	59.81	55.14	household Y/N	90.66	89.94	90.53	89.09
animal	67.82	62.07	67.43	60.92	location	56.93	52.16	57.14	53.03
animal Y/N	92.26	89.68	93.76	88.82	nature Y/N	83.52	83.68	84.31	83.52
appliance Y/N	96.40	91.37	96.40	92.81	objects Y/N	89.04	85.62	92.47	89.73
building Y/N	73.91	86.96	82.61	86.96	person Y/N	93.27	93.52	94.14	92.52
clothing Y/N	92.11	92.11	97.37	97.37	scene	95.00	93.30	93.80	93.00
color	43.34	43.78	43.63	41.80	sport Y/N	95.33	91.36	93.46	93.69
counting	79.67	79.19	80.67	78.25	street Y/N	85.71	88.57	87.86	89.29
electronic Y/N	88.89	94.44	87.04	87.04	structure Y/N	88.75	87.46	90.03	89.07
food	47.78	51.11	58.89	43.33	vehicle	70.98	66.32	70.98	60.62
food Y/N	94.50	94.16	94.50	94.50	vehicle Y/N	89.01	90.58	92.15	87.89
holding	61.01	49.54	55.05	52.29					

Table S5: NSD-VQA performance by category for Subject 1. Mean and standard deviation are reported. Bold indicates statistically significant improvement over MindLLM under paired bootstrap testing ($p < 0.05$).

Category	MindLLM	BRAIN-IT VQA	Category	MindLLM	BRAIN-IT VQA
action	49.53 ± 2.80	$\mathbf{61.99 \pm 2.74}$	household Y/N	85.86 ± 0.89	$\mathbf{90.66 \pm 0.75}$
animal	51.34 ± 3.09	$\mathbf{67.82 \pm 2.90}$	location	41.13 ± 1.61	$\mathbf{56.93 \pm 1.63}$
animal Y/N	87.31 ± 1.54	$\mathbf{92.26 \pm 1.24}$	nature Y/N	80.19 ± 1.58	$\mathbf{83.52 \pm 1.50}$
appliance Y/N	90.65 ± 2.48	$\mathbf{96.40 \pm 1.59}$	objects Y/N	83.56 ± 3.09	89.04 ± 2.60
building Y/N	73.91 ± 9.06	73.91 ± 9.26	person Y/N	90.27 ± 1.04	$\mathbf{93.27 \pm 0.88}$
clothing Y/N	86.84 ± 5.53	92.11 ± 4.33	scene	91.70 ± 0.87	$\mathbf{95.00 \pm 0.69}$
color	41.80 ± 1.33	43.34 ± 1.36	sport Y/N	93.22 ± 1.21	95.33 ± 1.03
counting	75.86 ± 0.59	$\mathbf{79.67 \pm 0.55}$	street Y/N	84.29 ± 2.19	85.71 ± 2.12
electronic Y/N	90.74 ± 3.94	88.89 ± 4.25	structure Y/N	84.89 ± 2.06	88.75 ± 1.80
food	44.44 ± 5.26	47.78 ± 5.28	vehicle	54.40 ± 3.64	$\mathbf{70.98 \pm 3.25}$
food Y/N	86.94 ± 1.96	$\mathbf{94.50 \pm 1.34}$	vehicle Y/N	86.77 ± 1.62	89.01 ± 1.48
holding	46.33 ± 3.38	$\mathbf{61.01 \pm 3.29}$			

D.4 NSD-VQA Question-Only Sanity Check

Table S6 reports a Question-Only Sanity Check obtained by evaluating the model without brain input, conditioning only on the textual question. Some binary (Y/N) categories achieve slightly above-chance performance because the evaluation split is not perfectly balanced between positive and negative answers, despite balancing being enforced at the full-dataset level. This allows the question-only baseline to partially exploit answer-frequency priors.

Table S6: NSD-VQA no fMRI baseline accuracy by category.

Category	Acc (%)	Category	Acc (%)	Category	Acc (%)
action	14.02	electronic Y/N	46.30	person Y/N	54.99
animal	10.34	food	18.89	scene	70.90
animal Y/N	45.81	food Y/N	54.98	sport Y/N	43.93
appliance Y/N	59.71	holding	1.83	street Y/N	52.86
building Y/N	60.87	household Y/N	50.56	structure Y/N	50.80
clothing Y/N	52.63	location	13.53	vehicle	25.39
color	31.92	nature Y/N	45.80	vehicle Y/N	51.12
counting	44.76	objects Y/N	52.74		

D.5 Failure Modes and Systematic Errors

This section expands the failure analysis summarized in Sec. 5.3. All results are pooled across subjects 1, 2 and 5, the subjects with the highest SNR and decoding performance, for which the error structure can be estimated most reliably.

Structured confusions in open-ended categories. When the model is wrong, it is wrong in a structured way: the predicted answer is typically close to the true answer in semantic space. Table S7 lists, for each frequent true answer, the single most likely incorrect prediction and the fraction of instances receiving it. The dominant errors are within-category substitutions of visually and semantically similar objects (dog/cat, cow/horse, truck/car, standing/walking). This indicates that the limiting factor is discrimination among similar alternatives rather than detection, and accounts for the gap between the reliable binary accuracies and the lower open-ended accuracies in Table 4. One interpretation is that similar objects may have similar fMRI representations. However, we cannot, rule out that the model also relies substantially on scene context as well.

Table S7: **Systematic confusions in open-ended categories.** Most likely confusion per true answer, pooled across subjects 1, 2 and 5. n is the number of questions with this true answer; rate is the fraction of those that received the given incorrect prediction.

Category	True answer	n	Most likely error	Rate
Action	standing	180	walking	24%
Action	biking	24	riding	25%
Action	eating	36	sitting	44%
Action	snowboarding	40	skiing	30%
Animal	dog	104	cat	19%
Animal	cat	104	dog	24%
Animal	cow	80	horse	19%
Animal	sheep	56	horse	20%
Food	donut	44	cake	11%
Food	hot dog	16	sandwich	50%
Food	broccoli	28	salad	11%
Vehicle	bus	84	car	20%
Vehicle	truck	36	car	47%
Vehicle	train	148	bus	9%

Counting: Counting reveals that the model captures presence and coarse quantity more reliably than exact enumeration. Performance is high for absence and single instances (88.1% and 84.9%), but drops markedly for *two* (37.6%), *few* (37.6%), and especially *three* instances (18.7%). Errors are predominantly downward, with *two* most often predicted as *one*, while *three* is most often confused with *one* or *two*. Likewise, *few* is frequently collapsed to *one*. Thus, the model appears to recover rough cardinality and presence, but has difficulty distinguishing fine-grained quantities.

Color: Color carries some signal in the fMRI response, but this information is strongly dependent on color type. Neutral colors are relatively recoverable, whereas vivid chromatic colors are not. Errors are also strongly biased toward neutral responses, suggesting that the model captures coarse color information but struggles to reliably distinguish specific chromatic categories. Color is known to be decodable from the visual cortex, but in these scenes it must be bound to a specific object within a complex visual context, which is substantially more difficult than decoding color information in isolation.

E ABLATIONS

Our main results use the predicted-fMRI data augmentation described in Sec. 5.2. We report results without it, and test whether its benefit is specific to the encoder we use by repeating it with two alternatives: a CLIP-based encoder (Radford et al., 2021), obtained by replacing the backbone of the encoder we use with CLIP, and the Transformer brain encoder of Adeli et al. (Adeli et al., 2025). Results are reported in Tables S8 and S9.

Even without augmentation, our model outperforms all prior methods on VQA-v2 and FSVQA, and leads on all captioning metrics but CIDEr, where it is second to Brain-language fusion (Bosch et al., 2025). The augmentation improves performance with all three encoders, one differing in its vision backbone and the other in its architecture. This indicates that image-to-fMRI encoders capture the structure of fMRI responses well enough to generate training data that improves decoding, independently of the particular encoder used.

Table S8: **Effect of the predicted-fMRI augmentation on captioning**(subject 1). "w/o external" removes the predicted-fMRI augmentation entirely; the CLIP and Adeli variants replace the image-to-fMRI encoder used to generate the external data.

Method	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	METEOR ↑	ROUGE ↑	CIDEr ↑	SPICE ↑
MindLLM (Qiu et al., 2025)	61.75	42.84	29.86	21.24	19.54	45.82	0.610	–
Brain-language fusion (Bosch et al., 2025)	62.10	43.70	29.80	20.40	–	46.00	0.659	–
Ours (w/o external)	66.62	47.56	33.12	22.86	23.85	47.03	0.633	14.60
Ours (w/ CLIP enc.)	68.06	**49.35**	**35.13**	**24.88**	24.87	48.16	**0.688**	16.07
Ours (w/ Adeli enc.)	67.72	49.03	34.84	24.45	**25.00**	**48.25**	0.684	**16.42**
Ours (full)	**68.11**	49.30	35.08	24.81	24.82	47.97	0.683	16.00

Table S9: **Effect of the predicted-fMRI augmentation on VQA-v2 and FSVQA accuracy** (subject 1). We report our model trained without the predicted-fMRI augmentation, and with it using three image-to-fMRI encoders: a CLIP-based encoder, the Transformer brain encoder of Adeli et al., and the encoder used in our full model.

Method	VQA-v2 Acc. ↑	FSVQA Acc. ↑
MindLLM (Qiu et al., 2025)	52.14	43.00
Ours (w/o external)	55.14	47.93
Ours (w/ CLIP enc.)	57.59	51.13
Ours (w/ Adeli enc.)	56.72	50.53
Ours (full)	**57.61**	**52.00**

To evaluate whether direct VQA decoding offers an advantage over image-based VQA, we compare Brain-IT VQA against InstructBLIP applied to Brain-IT reconstructed images. As shown in Table S10, Brain-IT VQA outperforms the image-based approach, suggesting that decoding answers directly from brain activity is more effective than first reconstructing the image and then applying a VQA model. Nonetheless, Brain-IT (Images) remains a strong approach, surpassing all previous brain-to-VQA methods. For reference, we also report InstructBLIP applied to the ground truth images, which provides an upper bound on the performance attainable through this pipeline.

Table S10: **VQAv2 Accuracy Comparison.** *Brain-IT VQA vs. InstructBLIP on reconstructed images vs. ground truth images (subject 1).*

Approach	VQAv2 Acc.
Brain-IT VQA	57.61
Brain-IT (Images)	52.79
Ground Truth Images	72.28

F QUALITATIVE RESULTS

Figs. S4 and S5 presents additional qualitative examples generated from fMRI signals using **Brain-IT-VQA**, including both image descriptions and answers to visual questions taken from NSD-VQA-FS.



Figure S4: Additional qualitative results for *Brain-IT-VQA*. The figure includes both generated image descriptions and answers to visual questions decoded directly from fMRI signals. Visual questions are drawn from the NSD-VQA-FS dataset.



Figure S5: Additional qualitative results for *Brain-IT-VQA*. The figure includes both generated image descriptions and answers to visual questions decoded directly from fMRI signals. Visual questions are drawn from the NSD-VQA-FS dataset.

G Decoding Contribution Analysis

G.1 Masking Procedure

Since Q-BIT processes fMRI inputs via a graph neural network, it naturally supports variable numbers of voxels. Masking is implemented by simply excluding the relevant voxels from the input graph. We use the Schaefer-1000 parcellation (Schaefer et al., 2018) for convenience, but any desired parcellation can be applied.

G.2 Additional Brain Contribution Maps: Subject 1

Additional voxel-cluster contribution maps for subject 1 are shown in Figs. S6, S7. They provide contribution patterns across a broader set of question categories than those illustrated in the main paper.

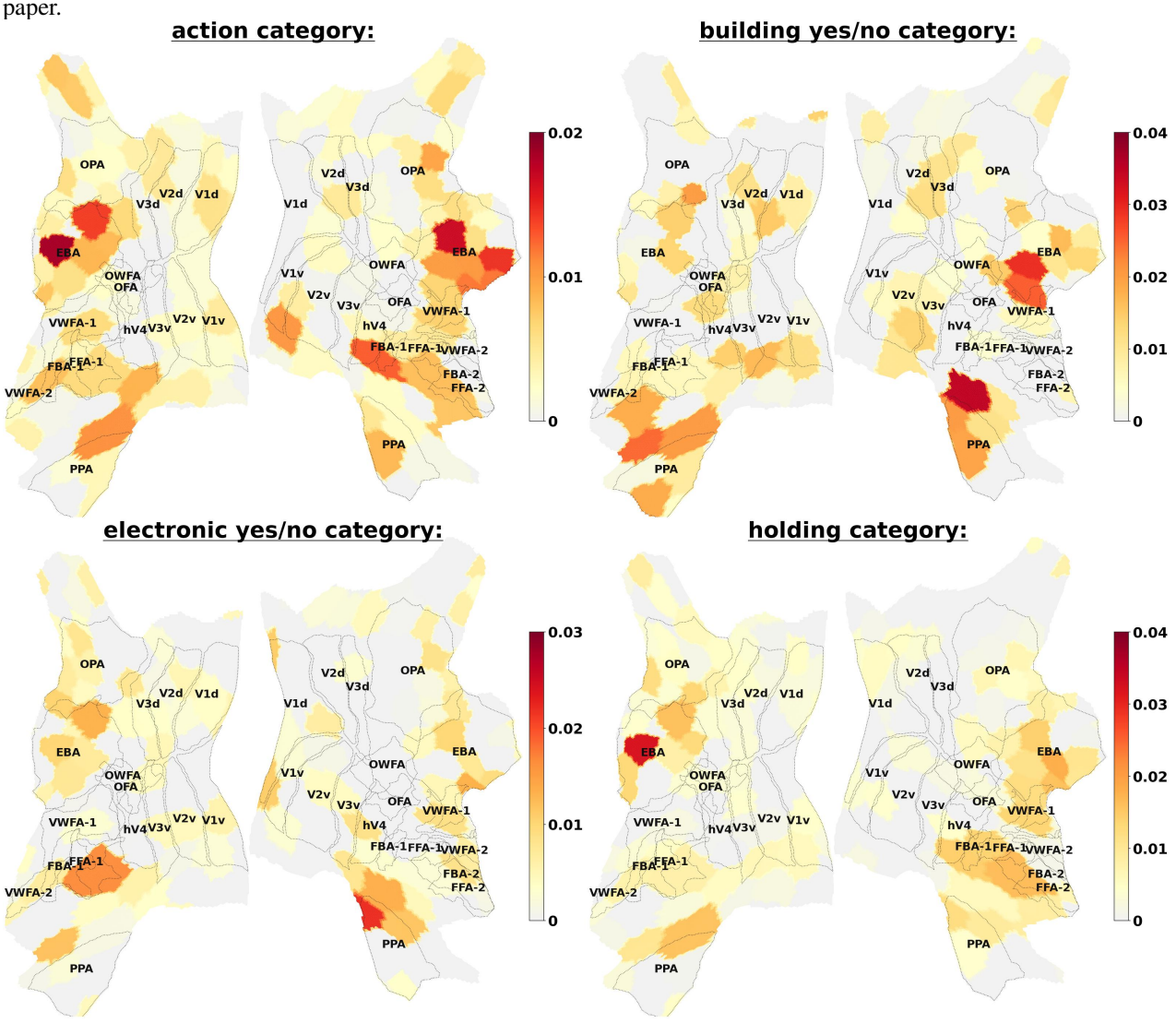


Figure S6: **Visualization of parcel marginal contributions across question categories, subject 1(a)**

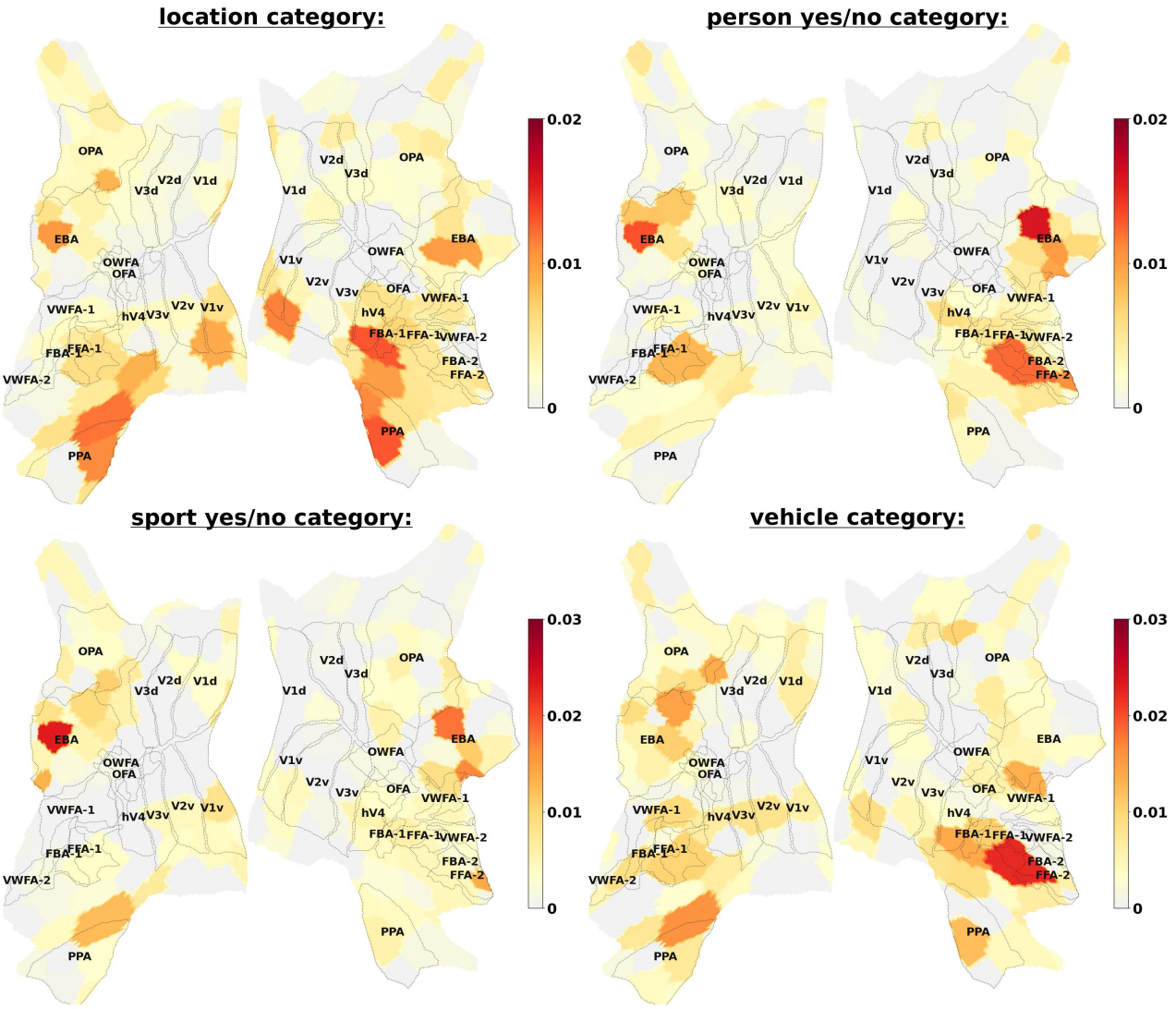


Figure S7: **Visualization of parcel marginal contributions across question categories, subject 1(b)**

G.3 BRAIN CONTRIBUTION MAPS: ADDITIONAL SUBJECTS

Additional contribution maps for subjects 2 and 5 are shown below. The maps exhibit broad consistency in the general location of contributing regions, while precise patterns vary across individuals.

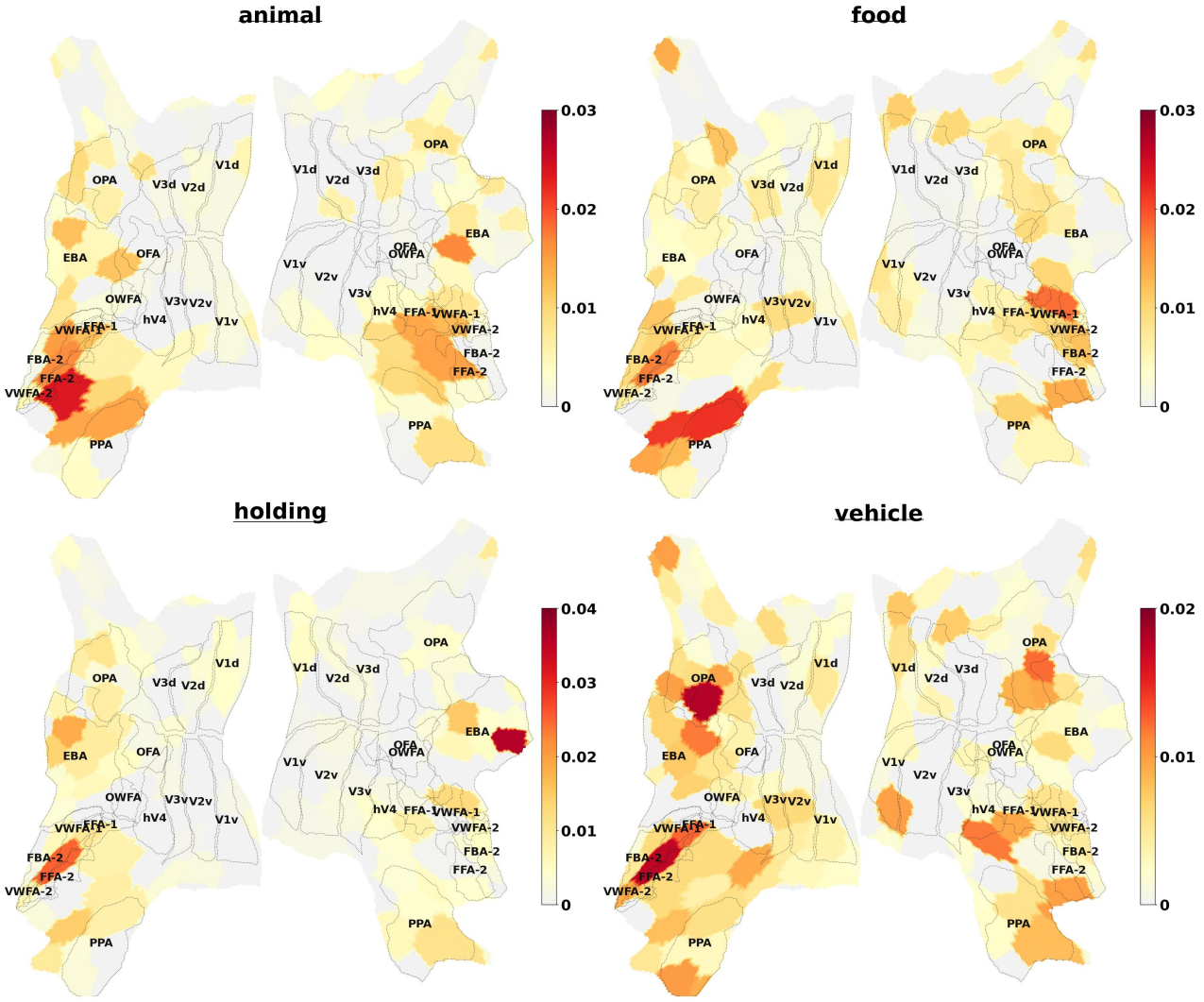


Figure S8: **Visualization of parcel marginal contributions across selected question categories, subject 2**

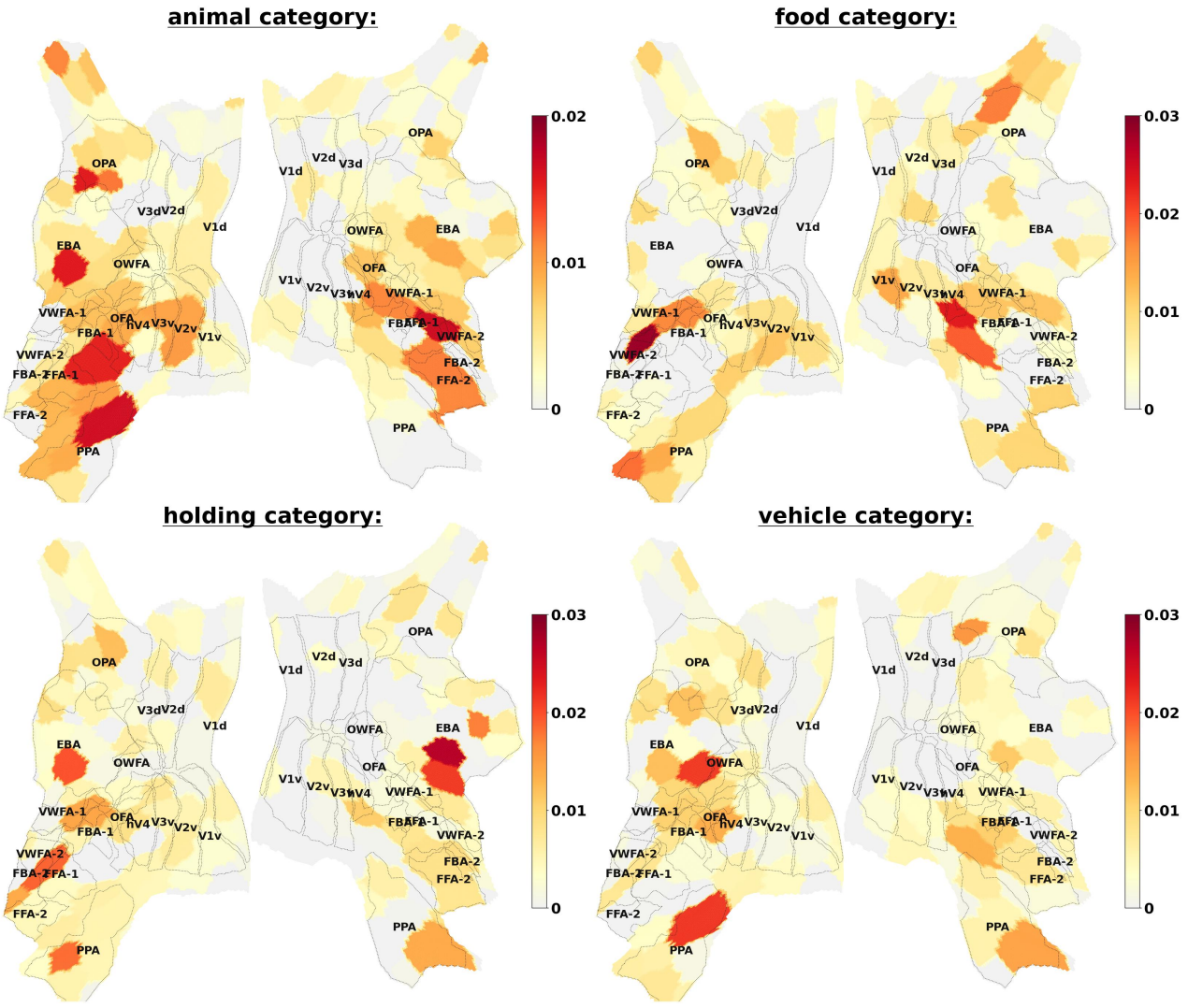


Figure S9: **Visualization of parcel marginal contributions across selected question categories, subject 5**

G.4 Extended Discussion of Contribution Patterns

The contribution maps reveal several recurring patterns across question categories. First, categories differ not only in the location of their contributing functional voxel groups, but also in the spatial distribution of these contributions. Human-object interaction categories such as *holding* tend to exhibit relatively concentrated contributions, whereas semantic categories such as *food* show more distributed patterns across ventral visual regions. Similar variation in spatial distribution is visible across the additional categories shown in App. G.2.

Second, the parcel-level analysis reveals distinctions that can be obscured by coarser anatomical parcellations. In particular, different functional voxel groups within the same anatomical ROI can contribute differently to different question categories; for example, the maps show distinct contribution patterns for *location* and *action* within the broader PPA region. This suggests that aggregating activity at the ROI level can mask finer-grained category-dependent structure.

Third, the qualitative organization of the contribution maps is broadly consistent across subjects, although the precise locations and relative magnitudes of individual contributions vary. This variation is expected given inter-subject differences in functional organization, while the recurring category-dependent patterns provide evidence that the observed structure is not restricted to a single subject (see App. G.3).

These observations are intended as evidence of the exploratory value of the proposed framework rather than as definitive functional localization claims. The contribution estimates depend on the underlying decoding model and on the available NSD stimuli; establishing correspondence with known neuroimaging functional maps, or determining whether any observed patterns are novel, would require a dedicated neuroscientific analysis beyond the scope of this work.

G.5 Functional ROI-Level Contributions

Compared to the cluster-level analysis, ROIs represent a coarser partition of brain activity, with some regions containing thousands of voxels. EBA dominates contributions across most question categories, though this may be partly attributed to its disproportionately large number of voxels rather than functional specificity. Overall, the ROI-level analysis is less informative than the cluster-level results, motivating the use of finer-grained functional parcellations.

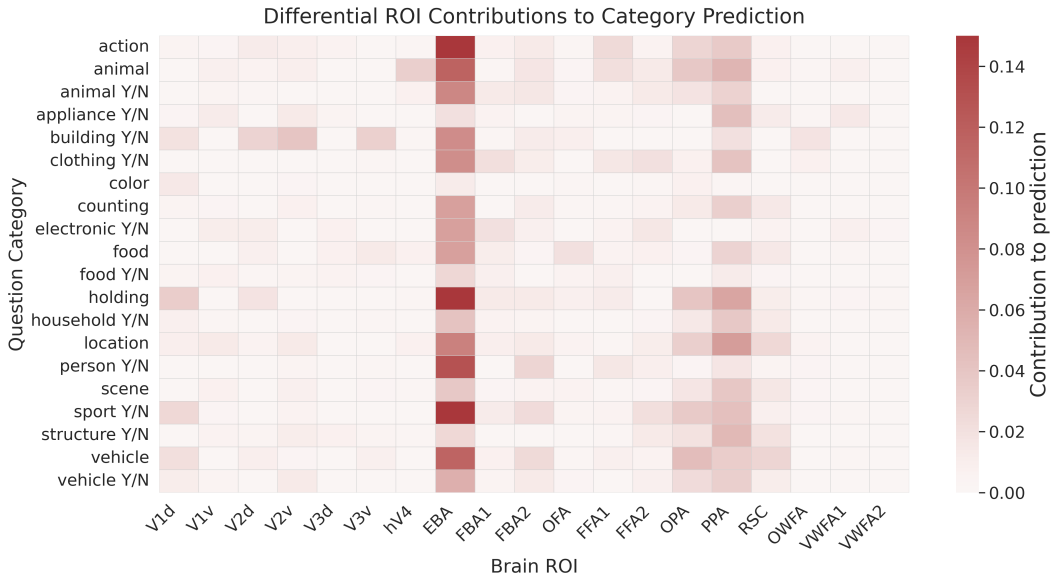


Figure S10: Visualization of voxel-cluster contributions across question categories. Different clusters show varying levels of importance depending on the type of question (e.g., object, attribute, relation), highlighting how distinct brain regions support different aspects of visual and semantic processing.