Words, Images, and the Mind: Bridging Neural Signals with Vision and Multimodal Semantics

Shachaf Haviv Shani Angel
Technion
{shachafhaviv@, shani.angel@}
campus.technion.ac.il

Abstract

How the human brain represents linguistic and conceptual meaning remains a central challenge in cognitive neuroscience and computational linguistics. This study investigates the alignment between brain activity and diverse semantic representations derived from static (GloVe, Word2Vec), contextual (BERT), visual (ResNet, DINOv2, BLIP-2), and vision-language models (CLIP, OpenCLIP, SigLIP, VDR). Using fMRI data from Pereira et al. (2018), we decoded isolated concepts into all embedding spaces, and sentence-level brain activity into textual spaces, comparing performance across models and semantic categories. While contextual embeddings offered moderate improvements over static baselines, image-based and multimodal representations showed modest alignment with neural data, often varying by category and model. Multimodal fusion, implemented through simple averaging of image and text embeddings, yielded mixed results, with benefits varying by model. Category-level analysis suggests that decoding success may depend on semantic domain, with no single modality excelling across all categories. These findings underscore the complementary roles of linguistic and visual information in neural meaning representation and highlight ongoing challenges in bridging these modalities. [1]

1 Introduction

Understanding how linguistic meaning is represented in the human brain is a central question in cognitive neuroscience and computational linguistics. A major line of research has focused on aligning distributional semantic models, such as word and sentence embeddings, with neural activity patterns measured during language processing. Early work demonstrated that it is possible to decode individual word or sentence meanings from brain signals using vector-based representations derived from large corpora (Mitchell et al. (2008); Pereira et al. (2018)). These studies established the utility of static embeddings, such as GloVe (Pennington et al., 2014), for mapping between linguistic and neural spaces.

Subsequent research has shown that contextualized word embeddings from deep neural models, such as BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019), can yield improvements in neural decoding and encoding tasks (Toneva and Wehbe, 2019). These models capture richer syntactic and semantic nuances by conditioning word representations on surrounding context, potentially leading to representations more aligned with cognitive processes. However, both static and contextual embeddings are limited to textual input and do not explicitly encode perceptual or grounded information.

Recent advances in multimodal models, particularly those trained jointly on visual and textual inputs, such as CLIP (Radford et al., 2021), open new possibilities for capturing grounded semantics. These models learn to align image and text representations in a shared latent space, potentially bridging the gap between linguistic meaning and perceptual grounding. Recent large-scale comparisons between models and brain activity suggest that training on diverse visual data and vision-language tasks leads to stronger alignment with high-level visual regions in the brain, emphasizing the importance of a model's "visual diet" in predicting neural responses (Conwell et al., 2022).

This paper explores whether embeddings from vision and multimodal models more closely align with neural representations of meaning. We build on the dataset of Pereira et al. (2018), which includes fMRI recordings of participants processing both full sentences and individual concepts. Each concept was presented along with representative images, enabling the study of how different

[1] Code is available at github.com/multimodal-decoding-and-the-brain.

semantic representations align with neural activity. Our work systematically evaluates how well static embeddings (GloVe, Word2Vec), contextual embeddings (BERT), and a diverse set of multimodal or vision-language models (CLIP, OpenCLIP, SigLIP, VDR, BLIP-2, DINOv2, ResNet) align with brain activity during language comprehension. Beyond replicating established decoding and encoding paradigms, we introduce an extensive analysis that incorporates both visual and combined multimodal representations.

We hypothesize that visual and multimodal embeddings capture complementary aspects of neural meaning representations, yielding the strongest alignment for perceptually grounded concepts.

Our contributions: We conduct a systematic comparison across models and present a fine-grained category-level analysis that uncovers how decoding performance varies across semantic domains. Additionally, we examine the impact of low-level visual features (image brightness and edge density) on decoding, highlighting potential perceptual confounds in aligning neural and representational spaces.

Our results show that visual and multimodal embeddings align with neural data in nuanced, domain-dependent ways. Some perceptual categories (e.g., food, people) benefit from visual grounding, while others do not, suggesting that visual information alone may be insufficient to capture more abstract or relational meanings. Moreover, combining visual and textual modalities leads to only modest gains, highlighting the persistent challenge of fully integrating perceptual and linguistic representations in neural decoding.

2 Data

We used three publicly available fMRI datasets from Pereira et al. (2018), each capturing whole-brain voxelwise BOLD responses (185,866 voxels) to linguistic stimuli presented to native English-speaking adults.

The first dataset includes 180 word-level concepts, each presented in three formats; sentence, definition, and six representative images, to elicit semantic processing. Neural imaging data and 300-dimensional GloVe embeddings are provided for each concept from a single participant.

The second dataset contains 384 sentences grouped into thematic passages, each paired with GloVe embeddings and fMRI responses from the same participant as the first dataset.

The third dataset consists of 243 independently constructed sentences, similarly organized into topics and passages, with GloVe embeddings and brain recordings from a different participant.

To enable the evaluation of multimodal representations, we utilized the image-based stimuli from the first dataset. Concepts were accompanied by six curated images designed to visually represent its meaning. We processed these images using multiple vision and vision-language models to obtain image-based and multimodal embeddings, which were then used to evaluate alignment with neural activity. Representative concept-image pairs are illustrated in Figure 1.

While many of the concepts are concrete and visually grounded (e.g., *plant*, *table*), others are more abstract or emotionally charged (e.g., *liar*, *unaware*), possibly posing greater challenges for visual representation. We address this variability in visualizability in our analysis and explore how it impacts decoding performance across embedding types.

3 Experiments and Results

We organize our results into two parts. First, the structured tasks, which replicate and extend the decoding and encoding experiments from Pereira et al. (2018) using static and contextual sentence embeddings. Second, the open-ended task, where we investigate the contribution of perceptual grounding through multimodal and visual representations.

3.1 Structured Tasks

3.1.1 Sentence Decoding

Methods The first stage of our analysis evaluates whether brain activity elicited by word concepts can be decoded into their semantic embeddings. Building on prior work (e.g., Pereira et al. (2018)), which showed reliable decoding for concrete concepts, this step establishes a baseline for assessing more complex embedding models in subsequent analyses.

The analyses in Pereira et al. (2018) were structured into three experiments:

- **Experiment 1 (Concept-level decoding):** One participant viewed 180 concepts presented as sentences, definitions, and image lists; a decoder mapped fMRI to GloVe embeddings.

- **Experiment 2 (Sentence decoding – in-domain):** The same participant read 384 sentences; the decoder predicted sentence-level GloVe embeddings from brain activity.
- **Experiment 3 (Sentence decoding – cross-domain):** A new participant read 243 novel-topic sentences; the decoder was tested for generalization across subjects and content.

In our study, we closely follow this structure. We first replicate concept-level decoding using the 180 concepts from Experiment 1. A linear decoder was trained to map fMRI signals to semantic space using ridge regression (McDonald, 2009). We employed 18-fold cross-validation, holding out 10 concepts per fold, training on the remaining 170, and evaluating decoding performance on the held-out items.

For each predicted embedding, we computed its cosine similarity with all 180 target embeddings, ranking the correct concept among the candidates. Lower ranks indicate better decoding performance. We conducted this procedure using both GloVe and Word2Vec embeddings to assess model-specific differences.

We then applied the trained decoder to sentence-level brain imaging data from Experiments 2 and 3. The goal here was to evaluate whether a decoder trained at the concept level could generalize to longer, sentence-level stimuli.

To assess finer-grained semantic effects, we grouped sentences by topic (available in the dataset metadata) and computed average decoding ranks within topics, allowing us to examine which semantic domains are more easily decoded from neural signals.

Results The average decoding rank across folds was 61.91 for GloVe and 61.08 for Word2Vec, indicating slightly better performance for Word2Vec. Word2Vec also yielded a greater number of concepts with an average rank below 90 (137 vs. 134), and demonstrated more consistent performance across folds. GloVe exhibited higher variance, with some folds performing substantially worse. These trends are visualized in Figure 2.

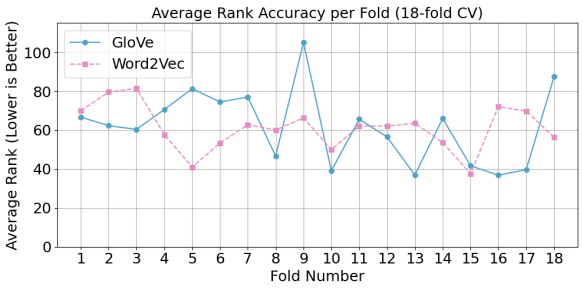

Figure 2: Average Rank Accuracy per Fold (18-fold CV) for GloVe and Word2Vec. Word2Vec demonstrates slightly better and more stable performance across folds.

Both embeddings showed overlapping and diverging strengths at the concept level. Concepts such as *laugh*, *food*, and *hair* were reliably decoded by both models. GloVe showed advantages for more abstract terms (e.g., *money*, *great*), while Word2Vec performed better on emotionally charged or action-related concepts (e.g., *liar*, *emotion*, *dig*).

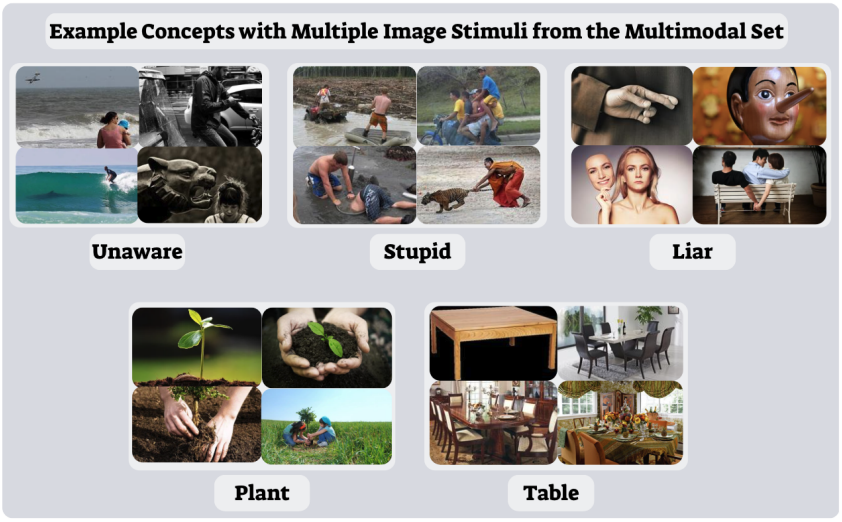

Figure 1: Examples of image stimuli used to represent concrete concepts in the multimodal subset. These images are part of the original fMRI dataset from Pereira et al. (2018) and were used to extract multimodal decoding.

When generalizing to sentence-level decoding, the GloVe-based decoder achieved average ranks of 158.93 (384 sentences) and 100.89 (243 sentences), outperforming chance levels (192 and 122, respectively). This suggests the model captures meaningful neural patterns, though still far from optimal decoding.

Figures 7 and 8 (Appendix) show variability in decoding performance across topics, but no single semantic category consistently predicts success or failure. Some abstract topics like *dreams* are well decoded, while others like *profession* perform poorly. Similarly, concrete topics such as *body part* are decoded accurately, while others like *owl* or *insect* are not. These results suggest that factors beyond simple concreteness or familiarity, such as linguistic variability or sentence structure may influence decoding performance, and further investigation is needed to identify consistent patterns.

3.1.2 Sentence Representations

Methods We next examined whether contextual embeddings better align with brain responses than static word embeddings. Contextual models like BERT encode word meaning in context, potentially offering a more brain-like representation of language. This analysis used the 384-sentence dataset (Experiment 2), where GloVe embeddings were computed by averaging word vectors per sentence. For contextual embeddings, we used `bert-base-uncased` from Hugging Face (Devlin et al., 2018), extracting the mean-pooled last hidden layer (768 dimensions) for each tokenized sentence. We trained linear decoders using 18-fold cross-validation to predict either GloVe or BERT embeddings from brain activity.

Results The BERT-based decoder outperformed GloVe, achieving a lower average rank of 96.94 compared to 124.54. Figure 3 shows the distribution of ranks for both models. BERT produced a lower median rank, a tighter distribution and fewer extreme outliers.

We also compute Top-k accuracy, defined as the proportion of test items for which the correct target appears within the top k predicted embeddings. Table 1 quantifies the improvement across different rank thresholds. For example, BERT achieved a top-5 rank accuracy of 9% compared to 4% for GloVe, and nearly doubled top-10 and top-50 accuracies.

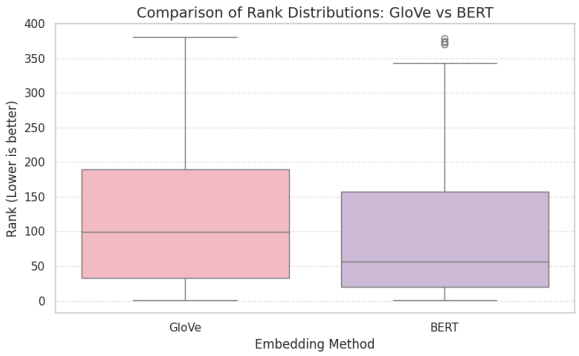

Figure 3: Distribution of decoding ranks for GloVe and BERT embeddings. BERT achieves lower median rank and tighter distribution.

Figures 9 and 10 (Appendix) indicate that both GloVe and BERT achieved their best decoding performance on sentences describing relatively simple and concrete items, such as apple, fork, and desk. This suggests that even contextual models like BERT struggle to robustly capture structured or abstract content at the fine-grained level of individual sentences. Moreover, while BERT offers modest gains on average, it still falls short of fully capturing how the brain encodes more abstract or relational semantics.

3.1.3 Brain Encoder Model

Methods Unlike prior sections focused on decoding, this analysis examined the reverse direction, predicting brain activity from sentence embeddings. This encoding approach assesses how well different embeddings explain variance in voxelwise fMRI responses.

Using the 243-sentence dataset from Pereira et al. (2018), we trained a separate linear regression model for each of the 185,866 voxels, predicting activation from sentence embeddings. We compared GloVe 300-dimensional static embeddings averaged over words and BERT 768-dimensional contextual embeddings from the mean-pooled final

Method	Top-1	Top-5	Top-10	Top-50	Top-100	Top-192
GloVe	0.01	0.04	0.07	0.33	0.50	0.76
BERT	**0.02**	**0.09**	**0.13**	**0.46**	**0.66**	**0.81**

Table 1: Top-k decoding accuracy for sentence embeddings, showing the proportion of test sentences for which the correct target appeared within the top-k ranked candidates (out of 192).

hidden layer of `bert-base-uncased` (Devlin et al., 2018).

Data were split 80/20 into training and test sets, and voxel-level performance was evaluated using the R^2 score. This process was repeated for both GloVe and BERT embeddings to compare their neural predictive power.

Results The R^2 distributions across all voxels are shown in Figure 4. Most voxels yielded negative R^2 scores, indicating poor predictive performance, but important differences emerged between embedding types.

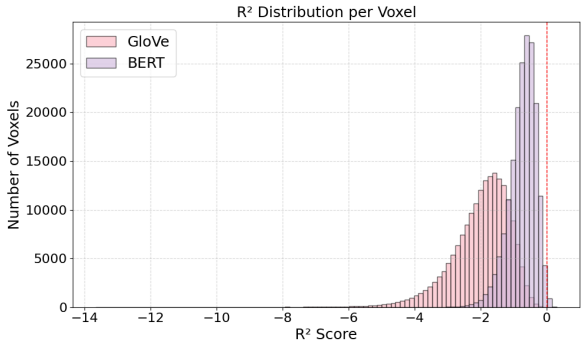

Figure 4: R^2 score distributions for voxelwise encoding models using GloVe and BERT embeddings. Only a small subset of voxels exceeded $R^2 > 0$ (dashed red line), primarily for BERT.

- **GloVe:** Only **2 voxels** achieved $R^2 > 0$, with a mean R^2 of 0.0310 across them.
- **BERT: 1,427 voxels** achieved $R^2 > 0$ (**0.77%** of the brain), with a higher mean R^2 of 0.0640 for those voxels.

These results suggest that contextualized sentence embeddings better capture features of brain activity than static embeddings, even when trained using a simple linear model. While the absolute R^2 scores remain low, the significantly larger set of positively predicted voxels using BERT highlights the added value of contextual information in encoding brain responses.

3.2 Open-Ended Task

3.2.1 Vision and Multimodal Decoding

The previous analyses demonstrated that both static and contextualized text-only embeddings can capture aspects of neural representations for linguistic stimuli. However, semantic processing in the brain is widely theorized to involve not just abstract linguistic features but also perceptual and grounded information, especially for concrete, imageable concepts (Kewenig et al., 2024; Binder et al., 2005). Vision-language models offer a promising approach to capturing this grounded semantics by integrating visual and textual data into shared representational spaces. In this section, we investigate whether such embeddings, derived from both images and text, better align with neural activity compared to purely text or image-only embeddings.

Methods We utilized the concept dataset from Pereira et al. (2018), which includes linguistic stimuli paired with six curated image embeddings for each concept. Visual embeddings were computed for each image and then averaged across the six images per concept to obtain a single representation. For multimodal evaluation, we explored three strategies:

1. Using only the **image-based embeddings** (from averaging over the six image embeddings).
2. Using only **text embeddings** by encoding each concept label.
3. Creating a **combined embedding** by averaging the image and text embeddings from the previous strategies, to integrate perceptual and symbolic information.

Models and Embedding Extraction We evaluated a diverse suite of modern vision and vision-language models, each representing a different type of grounding:

Vision-Language Models:
CLIP (ViT-B/32) (Radford et al., 2021): a canonical vision-language model jointly trained on image-text pairs.

OpenCLIP (ViT-H/14) (Ilharco et al., 2021): a large-scale extension of CLIP trained on LAION-2B data.

SigLIP (Zhai et al., 2023): a Google-trained model designed to improve on CLIP with better language-image similarity training objectives.

VDR - Vision-Domain Representation (LlamaIndex Team, 2024): a recent multi-modal embedding model designed to generalize across diverse modalities.

Additionally, we evaluated models for visual representation:

Vision Models:
ResNet-50 (Koonce, 2021), a classical convolutional baseline.

DINOv2 (Oquab et al., 2023), a self-supervised vision transformer.

BLIP-2 (Li et al., 2023), used here for its visual encoder outputs.

For each, we extracted embeddings for the images. For models supporting text (VLMs), we also computed embeddings for the concept labels themselves.

Decoder Training and Evaluation We used the same ridge regression approach (McDonald, 2009) as in earlier sections, training decoders to map voxel-wise brain data to these embedding spaces. We employed 18-fold cross-validation over concepts, holding out 10 concepts per fold.

We evaluated decoder performance using three complementary metrics: the **Average Rank** of the true concept embedding in similarity space; the **Top-5 Accuracy**, representing the fraction of cases where the true concept was among the top-5 most similar candidates; and the **Mean Reciprocal Rank (MRR)** (Craswell, 2016), which captures the average inverse rank of the correct answer, thereby rewarding predictions that rank the correct concept near the top more heavily.

We also computed 95% bootstrap confidence intervals (CI) on the mean rank, using 1,000 resamples with replacement, to better quantify the stability and variability of decoding across folds. Additionally we performed a detailed **Category-level Analysis**, comparing decoding performance across semantic domains such as *Objects*, *Cognitive States*, *Social Concepts*, and *Abstract Ideas*, to see where multimodal or purely visual grounding offered the biggest gains. A complete list of concept assignments to categories is provided in the Appendix Concept Category Lists.

Finally, to explore whether simple perceptual characteristics of the image stimuli could partly explain decoding differences, we correlated decoding ranks with two low-level **Visual Features**: mean image brightness and edge density. This allowed us to assess whether certain low-level visual biases might drive some of the observed differences across models.

Results

Image-only Decoding Across All Models Table 2 summarizes the decoding performance for image embeddings across all evaluated models. We found that all models performed broadly similarly, with slight advantages for more recent or larger-scale models.

Model	Average Rank	Top-5 Accuracy (%)	MRR
BLIP-2	78.99	**4.44**	0.04
CLIP	76.26	2.22	0.04
DINOv2	75.22	3.33	0.04
OpenCLIP	74.21	2.78	0.04
ResNet	76.94	**4.44**	0.04
SigLIP	**73.89**	2.78	0.03
VDR	74.72	2.22	0.03

Table 2: Decoding performance using image embeddings across all evaluated models.

We note that SigLIP achieved the lowest average rank, while ResNet and BLIP-2 had slightly better top-5 accuracy. For comparison, in the structured decoding task with text-only embeddings, GloVe and Word2Vec achieved lower average ranks (61.91 and 61.08), suggesting that despite advances in visual representation learning, language-based embeddings may still better align with neural representations of semantic concepts.

Multimodal Integration: Image and Text We next evaluated whether combining image and text embeddings improved decoding. For each encoder supporting both modalities, we averaged the image and text vectors ('combo') before running the same pipeline. We also computed 95% bootstrap CIs on the mean rank to assess statistical reliability. Results are shown in Table 3.

Among the evaluated models, VDR combo embeddings achieved the best overall performance, with the lowest average rank (67.19). OpenCLIP's combo also stood out, yielding the highest Top-5 accuracy (6.67%), suggesting synergy between image and text features. In contrast, CLIP's combo did not outperform its unimodal variants, highlighting that the benefits of multimodal fusion are model-dependent. Notably, although some multimodal representations outperformed their image-only counterparts, even the best combinations did not surpass the performance of static text embeddings in the structured task, suggesting that effective fusion of modalities remains a challenge despite recent progress.

Category Analysis with Image Embeddings To systematically assess the contribution of visual semantics, we conducted an analysis of category-level decoding performance using only the im-

Model	Modality	Average Rank	Top-5 Acc. (%)	MRR	Mean Rank CI
CLIP	image	76.26	2.22	0.04	[69.63, 83.65]
	text	81.53	4.44	0.04	[73.45, 89.13]
	combo	78.42	2.22	0.03	[70.84, 85.59]
OpenCLIP	image	74.21	2.78	0.04	[66.95, 80.88]
	text	78.55	3.33	0.04	[71.58, 85.53]
	combo	73.59	**6.67**	**0.05**	[66.82, 80.33]
SigLIP	image	73.89	2.78	0.03	[66.48, 80.51]
	text	89.84	3.33	0.03	[82.23, 97.99]
	combo	80.28	6.11	0.04	[72.28, 87.68]
VDR	image	74.72	2.22	0.03	[68.46, 81.61]
	text	71.46	3.89	0.04	[64.57, 78.26]
	combo	**67.19**	3.89	0.04	[60.83, 73.50]

Table 3: Decoding performance across image, text, and combined embeddings with 95% bootstrap confidence intervals on mean rank. Combining modalities sometimes led to modest improvements.

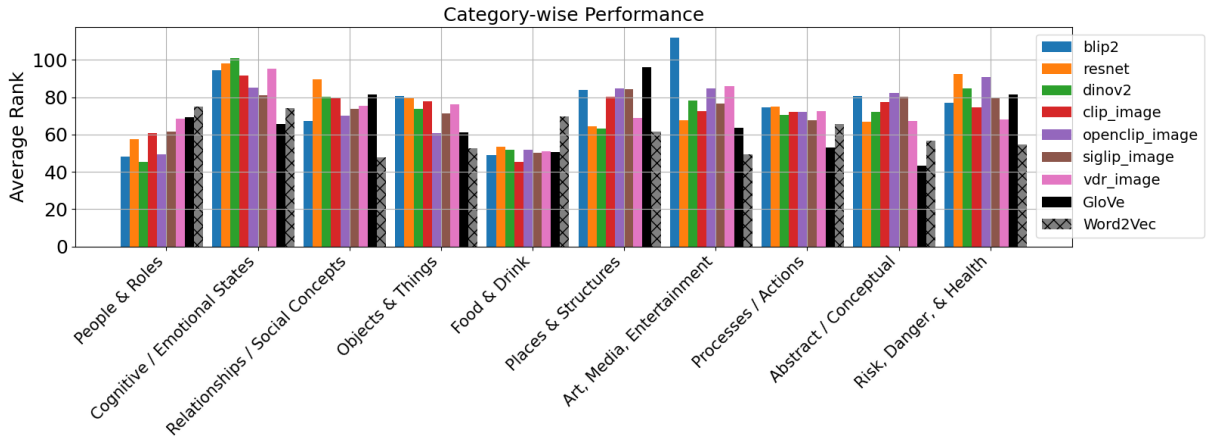

Figure 5: Category-wise decoding performance using image embeddings and textual baselines (GloVe, Word2Vec). Lower average ranks indicate better decoding.

age embeddings across all evaluated models, with GloVe and Word2Vec added as textual baselines.

As shown in Figure 5, some categories such as *Food & Drink* and *People & Roles* benefit from purely visual encoding, possibly due to their strong perceptual consistency. Interestingly, *Objects & Things*, despite being a highly perceptual category, did not perform as well, suggesting that visual appearance alone may not capture the necessary conceptual distinctions for decoding. In contrast, more abstract or relational categories, such as *Relationships*, *Processes*, and *Cognitive States*, exhibited weaker performance, underscoring that purely visual models may struggle to capture symbolic or socially constructed meanings. While GloVe and Word2Vec generally performed better than the image models, they did not consistently outperform them across all categories, indicating that the relative strength of each modality is domain-dependent.

Category-Level Analysis Across Modalities We then extended this analysis to examine how decoding performance varies not only by category, but also by modality, comparing image-only, text-only, and combined multimodal embeddings. Figure 11 (Appendix) shows average rank performance across concept categories, broken down by model and modality. This analysis highlights that no single modality universally outperformed the others across all concept categories. While image embeddings often provided advantages for perceptually grounded domains like *People & Roles* and *Places & Structure*, text or combined representations sometimes performed slightly better on more abstract categories such as *Art, Media, Entertainment* and *Cognitive / Emotional States*. These patterns suggest that different modalities capture complementary aspects of neural representations, and that the optimal representational strategy may vary depending on the underlying semantic domain.

Correlation with Visual Features To assess whether low-level visual properties contribute to decoding difficulty, we computed two summary statistics for each concept: average brightness and

edge density. Brightness was measured as the average grayscale intensity of the concept's images, reflecting how light or dark an image appears overall. Edge density was computed using the Canny edge detector (Canny, 1986), capturing the proportion of image pixels identified as edges, effectively quantifying the amount of visual texture or structural complexity.

As shown in Figure 6, we observed consistent negative Pearson correlations between decoding rank and brightness, suggesting that darker images tend to be decoded more accurately. In contrast, positive correlations with edge density imply that more visually complex images are associated with worse decoding performance. The consistency across models suggests that even simple perceptual features can influence how well visual embeddings align with brain activity, possibly due to differences in how cleanly visual concepts are represented or perceived.

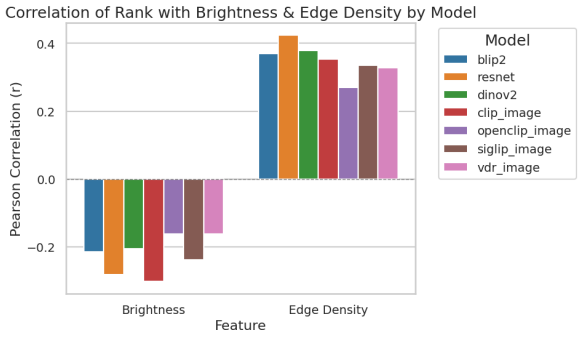

Figure 6: Pearson correlations between average decoding rank and average image brightness / edge density. Positive correlations indicate worse decoding for that feature.

4 Discussion and Conclusions

This study investigates the alignment between neural representations of semantic content and a wide range of embedding models, spanning from static word vectors to vision and multi-modal vision-language encoders.

In the structured decoding tasks, contextualized embeddings from BERT offered modest but consistent gains over static embeddings like GloVe and Word2Vec. These results support the view that context-sensitive representations are better aligned with how the brain encodes linguistic meaning, though the overall improvements remained relatively small. The encoding analysis further reinforced this conclusion, with BERT explaining variance in more voxels than GloVe.

Turning to multimodal and image-based models, we found that purely visual embeddings were able to predict neural responses with reasonable accuracy. However, their overall performance remained lower than text-based models, suggesting that visual similarity alone is insufficient to fully capture the conceptual distinctions reflected in neural activity. Interestingly, combining text and image embeddings led to modest improvements in some models (e.g., VDR, OpenCLIP), but not all. These results highlight that while multimodal fusion has potential, its benefits are highly model-dependent, and many current architectures still fall short of effectively integrating perceptual and symbolic information in ways that mirror brain representations.

Our category-level analyses offered deeper insight into when visual grounding helps or hinders decoding. While visual models performed well for some categories with consistent perceptual features, such as *Food & Drink*, other highly visual domains like *Objects & Things* performed surprisingly poorly. This suggests that visual complexity or ambiguity may interfere with decoding accuracy, even within concrete domains. In contrast, abstract categories involving social or relational meaning consistently showed weaker performance across all models. That said, GloVe and Word2Vec did not consistently outperform image models, highlighting category-specific strengths.

Our correlation analysis revealed that simple perceptual features, such as brightness and edge density, modulate decoding accuracy, with darker and less textured images yielding better results. This suggests that basic visual properties can affect how well image embeddings align with brain data, and that clear visuals play a role in decoding success.

Limitations and Future Work. Several limitations should be acknowledged. First, the datasets used in this study involve a limited number of participants and stimuli, which may restrict generalizability. Second, our decoders were linear, and while this improves interpretability, it may underestimate the representational capacity of deep embeddings. Third, our image-text combination strategy relied on simple averaging, which may not capture the nuanced interplay between modalities that more advanced fusion architectures could exploit.

Future work could explore non-linear decoding models, more sophisticated multimodal fusion techniques (e.g., cross-attention), and larger, more diverse neuroimaging datasets.

References

Jeffrey R Binder, Chris F Westbury, Kristen A McKiernan, Edward T Possing, and David A Medler. 2005. Distinct brain systems for processing concrete and abstract concepts. *Journal of cognitive neuroscience*, 17(6):905–917.

John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. 2022. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, pages 2022–03.

Nick Craswell. 2016. Mean reciprocal rank. In *Encyclopedia of database systems*, pages 1–1. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, et al. 2021. Openclip. 7.

Viktor Nikolaus Kewenig, Gabriella Vigliocco, and Jeremy I Skipper. 2024. When abstract becomes concrete, naturalistic encoding of concepts in the brain. *eLife*, 13:RP91522.

Brett Koonce. 2021. Resnet 50. In *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 63–72. Springer.

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

LlamaIndex Team. 2024. VDR-2B-Multi-V1: A multimodal embedding model. `https://huggingface.co/llamaindex/vdr-2b-multi-v1`. Accessed: 2025-06-30.

Gary C McDonald. 2009. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

A Appendix - Figures

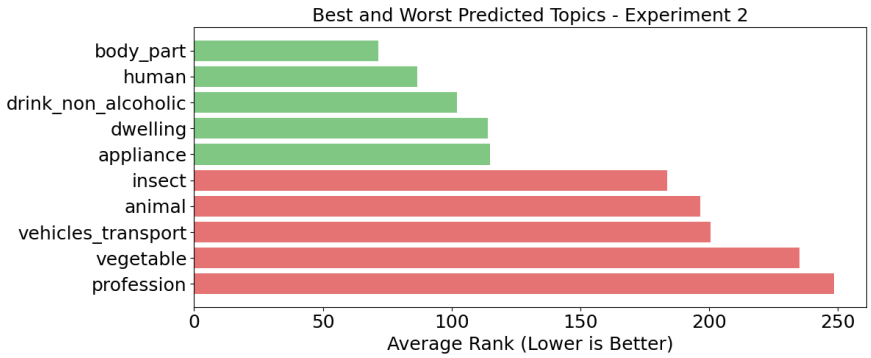

Figure 7: Best (top 5) and worst (bottom 5) predicted topics for sentence-level decoding in Experiment 2 (384 sentences).

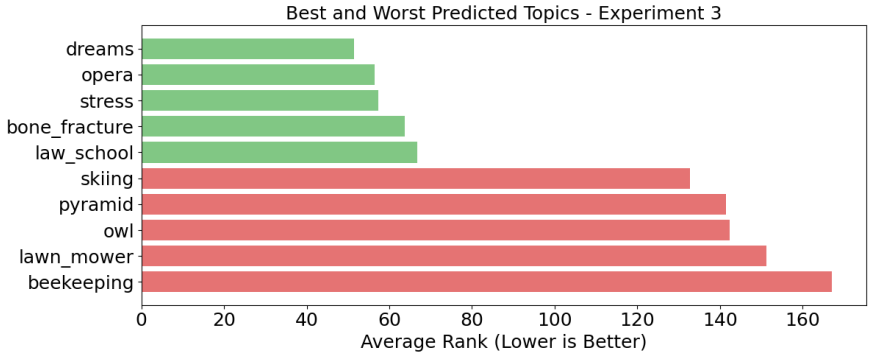

Figure 8: Best (top 5) and worst (bottom 5) predicted topics for sentence-level decoding in Experiment 3 (243 sentences).

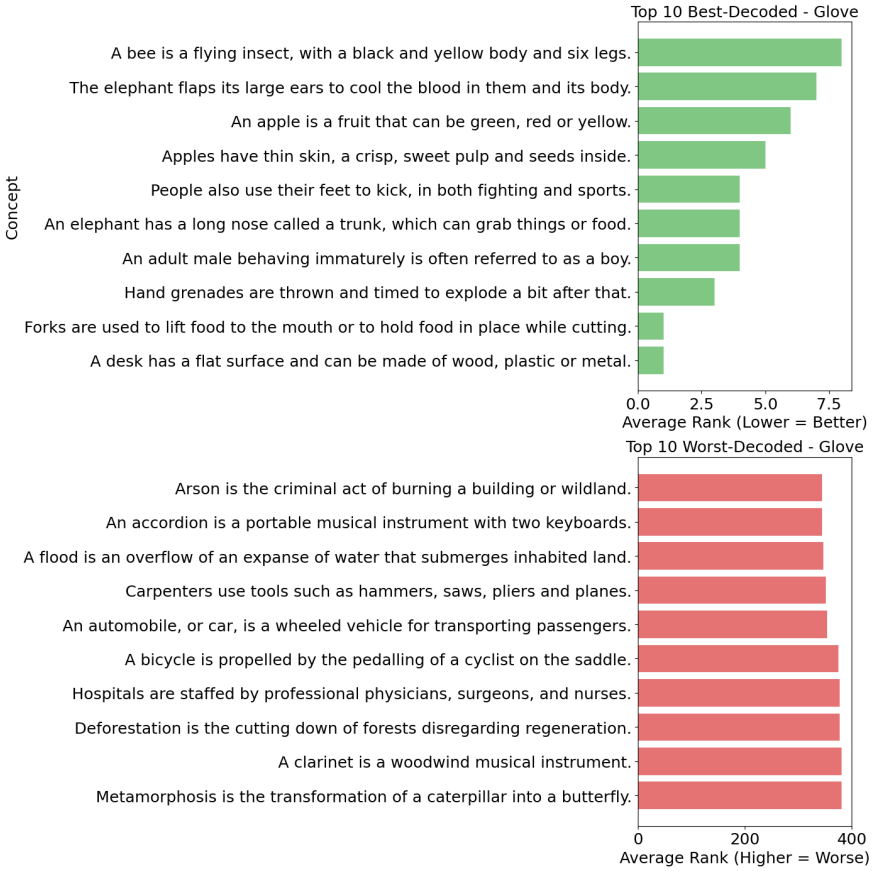

Figure 9: Top-10 best and worst decoded sentences using GloVe embeddings.

Figure 10: Top-10 best and worst decoded sentences using BERT embeddings.

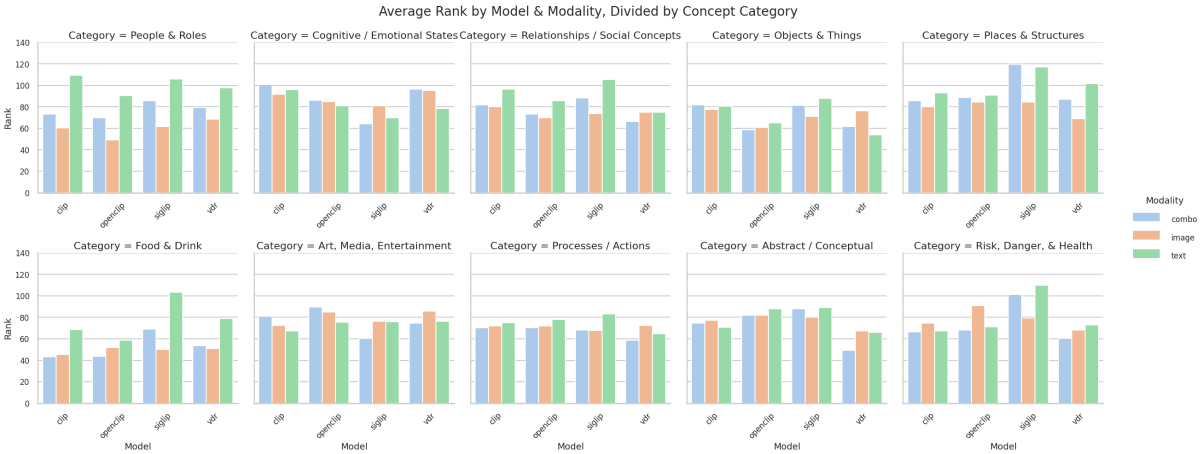

Figure 11: Average rank by model and modality, divided by concept category. Lower ranks indicate better decoding performance.

Concept Category Lists

- **People & Roles:** Doctor, Driver, Student, King, Lady, Professional, Liar, Suspect, Job
- **Cognitive / Emotional States:** Ability, Accomplished, Angry, Attitude, Emotion, Emotionally, Feeling, Great, Help, Hurting, Impress, Kindness, Laugh, Personality, Sad, Smart, Smiling, Stupid, Successful, Typical, Unaware, Willingly, Ignorance, Charming, Crazy, Silly, Pleasure, Poor
- **Relationships / Social Concepts:** Argument, Argumentatively, Charity, Marriage, Relationship, Religious, Sin, Protection, Obligation, Law, Team
- **Objects & Things:** Bag, Ball, Bar, Bed, Beer, Camera, Clothes, Cockroach, Computer, Device, Gold, Gun, Hair, Light, Medication, Money, Picture, Pig, Plant, Ship, Sign, Skin, Star, Sugar, Table, Tool, Toy, Tree, Bear, Bird, Dog, Engine
- **Places & Structures:** Apartment, Building, Construction, Jungle, Land, Mountain, Nation, Residence, Road, Prison, Vacation
- **Food & Drink:** Beer, Dessert, Dinner, Fish, Food, Garbage, Seafood, Sugar, Taste
- **Art, Media, Entertainment:** Art, Dance, Movie, Music, Show, Applause, News, Noise
- **Processes / Actions:** Beat, Burn, Challenge, Cook, Counting, Damage, Deceive, Dedication, Deliberately, Delivery, Dig, Dissolve, Disturb, Do, Movement, Plan, Play, Reaction, Read, Sell, Sew, Spoke, Trial, Tried, Wash, Wear, Carefully, Dressing, Fight, Flow, Left
- **Abstract / Conceptual:** Big, Business, Code, Collection, Economy, Election, Electron, Elegance, Event, Experiment, Invention, Investigation, Invisible, Level, Magic, Material, Mathematical, Mechanism, Philosophy, Quality, Science, Shape, Solution, Soul, Sound, Texture, Time, Usable, Useless, Word, Extremely, Sexy
- **Risk, Danger, & Health:** Blood, Body, Broken, Dangerous, Disease, Illness, Pain, War, Weak, Brain, Weather