

Discovering Semantic Subdimensions through Disentangled Conceptual Representations

Yunhao Zhang^{1,2}, Shaonan Wang^{1,2,*}, Nan Lin^{3,4,*},
Xinyi Dong⁵, Chong Li^{1,2}, Chengqing Zong^{1,2}

¹State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS, Beijing, China

²School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

³State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, CAS, Beijing, China

⁴Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

⁵State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University

zhangyunhao2021@ia.ac.cn; shaonan.wang@polyu.edu.hk; linn@psych.ac.cn

Abstract

Understanding the core dimensions of conceptual semantics is fundamental to uncovering how meaning is organized in language and the brain. Existing approaches often rely on predefined semantic dimensions that offer only broad representations, overlooking finer conceptual distinctions. This paper proposes a novel framework to investigate the subdimensions underlying coarse-grained semantic dimensions. Specifically, we introduce a **Disentangled Continuous Semantic Representation Model (DCSRM)** that decomposes word embeddings from large language models into multiple sub-embeddings, each encoding specific semantic information. Using these sub-embeddings, we identify a set of interpretable semantic subdimensions. To assess their neural plausibility, we apply voxel-wise encoding models to map these subdimensions to brain activation. Our work offers more fine-grained interpretable semantic subdimensions of conceptual meaning. Further analyses reveal that semantic dimensions are structured according to distinct principles, with polarity emerging as a key factor driving their decomposition into subdimensions. The neural correlates of the identified subdimensions support their cognitive and neuroscientific plausibility.

1 Introduction

Core dimensions are fundamental to structure mental representations, enabling systematic classification, context-sensitive interpretation, and generalization across novel situations (Allen, 1984; Shepard, 1987; Gardenfors, 2004). In perceptual domains like vision, core dimensions include color, shape, motion, depth, and texture (Ge et al., 2022; Palmer, 1999; Mapelli and Behrmann, 1997); in audition, they include pitch, loudness, timbre, spatial location, and rhythm (Poeppel and Assaneo, 2020; Bizley and Cohen, 2013; Temperley, 2004).

These dimensions support object recognition (e.g., a red, round, shiny, motionless object = apple) and flexible inference (e.g., a loud sound in a forest = danger). Crucially, this dimensional structure enables generalization by allowing novel stimuli to be interpreted based on their positions within a shared representational space.

Extending this framework to conceptual semantics is more challenging, as the underlying dimensions are more abstract and less perceptually grounded. One promising approach to addressing the above challenges is to empirically define a set of semantic dimensions (Binder et al., 2016; Fernandino et al., 2016; Diveica et al., 2023; Wang et al., 2023b). Binder et al. (2016) introduced 65 semantic dimensions grounded in neuroscience research on conceptual representations. However, recent studies have shown that these 65 dimensions exhibit substantial overlap and lack neurobiological plausibility (Wang et al., 2022b; Fernandino et al., 2022; Zhang et al., 2025a). To address this, Wang et al. (2023b) proposed six major semantic dimensions: vision, action, space, time, social, and emotion. Follow-up neuroimaging studies have validated the robustness of these six dimensions (Zhang et al., 2025a; Tang et al., 2025; Lin et al., 2024). Nevertheless, this set of interpretable dimensions offers a coarse-grained representation of semantic space, which is limited in its ability to capture finer semantic distinctions that are crucial for accurately modeling the conceptual meaning (Binder et al., 2016; Hoffman and Ralph, 2013).

This paper proposes a novel framework for investigating the subdimensions underlying six semantic dimensions. To establish neural validity, we examine their corresponding representational patterns in the brain. Specifically, we introduce a Disentangled Continuous Semantic Representation Model (DCSRM), which decomposes word embeddings from large language models into multiple semantic-specific sub-embeddings. Each sub-embedding is

*Corresponding authors.

learned using a multi-objective optimization approach to maximize the encoding of its target semantic while minimizing interference from others.

We then interpret the meaning of the semantic subdimensions captured by these sub-embeddings by inspecting words with high and low loadings. To assess their neural plausibility, voxel-wise encoding models (Huth et al., 2016) are applied to map these subdimensions onto brain responses during natural language comprehension.

Our work offers a more fine-grained and interpretable set of semantic subdimensions for representing conceptual meaning than the conventional six semantic dimensions. These subdimensions are shown to be structured according to distinct principles, with polarity emerging as a key factor driving their decomposition. Moreover, we identify neural correlates of these subdimensions, aligning with prior neuroimaging studies (Lindquist et al., 2012; Lin et al., 2024), thereby supporting their cognitive and neuroscientific validity.

To summarize, our main contributions include:

- We propose a disentangled continuous semantic representation model that separates distinct type of semantic information from large language models and extracts subdimensions within six semantic dimensions.
- We employ interpretable analysis method to define the meaning of each subdimension and identify key factors driving the decomposition of semantic subdimensions.
- We use neural encoding models to reveal the brain representations of the identified subdimensions, further supporting their cognitive plausibility.

2 Related work

2.1 Conceptual semantic dimensions

Conceptual semantic dimensions are typically defined using either experience-based or data-driven approaches. Experience-based approaches, grounded in neuroscience and psychology, offer a solid foundation for defining semantic dimensions. Binder et al. (2016) proposed 65 semantic dimensions, which have been shown to explain semantic-related behavior and brain activation (Anderson et al., 2017, 2019; Tong et al., 2022; Fernandino et al., 2022). However, the 65 semantic dimensions exhibit notable overlap and redundancy. Specifically, some pairs of dimensions exhibit Pearson correlation coefficients exceeding 0.8, indicating a

lack of independence between them (Wang et al., 2022b). Furthermore, these dimensions do not all contribute equally to the explanation of brain activation. For instance, Zhang et al. (2025a), using brain decoding methods, found that several non-sensorimotor dimensions (e.g., near, toward, away, number, benefit, needs) were not broadly represented across brain semantic networks. Building on these prior findings, Wang et al. (2023b) proposed six coarse-grained semantic dimensions (vision, action, space, time, social, and emotion) and developed the Six Semantic Dimension Database, which contains subjective ratings for 17,940 Chinese words. Follow-up neuroimaging studies have further validated the effectiveness of these six dimensions (Zhang et al., 2023b; Lin et al., 2024; Zhang et al., 2025a; Tang et al., 2025).

Data-driven approaches leverage the rich semantic information encoded in language models to uncover semantic dimensions by analyzing their word embeddings (Hollis and Westbury, 2016; Grand et al., 2022). Hollis and Westbury (2016) applied PCA to skip-gram embeddings and identified emotion-related dimensions like valence and dominance. However, even after PCA, each dimension of embeddings often entangle multiple types of information. Moreover, skip-gram models capture less semantic richness than large language models (e.g., LLaMA, Alpaca), limiting their ability to identify fine-grained semantic dimensions.

2.2 Disentangled methods

Disentangled word representations serve two main purposes. First, they enhance interpretability by revealing the information encoded in embeddings (Karwa and Singh, 2025; O'Neill et al., 2024; Liao et al., 2020); for instance, Liao et al. (2020) decompose dense embeddings into sub-embeddings tied to discrete attributes (e.g., animal, location, adjective). Second, they benefit downstream NLP tasks such as sentence representation (Chen et al., 2019), text generation (Iyyer et al., 2018), word sense disambiguation (Silva De Carvalho et al., 2023), and sentiment/style transfer (Zhu et al., 2024).

Unlike prior work, this paper addresses a key question in psycholinguistics and neuroscience: What fine-grained semantic subdimensions enable more accurate representation of conceptual meaning, and to what extent are they grounded in neural activity? A central challenge lies in the overlapping and intertwined relationships among different semantic components within conceptual representa-

tions, making disentanglement inherently difficult. To address this, we propose a disentangled continuous semantic representation model (DCSRM), which segments large language model embeddings into multiple low-dimensional sub-embeddings, each encoding specific semantic information. We then analyze these sub-embeddings to identify semantic subdimensions and use voxel-wise encoding models to examine their neural correlates.

3 Methods

To investigate the subdimensions of each semantic dimension and their neural representations, we propose a three-step framework (Figure A3): 1) use DCSRm to encode each semantic information into a low-dimensional sub-embedding; 2) analyze the sub-embedding dimensions to identify semantic subdimensions; and 3) perform voxel-wise encoding to examine their relationship with brain activity during natural story comprehension.

3.1 Disentangled continuous semantic representation model

To encode each semantic-specific information into low-dimensional sub-embeddings, we propose a disentangled continuous semantic representation model (DCSRM). Our goal is to transform M h -dimensional dense word vectors $V \in \mathbb{R}^{M \times h}$ into disentangled embedding $X \in \mathbb{R}^{M \times h}$ by leveraging N continuous semantic attributes $B = \{b_1, \dots, b_N\}$ labeled on words.

X is expected to have two properties. The first is retaining information encoded in V . More specifically, we require $VV^T \approx XX^T$ as pointed out by [Levy and Goldberg \(2014\)](#), to ensure that the global similarity structure remains stable after the transformation. The second property is that the h column vectors of X are decomposed into $N+1$ sub-embedding sets, $X_{b_1}, \dots, X_{b_N}, X_{\text{unseen}}$, where each sub-embedding encodes information specific to one semantic attribute. For instance, X_{b_1} is expected to represent information solely related to semantic attribute b_1 , independent of the other semantic attributes $b_2, \dots, b_N$. To achieve these targets, we employ a multi-objective learning framework.

Orthogonal constraint ($\mathcal{L}_{\text{ORT}}$). We transform the original embedding V using a learnable projection matrix $W \in \mathbb{R}^{h \times h}$, yielding $X = VW$. To preserve the global semantic structure after transformation, we impose an orthogonality constraint

$W^TW \approx I$, minimizing:

$$\mathcal{L}_{\text{ORT}} = \|W^TW - I\|_2 \quad (1)$$

We have $XX^T = (VW)(VW)^T = V(WW^T)V^T = VV^T$ if $WW^T = I$ holds.

Continuous attribute prediction ($\mathcal{L}_{\text{SL}}$). For modeling the relationship between sub-embeddings and continuous semantic attributes, we minimize the conditional expectation of prediction error:

$$\mathbb{E}_{x_{b,j} \sim X_b, y_j \sim Y} [|y_j - q_\theta(x_{b,j})|] \quad (2)$$

where $x_{b,j}$ denotes the j -th word representation in the sub-embedding matrix X_b for attribute b , and y_j is the corresponding ground-truth rating for that attribute. A parametric regression model $q_\theta(\cdot)$ is trained to predict y_j from $x_{b,j}$ using the Smooth L1 loss:

$$\mathcal{L}_{\text{SL}} = \sum_{j=1}^N \text{SmoothL1}(y_j, q_\theta(x_{b,j})) \quad (3)$$

Semantic contrastive loss ($\mathcal{L}_{\text{CE}}$). To help the sub-embeddings capture more attribute-specific information, for each semantic attribute b , we treat words with ratings greater than a threshold as positive examples, and the others as negative examples. We minimize the contrastive cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\frac{x_i^T x_i^+}{\tau})}{\sum_{j=1}^N \exp(\frac{x_i^T x_j}{\tau})} \quad (4)$$

where x_i and x_j denote the i -th and j -th words in the embedding space X , x_i^+ is the positive sample of x_i , and τ is a temperature parameter.

Reconstruction loss ($\mathcal{L}_{\text{REC}}$). To preserve input information while supporting semantic disentanglement, we let X_b be features to reconstruct original vectors for words having attribute b by minimize the reconstruction error:

$$\mathcal{L}_{\text{REC}} = \|v_j - \varphi(x_{b,j})\|^2 \quad (5)$$

where φ is a single fully connected layer mapping sub-embeddings back to the original space.

KL-based sparsity constraint ($\mathcal{L}_{\text{KL}}$). To promote feature selection and disentanglement, we apply variational dropout ([Molchanov et al., 2017](#); [Liao et al., 2020](#)) to each dimension of X . In the training process, we inject multiplicative noise on X :

$$\xi \sim \mathcal{N}\left(1, \alpha_b = \frac{p_b}{1 - p_b}\right) \quad (6)$$

where $p_b = \text{sigmoid}(\log \alpha_b)$ is the h -dimension dropout rates. For each attribute b in B , the dimensions with dropout rates lower than 40% are normally regarded as X_b .

Distribution alignment loss ($\mathcal{L}_{\text{DIS}}$). To encourage disentanglement when handling multiple attributes, we include a loss function on dropout rates. Let a N -dimensional vector P be $1 - p_b$ for all b in B in a specific dimension. The idea is to minimize $\prod_{j=1}^N p_j$ with constraint $\sum_{j=1}^N p_j = 1$. The optimal solution is that the dimension is relevant to only one attribute b' where $1 - p'_{b'} \approx 1$. In implementation, we minimize the following loss function:

$$\mathcal{L}_{\text{DIS}} = \sum_{j=1}^N \log P_j + \beta \left\| \sum_{j=1}^N P_j - 1 \right\|^2 \quad (7)$$

We empirically set $\beta = 1$ following Liao et al. (2020) to balance the trade-off between encouraging sparsity and maintaining normalized dimension weights, ensuring effective disentanglement.

3.2 Sub-embedding analysis

To identify the semantic subdimensions encoded in each sub-embedding, we first apply Principal Component Analysis (PCA) to orthogonalize the dimensions within each sub-embedding, yielding transformed sub-embeddings $X'_{b_1}, X'_{b_2}, \dots, X'_{b_N}$. We then compute the Pearson correlation and pairwise order consistency between each dimension and the rating data, selecting those with significant correlations. For each, we instruct multiple large language models (e.g., GPT-4o, Grok3, Claude 3.7 Sonnet) as linguists to annotate the corresponding semantic subdimension based on the top-ranked words. Finally, we combine the results from these models to derive the final subdimension labels.

3.3 Voxel-wise encoding models

Voxel-wise encoding models map each transformed sub-embedding to its corresponding fMRI signal. To align with BOLD signal delay, word vectors are convolved with a canonical hemodynamic response function (HRF)¹ and downsampled to match the fMRI sampling rate. During training, 14 additional regressors—capturing low-level stimulus properties such as word rate, word length, part-of-speech (adverb, noun, particle, verb), sound envelope,

word frequency, and six head motions—are included but excluded during prediction.

We employ ridge regression and perform 5-fold nested cross-validation to ensure robust evaluation. Model performance is assessed by computing the Pearson correlation between predicted and observed fMRI signals. Group-level statistical significance is determined by comparing the estimated correlations to a null distribution of correlations derived from two independent Gaussian random vectors of equivalent length (Huth et al., 2016).

Next, we explored the neural correlates of the semantic subdimensions within the brain regions significantly associated with the transformed sub-embeddings. We compute the average weight matrix by averaging the cross-validation weight matrices. To ensure consistent interpretability of weight direction, we apply sign correction to the weight matrix based on the correlation between each transformed sub-embedding dimension and rating data, as described in Section 3.2. Finally, for each voxel, we assign the semantic subdimension with the highest weight as its representative subdimension.

4 Experimental Setup

Brain imaging data. We use the Chinese fMRI dataset from the SMN4Lang (Wang et al., 2022a), which was collected from 12 native speakers as they listened to 60 stories from the Renmin Daily Review². These stories covered a broad range of topics, with each story lasting between 4 to 7 minutes, resulting in approximately 5 hours of audio content. The text and audio for all stories were downloaded from the Renmin Daily Review website, and the text was manually verified for consistency with the audio. The total word count across all stories was 43,326, forming a vocabulary of 9,153 unique words. After data collection, the fMRI data were preprocessed following the Human Connectome Project (HCP) pipeline (Glasser et al., 2013).

Semantic rating data. Our study utilized the rating dataset from the SSDD (Wang et al., 2023b), which includes subjective ratings for 17,940 Chinese words. It focuses on six semantic dimensions: vision, action, social, emotion, space, and time. These 17,940 words encompass nearly all commonly used Chinese words, including verbs, nouns, adjectives, adverb, and quantifiers, etc. Table A4 presents the definitions for each semantic dimension. Thirty human raters evaluated each word on

¹The canonical HRF models the expected BOLD response to a neural event.

²<https://www.ximalaya.com/toutiao/30917322/>

	DCSRM _{vis}		DCSRM _{act}		DCSRM _{soc}		DCSRM _{emo}		DCSRM _{time}		DCSRM _{spc}		Average	
	vis ↑	non_vis ↓	act ↑	non_act ↓	soc ↑	non_soc ↓	emo ↑	non_emo ↓	time ↑	non_time ↓	spc ↑	non_spc ↓	target ↑	non_target ↓
GloVe	0.602	0.243	0.535	0.196	0.592	0.153	0.450	0.141	0.463	0.149	0.622	0.127	0.544	0.168
Word2Vec	0.826	0.228	0.702	0.245	0.792	0.189	0.654	0.149	0.649	0.225	0.790	<u>0.161</u>	0.736	0.200
MacBERT-large	0.840	0.172	0.728	0.190	0.858	0.127	0.772	0.114	0.807	0.160	0.855	0.213	0.810	0.163
LLaMA2-1.3b	0.881	<u>0.204</u>	0.750	0.325	0.854	<u>0.058</u>	0.770	0.118	<u>0.806</u>	0.123	0.874	0.168	0.823	0.166
Alpaca2-1.3b	0.867	0.266	0.750	0.314	0.851	0.059	<u>0.771</u>	0.116	0.796	0.136	0.861	0.174	0.816	0.177
LLaMA2-7b	0.886	0.219	<u>0.755</u>	0.168	<u>0.856</u>	0.056	0.772	0.121	0.804	0.130	<u>0.877</u>	0.170	<u>0.825</u>	0.144
Alpaca2-7b	<u>0.892</u>	0.209	0.760	0.158	0.852	0.081	0.772	0.132	<u>0.806</u>	<u>0.128</u>	0.880	0.176	0.827	<u>0.147</u>
LLaMA3-8b	0.896	0.223	0.730	<u>0.166</u>	0.854	0.056	0.767	<u>0.115</u>	<u>0.771</u>	0.148	0.851	0.172	0.812	<u>0.147</u>

Table 1: **Semantic prediction performance of DCSR sub-embeddings.** For each semantic dimension (e.g., social), the “target” column (e.g., soc) reports the Pearson correlation between sub-embedding predictions and ground-truth ratings for the corresponding semantic dimension. The “non_target” column (e.g., non_soc) shows the average correlation with ratings from all other dimensions. Higher target and lower non_target scores indicate better semantic disentanglement. Bolded values highlight the top-performing models, while underlined values indicate the second-best.

a 1–7 scale (7 = very high, 1 = very low) for all six semantic dimensions, based on the given definitions. For each word on each dimension, the final rating was obtained by averaging the 30 individual scores. These semantic ratings accurately reflect the extent to which a concept involves information related to each dimension. Table A5 shows six-dimension semantic ratings for several concepts. For instance, the word “justice” received a score of 1 on the time dimension and 1.133 on the space dimension, indicating that it lacks temporal information and contains only a minimal amount of spatial information.

Word representations. We utilize eight widely adopted Chinese computational language models, categorized into two groups: context-independent models, including Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), and context-aware models, comprising MacBERT-large (Cui et al., 2020), LLaMA2 (1.3B and 7B) (Touvron et al., 2023), Alpaca2 (1.3B and 7B) (Taori et al., 2023), and LLaMA3 (8B) (Grattafiori et al., 2024).

In line with prior research (Wang et al., 2024; Zhang et al., 2023a, 2025b), for context-independent models, Word2Vec and GloVe embeddings were trained on the Xinhua News corpus (19.7 GB)³ using identical model parameters. For context-aware models, we randomly sampled up to 1,000 sentences per target word from the Xinhua News corpus. These sentences were input into the models, and word vectors were extracted from the final layer. The vectors for each target word were then averaged to derive its word representation. To address substantial variations in hidden layer dimensions of context-aware models, we applied PCA to their word representations, reducing dimensionality while retaining at least 80%

explained variance.

Evaluation. We evaluate DCSR using a semantic prediction task, where ridge regression predicts semantic ratings from sub-embeddings X_b . The Pearson correlation between predicted and ground-truth values is computed, and performance is assessed via five-fold nested cross-validation. To examine the impact of loss functions, we evaluate sub-embeddings obtained by removing each loss function individually. Additionally, we compare the performance of the original and disentangled embeddings to demonstrate that DCSR retains the original information.

5 Results and Analysis

5.1 DCSR result

To assess the models’ capacity for semantic disentanglement, we report the prediction performance of DCSR sub-embeddings across six semantic dimensions in Table 1. As shown, DCSR consistently produces well-disentangled representations across all language models: each sub-embedding captures rich, dimension-specific information while suppressing unrelated semantics. While Alpaca2-7B and LLaMA2-7B achieve the best average performance, the top-performing model differs by dimension. For instance, in the vision dimension, LLaMA3-8B better captures vision-specific semantics. We therefore select the best-performing model for each dimension in the subsequent subdimension analysis.

Additionally, we observe that 7B models often produce sub-embeddings with richer target-specific semantic content than their 1.3B counterparts, and Alpaca2 achieves performance comparable to LLaMA2. This suggests that moderate parameter increases enhance semantic disentanglement, and that models retain fine-grained semantic capabili-

³<http://www.xinhuanet.com/whxw.htm>

	DCSRM _{vis}		DCSRM _{act}		DCSRM _{soc}		DCSRM _{emo}		DCSRM _{time}		DCSRM _{spc}		Average	
	vis ↑	non_vis ↓	act ↑	non_act ↓	soc ↑	non_soc ↓	emo ↑	non_emo ↓	time ↑	non_time ↓	spc ↑	non_spc ↓	target ↑	non_target ↓
DCSRM- $\mathcal{L}_{DIS}$	<u>0.879</u>	0.820	0.770	0.842	0.863	0.824	0.794	0.837	0.805	0.835	0.870	0.822	0.830	0.830
DCSRM- $\mathcal{L}_{KL}$	0.866	<u>0.207</u>	<u>0.734</u>	0.168	<u>0.854</u>	<u>0.056</u>	0.766	0.115	0.771	<u>0.144</u>	0.849	<u>0.171</u>	0.807	0.144
DCSRM- $\mathcal{L}_{SL}$	\	\	0.461	0.572	\	\	0.162	0.171	\	\	\	\	\	\
DCSRM- $\mathcal{L}_{REC}$	0.864	0.227	0.723	0.156	0.853	0.055	0.766	<u>0.133</u>	<u>0.772</u>	0.139	0.849	0.170	0.804	<u>0.147</u>
DCSRM- $\mathcal{L}_{CE}$	0.866	0.206	0.725	<u>0.160</u>	0.851	0.060	0.762	0.115	<u>0.772</u>	0.150	0.849	0.197	0.804	0.148
DCSRM	0.896	0.223	0.730	0.166	<u>0.854</u>	<u>0.056</u>	<u>0.767</u>	0.115	0.771	0.148	<u>0.851</u>	0.172	<u>0.812</u>	<u>0.147</u>

Table 2: **Ablation study of DCSRМ training on LLaMA3-8B.** The table reports semantic prediction performance across six semantic dimensions. For each semantic dimension (e.g., *soc*), the “target” column reports the Pearson correlation between the predicted and ground-truth ratings for that dimension, based on its corresponding sub-embedding obtained without a specific loss (e.g., $\mathcal{L}_{DIS}$). The “non_target” column reports the average correlation with ratings from all other dimensions. A slash (\) indicates that no valid sub-embeddings are formed under the corresponding dropout setting. Higher target and lower non_target scores indicate better semantic disentanglement. Bolded values denote the best results, and underlined values indicate the second-best. Loss terms: $\mathcal{L}_{DIS}$ = distribution alignment loss, $\mathcal{L}_{KL}$ = KL-based sparsity constraint, $\mathcal{L}_{SL}$ = continuous attribute prediction loss, $\mathcal{L}_{ORT}$ = orthogonality constraint, $\mathcal{L}_{REC}$ = reconstruction loss, and $\mathcal{L}_{CE}$ = semantic contrastive loss.

ties even after supervised fine-tuning. Furthermore, context-aware models (e.g., LLaMA2, Alpaca2) consistently outperform context-independent models (e.g., Word2Vec, GloVe). This highlights their advantage in capturing dimension-specific semantic nuances and effectively suppressing irrelevant information, resulting in clearer semantic disentanglement.

Table 2 reports the performance of LLaMA3-8b semantic-specific sub-embeddings on the semantic prediction task after ablating each loss function in DCSRМ. The results show that removing either the $\mathcal{L}_{DIS}$ or the $\mathcal{L}_{SL}$ leads to the disappearance of sub-embeddings under certain dropout conditions or the emergence of entangled representations that mix target and non-target semantics. These findings indicate that $\mathcal{L}_{SL}$ (continuous attribute prediction loss) preserves magnitude differences within sub-embeddings, aligning attribute strengths with ground truth and facilitating effective subspace extraction via dropout. $\mathcal{L}_{DIS}$ (distribution alignment loss) promotes decorrelation among sub-embeddings, and its removal results in overlapping representations and weakened semantic boundaries. Moreover, the inclusion of $\mathcal{L}_{KL}$, $\mathcal{L}_{REC}$, and $\mathcal{L}_{CE}$ further improves the separation between target and non-target semantics within the sub-embedding space.

We also observe that the original embeddings and the disentangled embeddings achieve comparable performance on the semantic prediction task, demonstrating that DCSRМ preserves the information from the original vectors⁴.

Overall, these results demonstrate that DCSRМ effectively separates different types of semantic information into distinct sub-embeddings. While not

aiming for complete disentanglement, our goal is to maximize the separation of semantic dimensions within word representations.

5.2 Sub-embedding analysis result

All principal components (PCs) showed significant correlations with their corresponding semantic ratings ($p < 0.05$), indicating relevance to their target domains⁵. However, PCs with weak correlations ($r < 0.1$), accounting for 13.04% of the total PCs, were considered to encode minimal semantic-relevant information and excluded from subsequent subdimension analyses.

To interpret the meaning of each subdimension, we present representative words from each PC in Table 3. As shown, each semantic dimension is subdivided into two or more subdimensions, indicating the finer-grained organization within each semantic dimension. Some subdimensions align with prior research, such as the presence of negative valence and positive valence within the **emotion** dimension (Lindquist et al., 2012; Russell, 2003). We also reveal previously unexplored semantic subdimensions, such as the dynastic eras and history change subdimensions in the **time** dimension. Our work provides a more detailed, fine-grained description of conceptual semantics compared to the conventional six semantic dimensions.

A cross-dimension comparison reveals that semantic dimensions vary in the number, polarity, and interpretability of their subdimensions. The **action** dimension exhibits the most diverse structure, comprising six subdimensions related to affective outburst, physical motion, and ritual behavior. In contrast, **vision**, **social**, and **emotion** show more well-defined relationships between orthogo-

⁴See Table A6 for details.

⁵See Figure A4 for details.

Dimension	PC	Representative Words	Semantic subdimension
vision	PC1	glasses, umbrella, parrot, tortoise, toaster, camellia, pepper, mask	Static vision
	PC2	hook, chisel, search, leap, drill, break in, knock, patch	Dynamic vision
action	PC1	cry, weep, shout, yell, tears, roar, laugh, shed tears	Outburst acts
	PC2	blink, cry, flicker, tremble, jump, shed tears, twinkle, gasp	Micro-movements
	PC3	push-up, dive, handstand, fall, capture, karate, retreat, registration	Forceful acts
	PC4	kneel, take down, issue, kneel on ground, sink, bury, order, play chess	Downward acts
	PC5	applaud, cheer, celebrate, run, ski, beg, swim, applause	Functional body acts
	PC6	bow, push-up, bend over, kowtow, handstand, squat, lower head, salute	Bending/ritual acts
social	PC1	duel, revolution, military, alliance, comrades, dispute, opponent, arbitration	Conflict
	PC2	pass, goal, submission, vote, assist, delivery, community, bidding	Collaboration & exchange
emotion	PC1	obsession, love, affection, hate, disappointment, happiness, passion, addiction	Emotional load
	PC2	death, disease, corruption, murder, tragedy, grief, violation, suffering	Negative valence
	PC3	love, affection, romance, passion, care, devotion, fondness, attachment	Positive valence
time	PC1	tomorrow, next year, recent, millennium, long-term, era, ancient, years	Temporal span
	PC2	Northern Wei, Eastern Jin, Western Jin, Jin Dynasty, military governor, prefect, poverty alleviation, Friday	Historical change
	PC4	anniversary, annual, Christmas, birthday, holiday, winter, same day, that day	Commemorative events
	PC5	Qing Dynasty, late Qing, Yuan Dynasty, Song Dynasty, Southern Song, early Qing, late Qing, Ming Dynasty	Dynastic eras
space	PC1	railway, suburban, frigate, destroyer, southeast, northwest, capital, total area	Regional locations
	PC2	urban-rural, north-facing, nomadic, assembly hall, launch site, carrier rocket, rural, touring	Sites & orientation
	PC3	epicenter, climbing, Arctic, cliff, bridge surface, canyon, plateau, outer space	Extreme spaces

Table 3: **Semantic subdimensions identified across six semantic dimensions.** For each PC, we present representative words selected from one end of the dimension—specifically, 8 out of the top 20 highest-loading words that strongly reflect the target semantic subdimension. The opposite end of the PC typically contains words unrelated to that subdimension. PCs with limited semantic relevance (Action PC7, Time PC3, and Social PC3) are excluded from this analysis (see Section 5.2 for details).

nal subdimensions within their respective semantic dimensions, which are structured along clear semantic axes that indicate polarized relationships (e.g., static vs. dynamic vision, conflict vs. collaboration, positive vs. negative valence). This suggests that polarization serves as an important factor driving the decomposition of semantic dimensions into finer-grained subdimensions.

We also find that individual subdimensions span multiple semantic dimensions. For instance, the dynamic vision subdimension in the **vision** dimension encodes action-related information; some high-loading words in subdimensions of the **space** dimension also encode aspects of visual information. These findings align with prior research that suggests the original rating data for vision, action, and space dimensions exhibit relatively high correlations (Wang et al., 2023b; Lin et al., 2024).

Moreover, the data-driven decomposition reveals semantic structures that deviate from intuitive or traditional expectations. For instance, the **time** blends abstract temporal spans (e.g., millennium, recent) with culturally grounded markers (e.g., Qing Dynasty, Christmas), while the **space** ranges from concrete geographical references (e.g., suburban, north-facing) to symbolic or extreme locations (e.g., outer space, cliff). These findings suggest that time and space may not be structured solely

along geometric or chronological lines, but are enriched by cultural and experiential semantics. This may also reflect the nature of large language models trained purely on text, which tend to capture more abstract and culturally embedded semantic patterns.

Overall, these patterns reveal that conceptual semantic dimensions are not flat or homogeneous, but instead exhibit structured subspaces organized along psychologically meaningful dimensions. This supports the view that human semantic knowledge is shaped by a small set of interpretable principles, such as polarity and hierarchy (Rosch, 1975; Lakoff, 2008; Gardenfors, 2004).

5.3 Voxel-wise encoding result

Figure 1 illustrates both distinct and overlapping neural correlates of SSDD ratings and transformed sub-embeddings. In embodied-related dimensions (e.g., vision, action), SSDD rating data predicts a broader range of brain regions compared to the transformed sub-embeddings. However, in abstract dimensions (e.g., emotion and time), the transformed sub-embeddings predict a wider array of neural activation patterns than the SSDD rating data. These findings suggest that language models trained on vast amounts of pure text data capture human-level conceptual representation in

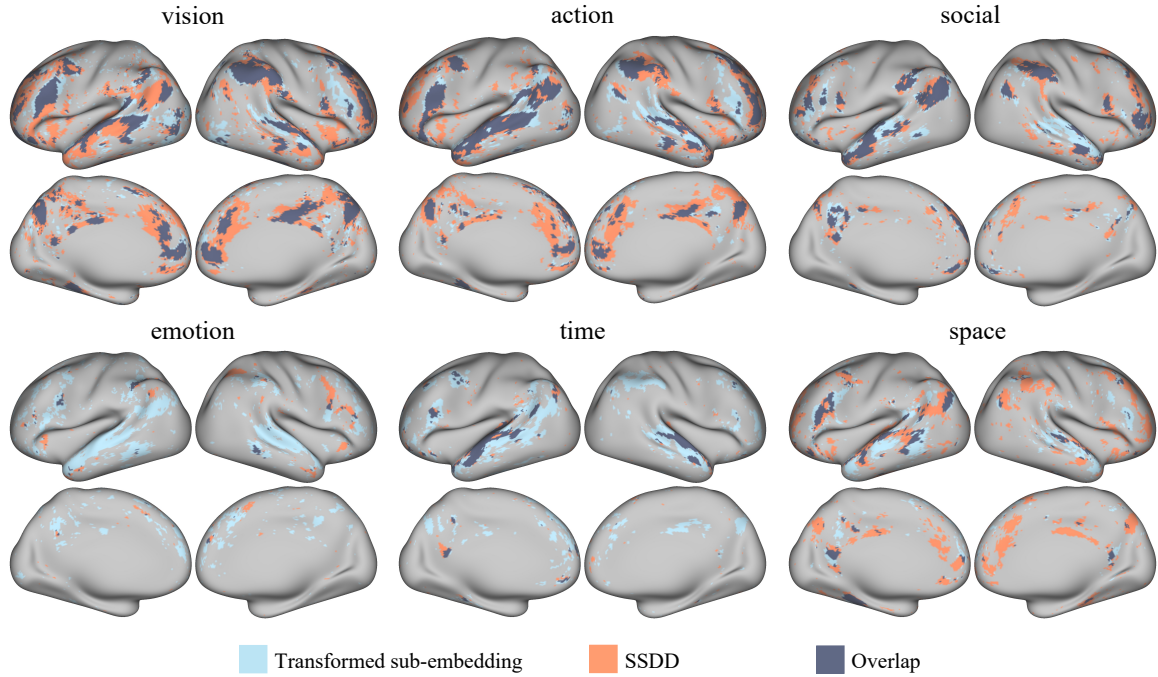


Figure 1: **Distinct and overlapping neural correlates of SSDD ratings and transformed sub-embeddings across semantic dimensions.** Blue regions indicate voxels where BOLD responses are significantly predicted by transformed sub-embeddings, while orange regions show the same for SSDD rating data (one-tailed t-test, $-\log(p)$, $p < 0.001$). Slate-gray regions denote overlap between the two. Transformed sub-embeddings denote the result of applying PCA separately to each semantic-specific sub-embedding generated by DCSRm (see Section 3.2 for details). This figure integrates results from Figure A6 and Figure A7.

non-sensorimotor dimensions, but they capture relatively less sensorimotor information, consistent with recent studies (Xu et al., 2025). Moreover, the sub-embeddings activate brain regions including the inferior frontal gyrus (IFG), superior temporal gyrus (STG), posterior superior temporal sulcus (pSTS), middle temporal gyrus (MTG), inferior temporal gyrus (ITG), precuneus (Pcun), cingulate gyrus (CG), fusiform gyrus (FG), and angular gyrus (AG). These regions are largely consistent with prior findings on language-processing networks in the brain⁶ (Binder et al., 2009; Huth et al., 2016; Yang et al., 2019), thereby supporting the validity of our computational framework for studying neural language comprehension. We also find that brain regions activated by different sub-embeddings has substantial overlap. A possible explanation is the multi-functionality of brain regions, where the same region represent multiple semantic subdimensions.

Figure 2 shows neural correlates of semantic subdimensions. As shown, semantic subdimensions are distributively represented across semantic networks, indicating that these subdimensions rely heavily on distributed brain networks rather than

localized brain regions. This finding aligns with the widely held view in cognitive neuroscience that the brain employs distributed representations to encode all types of information even primitive features (Zhang et al., 2022; Lin et al., 2024).

Several subdimensions across different dimensions converge on similar brain regions, particularly the CG, IFG, MTG, and AG. For instance, collaboration & exchange (social), positive valence (emotion), and dynastic eras (time) all activate regions involved in autobiographical memory, value processing, and social reasoning (D’Argembeau et al., 2014; Singer and Bluck, 2001; Lin et al., 2020). This suggests shared cognitive mechanisms underlying semantically rich, socially meaningful concepts. Moreover, we find that the left AG and left STG are involved not only in representing sensorimotor information (e.g., static vision, micro-movements) but also in encoding higher-level social and non-sensorimotor information (historical change, conflict, emotional valence). This supports a popular research view that the left AG and STG function as semantic hubs, integrating associations between sensory and social knowledge (Lin et al., 2018; Skipper et al., 2011).

Our results are also consistent with prior neu-

⁶See Figure A5 for details.

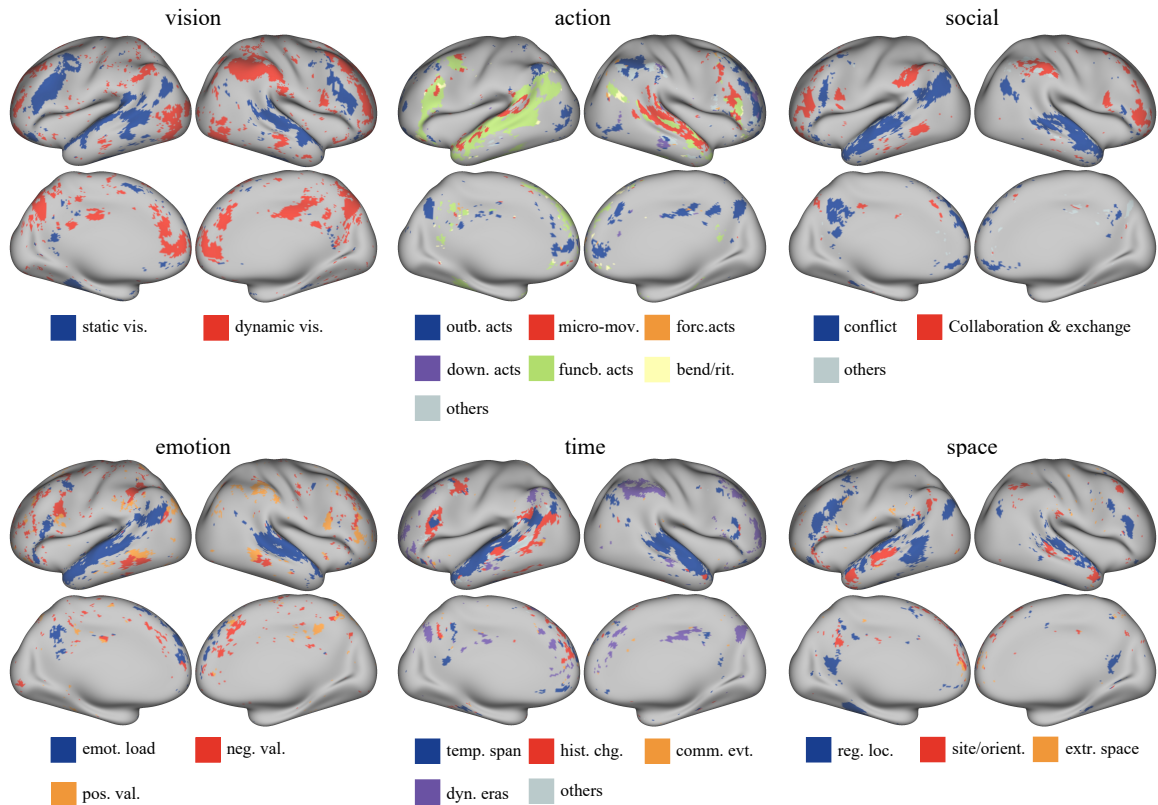


Figure 2: **Neural correlates of semantic subdimensions across six semantic dimensions.** Others refers to subdimensions from the transformed sub-embeddings that carry minimal information related to the corresponding semantic dimension (see Section 5.2 for details).

roimaging studies. We find that the Pcc, right MTG, and right AG are involved in representing collaboration & exchange, corroborating previous findings on the neural representation of collaborative behavior (Xie et al., 2020). Brain regions associated with positive and negative valence also align with prior research on emotional processing (Kragel and LaBar, 2016; Wager et al., 2015; Saarimäki et al., 2016). The subdimension of static vision engages the AG, Pcc, posterior CG, and FG, consistent with known visual-semantic areas (Lin et al., 2018; Sabsevitz et al., 2005).

Beyond these, we also identify brain regions associated with certain semantic subdimensions that, to our knowledge, have not been reported in previous work. For instance, the dynastic eras activates the Pcc, MTG, and CG. The historical change activates the dorsal IFG, ITG, and AG. For Action, the micro-movements activates the right STG and right IFG, while functional body acts engage the left STG, left MTG, left IFG, and left Pcc. These results extend prior semantic neuroscience by identifying novel neural correlates for semantic subdimensions, advancing our understanding of how fine-grained conceptual semantics are encoded in the brain. Future work will further validate the reliability of these findings and explore their cognitive and neural significance in greater depth.

6 Conclusions

Our primary goal is to investigate the fine-grained structure of conceptual semantics. A key challenge lies in defining and quantifying semantic subdimensions. To address this, we propose a Semantic Representation Disentangling Model with an interpretable framework that decomposes word embeddings into multiple sub-embeddings, each capturing distinct semantic content and corresponding to a specific subdimension. These subdimensions are shown to be structured according to distinct principles, with polarity emerging as a key factor driving their decomposition. Neural encoding analyses further confirm their representation in the brain. Compared to traditional approaches, our method automatically uncovers more fine-grained semantic dimensions from large-scale data without manually specifying their number or type, offering a data-driven foundation for constructing a conceptual semantic representation system that supports systematic classification, context-sensitive interpretation, and generalization to novel situations.

Limitations

First, our analysis currently focuses only on the six semantic dimensions included in the SSDD database. Other dimensions, such as auditory and tactile, may also play a significant role in conceptual semantics. In future work, when large-scale rating data for additional semantic dimensions become available, we will expand the fine-grained semantic space we have developed and explore the neural correlates of a broader range of semantic dimensions in natural language understanding.

Second, we rely solely on large language models trained on text data to explore the subdimensions of each semantic dimension. However, human conceptual knowledge is acquired not only through abstract symbolic interactions but also through sensory and motor experiences (Bi, 2021; Paivio, 1990). Recent advances in multimodal models, which integrate both linguistic and sensory inputs, have demonstrated that these models more closely resemble human cognitive representations (Tang et al., 2021; Wang et al., 2023a). Moreover, recent studies have demonstrated that multimodal models can successfully accomplish many tasks that purely text-based models fail to handle, and they consistently achieve superior performance on these tasks (Jian et al., 2024, 2025; Ren et al., 2025; Guan et al., 2025; Guo et al., 2025). This suggests that these models encode richer and more comprehensive semantic information. In future work, we plan to integrate multimodal large models to further explore the fine-grained semantic space, enabling a more comprehensive investigation of semantic subdimensions.

Third, our findings are based on Chinese data, and it remains unclear whether these results can be generalized to other languages, such as English or German. In future studies, we plan to use large-scale rating data from other languages to further validate the stability of our conclusions.

Finally, our current work primarily utilizes the semantic subdimension framework to explore the neural representation patterns of fine-grained conceptual semantics. In future studies, we plan to investigate the broader potential of this framework in other domains, such as brain decoding (Sun et al., 2023; Ye et al., 2025), model interpretability (Li et al., 2023; Duan et al., 2025), and the diagnosis and auxiliary treatment of neurological disorders (Dong et al., 2024; Wang et al., 2025).

Ethical Statements

We used preprocessed data from a publicly available dataset. All cognitive data were anonymized to ensure that no personally identifiable information was retained. The dataset was provided by the Institute of Automation, Chinese Academy of Sciences. All experimental procedures were approved by the Institutional Ethics Committee of the Institute of Psychology, Chinese Academy of Sciences, and the Institutional Review Board of Peking University, in accordance with established ethical guidelines and regulations.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful discussions and valuable comments. This research was supported by grants from the National Natural Science Foundation of China (No. 62036001) and the STI2030-Major Project (No. 2021ZD0204105).

References

- James F Allen. 1984. Towards a general theory of action and time. *Artificial intelligence*, 23(2):123–154.
- Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev DS Raizada. 2017. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, 27(9):4379–4395.
- Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Rajeev DS Raizada, Feng Lin, and Edmund C Lalor. 2019. An integrated neural decoder of linguistic and experiential meaning. *Journal of Neuroscience*, 39(45):8969–8987.
- Yanchao Bi. 2021. Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, 25(10):883–895.
- Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. 2016. Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, 33(3-4):130–174.
- Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. 2009. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, 19(12):2767–2796.

- Jennifer K Bizley and Yale E Cohen. 2013. The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, 14(10):693–707.
- Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. [A multi-task approach for disentangling syntax and semantics in sentence representations](#). 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 2453–2464, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. [Revisiting pre-trained models for Chinese natural language processing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.
- Veronica Diveica, Penny M Pexman, and Richard J Binnie. 2023. Quantifying social semantics: An inclusive definition of socialness and ratings for 8388 english words. *Behavior Research Methods*, 55(2):461–473.
- Xinyi Dong, Bing Liu, Weijie Huang, Haojie Chen, Yunhao Zhang, Zeshan Yao, Amir Shmuel, Aocai Yang, Zhengjia Dai, Guolin Ma, and 1 others. 2024. Disrupted cerebellar structural connectome in spinocerebellar ataxia type 3 and its association with transcriptional profiles. *Cerebral Cortex*, 34(6):bhae238.
- Xufeng Duan, Zhaoqian Yao, Yunhao Zhang, Shaonan Wang, and Zhenguang G Cai. 2025. How syntax specialization emerges in language models. *arXiv preprint arXiv:2505.19548*.
- Arnaud D’Argembeau, Helena Cassol, Christophe Phillips, Evelyne Baeteau, Eric Salmon, and Martial Van der Linden. 2014. Brains creating stories of selves: the neural basis of autobiographical reasoning. *Social cognitive and affective neuroscience*, 9(5):646–652.
- Leonardo Fernandino, Jeffrey R Binder, Rutvik H Desai, Suzanne L Pendl, Colin J Humphries, William L Gross, Lisa L Conant, and Mark S Seidenberg. 2016. Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral cortex*, 26(5):2018–2034.
- Leonardo Fernandino, Jia-Qing Tong, Lisa L Conant, Colin J Humphries, and Jeffrey R Binder. 2022. Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences*, 119(6):e2108091119.
- Peter Gardenfors. 2004. *Conceptual spaces: The geometry of thought*. MIT press.
- Yunhao Ge, Yao Xiao, Zhi Xu, Xingrui Wang, and Laurent Itti. 2022. Contributions of shape, texture, and color in visual recognition. In *European Conference on Computer Vision*, pages 369–386. Springer.
- Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, and 1 others. 2013. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124.
- Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Boyu Guan, Yining Zhang, Yang Zhao, and Chengqing Zong. 2025. [TriFine: A large-scale dataset of vision-audio-subtitle for tri-modal machine translation and benchmark with fine-grained annotated tags](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8215–8231, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jiawei Guo, Feifei Zhai, Pu Jian, Qianrun Wei, and Yu Zhou. 2025. Crop: Contextual region-oriented visual token pruning. *arXiv preprint arXiv:2505.21233*.
- Paul Hoffman and Matthew A Lambon Ralph. 2013. Shapes, scents and sounds: quantifying the full multi-sensory basis of conceptual knowledge. *Neuropsychologia*, 51(1):14–25.
- Geoff Hollis and Chris Westbury. 2016. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review*, 23:1744–1756.
- Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.
- Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. [Adversarial example generation with syntactically controlled paraphrase networks](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.
- Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. 2025. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3619–3638.

- Pu Jian, Donglei Yu, and Jiajun Zhang. 2024. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for vqa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10939–10956.
- Saniya Karwa and Navpreet Singh. 2025. [Disentangling linguistic features with dimension-wise analysis of vector embeddings](#). In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 461–488, Albuquerque, New Mexico. Association for Computational Linguistics.
- Philip A Kragel and Kevin S LaBar. 2016. Decoding the nature of emotion in the brain. *Trends in cognitive sciences*, 20(6):444–455.
- George Lakoff. 2008. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.
- Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27.
- Chong Li, Shaonan Wang, Yunhao Zhang, Jiajun Zhang, and Chengqing Zong. 2023. [Interpreting and exploiting functional specialization in multi-head attention under multi-task learning](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16460–16476, Singapore. Association for Computational Linguistics.
- Keng-Te Liao, Cheng-Syuan Lee, Zhong-Yu Huang, and Shou-de Lin. 2020. [Explaining word embeddings via disentangled representation](#). In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 720–725, Suzhou, China. Association for Computational Linguistics.
- Nan Lin, Xiaoying Wang, Yangwen Xu, Xiaosha Wang, Huimin Hua, Ying Zhao, and Xingshan Li. 2018. Fine subdivisions of the semantic network supporting social and sensory-motor semantic processing. *Cerebral Cortex*, 28(8):2699–2710.
- Nan Lin, Yangwen Xu, Huichao Yang, Guangyao Zhang, Meimei Zhang, Shaonan Wang, Huimin Hua, and Xingshan Li. 2020. Dissociating the neural correlates of the sociality and plausibility effects in simple conceptual combination. *Brain Structure and Function*, 225:995–1008.
- Nan Lin, Xiaohan Zhang, Xiuyi Wang, and Shaonan Wang. 2024. The organization of the semantic network as reflected by the neural correlates of six semantic dimensions. *Brain and Language*, 250:105388.
- Kristen A Lindquist, Tor D Wager, Hedy Kober, Eliza Bliss-Moreau, and Lisa Feldman Barrett. 2012. The brain basis of emotion: a meta-analytic review. *Behavioral and brain sciences*, 35(3):121–143.
- Daniela Mapelli and Marlene Behrmann. 1997. The role of color in object recognition: Evidence from visual agnosia. *Neurocase*, 3(4):237–247.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. 2017. Variational dropout sparsifies deep neural networks. In *International conference on machine learning*, pages 2498–2507. PMLR.
- Charles O’Neill, Christine Ye, Kartheik Iyer, and John F Wu. 2024. Disentangling dense embeddings with sparse autoencoders. *arXiv preprint arXiv:2408.00657*.
- Allan Paivio. 1990. *Mental representations: A dual coding approach*. Oxford university press.
- Stephen E Palmer. 1999. *Vision science: Photons to phenomenology*. MIT press.
- 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.
- David Poeppel and M Florencia Assaneo. 2020. Speech rhythms and their neural foundations. *Nature reviews neuroscience*, 21(6):322–334.
- Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. 2025. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*.
- Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.
- James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.
- Heini Saarimäki, Athanasios Gotsopoulos, Iiro P Jääskeläinen, Jouko Lampinen, Patrik Vuilleumier, Riitta Hari, Mikko Sams, and Lauri Nummenmaa. 2016. Discrete neural signatures of basic emotions. *Cerebral cortex*, 26(6):2563–2573.
- David S Sabsevitz, David A Medler, Mark Seidenberg, and Jeffrey R Binder. 2005. Modulation of the semantic system by word imageability. *Neuroimage*, 27(1):188–200.
- Roger N Shepard. 1987. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323.
- Danilo Silva De Carvalho, Giangiacomo Mercatali, Yingji Zhang, and André Freitas. 2023. [Learning](#)

- disentangled representations for natural language definitions. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1371–1384, Dubrovnik, Croatia. Association for Computational Linguistics.
- Jefferson A Singer and Susan Bluck. 2001. New perspectives on autobiographical memory: The integration of narrative processing and autobiographical reasoning. *Review of General Psychology*, 5(2):91–99.
- Laura M Skipper, Lars A Ross, and Ingrid R Olson. 2011. Sensory and semantic category subdivisions within the anterior temporal lobes. *Neuropsychologia*, 49(12):3419–3429.
- Jingyuan Sun, Mingxiao Li, Zijiao Chen, Yunhao Zhang, Shaonan Wang, and Marie-Francine Moens. 2023. Contrast, attend and diffuse to decode high-resolution images from brain activities. *Advances in Neural Information Processing Systems*, 36:12332–12348.
- Jerry Tang, Amanda LeBel, and Alexander G Huth. 2021. Cortical representations of concrete and abstract concepts in language combine visual and linguistic representations. *bioRxiv*, pages 2021–05.
- Xiangrong Tang, James R Booth, Yunhao Zhang, Shaonan Wang, and Guosheng Ding. 2025. Neural trajectories reveal orchestration of cortical coding underlying natural language composition. *bioRxiv*, pages 2025–04.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.
- David Temperley. 2004. *The cognition of basic musical structures*. MIT press.
- Jiaqing Tong, Jeffrey R Binder, Colin Humphries, Stephen Mazurchuk, Lisa L Conant, and Leonardo Fernandino. 2022. A distributed network for multimodal experiential representation of concepts. *Journal of Neuroscience*, 42(37):7121–7130.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Tor D Wager, Jian Kang, Timothy D Johnson, Thomas E Nichols, Ajay B Satpute, and Lisa Feldman Barrett. 2015. A bayesian model of category-specific emotional brain responses. *PLoS computational biology*, 11(4):e1004066.
- Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. 2023a. 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.
- Shaonan Wang, Jingyuan Sun, Yunhao Zhang, Nan Lin, Marie-Francine Moens, and Chengqing Zong. 2024. Computational models to study language processing in the human brain: A survey. *arXiv preprint arXiv:2403.13368*.
- Shaonan Wang, Xiaohan Zhang, Jiajun Zhang, and Chengqing Zong. 2022a. A synchronized multimodal neuroimaging dataset for studying brain language processing. *Scientific Data*, 9(1):590.
- Shaonan Wang, Yunhao Zhang, Weiting Shi, Guangyao Zhang, Jiajun Zhang, Nan Lin, and Chengqing Zong. 2023b. A large dataset of semantic ratings and its computational extension. *Scientific Data*, 10(1):106.
- Shaonan Wang, Yunhao Zhang, Xiaohan Zhang, Jingyuan Sun, Nan Lin, Jiajun Zhang, and Chengqing Zong. 2022b. An fmri dataset for concept representation with semantic feature annotations. *Scientific Data*, 9(1):721.
- Yifan Wang, Jingyuan Sun, Jichen Zheng, Yunhao Zhang, Chunyu Ye, Jixing Li, Chengqing Zong, and Shaonan Wang. 2025. Bridging brains and models: Moe-based functional lesions for simulating and rehabilitating aphasia. *arXiv preprint arXiv:2508.04749*.
- Hua Xie, Iliana I Karipidis, Amber Howell, Meredith Schreier, Kristen E Sheau, Mai K Manchanda, Rafi Ayub, Gary H Glover, Malte Jung, Allan L Reiss, and 1 others. 2020. Finding the neural correlates of collaboration using a three-person fmri hyperscanning paradigm. *Proceedings of the National Academy of Sciences*, 117(37):23066–23072.
- Qihui Xu, Yingying Peng, Samuel A Nastase, Martin Chodorow, Minghua Wu, and Ping Li. 2025. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature human behaviour*, pages 1–16.
- XiaoHong Yang, HuiJie Li, Nan Lin, XiuPing Zhang, YinShan Wang, Ying Zhang, Qian Zhang, XiNian Zuo, and YuFang Yang. 2019. Uncovering cortical activations of discourse comprehension and their overlaps with common large-scale neural networks. *NeuroImage*, 203:116200.
- Chunyu Ye, Yunhao Zhang, Jingyuan Sun, Chong Li, Chengqing Zong, and Shaonan Wang. 2025. [Decoding the multimodal mind: Generalizable brain-to-text translation via multimodal alignment and adaptive routing](#). *Preprint*, arXiv:2505.10356.
- Xiaohan Zhang, Shaonan Wang, Nan Lin, Jiajun Zhang, and Chengqing Zong. 2022. Probing word syntactic representations in the brain by a feature elimination method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10, pages 11721–11729.
- Yunhao Zhang, Chong Li, Xiaohan Zhang, Xinyi Dong, and Shaonan Wang. 2023a. A comprehensive neural and behavioral task taxonomy method for transfer learning in nlp. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 233–241.

Yunhao Zhang, Shaonan Wang, Xinyi Dong, Jiajun Yu, and Chengqing Zong. 2023b. Navigating brain language representations: A comparative analysis of neural language models and psychologically plausible models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Yunhao Zhang, Shaonan Wang, Nan Lin, Lingzhong Fan, and Chengqing Zong. 2025a. A simple clustering approach to map the human brain’s cortical semantic network organization during task. *NeuroImage*, 309:121096.

Yunhao Zhang, Xiaohan Zhang, Chong Li, Shaonan Wang, and Chengqing Zong. 2025b. Mulcogbench: a multi-modal cognitive benchmark dataset for evaluating chinese and english computational language models: Y. zhang et al. *Language Resources and Evaluation*, pages 1–24.

Kangchen Zhu, Zhiliang Tian, Jingyu Wei, Ruifeng Luo, Yiping Song, and Xiaoguang Mao. 2024. [StyleFlow: Disentangle latent representations via normalizing flow for unsupervised text style transfer](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15384–15397, Torino, Italia. ELRA and ICCL.

A Overview of the framework

Figure A3 shows the overview of the framework to identify the subdimensions within each semantic dimension, and use the voxel-wise encoding method to assess their neural plausibility.

B Validation result

Table A6 shows the semantic prediction accuracies on the original and disentangled embeddings.

C Semantic alignment of transformed sub-embddings.

Figure A4 illustrates the semantic alignment between the semantic rating data and each dimension of the transformed sub-embedding within the corresponding semantic dimension.

D Anatomically defined brain regions

Figure A5 shows the anatomically defined brain regions that are commonly associated with language processing.

E Voxel-wise encoding results for SSDD and transformed sub-embeddings

Figure A6 and Figure A7 show the group-level voxel-wise encoding performance for different semantic dimensions, using human rating data from SSDD and transformed sub-embeddings derived from DCSRm, respectively.

F Licenses of scientific artifacts

We follow and report the licenses of scientific artifacts involved in Table A7.

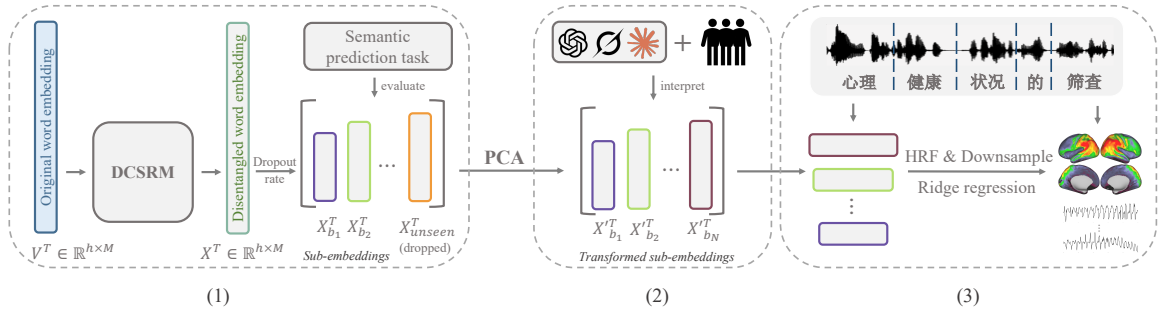


Figure A3: Overview of the proposed framework. The framework consists of three key stages: (1) Disentangled continuous semantic representation model (DCSRM). Original word embeddings $V \in \mathbb{R}^{M \times h}$ extracted from language models are input to DCSR, producing disentangled embeddings $X \in \mathbb{R}^{M \times h}$ and a dropout rate matrix $D \in \mathbb{R}^{N \times h}$, where M is the number of words, h is the embedding dimension, and N is the number of semantic attributes. Each element (i, j) in D indicates the selectivity of the j -th dimension of X for the i -th attribute, with lower dropout rates denoting higher semantic relevance. Thresholding D yields multiple attribute-specific sub-embeddings. (2) Sub-embedding analysis. Each attribute-specific sub-embedding $X_{b_1}, X_{b_2}, \dots, X_{b_N}$ is orthogonalized via PCA to obtain transformed sub-embeddings $X'_{b_1}, X'_{b_2}, \dots, X'_{b_N}$. For each dimension within each transformed sub-embedding, we prompt multiple large language models (e.g., GPT-4o, Grok-3, Claude 3.7 Sonnet) to infer the underlying semantic meaning based on the top-ranked words. Their responses are aggregated to determine the final subdimension labels. (3) Voxel-wise encoding. We map each transformed sub-embedding to brain activity during natural story comprehension using voxel-wise encoding, resulting in a weight matrix $U \in \mathbb{R}^{f \times g}$, where f is the number of subdimensions and g is the number of cortical voxels. Each voxel is assigned the subdimension with the highest weight, indicating its strongest semantic preference. The example stimulus "心理健康状况的筛查" is taken from the fMRI dataset used in this study, and its English translation is "Screening for mental health conditions".

Dimension	Definition
Vision	The extent to which the meaning of a word can easily and quickly trigger corresponding visual images in your mind
Action	The extent to which the meaning of a word can easily and quickly trigger corresponding body actions in your mind
Social	The extent to which the meaning of a word relates to relationships or interactions between people
Emotion	The extent to which the meaning of a word relates to positive or negative emotions
Time	The extent to which the meaning of a word relates to time, including early or late, length, sequence, frequency, etc.
Space	The extent to which the meaning of a word relates to spatial information, including location, direction, distance, path, scene, etc.

Table A4: Definition of each semantic dimension.

Word	Vision	Action	Social	Emotion	Time	Space
justice	1.600	1.667	2.400	3.867	1.000	1.133
sea	6.033	2.767	1.133	1.500	1.000	5.933
june	2.333	2.033	1.067	1.167	6.733	1.267
football	6.500	5.433	4.133	1.533	1.300	3.200
of	1.067	1.000	1.100	1.000	1.033	1.033

Table A5: Six-dimension semantic ratings for the concepts "justice (正义)", "sea (海洋)", "june (六月)", "football (足球)" and "of (的)".

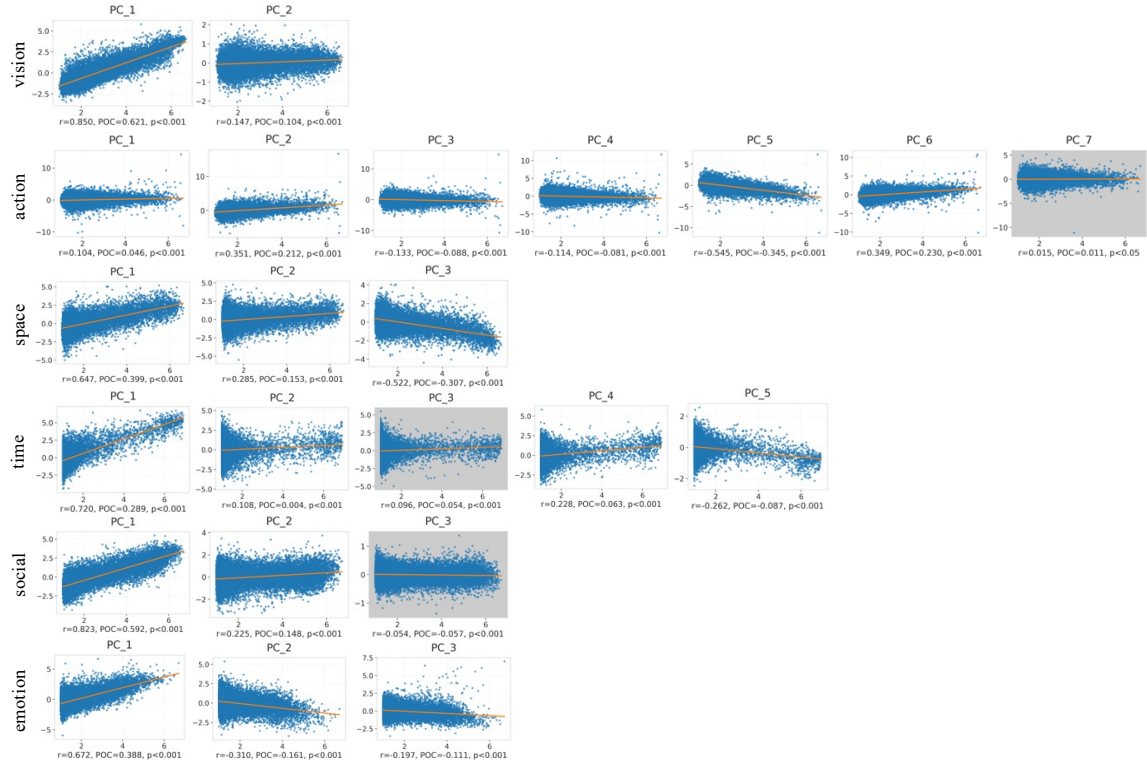


Figure A4: **Semantic alignment of transformed sub-embeddings.** Pearson correlation (r) and pairwise order consistency (POC) between the semantic rating data (x-axis) and each dimension of the transformsub-embedding within the corresponding semantic dimension (y-axis). We only annotate $p < 0.001$ on the figure when both correlation results are statistically significant at $p < 0.001$. Significant correlations ($p < 0.001$) indicate that the respective sub-embedding dimension encodes the corresponding semantic information. Dimensions with a correlation below 0.1 are shown with a gray background. Transformed sub-embeddings refer to the outputs obtained by applying PCA individually to each semantic-specific sub-embedding produced by DCSRM (see Section 3.2 for details).

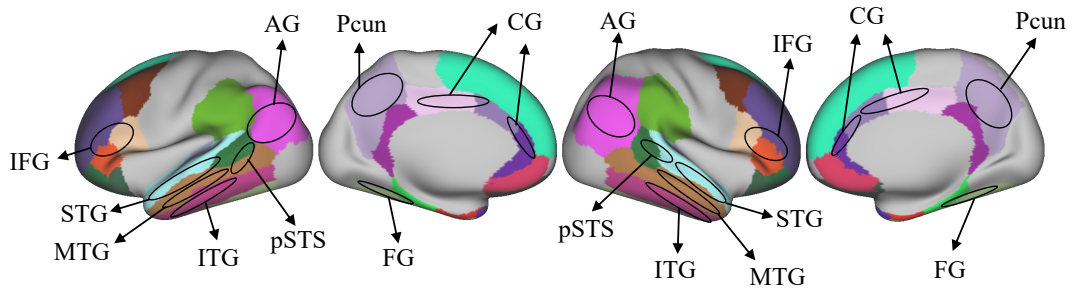


Figure A5: **Anatomically defined brain regions involved in language processing.** The labeled regions are widely implicated in language comprehension and production, including the inferior frontal gyrus (IFG), superior temporal gyrus (STG), posterior superior temporal sulcus (pSTS), middle temporal gyrus (MTG), inferior temporal gyrus (ITG), precuneus (Pccun), cingulate gyrus (CG), fusiform gyrus (FG), and angular gyrus (AG).

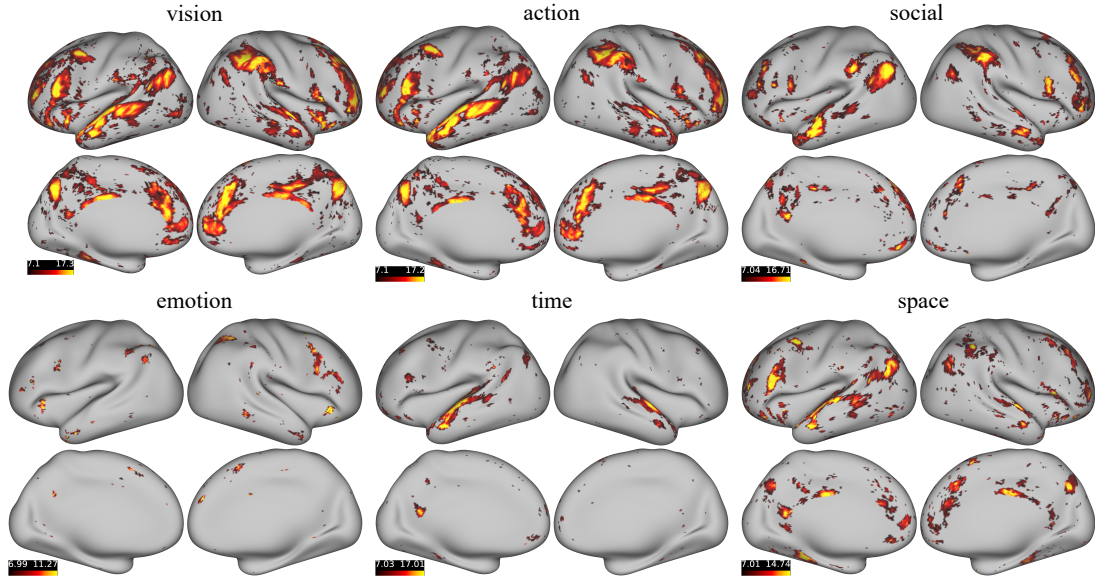


Figure A6: **Group-level voxel-wise encoding results for SSDD.** Model performance was evaluated by computing the Pearson correlation between predicted and observed BOLD responses at each voxel. Color-highlighted regions indicate areas where BOLD responses predicted from SSDD rating data significantly exceeded a random baseline (one-tailed t-test, $-\log(p)$, $p < 0.001$).

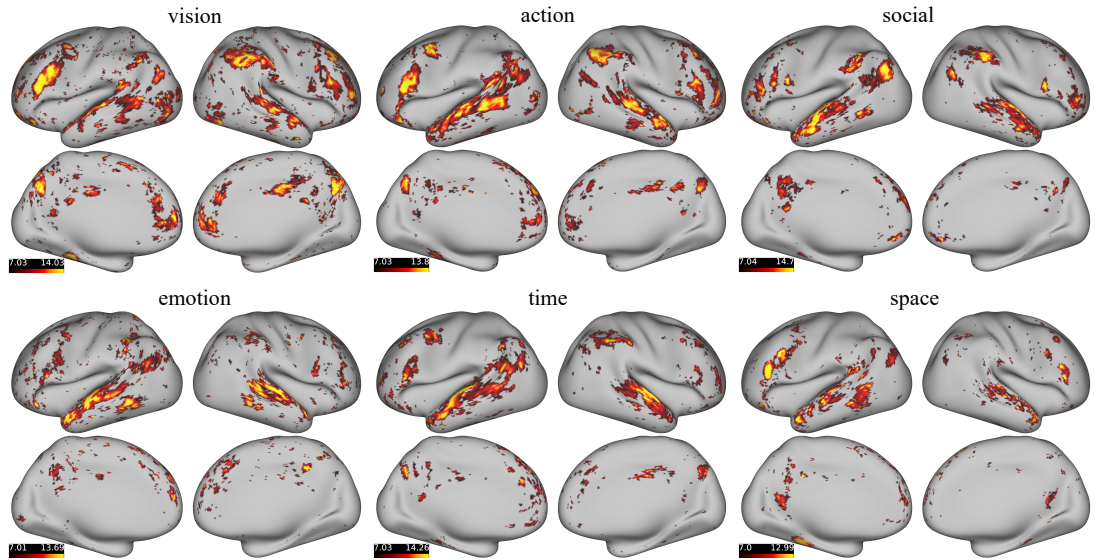


Figure A7: **Group-level voxel-wise encoding results for transformed sub-embeddings.** Model performance was evaluated by computing the Pearson correlation between predicted and observed BOLD responses at each voxel. Color-highlighted regions indicate areas where BOLD responses predicted from DCSRm transformed sub-embeddings significantly exceeded a random baseline (one-tailed t-test, $-\log(p)$, $p < 0.001$). Transformed sub-embeddings refer to the outputs obtained by applying PCA individually to each semantic-specific sub-embedding produced by DCSRm (see Section 3.2 for details).

	vis		act		soc		emo		time		spc		Average	
	origin	disent.	origin	disent.	origin	disent.	origin	disent.	origin	disent.	origin	disent.	origin	disent.
GloVe	0.647	0.646	0.604	0.602	0.651	0.650	0.561	0.559	0.534	0.531	0.681	0.679	0.613	0.611
Word2Vec	0.854	0.853	0.758	0.758	0.823	0.822	0.746	0.746	0.697	0.697	0.822	0.822	0.783	0.783
MACBERT	0.878	0.878	0.795	0.795	0.875	0.875	0.836	0.836	0.843	0.843	0.885	0.885	0.852	0.852
LLaMA2-1.3b	0.906	0.900	0.835	0.831	0.885	0.885	0.849	0.857	0.876	0.872	0.900	0.898	0.875	0.874
Alpaca2-1.3b	0.893	0.893	0.816	0.816	0.879	0.879	0.842	0.842	0.869	0.869	0.894	0.894	0.865	0.865
LLaMA2-7b	0.905	0.904	0.831	0.831	0.884	0.884	0.850	0.850	0.876	0.877	0.900	0.900	0.874	0.874
Alpaca2-7b	0.906	0.906	0.835	0.835	0.885	0.884	0.849	0.849	0.876	0.876	0.900	0.900	0.875	0.875
LLaMA3-8b	0.908	0.908	0.799	0.799	0.874	0.874	0.834	0.834	0.857	0.858	0.885	0.885	0.860	0.860

Table A6: **Semantic prediction accuracies on original and disentangled embeddings.** For each semantic dimension (e.g., vision), columns labeled “origin” and “disent.” report the Pearson correlation between predicted and ground-truth ratings using original and disentangled embeddings, respectively. “Origin” refers to the original embeddings $V \in \mathbb{R}^{M \times h}$ extracted from language models, and “disent.” refers to the disentangled embeddings $X \in \mathbb{R}^{M \times h}$ obtained by inputting V into the DCSR model, where M is the number of words and h is the embedding dimension.

Name	License
Transformers	Apache 2.0 license
Connectome Workbench	GNU General Public license
NiBabel	MIT license
Matplotlib	PSF license
LLaMA3-8b	Apache 2.0 license
Alpaca2-7b	Apache 2.0 license
Alpaca2-1.3b	Apache 2.0 license
LLaMA2-7b	Apache 2.0 license
LLaMA2-1.3b	Apache 2.0 license
MacBERT-large	Apache 2.0 license
SMN4Lang	CC BY-SA 4.0 and CC0 license
SSDD	CC BY-SA 4.0 license

Table A7: Licenses of scientific artifacts involved in this work.