MAKING SENSE OF SOFT TOKENS: BRAIN CAPTIONING WITH CAUSALLY TESTABLE SEMANTIC CHANNELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain captioning models map fMRI responses to continuous soft tokens that a frozen language model consumes to generate natural language. While effective for producing captions, these brain-derived tokens are primarily shaped to *fit the language model's output space*, without an explicit, text-like semantic interface connecting brain activity to language. Consequently, it remains unclear how semantic information is organized within these tokens, what individual tokens represent, and which representations the language model actually uses. We introduce **MindLens**, a lightweight semantic interface that transforms a subset of LLM-facing soft tokens into named semantic channels. Each channel is assigned a human-interpretable concept group and directly scores its concepts in the LLM's own token-embedding space, requiring only two additional scalar parameters in total and no architectural modification to the language model. By assigning each concept to a single channel, MindLens encourages channels to specialize in coherent yet distinct semantic content, transforming an opaque neural prefix into an explicit and structured semantic interface. Rather than adding information, MindLens *realigns the semantic information already encoded in the fMRI tokens* into interpretable directions that can be directly read and causally tested. On the Natural Scenes Dataset, MindLens achieves a concept mAP of 0.410 without a learned probe, closely approaching the 0.420 obtained by a trained probe from the unstructured backbone, while preserving caption quality (CIDEr: 64.80 vs. 64.21). Targeted interventions further show that individual channels selectively control their associated concepts during generation. These results establish MindLens as an interpretable intermediate layer between neural activity and language, enabling brain captioning to move beyond fitting final language outputs toward explicitly representing, validating, and causally testing the semantic information carried by brain-derived tokens.

1 INTRODUCTION

Decoding brain activity into natural language offers a promising window into how the brain represents the visual world, with brain captioning emerging as a powerful approach for probing these neural representations (Wang et al., 2023; Kozlov, 2025; Qiu et al., 2025; Li et al., 2026). Functional magnetic resonance imaging (fMRI) is the dominant modality for such studies, non-invasively recording the distributed patterns of cortical activity elicited as a subject views a visual stimulus (Allen et al., 2022). The dominant paradigm maps fMRI responses to a sequence of continuous *soft tokens*, which form a **prefix** provided to a pretrained large language model (LLM) alongside a text question or instruction, following prefix-tuning (Li & Liang, 2021) and multimodal prefix methods such as LLaVA (Liu et al., 2023) and BLIP-2 (Li et al., 2023). Yet, progress is typically measured by the quality of the generated caption, leaving a fundamental gap in understanding what information the model actually extracts from the brain. A caption "a person riding a horse" is more than a plausible linguistic description: it implicitly suggests that the brain-derived prefix contains neural information about *person* and *horse*. However, a successful caption alone cannot reveal whether these concepts are explicitly represented in the prefix, where they are encoded, or which information the LLM actually uses to generate the output. The key question, therefore, is not simply whether an LLM can transform brain activity into a plausible caption, but whether the prefix truly encodes *what the brain sees* and whether the LLM actually uses those brain-derived representations to generate the caption.

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.

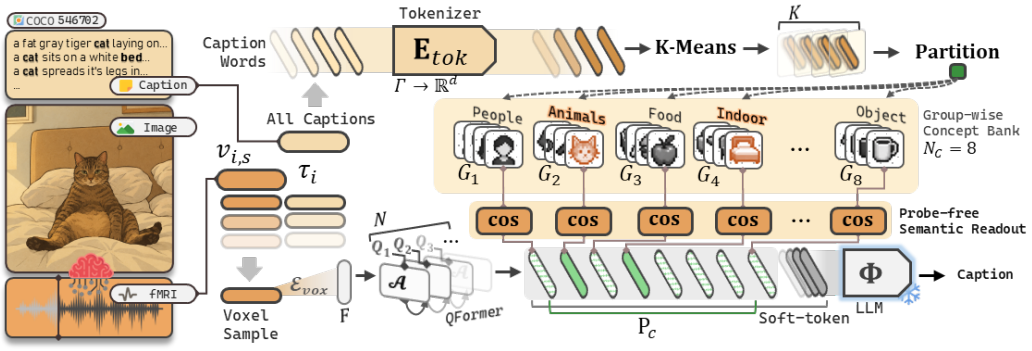

Figure 1: Overview of MindLens, whose first N_c fMRI-derived soft tokens are semantic channels.

This framing exposes a fundamental gap that caption quality alone cannot close. The soft-token prefix is trained primarily through the caption it ultimately produces, so supervision constrains the prefix as a whole without specifying what any individual token represents. Any allocation of semantic content across tokens is equally acceptable as long as it leads to the correct caption. The resulting representation is therefore fundamentally underdetermined: a generated caption cannot be traced back to the neural evidence that supports it. Nothing tells us which part of the prefix encodes *person*, which part encodes *riding*, or whether these concepts are represented at all. This distinction is critical. A model that generates a plausible sentence is not necessarily a model that faithfully reports what the brain encodes. Yet existing brain captioners (Qiu et al., 2025; Chen et al., 2025a; 2026) optimize the final caption while leaving the semantic meaning of the intermediate neural-to-language tokens entirely unconstrained and unexplored. (See Appendix A.1 for a detailed related work.)

A natural next step is to probe the prefix itself and ask: *what information do these fMRI-derived tokens encode before entering the LLM?* Empirical results from our base model reveal an intriguing pattern: tokens from the same input are highly similar, while corresponding tokens across different inputs remain discriminative. Yet, these neural tokens show no clear semantic alignment with the language-token space. This limits our ability to validate and interpret the latent neural representations before language generation. Rather than relying on the LLM's final output as the sole window into neural signals, we aim to directly interrogate these intermediate representations, revealing how semantic information transforms into meaningful concepts. These limitations raise a central question:

> *Can the soft tokens of a brain captioner be organized so that their neural semantic content is directly readable and their contribution to the caption can be causally tested, without degrading the caption or altering the language model?*

To address this challenge, we introduce **MindLens**, a semantic interface that transforms fMRI-derived prefix tokens into interpretable semantic channels as shown in Figure 1. Specifically, we assign a small number of prefix positions as named semantic channels, each linked to a set of related but distinct human-interpretable concepts, such as various types of "food" defined in the LLM's token-embedding space. We encourage each channel to capture coherent yet non-redundant concepts, ensuring both semantic consistency and diversity. This design provides a direct way to examine what information is encoded in fMRI representations, moving beyond the final language output to track, validate, and disentangle semantic information directly from neural signals. We evaluate MindLens on the Natural Scenes Dataset (Allen et al., 2022) and organize our findings around the tokens, not only the caption. To sum up, our contributions are highlighted as follows:

- **Reframing soft token interpretability.** We show that the prefix of a competitive brain captioner is not uninformative, but *unaddressable*: its semantic content can be decoded from individual positions by trained probes, while direct reading recovers substantially less, and a single token captures nearly as much information as the full prefix. This reveals a fundamental limitation of probing. It can expose what is encoded in the prefix, but not what the LLM actually uses. We therefore turn to *intervention* for causal analysis.
- **MindLens as an interpretable semantic interface.** We introduce MindLens, which organizes the latent prefix into named semantic channels with only two additional scalar parameters and no change to the LLM's input format. The resulting channels are directly readable without a probe and approach the probe-based performance of the unstructured

backbone. Instead, MindLens reorganizes existing neural semantics into explicit, interpretable directions that can be directly accessed and analyzed.

- **Causal characterization of brain-derived semantics.** Through targeted interventions, we show that the information used by the LLM is concentrated in the semantic channels. Suppressing a single channel selectively reduces its corresponding concept group in generated captions, an effect not observed for the unstructured prefix. Moreover, channel evidence tracks the category-selective cortical responses associated with the concepts they represent, mirroring the dependence observed in the backbone.

2 METHODS

2.1 PROBLEM FORMULATION

Consider a neural dataset collected from subjects $\mathcal{S}$, where each stimulus is associated with an fMRI voxel response and a textual description. For subject $s \in \mathcal{S}$, let $\mathbf{v}_s \in \mathbb{R}^{L_s}$ denote the voxel response for the stimulus, where L_s varies across subjects, and let $\boldsymbol{\tau} \in \mathcal{T}$ denote its corresponding caption.

As a neural decoding task, **brain captioning** aims to recover the semantic content of a subject's visual experience in natural language. A prevailing approach maps $\mathbf{v}_s$ to N continuous **soft tokens** $\mathbf{P} \in \mathbb{R}^{N \times d}$ that serve as a **prefix** to the prompt of a frozen LLM. These tokens are optimized mainly through the caption they produce, with no objective acting on any individual token. Consequently, no position is assigned a meaning, and a generated caption cannot be traced back to the neural evidence behind it. This motivates a prefix whose semantic content is instead carried by designated positions that can be read directly and tested for their contribution to the caption.

2.2 FRAMEWORK OVERVIEW

MindLens realizes such a prefix by augmenting the brain captioning backbone with a **semantic interface**: a mechanism that designates a small subset of soft tokens as *readable and testable* positions and anchors them to fixed concept references, while leaving the backbone's original LLM input and forward pathway unchanged. The backbone maps a voxel response to a soft-token prefix $\mathbf{P} = [\mathbf{P}_c,\ \mathbf{P}_r]$, which is prepended to the prompt and fed into the LLM to generate a caption. The semantic interface provides a second output, a *probe-free concept readout* $\hat{\mathbf{y}}$, obtained by comparing the designated concept tokens $\mathbf{P}_c$ against the LLM's tokenizer space.

$$\underbrace{\mathbf{v}_s \in \mathbb{R}^{L_s}}_{\text{voxels}} \longrightarrow \underbrace{\mathbf{P} = [\mathbf{P}_c,\ \mathbf{P}_r]}_{\text{soft tokens}} \begin{cases} \left[\mathbf{P};\ \mathcal{T}_{\text{prompt}}\right] \xrightarrow{\text{frozen LLM}} \underbrace{\boldsymbol{\tau} \in \mathcal{T}}_{\text{caption}} \\ \mathbf{P}_c \longrightarrow \underbrace{\hat{\mathbf{y}} \in [0,1]^K}_{\text{concept readout}} \end{cases} \tag{1}$$

The design rests on three principles. First, **Unconstrained Prefix**: without an objective on individual tokens, caption supervision alone leaves the distribution of content across the prefix undetermined, which is the condition the interface removes for the channels. Second, **Language-Native Reference**: the frozen LLM's own embedding space supplies a fixed semantic reference, so the channels can be read without training a separate decoder. Third, **Exclusive Ownership**: assigning each concept to exactly one owner channel breaks the permutation symmetry between channels, giving each a distinct semantic role rather than letting the objective distribute concepts arbitrarily.

2.3 BRAIN CAPTIONING BACKBONE

Our model progressively transforms subject-specific voxel responses into a fixed set of LLM-facing soft tokens. This design addresses three characteristics of fMRI-to-language mapping: subjects have different functional organizations, fMRI responses are high-dimensional and noisy, and the language model requires a compact interface with a fixed embedding dimension.

Subject-specific Projection. Following the shared-subject functional alignment in MindEye2 (Scotti et al., 2024), we first map each subject's voxel response into a common latent space using a subject-specific linear projection:

$$\mathbf{f} = \mathbf{W}_s^\top \mathbf{v}_s + \mathbf{b}_s \in \mathbb{R}^{d_{\text{share}}}, \tag{2}$$

where $\mathbf{v}_s \in \mathbb{R}^{L_s}$ and $\mathbf{W}_s \in \mathbb{R}^{L_s \times d_{\text{share}}}$. This avoids forcing subjects with different voxel counts and functional organizations into a common voxel space, while allowing the projection to learn a shared functional representation directly from the data. In particular, each shared dimension can aggregate responses from functionally related voxels, providing a data-driven basis for forming hypervoxel-like units without requiring anatomical correspondence across subjects. The resulting common representation can then be processed by the same neural encoder for all subjects.

Shared Multi-Head Neural Encoding. Like a convolutional layer that applies many kernels to the same input to extract a diverse bank of feature maps, we do not commit the shared representation to a single high-dimensional vector. Instead, a shared multi-head MLP encoder $\mathcal{E}_{\text{vox}} : \mathbb{R}^{d_{\text{share}}} \to \mathbb{R}^{N_{\text{vox}} \times d_{\text{vox}}}$ passes the single vector $\mathbf{f}$ through parallel heads, each free to emphasize a different feature subspace of the neural signal. This expands $\mathbf{f}$ into a set of richer neural tokens $\mathbf{F}$:

$$\mathbf{F} = \mathcal{E}_{\text{vox}}(\mathbf{f}) \in \mathbb{R}^{N_{\text{vox}} \times d_{\text{vox}}}. \tag{3}$$

Rather than requiring a single vector to carry all neural information, it provides multiple slots from which later queries can selectively extract information. The encoder is shared across subjects, encouraging the resulting neural tokens to remain in a common representational space.

Queryable fMRI Token Extraction. Even after fMRI encoding, fMRI representations remain noisy or redundant (Chen et al., 2023) and carry far more information than is needed to describe the viewed scene. We therefore extract a fixed set of captioning-relevant representations with a querying transformer (QFormer) of N learnable queries, following BLIP-2 (Li et al., 2023). Unlike fixed ViT-style patchifying (Dosovitskiy et al., 2020), the learnable queries adaptively attend to different parts of the neural representation $\mathbf{F}$ under the captioning objective, focusing on the signal useful for language generation while ignoring noise. Each query layer applies self-attention among the queries, cross-attention to $\mathbf{F}$, and a feed-forward network. The refined queries are then linearly projected into the LLM embedding space to form the soft-token prefix $\mathbf{P} \in \mathbb{R}^{N \times d}$ consumed by the frozen LLM:

$$\mathbf{P} = \mathsf{QFormer}(\mathbf{F}) \in \mathbb{R}^{N \times d} \tag{4}$$

Language Decoding. The prefix $\mathbf{P}$ serves as the extracted neural context for a frozen LLM Φ. It is prepended to the fixed prompt "*Scene description:*", and the brain-side modules are trained using caption supervision. At each training step, one caption is randomly sampled from the ground-truth caption pool associated with each stimulus. Given the target token sequence $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$, we optimize the autoregressive captioning loss:

$$\mathcal{L}_{\text{cap}} = -\frac{1}{n} \sum_{j=1}^{n} \log p_\Phi \left(\tau_j \mid \mathbf{P}, \boldsymbol{\tau}_{<j} \right). \tag{5}$$

Since the LLM remains frozen, this objective trains only the preceding neural modules to transform fMRI-derived information into a representation that is compatible with the fixed language space. The resulting soft-token prefix therefore serves as the interface between the neural representation and the LLM, providing the basis for imposing the semantic channels described next.

2.4 Semantic Interface

We propose **MindLens**, a semantic interface design that resolves the learned prefix into semantically readable components: a subset of the soft tokens is given explicit semantic roles and anchored to fixed concept references in the LLM's input space, so that their content can be read directly and tested. MindLens acts as a lens through which the prefix can be examined, providing interpretability and semantic grounding without imposing direct control over the underlying representations.

Concept bank and grouping. From the training captions, we extract key concepts (primarily nouns) and embed each with the LLM's tokenizer $\mathbf{E}_{\text{tok}} : \mathcal{T} \to \mathbb{R}^d$, giving a fixed concept bank $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K) \in \mathbb{R}^{K \times d}$. Because different captions name similar concepts differently (e.g., *dog*, *puppy*, *canine*), we cluster the K concept embeddings with KMeans and manually adjust the clusters into N_c coherent, nameable groups $\mathbf{G} = (G_1, G_2, \ldots, G_{N_c})$ (full procedure in Appendix A.5), and define the fixed ownership map:

$$\pi : \{1, \ldots, K\} \to \{1, \ldots, N_c\}, \qquad \pi(k) = j \ \text{ if } \ \mathbf{c}_k \in G_j, \tag{6}$$

which assigns each concept to exactly one group. The first N_c soft tokens are the **semantic channels** $\mathbf{P}_c = (\mathbf{p}_c^1, \mathbf{p}_c^2, \ldots, \mathbf{p}_c^{N_c})$, with channel j owning group G_j. The remaining N-N_c tokens form the

residual prefix $\mathbf{P}_r$, which stays purely unconstrained and captures information not covered by the groups. Each channel is trained to distinguish the individual concepts within its group rather than merely detect the group's presence: the food channel, for instance, separates *pizza*, *bowl*, and *plate* rather than only signaling that food appears.

Probe-free readout. For the k-th concept, the semantic score comes only from its owner channel $\alpha_k = \cos\left(\mathbf{p}_c^{\pi(k)}, \mathbf{c}_k\right)$ with cosine similarity, and the full readout applies a learnable scale γ and bias β through a sigmoid activation $\sigma(\cdot)$:

$$\hat{\mathbf{y}} = \sigma(\gamma\boldsymbol{\alpha} + \beta), \quad \text{where} \quad \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_K). \tag{7}$$

Because the concept references are fixed in the LLM's input space, $\hat{\mathbf{y}}$ is obtained directly from the channels with no trained probe. It is used for semantic supervision and analysis, while the complete prefix $\mathbf{P}$ is still passed to the frozen LLM unchanged.

Exclusive ownership. Assigning each concept to exactly one channel, rather than to its best-matching channel, is what makes the channels specialize: matching to the best channel is invariant to permuting the channels, so nothing would break the symmetry and the channels would converge onto the same concepts. Fixed ownership breaks this symmetry and gives each channel a distinct role. Supervision nonetheless remains at the individual-concept level, channel j is responsible for the concepts in G_j, while concepts in other groups are left unconstrained by its objective.

2.5 Model Training

The semantic interface is trained jointly with the brain captioning backbone while keeping the LLM frozen. For each stimulus, we construct a K-dimensional multi-label target $\mathbf{y}$ from the ground-truth caption: $y_k = 1$ if concept k is present in the caption and $y_k = 0$ otherwise. Thus, the concept bank defines the K possible labels, while the concept groups determine which semantic channel is responsible for each label. We optimize a binary cross-entropy loss $\mathcal{L}_{\text{cpt}}$ over the K concept labels:

$$\mathcal{L}_{\text{cpt}} = \mathcal{H}_{\text{BCE}}\left(\hat{\mathbf{y}}, \mathbf{y}\right), \tag{8}$$

where $\hat{\mathbf{y}}$ is the predicted semantic readout in equation 7. The total training objective combines caption generation and semantic supervision, with $\lambda > 0$ controlling the strength of the semantic constraint:

$$\mathcal{L} = \mathcal{L}_{\text{cap}} + \lambda\mathcal{L}_{\text{cpt}}. \tag{9}$$

The semantic loss therefore organizes selected soft tokens around predefined concept references without changing the LLM or introducing a separate semantic decoder. Training details, including the initialization of the semantic score scale and bias, are provided in Appendix A.2.

3 Experiments

3.1 Experimental Setup

Dataset. We evaluate our MindLens on the Natural Scenes Dataset (NSD) (Allen et al., 2022) following the standard brain-captioning protocol of prior work. NSD comprises fMRI responses collected while subjects viewed images from MS-COCO, each associated with approximately four human-written captions. Each subject viewed 10,000 images, including 1,000 images shared across subjects and 9,000 subject-specific images. We use *subj01*, *subj02*, *subj05*, and *subj07*, which completed all sessions and have the corresponding full recordings publicly available. A single model is jointly trained across these four subjects, with subject-specific input projections mapping their voxel spaces into a shared neural representation. Following prior protocol (Takagi & Nishimoto, 2023a; Gu et al., 2024; Scotti et al., 2024; Yin et al., 2025), we adopt the 1,000 shared images as the common test set, while the remaining subject-specific data are used for training.

Concepts and channels. We derive K = 132 noun-based visual concepts from the stimulus captions, and partition them by hand into $N_c = 8$ nameable groups ("people," "animals," "food," "indoor," "outdoor," "vehicle," "sport," and "object"); an image is labeled with a concept when any of its human captions mentions it. Each concept is represented by the frozen GPT-2 input embedding of its name, and each of the eight semantic channels is assigned one group. We report concept decoding as mean average precision (mAP) over the 132 concepts (chance 2.89%) and caption quality with the standard n-gram and semantic metrics. Additional details on preprocessing, the concept taxonomy, and model implementation are in Appendix A.2.

Table 1: Brain captioning performance on NSD, evaluated jointly over subj01, subj02, subj05, and subj07. The backbone is our MindLens without the semantic interface constraint.

	BLEU-n Reference				Content & Semantic Agreement			
Methods	**B@1**	**B@2**	**B@3**	**B@4**	**ROUGE**	**METEOR**	**CIDEr**	**SPICE**
BrainCap (Ferrante et al., 2023)	54.82	35.14	21.96	14.04	40.39	16.24	38.86	8.84
Neuro2Lan (Shen et al., 2024)	57.19	37.17	23.78	15.85	36.67	18.60	49.51	12.39
UMBRAE (Xia et al., 2024)	57.38	37.98	25.18	17.11	41.99	18.27	51.68	11.82
MindArt (Chen et al., 2024)	57.26	38.62	25.17	17.71	42.86	21.93	45.95	12.27
MindLLM (Qiu et al., 2025)	56.23	36.60	23.66	15.91	41.56	16.16	39.97	8.82
MindGPT (Chen et al., 2025a)	56.91	36.95	23.41	15.87	38.41	20.04	43.56	10.97
Mind-SA (Chen et al., 2025b)	59.76	39.81	27.05	18.86	42.72	38.08	54.03	12.25
Yo'Mind (Chen et al., 2026)	**61.81**	**42.75**	**29.76**	**20.88**	43.63	38.40	56.07	12.42
Backbone (Ours)	59.62	41.14	26.89	17.48	44.83	40.01	64.21	13.46
MindLens (Ours)	59.56	41.41	27.49	18.05	**45.08**	**40.33**	**64.80**	**13.77**

3.2 Brain Captioning Performance

We position **MindLens** within the classic brain captioning task and first evaluate it numerically against recent brain captioning models. We adopt eight standard metrics: n-gram overlap metrics BLEU-1–4 (Papineni et al., 2002) and ROUGE-L (Lin & Hovy, 2003), and semantic metrics METEOR (Denkowski & Lavie, 2014), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). Table 1 reports results computed jointly over subj01, subj02, subj05, and subj07.

The semantic interface adds readable structure at no cost to caption quality. MindLens designates 8 of the 48 soft tokens as semantic channels so that their content can be read directly and switched off to test its contribution to the caption. These properties motivate the analyses in the following sections. For those analysis to be meaningful, the added structure must not alter the captions themselves, since a mechanism that changed caption quality would confound any later claim about what the channels carry. Table 1 confirms that it does not. The backbone and MindLens share the same architecture and the same LLM input interface, differing only in whether eight tokens additionally receive per-concept supervision, and the two match across all eight metrics, differing by at most 0.59 CIDEr points (64.80 *vs.* 64.21). Caption quality is thus preserved at the backbone's level, and the analyses that follow are carried out on a model that captions no worse than the backbone from which it is built.

The comparison is made on a captioner at the level of the field. This preservation matters only if the underlying captioner is itself competitive, so that the interface's structure is demonstrated on a representative model rather than a weak one. Among the methods in Table 1, the backbone obtains the highest CIDEr (64.21) and SPICE (13.46), while Yo'Mind (Chen et al., 2026) attains the highest n-gram overlap, including BLEU-4 (20.88 *vs.* 17.48). The two orderings track the distinction between the metrics: CIDEr and SPICE reward semantic agreement with the references, whereas BLEU rewards lexical overlap. Both Yo'Mind and Mind-SA (Chen et al., 2025b) use the same GPT-2 backbone as ours, providing a comparable language-modeling foundation for this comparison.

3.3 The Prefix Is Unaddressable; Semantic Channels Make It Readable

We distinguish semantic *presence* from semantic *addressability*. A trained linear probe asks whether semantic information can be recovered from the soft-token prefix after learning an external readout, whereas direct cosine similarity against fixed concept embeddings in the LLM's input space asks whether the same information is readable without such a probe. We evaluate both using $K=132$ visual concepts over the 4,000 test rows from the four subjects (chance 0.029). For each concept, the probe-free score is the cosine between its concept embedding and the soft token at the position that carries it. For MindLens, this is the concept's owner channel. For the unconstrained backbone, which has no ownership map, we use the most permissive readout available: the maximum cosine over the eight designated positions per concept.

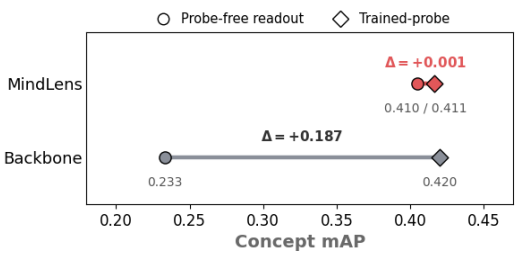

Figure 2: Probe-free readout versus trained probe.

The backbone contains semantic information, but it is not directly addressable. A linear probe trained on the mean-pooled prefix recovers concepts far above chance, achieving 0.420 mAP for the backbone and 0.411 for MindLens. The semantics needed to describe the perceived scene are thus already in the prefix and are not introduced by the interface. Direct cosine readout from the backbone, however, reaches only 0.233 mAP, far below the 0.420 as shown in Figure 2. The information is present but cannot be reached through the LLM's embedding space. It is also unlocalized: a probe trained on any single one of the 48 tokens reaches 0.405 to 0.416 mAP, nearly the full-prefix probe, and across the 132 concepts the most informative token occupies 43 of the 48 positions. The backbone therefore carries the semantics without assigning them any stable, predefined address.

Semantic channels make the same information directly readable. With MindLens, direct cosine readout rises to 0.410 mAP, essentially matching the 0.411 a trained probe recovers. Once the channels are designated, little is gained by learning an external readout or searching over positions, as the content is read directly from its assigned channel. This holds group by group (Figure 3). Because no backbone position owns a group, the fair baseline is the mean readout of that group across all 48 positions, the expected readout when no position is privileged; the owner channel exceeds this baseline in every group, by 0.29 to 0.47 mAP. The content the backbone spreads across positions with none reading it preferentially, the interface concentrates on a single readable channel. The channels are legible in the LLM's vocabulary as well, by a method independent of the concept bank. Decoding each channel to its nearest vocabulary tokens, after centering across images, returns words from the channel's own group and rarely from any other: own-group hit rates range from 0.28 to 0.78 across the eight channels, against at most 0.09 for words of any non-owned group. The modal words per channel are recognizably the group, e.g., *man*, *person*, *woman* for *people*; *dog*, *bird*, *cat* for *animals*; and *surfing*, *skiing*, *tennis* for *sport*. Hit rate and word quality can diverge: the *sport* channel decodes to clean sport terms yet matches the reference captions' content words only 0.08 of the time, because those captions often describe the action indirectly ("a man on a wave" rather than "surfing"). That nearest-token decoding recovers group-appropriate words at all is notable, since continuous prompts are generally not interpretable this way (Khashabi et al., 2022). Here, the channels are readable even under this stricter, vocabulary-level test.

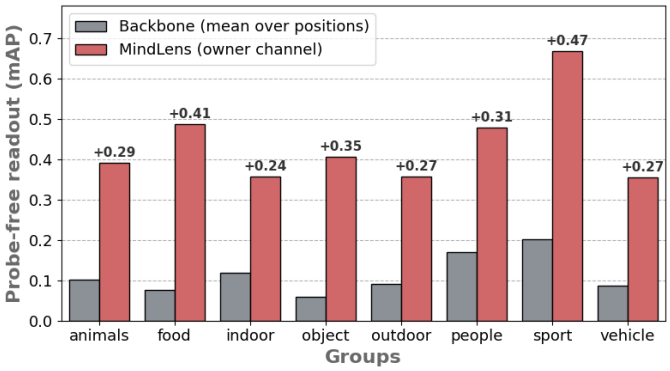

Figure 3: Per-group probe-free readout.

The interface reorganizes semantic information rather than adding it. A probe recovers essentially the same amount from both models, so the interface does not increase the semantic information the prefix contains, and what changes is only whether that information can be read directly. The interface reorganizes content the backbone already encodes into predefined positions legible in the LLM's native embedding space, turning distributed, probe-recoverable semantics into directly readable channels without an added decoder. Whether the language model actually uses these channels during caption generation will be investigated in the next subsection.

3.4 LLM Uses the Channels, and Each Is Necessary for Its Own Concepts

Readability shows the channels can be read; it does not show the language model reads them. A trained probe recovers concepts from every one of the 48 positions at 0.40 to 0.42 mAP in both models, so probing cannot distinguish the two prefixes at all. Only intervention can. We test what the language model uses, first by isolating the channels, then by removing them one at a time.

The channels carry the caption. Keeping only the eight semantic channels, with every other position replaced by its across-image mean, sustains a CIDEr of 49.19, while keeping only the forty residual tokens yields 23.40. At the same eight positions the backbone shows the opposite ordering (41.64 versus 50.86), so the concentration is created by the interface, not by position. To check that these captions describe the specific image rather than scoring through generic text, we report image-specific CIDEr: the CIDEr against the image's own references minus the mean CIDEr against other random

images' references, which is near zero for a generic caption and large only for one tied to its image. The channels reach 43.63, against 1.02 to 8.15 for eight random residual tokens.

Removing a single channel selectively removes its own concepts. We encode each test image into its 48 soft tokens, replace one channel with its mean over the images that do not contain the concept, and regenerate the caption. Every other position and the prefix length are held fixed, so the only change is that this position no longer carries the concept's information. For each caption we detect which of the 132 concepts it mentions by matching the member words that define each concept against the generated text; validated against the human labels, this detector reaches a concept F1 of 96.97%, so the mention counts are a reliable check (detailed in Appendix A.3). For each removed channel, we measure the change in every group's recall, i.e., the rate at which a group is mentioned given that the image contains it. This yields an 8×8 matrix whose diagonal is the effect on the channel's own group and whose off-diagonal is its *spillover* onto the other groups (Figure 4).

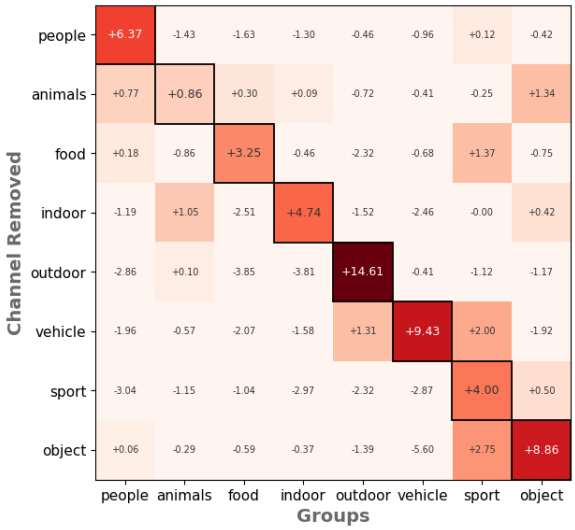

Figure 4: Reduction in concept group recall %. Positive values indicate lower occurrence of concepts in the generated captions following channel removal.

On the diagonal, the reductions are largest for the most visually and lexically distinct groups, reaching 14.61 points for *outdoor*, 9.43 for *vehicle*, and 8.86 for *object*, and smallest for *animals* (0.86 points), the most frequently mentioned group; for the remaining seven channels, the largest reduction in each row falls on the channel's own group, while off-diagonal entries are uniformly small. Across all eight channels, removing a channel reduces its own group's recall by 6.52 points on average, compared with a mean spillover of 1.35 points onto the other groups; removing a residual position instead reduces recall by only 0.55 points. To test whether this diagonal structure could arise by chance, we use the *selectivity* of the removals as our statistic: the fraction of the 56 (channel, other-group) pairs for which a channel's removal reduces its own group's recall more than that of the other group. The observed selectivity is 0.982 (55 of 56 pairs). Under the null hypothesis that channels are not tied to particular groups, the assignment of channels to groups is arbitrary, so we recompute selectivity under each of the $8! = 40{,}320$ permutations of that assignment; only the true assignment attains the observed value, giving an exact $p \approx 2.48 \times 10^{-5}$ against a mean of 0.49 across permutations. Applying the identical procedure to the backbone's first eight positions, with the same ownership map imposed after the fact, yields a selectivity of 0.268 ($p \approx 0.99$), with own-group reduction no larger than spillover, so the selective structure is produced by the semantic interface, not inherited from token position or model architecture.

Figure 5 shows this effect on individual captions. Removing the *outdoor* channel from the surfing image drops *ocean* while the surfboard and rider remain, and removing the *people* channel turns the surfer into "a bird" with the scene otherwise preserved; in the second image, removing *animals* eliminates *cat* and removing *people* drops the woman. These changes are selective but not universal: a single channel's removal alters its group's word on 1.6% to 28.3% of eligible images depending on the group, because the residual prefix still carries much of the scene. This is also why the per-image effect is modest while removing all eight channels collapses the concepts entirely.

3.5 Semantic Channels Recover Category-Selective Neural Grounding

We next ask whether the semantic channels are grounded in the neural responses that give rise to the captions, rather than merely aligned with semantic labels by the training objective. We first identify region-group correspondences from the recorded neural responses alone, then test whether perturbing the corresponding cortical regions selectively disrupts their paired channels, and finally examine whether this disruption propagates to generated language.

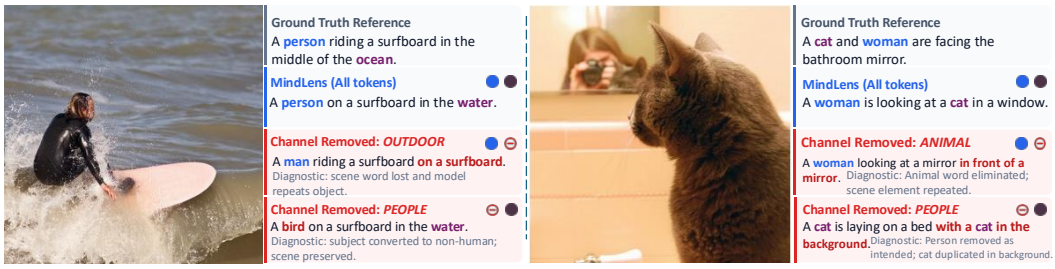

Figure 5: Qualitative effect of removing a single channel. For two test images, we show the ground-truth reference, the caption from the full MindLens prefix (all tokens), and captions after mean-replacing one channel. Removing the owner channel drops its group's word from the caption.

Category-selective cortex identifies the semantic correspondences. For each category-selective region–faces, bodies, places, words (Stigliani et al., 2015), and food (Jain et al., 2023)–and each concept group, we compute Cohen's d between the region's mean response to images containing the group and those that do not. Each region is assigned to the group with the largest d, retaining only assignments with $d > 0.3$ consistently across all four subjects. Four region-group pairs satisfy this criterion: *faces*↔*people* ($d = 0.91$), *bodies*↔*sport* ($d = 1.54$), *places*↔*indoor* ($d = 0.70$), and *food*↔*food* ($d = 0.86$). The words region has no corresponding group because our concept taxonomy contains no text category. Importantly, these correspondences are not simply imposed by the expected anatomical labels: bodies prefer *sport* over *people*, while places prefer *indoor* over *outdoor*.

Perturbing cortex selectively disrupts the corresponding channels. For each category-selective region, we mean-replace its voxels in all four subjects and re-encode the 4,000 test rows. For each of the four predicted region-group pairs, we compare the drop in the paired channel's own-group concept mAP with the mean drop of the other seven channels at the same region. Averaging these four within-region contrasts gives an excess drop of 8.4% mAP. We test whether this correspondence could arise from arbitrary channel assignments by exhaustively enumerating all 1,680 ways to assign four of the eight channels to the four regions. It turns out only **three** reach the observed effect ($p \approx 1.79 \times 10^{-3}$), and the median assignment gives essentially zero, showing that the region-channel correspondence does not arise from arbitrary channel assignments. Imposing the same group-to-position map on the backbone's first eight positions after the fact, where nothing was ever assigned, and repeating the perturbation gives 8.6 points. The cortical dependence is therefore already present in the neural representation rather than introduced by the interface.

The neural dependence propagates to generated language. The cortical dependence is not confined to the semantic readout. For images containing a given concept group, ablating its paired category-selective region reduces the probability that the corresponding group is mentioned in the generated caption. Across the four region-group pairs, the paired groups show an additional 10.2% drop in mention rate relative to the other groups at the same region ($p \approx 3.57 \times 10^{-3}$). The neural perturbation propagates from category-selective cortex through the semantic channel and into generated language. MindLens therefore **does not** create the cortical grounding. It **exposes** brain-derived semantic information at explicit channels whose neural origin can be tested through targeted perturbation.

4 Conclusion

We revisited the soft-token prefix that brain captioners feed to a frozen language model and argued that its central limitation is not a lack of information but a lack of addressability: a trained probe recovers concept content from almost any position, yet the content cannot be read directly, and its contribution to the caption cannot be tested. We introduced **MindLens**, a semantic interface that designates a few prefix positions as named semantic channels, each tied to a human-interpretable concept group through the language model's own token embeddings, at a cost of two scalar parameters and no change to the language model's input. The interface leaves caption quality intact and adds no semantic information: a probe recovers the same amount from the constrained and unconstrained models, so the channels expose the neural semantics the backbone already encodes rather than creating them. What the interface changes is that this content becomes directly readable, that the information the language model relies on is concentrated in the channels, and that removing a single channel selectively suppresses its own concept group in the generated caption.

Reproducibility Statement

A complete public repository containing the full source code, trained model weights, and scripts to replicate all experiments will be released upon acceptance of this paper.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments and to assist with implementing methods. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing proofs, propose or refine hypotheses, design the research methodology, assist with translation, clean or reformat datasets, support qualitative or thematic data analysis, or interpret research results. We reviewed and verified all AI-assisted work, including checking methodological suggestions against the relevant literature and verifying and testing AI-assisted implementations for correctness. We take responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with the aid of generative AI.

References

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

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pp. 382–398. Springer, 2016.

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

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

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

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

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

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pp. 376–380, 2014.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

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

Zijin Gu, Keith Jamison, Amy Kuceyeski, and Mert R. Sabuncu. Decoding natural image stimuli from fmri data with a surface-based convolutional network. In Ipek Oguz, Jack Noble, Xiaoxiao Li, Martin Styner, Christian Baumgartner, Mirabela Rusu, Tobias Heinmann, Despina Kontos, Bennett Landman, and Benoit Dawant (eds.), *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pp. 107–118. PMLR, 10–12 Jul 2024. URL https://proceedings.mlr.press/v227/gu24a.html.

Nidhi Jain, Aria Wang, Margaret M Henderson, Ruogu Lin, Jacob S Prince, Michael J Tarr, and Leila Wehbe. Selectivity for food in human ventral visual cortex. *Communications Biology*, 6(1):175, 2023.

Daniel Khashabi, Xinxi Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, et al. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3631–3643, 2022.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. Pmlr, 2020.

Max Kozlov. Ai decodes brain activity to turn thoughts into text. *Nature*, 647:297, 2025.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3045–3059, 2021.

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.

Shurui Li, Zheyu Jin, Shi Gu, Ru-Yuan Zhang, and Yuanning Li. A large-scale fmri dataset for vision-language semantic association. *Scientific Data*, 2026.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 4582–4597, 2021.

Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pp. 150–157, 2003.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

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.

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.

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *Forty-first International Conference on Machine Learning*, 2024.

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

Anthony Stigliani, Kevin S Weiner, and Kalanit Grill-Spector. Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, 35(36):12412–12424, 2015.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14453–14463, 2023a.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, 2023.

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.

Zixiang Yin, Jiarui Li, and Zhengming Ding. i mind: Insightful multi-subject invariant neural decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

A Appendix

A.1 Related Work

Brain-to-Language Decoding Decoding language from brain activity has progressed from classifying or retrieving among fixed candidates to generating open-ended text. In the visual setting, the dominant recipe pairs fMRI responses with the captions of the viewed images and learns a mapping into the input space of a frozen language model, so that neural activity is rendered as a soft-token prefix the model reads (Ferrante et al., 2023; Xia et al., 2024; Qiu et al., 2025; Chen et al., 2025a). These systems differ mainly in how the neural signal is encoded and aligned: some regress toward image features or CLIP embeddings before decoding (Takagi & Nishimoto, 2023b; Scotti et al., 2024), others train querying modules that extract a fixed set of language-relevant tokens in the style of multimodal prefixes (Mokady et al., 2021; Li et al., 2023), and recent work reports steadily improving caption quality on the Natural Scenes Dataset (Allen et al., 2022; Chen et al., 2024; 2025b; 2026). A related line structures the *input* side of this mapping using neuroanatomy, for example by biasing attention with region-of-interest information so that anatomically meaningful voxels drive the prefix (Qiu et al., 2025). Our work is orthogonal to these advances in two respects. First, all of these methods are evaluated by the caption they produce and treat the prefix as an intermediate whose token-level content is never examined. We instead evaluate the tokens themselves. Second, where anatomy-based methods structure the input side of the prefix, we structure its *output* side by language, tying a few positions to concept directions the frozen model already reads. We treat caption quality as a precondition rather than a contribution, and position our backbone as competitive with recent methods rather than as a new state of the art.

Soft Prompts and the Interpretability of Continuous Prefixes Continuous prompts and prefixes are a standard way to steer a frozen language model without updating its weights (Li & Liang, 2021; Lester et al., 2021; Mokady et al., 2021; Li et al., 2023). Their effectiveness, however, has not translated into interpretability. A learned continuous prompt can be highly effective for a task while its nearest discrete-token projection is unrelated or even contradictory to that task, a phenomenon termed prompt waywardness (Khashabi et al., 2022). Consequently, mapping a soft prefix back to vocabulary tokens is not a reliable way to read it, and the semantic content of such prefixes is typically inferred only indirectly, if at all. This is the difficulty our setting inherits: the brain captioning prefix is accepted by the language model yet resists reading as words. Rather than forcing tokens to correspond to words, our semantic interface ties a small number of positions to fixed concept directions in the model's own embedding space, so that those positions can be read directly. In spirit this connects to concept bottleneck models (Koh et al., 2020), which route predictions through human-interpretable concepts, but the interface is not a bottleneck: the residual prefix remains, generation does not pass through the concept predictions, and the language model's input is left unchanged.

A.2 Implementation Details

Data. We use the betas averaged over the three repetitions of NSD subj01, subj02, subj05 and subj07, restricted to each subject's `nsdgeneral` mask, which gives $15{,}724$, $14{,}278$, $13{,}039$ and $12{,}682$ voxels respectively. Because voxel counts differ, each subject keeps its own input projection and no anatomical correspondence is imposed. The $1{,}000$ images shared by all four subjects form the test split and the $9{,}000$ subject-specific images of each subject form the training split, giving $36{,}000$ training rows and $4{,}000$ test rows, where one row is one subject's response to one image. Every number reported in this paper is computed on all $4{,}000$ test rows. No analysis uses a subsample.

Reference captions. Our copy of the COCO annotations for the NSD stimuli provides four reference captions for $72{,}809$ of the $73{,}000$ images, five for 189 and six for 2. All available references are used, both for caption metrics and for the concept labels, and a concept is labeled present for an image when any of its references mentions it.

Architecture. The subject projection maps $\mathbf{v}_s$ to $d_{\text{share}} = 2048$. The shared encoder is a 3-layer MLP with hidden width 4096, GELU activations, layer normalization, and dropout 0.1, producing $N_{\text{vox}} = 64$ neural tokens of width $d_{\text{vox}} = 512$. The querying transformer has $N = 48$ learnable queries, 4 pre-norm layers, width $d_q = 512$, and 8 attention heads. Each layer applies query self-attention, cross-attention to the neural tokens, and a feed-forward block. A linear map lifts the refined

Table 2: Trainable parameters. The language model is frozen and excluded.

Component	Parameters
Subject-specific projections (4 subjects)	114.13M
Shared multi-head encoder	159.44M
Querying transformer	16.84M
Projection to the LLM width	0.39M
Semantic interface (γ, β)	2
Total trainable	290.81M
Frozen GPT-2	124.44M

queries to the LLM width $d = 768$. The language model is a frozen GPT-2 (124.44M parameters), prompted with "*Scene description:*". Table 2 lists the trainable parameter budget.

Optimization. All runs use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.01), a peak learning rate of 2×10^{-4} with 3 warmup epochs followed by cosine decay to 1×10^{-6}, batch size 128, gradient clipping at 1.0, mixed precision, 20 epochs and seed 2026. One caption is sampled at random from the image's references at each training step. A run takes about 40 minutes on a single NVIDIA RTX 6000 Ada GPU. The backbone and MindLens differ only in the presence of $\mathcal{L}_{\text{cpt}}$. Every other setting, including the seed, is identical, so the two models are directly comparable.

Semantic interface parameters. The interface adds exactly two trainable scalars for the whole model, a scale and a bias shared by all channels and all concepts. The scale is parameterized as $\gamma = \min(e^{\theta}, 100)$ and initialized at $\gamma = 10$. The bias is initialized at $\beta = \text{logit}(0.028) = -3.55$, the marginal rate at which a concept is present, so that the readout starts calibrated to the base rate. After training, $\gamma = 16.86$ and $\beta = -3.56$. The concept loss is a per-concept binary cross-entropy over all $K = 132$ concepts with a positive class weight of 10, which compensates for that base rate, and it enters the total objective with $\lambda = 1$. Concept supervision is therefore 132 independent binary terms; the grouping only determines which channel supplies the score for each term.

Interventions. All interventions replace tokens or voxels with a constant and never delete them, which keeps the prefix length and every position index unchanged. Deleting a token would shift the language model's positional embeddings and confound content with geometry. For the readout analysis, a prefix position is set to its mean across the test images. For the single-channel removals in Section 3.4, a channel is set to that position's mean over the images that do **not** contain its group, the counterfactual of the group being absent, and the residual-token control uses the identical construction at a non-channel position. The replacement vector stays on the manifold the model expects: it retains 0.95 of a typical token's norm and a cosine of 0.92 to the token it replaces. For the cortical ablations, a region's voxels are set to that subject's mean over its test rows.

A.3 Concept Detector

Every caption-level result rests on deciding which of the $K = 132$ concepts a generated caption mentions. We do this by string matching rather than by a learned classifier, so that the measurement is independent of any model being evaluated, and we validate the rule against the ground-truth labels before using it.

Matching rule. Each concept is defined by a list of member words, the same lists from which the ground-truth labels were built (for *person*: people, person, player). A caption is lowercased and stripped of punctuation, and matching proceeds in three stages. Multi-word members are matched first, longest first, and claim their tokens, so that "hot dog" fires *hot dog* and not *dog*, and "traffic light" does not fire *traffic*. Each remaining token is then matched against the member words in surface form, so that "glasses" stays eyewear rather than becoming *glass*. Tokens that still do not match are reduced to their noun lemma with the WordNet lemmatiser, which recovers "dishes" as *dish*, and a small table supplies the irregular plurals WordNet misses (men, women, children, feet, teeth, mice, geese, knives, wolves, leaves). A concept group counts as mentioned when any of its concepts fires.

Validation. We apply the identical rule to the human reference captions of the $1{,}000$ unique test images, using the first row of each image so that no image is counted four times, and mark a concept

Table 3: Concept detector against the ground-truth labels of the $1{,}000$ test images, micro-averaged over all $1{,}000 \times 132$ image-concept cells. A concept is counted as present when the detector fires in at least n of that image's reference captions; the first row is the rule that defines the labels.

References required	Precision	Recall	F1
At least one (the labeling rule)	94.97	99.05	**96.97**
At least two	98.70	59.62	74.34
At least three	99.31	37.57	54.51

Table 4: Per-group F1 of the detector under the labeling rule, against the same ground truth. A group counts as mentioned when any of its concepts fires.

	people	animals	food	indoor	outdoor	vehicle	sport	object
F1	98.82	99.24	91.55	99.45	98.83	98.64	99.25	96.69

present when it fires in at least one reference. Comparing these $1{,}000 \times 132$ decisions against the stored ground-truth labels, micro-averaged over all image-concept cells, gives precision 94.97, recall 99.05 and F1 96.97 (Table 3). Requiring agreement from two or three references raises precision above 98 but collapses recall, confirming that the labels follow the "mentioned in at least one reference" rule. Per group, F1 ranges from 91.55 for *food*, where container and compound words are the main source of disagreement, to 99.45 for *indoor*.

Why member words rather than concept names. An earlier version of this detector matched only the concept *names*, which are cluster labels rather than the words people actually write. That rule missed "people", "player", "boy" and "girl" for *people* and the corresponding surface forms elsewhere, reaching a group-level F1 of only 82.32 for *people* and 82.08 for *sport* against the same ground truth. All caption-level numbers in this paper are computed with the member-word detector described here.

Recall. Throughout the paper, a group's recall is $P(\text{group mentioned in the generated caption} \mid \text{group labeled present for that image})$, where presence comes from the ground-truth labels and the mention comes from the detector applied to the generated text. Conditioning on presence removes the dilution from images where the group cannot appear, and the model's own scores are never consulted, so an intervention has to change the text to change the measurement.

A.4 Ablations

We ask three questions of the design: whether exclusive ownership is what makes the channels specialize, whether the conclusions depend on the particular taxonomy and prefix width, and whether the causal result depends on choices made in the analysis rather than in the model.

Exclusive ownership is what prevents collapse. Scoring each concept against its best-matching channel rather than its owner makes the objective invariant to permuting the channels, so nothing distinguishes one channel from another and duplicating a working channel is a valid solution. We observed exactly this in an earlier configuration of the model: with max pooling the anchored tokens became near-duplicates of one another, the largest pairwise correlation between them reaching 0.9999 against 0.534 under ownership, only 3.43 of 8 tokens claimed any concept against 6.68 of 8, and caption quality was lower (58.89 against 61.17 CIDEr, with the backbone of that configuration at 64.36). Those runs used a longer schedule and an earlier concept set, so we report them here rather than as a row of Table 5, and we did not repeat the comparison at the current setting.

Group count and partition. Table 5 sweeps $N_c \in \{6, 8, 10\}$ and compares the hand-adjusted taxonomy against pure k-means clusters over the same 132 concept embeddings. Three things hold across the sweep. Caption quality is flat, spanning 63.57 to 65.00 CIDEr around the backbone's 64.21, so no variant buys or costs captions. The probe-free readout is equally flat, 0.402 to 0.412, and every variant is far above the backbone's 0.233, so readability follows from having owners at all rather than from a particular partition. Selectivity, by contrast, varies with granularity, falling as the concepts are split more finely (0.967 at $N_c = 6$, 0.929 at 8, 0.889 at 10 for the hand-built

Table 5: Interface design sweep. All rows use the same schedule, seed, and data, and all are measured with the across-image mean replacement; the ablation used before Section 3.4 adopted the group-absent baseline; selectivity values are therefore comparable within this table but not with the 0.982 reported in Section 3.4. Chance selectivity is 0.49. The backbone rows apply the same ownership map to its first N_c positions after the fact.

Model	Partition	N_c	Queries	CIDEr	Probe-free mAP	Selectivity
MindLens	manual	8	48	64.80	0.410	0.929
MindLens	manual	10	48	63.57	0.406	0.889
MindLens	k-means	6	48	64.64	0.406	**0.967**
MindLens	k-means	8	48	64.65	0.402	0.875
MindLens	k-means	10	48	65.00	0.406	0.789
MindLens	manual	8	32	63.26	0.412	0.964
Backbone	(none)	–	48	64.21	0.233	0.464
Backbone	(none)	–	32	61.07	–	–

taxonomy), and at matched $N_c = 8$ the hand-built partition exceeds k-means (0.929 against 0.875). Every variant sits far above the backbone's 0.464 under the identical test. We therefore prefer the hand-built taxonomy because its groups can be named, which is what makes a channel interpretable, and not because it decodes or controls better.

Prefix width. Narrowing the prefix from 48 to 32 tokens costs the backbone 3.14 CIDEr (64.21 to 61.07) and the interface 1.54 (64.80 to 63.26), while the interface's readout is unchanged at 0.412 and its selectivity rises to 0.964. The interface therefore degrades more gracefully than the backbone at this width, though with one seed per cell we treat the 1.6-point difference in degradation as suggestive rather than established. A ten-group interface at 32 queries reaches the highest caption score of any variant we trained (65.54 CIDEr). We did not measure its readout or selectivity and so leave it out of the table.

Channel position. Because the channels occupy the first positions of the prefix, their effect could in principle come from that location rather than from the interface. We retrained the identical model with the channels moved to the last eight positions and to eight positions drawn at random (Table 6, top). Readability is unchanged (0.409 and 0.406 against 0.410) and the channels remain selective wherever they sit (0.911 in both cases), while the backbone read at those same positions stays unselective throughout (0.268 to 0.607). One effect is worth noting. Moved channels are individually more consequential: removing one costs its own group 16.17 and 15.36 points of recall against 6.52 when the channels are first, which suggests the early positions are partly buffered by the rest of the prefix.

Replacement value. The causal test replaces a channel with a constant, and the choice of that constant is an analysis decision rather than a property of the model, so we repeat the measurement under four choices (Table 6, bottom). The interface is selective under all four, from 0.839 to 0.982, with no sign reversal in any of them, and the backbone is unselective under the two that leave the captions intact. The two image-specific replacements are harsher interventions than they first appear: substituting the image's own residual mean destroys the caption at position 0 (CIDEr 31.56 for the interface and 0.15 for the backbone, against 64.80 and 64.21 unablated) and raises spillover fourfold, and the backbone's apparent selectivity under those two rows is produced by that one collapsed row rather than by any ownership.

None of these variants changes the qualitative conclusion. Across every partition, width, channel position and replacement value we measured, the interface is readable without a probe and selective under intervention, and the backbone subjected to the identical analysis is neither.

Table 6: Robustness of the causal result, measured with the group-absent mean replacement of Section 3.4. Top: the channels retrained at different prefix positions, with the backbone read at those same positions. Bottom: the primary model under four choices of the replacement value, with own-group drop and spillover in points of recall. Chance selectivity is 0.49.

Channel positions	CIDEr	Probe-free mAP	Own-group drop	Selectivity	Backbone
First eight (0–7)	64.80	0.410	6.52	**0.982**	0.268
Last eight (40–47)	65.00	0.409	16.17	0.911	0.464
Eight random	62.91	0.406	15.36	0.911	0.607
Replacement value	Own-group drop		Spillover	Selectivity	Backbone
Group-absent mean (used in the paper)	6.52		1.35	**0.982**	0.268
Across-image mean	5.05		1.17	0.929	0.464
Image's residual mean	9.59		5.47	0.839	0.607
Residual direction at matched norm	5.96		2.69	0.839	0.643

A.5 CONCEPT BANK AND GROUPING

The concept vocabulary is a fixed artifact, built once from the stimulus captions before any model was trained and released with the code. It is used in three places: to define the $K = 132$ supervision targets, to build the concept bank $\mathbf{C}$ in the LLM's embedding space, and to decide which concepts a generated caption mentions.

Candidate nouns. We parse the captions of the stimulus images with spaCy 3.8.14 and the `en_core_web_sm` model (v3.8.0), using only its tagger and lemmatiser, with the parser and named-entity components disabled. A candidate is any token tagged `NOUN` or `PROPN`, reduced to the lemma the model predicts. Two deliberate exceptions to plain part-of-speech filtering were made after measuring, for each candidate, how often the tagger missed it relative to its raw string occurrence in the captions. Five words are admitted regardless of their predicted tag because those miss rates were high: *hydrant* (35%), *vase* (18%), *surfboard* (12%), *skateboard* (8%) and *snowboard* (3%). The token *hot*, correctly tagged as an adjective, is retained only so that the lexicalized bigram *hot dog* can be formed.

Normalization. Spelling variants and synonymous forms are then merged through a hand-authored lookup table, written after inspecting the document frequencies of candidate variant pairs, for example *doughnut* into *donut*, *surf board* into *surfboard*, *phone* into *cell phone* and *hydrant* into *fire hydrant*. This step uses no embedding and no clustering algorithm. Compounds are kept as units rather than split into their parts, which is why 21 of the released surface forms are multi-word (*hot dog*, *traffic light*, *train track*, *living room*), and why the detector in Appendix A.3 must match multi-word forms before single words.

Vocabulary size. The vocabulary is fixed by rank rather than by a frequency cutoff: we keep the 200 most frequent entries by image document frequency, once compounds and normalization are resolved. The resulting minimum retained frequency, 571 of the $73{,}000$ images or 0.78%, is therefore a consequence of that rank cutoff and not a threshold we chose.

Concepts. Normalized forms that name the same thing are finally collected into $K = 132$ concepts, each with a name and a list of member words. The mapping is a partition: the released file records 184 distinct surface forms and no form belongs to two concepts, so no caption word can fire two concepts. Most concepts are a single form (100 of 132, such as *man* or *pizza*); the rest group inflectional variants and near synonyms, the largest being *tennis*, which covers court, racket, tennis, tennis ball, tennis court, tennis player and tennis racket. Frequencies span 571 to $17{,}470$ images with a median of $1{,}432$.

Groups. The 132 concepts are partitioned by hand into the $N_c = 8$ groups that the semantic channels own, and the partition is fixed before training and released as a file. The criterion is that every group must carry a single natural name, since a channel whose group cannot be named is not interpretable, whatever its statistics. Group sizes are consequently unequal, from 7 concepts for *people* to 26 for *outdoor* (Table 7), because the taxonomy follows the structure of scene descriptions rather than a balance constraint. One consequence is worth stating: the concept loss weights every concept equally, so a channel's share of the objective is proportional to its group's size, and the *people* channel receives roughly a quarter of the gradient that the *outdoor* channel does even though *people* concepts are the most frequently mentioned in the captions. Partitions obtained by k-means over the same concept embeddings are reported as ablations in Appendix A.4.

Scope and labels. Vocabulary selection, normalization and labeling were all run over the captions of all $73{,}000$ stimulus images, with no train and test split at any stage. The vocabulary is therefore not independent of the test images, although it is a frequency ranking over human captions and involves no model, no fMRI data, and no per-image supervision beyond the labels themselves. For each image, $y_k = 1$ when any of its reference captions mentions a member word of concept k. We verified that the column sums of the released label matrix equal the per-concept image frequencies exactly, so the labels implement that rule and no other, and the member-word detector of Appendix A.3 reproduces them at F1 96.97. An image carries 3.73 concepts on average (median 4), and 147 of the $73{,}000$ images carry none, in which case the concept loss supplies only negatives for that image.

Coverage. Figure 6 lists all 132 concepts with the number of images containing each, colored by group. The distribution is heavy-tailed within every group: *person* and *man* alone account for more images than the whole of *animals*, while 60 concepts appear in fewer than $1{,}200$ images each.

Table 7: The eight concept groups and their concepts, in order of the number of stimulus images containing each. Group sizes are given in parentheses.

Group	Concepts, in order of image frequency
people (7)	person, man, woman, child, body part, crowd, baby
animals (13)	dog, cat, animal, giraffe, horse, bird, elephant, bear, zebra, zoo, cow, teddy bear, sheep
food (23)	plate, food, drink, pizza, kitchen, counter, bowl, vegetable, cake, fruit, banana, sandwich, stove, meat, broccoli, donut, refrigerator, bottle, cheese, restaurant, pan, hot dog, orange
indoor (17)	table, wall, bathroom, chair, window, bed, sink, toilet, floor, cabinet, living room, mirror, couch, door, tv, bedroom, blanket
outdoor (26)	street, field, building, water, tree, grass, mountain, snow, ocean, beach, sidewalk, sky, park, plant, fence, city, pole, flower, house, rock, river, dirt, tower, forest, lake, store
vehicle (13)	car, train, airplane, bus, train track, motorcycle, truck, boat, bicycle, train station, airport, luggage, vehicle
sport (9)	skiing, ball, surfing, tennis, skateboarding, baseball, kite, frisbee, snowboarding
object (24)	sign, clock, shirt, computer, bench, light, umbrella, cell phone, camera, suit, hat, box, bag, eyewear, vase, fire hydrant, stop sign, jacket, tie, traffic light, book, paper, keyboard, video game

This imbalance is visible in the results. The groups whose channels show the largest causal effects, *outdoor*, *vehicle* and *object*, are groups of many moderately frequent concepts, whereas *animals*, whose channel is the weakest in both directions, is dominated by a few highly distinctive concepts that the rest of the prefix also encodes.

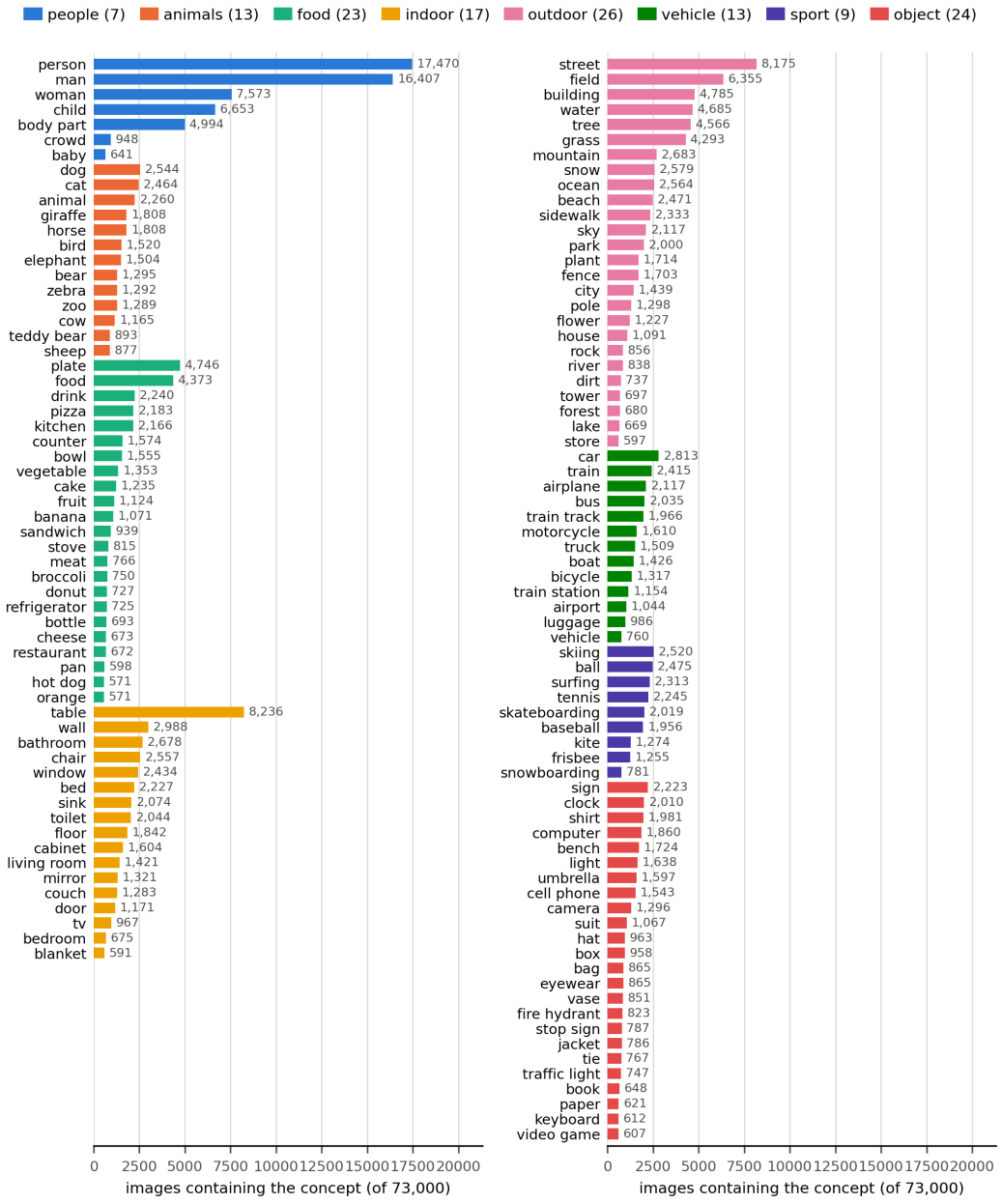

Figure 6: The 132 concepts and the number of the 73,000 stimulus images whose captions mention each, grouped by semantic channel and ordered by frequency within each group.